CASCADING BEHAVIOR
IN SOCIAL NETWORKS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Justin Cheng
August 2017

This dissertation is online at: http://purl.stanford.edu/dm101hm5137

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Jure Leskovec, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Michael Bernstein

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Jon Kleinberg

Approved for the Stanford University Committee on Graduate Studies.

Patricia J. Gumport, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.

Abstract

Cascades occur in social networks when information and behavior spreads from person to person. However, as these cascades grow through complex macro- and micro-level processes, their future behavior is difficult to predict.

In this thesis, we study the mechanisms through which cascades propagate in networks. Specifically, we present a framework for predicting a cascade's future trajectory, and develop methods that combine data mining and crowdsourcing techniques to explain how behavior spreads from individual to individual.

First, we examine the growth and recurrence of information cascades on Facebook. In contrast to prior work that argued that such cascades are unpredictable, we show how the size, structure, and recurrence of a cascade can be predicted, even over long periods of time.

Second, we study how behavior cascades by looking at how antisocial behavior such as trolling may spread from person to person. While past literature has characterized such behavior as confined to a vocal, antisocial minority, we instead demonstrate that ordinary people can become trolls under the right circumstances, and that such behavior can percolate and escalate through a community.

Altogether, this research explores a future where systems can better mediate information sharing and interpersonal interaction, and thus promote the development of prosocial communities.

Acknowledgements

This thesis would not have been possible without:

- Jure Leskovec. Jure has been an endless wellspring of ideas and intense motivation, and many of findings in this thesis were motivated by conversations with him.
- Michael Bernstein. Michael has been an incredible source of support for me, both in the projects that we've worked on together, as well as during periods when I felt discouraged and unmotivated.
- Jon Kleinberg. In addition to being an amazing mentor and teacher, Jon is one of the smartest people that I know, and his research sparked my interest in social networks.
- Lada Adamic. Lada is my role model for being successful in both academia and in industry, and many of our projects together at Facebook would not have even been conceived without her.
- Cristian Danescu-Niculescu-Mizil. Mentor, collaborator, and friend. Through the thought and care that he puts into projects he works on (including ones that we worked on together), Cristian has also made me a more careful and thoughtful researcher.
- Jeff Hancock and James Landay. Jeff's and James's work is worthy of much admiration, and I feel incredibly fortunate for them to be part of my orals

committee.

- Dan Cosley, Gilly Leshed, Lillian Lee, Daniel Romero, and Leo Kang. My fateful meeting with Dan in my first year of college put me on the track towards a career in research, and I'm very grateful for his advice over the years. Gilly opened my eyes to the wide world of human-computer interaction and also helped support my first visit and presentation at CHI. Lillian also supported my initial forays into research. And while Daniel lent me the first book I had ever read about machine learning, Leo showed me how research connects to art.

- Scott Klemmer and Chinmay Kulkarni. Scott and Chinmay shaped my formative years as a PhD student, and taught me a lot about presentation, aesthetics and experimental methods.

- The Stanford HCI Group and SNAP Group. Especially Jon Bassen, Ethan Fast, and Will McGrath – I am glad to have met each of you! I'm also deeply appreciative of Ali Alkhatib for continuing to run the "Reading with Friends" reading group. Ashton Anderson, Tim Althoff, Austin Benson, Christina Brandt, David Hallac, Will Hamilton, Tim Hsieh, Sanjay Kairam, Chloe Kliman-Silver, Nicolas Kokkalis, Geza Kovacs, Ranjay Krishna, Srijan Kumar, Kim Myunghwan, Himabindu Lakkaraju, Caroline Lo, Seth Myers, Niloufar Salehi, Yang Jaewon, and Bob West, thank you for the memories. Special thanks to Rok Sosic, Andrej Krevl, and Peter Kacin for help with SNAP and managing the group's data infrastructure. I'd also like to thank Helen Buendicho, Yesenia Gallegos, Andrea Kuduk, Jillian Lentz, Marianne Siroker, and Jay Subramanian for their administrative help.

- Facebook, Microsoft Research, Pinterest, Disqus, and the Stanford VPGE. Alex Dow, Karthik Subbian, and Bogdan State were great colleagues during my multiple stints at Facebook. Jaime Teevan, Shamsi Iqbal, Scott Counts, Emre Kiciman, and Eric Horvitz are only a few of the great researchers from MSR that I've met over the years. I also had the pleasure of working with multiple talented engineers and UX researchers at Pinterest. I'm especially grateful to Stanford

Office of the Vice Provost for Graduate Education and Microsoft Research for funding my research.

- My family. My parents, Gerard and Anita Cheng, who have *always* been there for me and supported my goals, and who I'm probably not grateful enough for. I'm also very proud of my sister, Chris-Tin, who's an actual doctor.

- Joy Kim, who probably deserves more than one bullet point. She's supported me in so many ways that I've lost count.

Parts of this work were also supported by an Alfred P. Sloan Fellowship, a Google Research Faculty Award, a Microsoft Faculty Fellowship, a Simons Investigator Award, ARO MURI, Boeing, DARPA NGS2, Lightspeed, NSF Grants CNS-1010921, IIS-1149837 and IIS-1159679, SAP, SDSI, the Stanford Data Science Initiative, and Volkswagen.

Finally, I'd like to thank you, the reader, for actually taking the time to read through at least some portion of this work.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Online social networks are increasingly where people spend their time [276] and where they get information about the world around them [128]. On these platforms, information and behavior can spread from person to person as content is shared, discussed, or voted on.

The *cascades* of information and behavior that result are both prevalent and influential. Cascades can be used to understand the spread of technological innovation [256], collective action [129], and even the propagation of negative health behaviors [78]. Two recent examples of cascades that occured on social media include the Ice Bucket Challenge on Facebook and Gamergate on Twitter. While the former raised awareness about Amyotrophic Lateral Sclerosis (or ALS) [82], the latter, which resulted in the spread of hateful messages [21], showed how cascades can also propagate negative outcomes.

These last examples exemplify two significant challenges that a better understanding of cascading behavior can help tackle – that of understanding virality, or how things become popular, and of understanding how antisocial behavior has become prevalent online [101]. In the former case, how content online "goes viral" remains less well-understood, with a recent line of work even having cautioned that virality may be

inherently unpredictable [257, 220]. In the latter case, while prior work has characterized antisocial behavior as propagated by a small but vocal minority [22, 143, 262], its prevalence suggests that other causes exist.

Correspondingly, this thesis seeks to answer two research questions:

1. Can we predict when information will "go viral"?
2. Can we predict when antisocial behavior will "go viral"?

Together, these questions aim to identify and understand the mechanisms of how people interact with one another to transmit a particular *contagion* (e.g., information or behavior) in a network, and to understand where these individual interactions can go on to have cascading effects.

Thus, first question is really a question about information cascades, where information is transmitted from person to person in a network. By understanding the mechanisms of the spread of information, we can then begin to predict this spread, and better understand virality and popularity. To answer this question, we demonstrate how a cascade's future behavior can be predicted from observing its initial spread, and show how a cascade's future size, network structure, and its recurrence are predictable.

In contrast, the second question focuses on behavioral cascades, where people influence one another through their interactions. And if these interactions affect a person's future behavior in a certain way, then behavior can similarly cascade through a network. We address this question by studying the propagation of antisocial behavior in online discussions. We first identify the causal factors that result in ordinary people engaging in trolling behavior, and then show how voting mechanisms can lead to antisocial behavior percolating through a community.

Overall, a better understanding of human behavior in online social networks provides us with a principled approach to improving the design of social systems. Knowing how information spreads, we can both estimate and maximize the potential spread of new content, or curb the spread of undesirable content. Knowing the mechanisms of

antisocial behavior, we can then design interventions that minimize the factors that lead to people behaving badly online, and build healthier online communities.

1.1 Thesis Overview

This thesis will examine the two questions above in greater detail. Chapters 3 and 4 focus on information cascades in social networks, and on predicting the future virality of such cascades. Chapters 5 and 6 focus on behavioral cascades, and in particular, antisocial behavior in online discussions and the mechanisms that can lead to such behavior becoming prevalent.

1.1.1 Information Cascades in Social Networks

This chapter addresses fundamental questions around the predictability of information cascades in social networks.

Information can cascade or diffuse in a network when a person shares content with their friends. And as that person's friends share that content with their friends in turn, an information cascade can result.

While it would be valuable to be able to predict the spread of information in networks to predict the eventual reach of a piece of content, prior work has suggested that these cascades are inherently unpredictable because popularity tends to arise from social influence [257, 220], which itself tends to be unpredictable. By viewing a cascade as continuously evolving, this chapter develops a framework for addressing cascade prediction problems. This framework first allows us to predict the future growth of a cascade based on its initial trajectory. It also allows us to predict a cascade's future structure. Will its network structure look more like a star, with a single node directly influencing many others? Or will it look more like a tree, with individual nodes influencing a small number of others, who go on to influence others in turn?

Another aspect relating to the predictability of cascades stems from a commonly-held perception that cascades tend to follow a basic rising-and-falling pattern – a rapid increase in popularity, followed by a gradual tapering in popularity. Indeed, such perceptions are perhaps reinforced by the rapid ascent and decline of trends in online social media such as "Gangnam Style" in 2012 or "Pokemon Go" in 2016. But by studying cascades over long periods of times, this chapter shows how many cascades might in fact recur, where cascades may experience multiple resurgences in popularity separated by significant periods of quiescence. By predicting whether cascades recur, we can then develop an even longer-range view of cascade growth, even after a cascade has seemingly stopped growing.

1.1.2 Antisocial Behavior in Online Discussions

While the previous chapter studied how information cascades at macro-scale, with a cascade being the unit of analysis, this chapter instead studies it at micro-scale, and tries to understand how individuals may influence adjacent others in a social network through their actions. In particular, this chapter focuses on the mechanisms that lead to negative behavior spreading in online communities.

While the social web has brought us closer together, it has also become a conduit for harassing others (for example, in the form of trolling). Negative behavior comprises a substantial fraction of user activity on many web sites [94, 112]. In fact, 40% of Internet users have experienced harassment in one form or another [101], while 73% have seen others harassed [101]. Anecdotally, many of us can easily recall instances of others behaving badly in online settings. But what has led to this prevalence of negative behavior online?

While past literature has suggested that negative behavior is perpetuated by a small but vocal antisocial minority [22, 143, 262, 51], this chapter will demonstrate how ordinary individuals can act like trolls under the right circumstances. Combining data analysis of a large discussion community (CNN.com) with a controlled online

laboratory experiment, we establish the causal factors that lead to people behaving like trolls.

To manage bad content, many social sites have turned to voting as a way to curate the content that is shown to users. But despite good intentions, voting, and in particular, downvoting, can instead cause negative behavior to percolate. Not only are authors of downvoted content more likely to write content of worse quality in the future, but they are also more likely to be perceived worse by the community at large.

1.2 Research Approach and Impact

While the web started out as a way to connect people to vast amounts of information, today, our interactions online are increasingly oriented towards interacting with other people. To this end, this thesis explores the impact and influence that that people have on each others' behavior.

The approach taken primarily in this thesis is to use computational techniques to understand human behavior. It combines large-scale data analysis with online experiments, while building on existing theory from sociology and social psychology.

To study information cascades in social networks, we conducted analyses on the entirety of the Facebook social graph, in contrast to past work that primarily studied synthetic networks or smaller-scale social networks. To understand behavioral cascades, we study the spread of antisocial behavior in online discussions. We take a quantitative and data-driven approach, and combine longitudinal analyses of multiple discussion communities with online experiments on Mechanical Turk, as opposed to prior work that has been largely qualitative and interview- or survey-driven (e.g., [98, 262, 189]).

This mixed-methods approach allows us to use large-scale data to identify generalizable insights, and then combine it with online experiments enabled via crowdsourcing that can then establish causality.

Practically, the work in this thesis can be used to inform the design of social systems. Predictive models of cascade growth and recurrence can not only help advertisers estimate content reach, but also help temper the spread of undesirable content (e.g., rumors). Our findings on the mechanisms of antisocial behavior also have immediate implications on how interactions between users should take place. And by establishing the causal factors that result in antisocial behavior also allows us to design interventions that can moderate these factors. Finally, models for predicting negative behavior can enable better moderation of discussion communities through increased automation.

Chapter 2

Related Work

The study of cascades and cascading behavior has its roots in multiple disciplines including sociology, psychology, physics, and epidemiology. More recently, cascades have also become a keen area of interest in computer science as well as the developing fields of network science and computational social science. Cascades can explain how practices and technology spread in organizations, how disease epidemics develop [160], how social movements arise, how financial markets work [147], or even how obesity might spread [78]. Online and on the social web, cascades can be used to characterize the flow of information (e.g., across websites [132]) and how recommendations spread from person to person [192].

In this chapter, we begin with a brief overview of cascades, and then broadly survey research on cascading behavior in two parts:

- We survey work on cascades as they relate to networks as a whole, and where our focus is primarily on abstract representations of these cascades, including theoretical models and algorithms.
- We then turn our attention to cascades as they relate to individuals, and delve into how individuals are influenced by other individuals and their environment. In other words, can we understand the direct effects between individuals that

lead to cascades?

2.1 An Introduction to Cascades

Suppose that a group of friends are deciding which of two instant messaging platforms, Facemag or Chatter, to use. The earliest adopter in the group, Jack, happens to meet an acquaintance that uses Facemag, so he decides to use Facemag too. The next person in the group to decide, Jill, while previously having no preference between platforms, now also picks Facemag so that she can communicate with Jack. The third person to join, Jack, now has an even stronger reason to pick Facemag over Chatter, given that both Jack and Jill are already using Facemag. Eventually, the whole group of friends is using Facemag.

Illustrated in the above example is a *cascade*, where the use of a certain technology (Facemag) spread from person to person. In fact, the earliest studies of cascades focused on the *diffusion of innovations* [247], or how technology adoption spreads through communities. One study on the adoption of hybrid seed corn in an Iowan farming community [256, 278] found that farmers tended to adopt the new agricultural innovation after hearing about its success from neighboring farms. Similarly, another study demonstrated the role of social networks in doctors' adoption of a novel drug [84].

Similar cascading effects have also been demonstrated experimentally. In an experiment where participants had to decide in turn which urn a ball belonged to, participants tended to base their decisions on those of earlier participants [12].

Concretely, a cascade can be broadly defined as the result of a *contagion*'s propagation in a network. A contagion is defined as the spreading of information (e.g., a new drug), behavior, or even a disease. Figure 2.1 illustrates how this might take place on a network. In the context of a social network, nodes represent users, and edges between nodes represent friendships. Suppose that a user A shares an interesting news

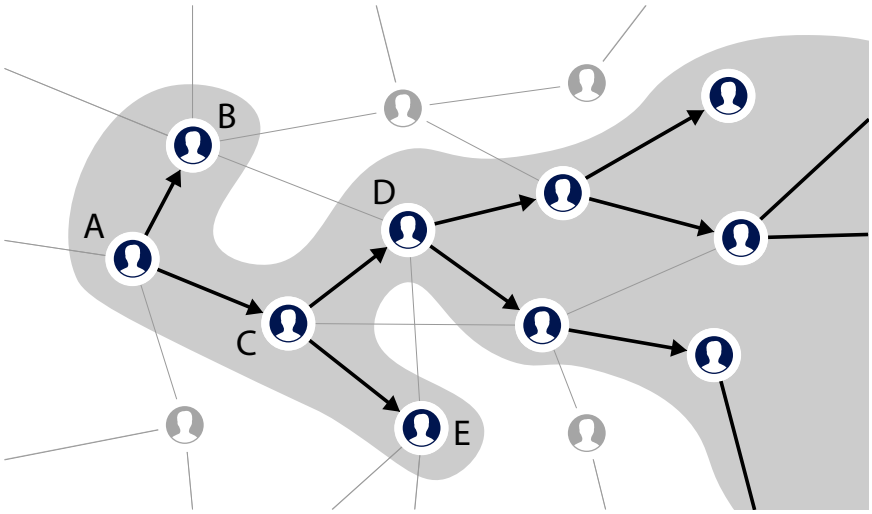


Figure 2.1: An example of a cascade in a network, highlighted in gray. The cascade begins with node A, spreads to some adjacent nodes B and C, and continues to grow from these adjacent nodes in turn.

article. Two of A's friends (B and C) see this news article, and decide to share it with their friends. Of B and C's friends, D and E share this article in turn, propagating the content further in the network, and so on. This propagation results in a cascade, highlighted in gray.

Thus, whenever individuals interact with one another in a network, information and behavior can spread rapidly through the network as any individual has the potential to influence every other individual they are connected to, with those individuals being able to influence their connections in turn. And as these interactions lead to contagion spreading from person to person in a network, cascades can develop. Given the general branching structure that develops as cascades grow, they are also known as diffusion trees.

		Player 1	
		x	y
Player 2	x	q,q	0,0
	y	0,0	(1-q),(1-q)

Table 2.1: The two-player coordination game. If Player 1 picks x and Player 2 picks x, they both receive a payoff of q. If they both pick y, they each receive a payoff of $(1 - q)$. However, if they pick different actions, they receive no payoff.

2.2 Cascades and Networks

This section examines research on cascades as they relate to networks as a whole. First, we introduce several models for studying cascades, then explore how recent research has led to a richer understanding of cascades (e.g., their temporal dynamics). Next, we visit the subfield of influence maximization, then turn our attention to predictive models of cascades. Finally, we present a abbreviated overview of the different domains that cascading behavior can be observed in.

2.2.1 Models

Game-Theoretic Models

The cascading behavior exhibited in our instant messaging example from earlier can actually be modeled as a two-player coordination game (Table 2.1). Players pick between one of two actions (e.g., x or y) to adopt, but only receive a payoff when they pick the same action. This payoff quantifies how beneficial a given action is to the player. Without any loss of generality, we define the payoff from coordinating on x as q, and that from coordinating on y as $1 - q$.

In our example, x and y correspond to Facemag and Chatter respectively. If Jack and Jill both decide to use Facemag, they each receive a payoff of q; if they both decide to use Chatter, they each receive a payoff of $1 - q$. Notice that if they decide to use different instant messaging platforms, they both receive a payoff of 0 (i.e.,

corresponding to not being able to communicate at all).

In the two-player instance, it is optimal for both to pick the action with the better payoff. For instance, if $1 - q > q$, both players should choose action y. Thus, if only John and Jill existed in the world, and Chatter provided a superior user experience (i.e., that $1-q > q$), they would both agree on using Chatter. However, if each player is playing the same game with multiple other players, but can only pick the same action, then they will pick the one that provides overall greater total utility, even if that action is the choice that individually provides less utility. For example, if a player is playing against $N-n_y$ players who are going to pick x, and n_y players who are going to pick y, then that player's optimal action is to pick y only if $n_y(1-q) > (N-n_y)q$ or that $n_y > Nq$. As such, since John initially had an acquaintance using Facemag, he also decided to use Facemag because $0 \not> 1 \cdot q$.

From this starting point, several theorems about the specific conditions under which cascades can successfully spread in a network have been developed. Consider q, the contagion threshold of y (i.e., the minimum fraction of neighbors needed for y to be adopted since y is picked only if $\frac{n_y}{N} > q$). As we vary q, is there a maximum value that limits y's spread in the network? We find that $q = 1/2$ is an upper bound for the contagiousness of y in all networks (or graphs) [219]. In other words, any contagion cannot spread very far if it requires a strict majority of neighbors for adoption.

Linear Threshold Models

The simple model above can be further generalized as what are known as linear threshold models.

Linear threshold models were initially proposed to explain how collective action occurs. They postulated that thresholds for taking actions exist, whether it is how the number of previous buyers influences a consumer's likelihood of purchasing a new product [32], or more generally, if there exists a point where the benefit to a given

person to participate exceeds the costs of participation [129]. Given a node in a network, these models state that it will adopt a given contagion only if more than a given fraction q of its neighbors have already adopted it. But unlike the previous model, this model allows each individual node to have its own value of q.

Independent Cascade Models

Further generalizing these linear threshold models is the independent cascade model [120], which presents a probabilistic view of cascading behavior. In this model, a node v transmits a contagion to each of its neighbors w with some given probability $p_{v,w}$.

In most cases, this probability $p_{v,w}$ is constant, but can also be modeled as an incremental function $p_v(u, X)$ for some node v, where u is an adjacent node and X is the set of neighbors of v not containing u. $p(u, X)$ is the probability that node u will activate v, given that all nodes in X has already failed.

Epidemic Models. Also closely related to the independent cascade model are compartmental models in epidemiology, the most commonly used being the SIR model [165], where S, I, and R represent different stratifications of a population (i.e., susceptible to infection, infected, and recovered respectively). In contrast to the previous models, these compartmental models were borne out of a desire to model the dynamics of disease transmission. In this model, nodes start out susceptible to being infected, become infected, and then recover. Other variations exist, such as SIS in the case of modeling the common cold, since one usually cannot become immune to colds. More recently, such models have been used in the context of online media to explain how blogs link to each other [196]. They can also be used to model certain types of recurrences such as seasonality effects (e.g., through introducing periodicity in contagion fitness [117]). Motivated by these applications, this thesis adapts an SIR model to simulate non-periodic cascade recurrence.

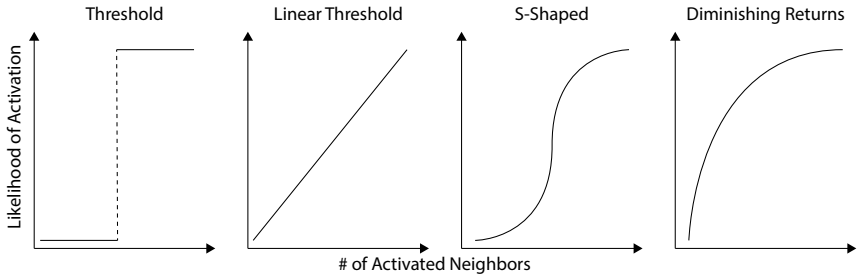


Figure 2.2: Different models suggest different ways in which the likelihood of a node's activation varies with the number of activated neighbors.

Model Extensions

Non-Linear Thresholds. One important early experiment on cascading behavior involved stationing crowds of varying size at a street corner looking up at the sky [214]. As the size of the crowd increased, the likelihood of passersby stopping to look up also increased, demonstrating both how behavior was more likely to be adopted as more individuals adopt such behavior. This suggests that at least in some cases, that the likelihood that a certain behavior is adopted varies with the number of other individuals who have already adopted the behavior.

Building on both the linear threshold and independent cascade models, an important question we might ask is how the likelihood of a node's activation (or infection) changes with the number of adjacent neighbors that are already activated.

Beyond thresholds where the probability of activation varies stepwise or linearly, do these threshold plots look more like curves (Figure 2.2)? Further, do these curves exhibit a transition point (resulting in a S-shaped adoption likelihood curve), or do they exhibit diminishing returns (increasing numbers of friends adopting have a decreasing effect on your likelihood of adoption)? Initial evidence from an analysis of LiveJournal [192] is suggestive: that the latter is more likely.

Complex Contagion. Complex contagion further builds on the idea of a contagion not having a single transmission probability. In contrast to *simple contagion* such as disease, where only one exposure is necessary for successful transmission, *complex*

contagion such as behavior may require multiple exposures to induce adoption [58]. Recent empirical work has demonstrated how the adoption of hashtags on Twitter behaves like complex rather than simple contagion [248]; other work has studied how the actual structure of an individual's network neighborhood affects the adoption likelihood [291].

Alternative Models

General Threshold Model. We can even further generalize the cascade models described thus far with a general threshold model [170]. We do so by defining an arbitrary function $f_v(N) \in (0, 1)$ for every node v in the network, where f_v is defined for any set of v's neighbors N, as well as some threshold $q_v \in (0, 1)$. In this model, a node is activated if $f_v(N) > q_v$. The generality of $f_v(N)$ allows us to encode complex rules such as adoption only in the case that a specific subset of neighbors also adopts the contagion.

Bayesian Models. Finally, some work has also developed Bayesian models [28] to model and explain decision-making cascades (e.g., in the ball-and-urn experiment).

2.2.2 Characterizing Cascade Growth

Perhaps owing to the proliferation of large online social networks and large amount of data on activity on these networks that has become available, recent work has seen a move from studying only synthetic models of cascading behavior to studying properties of cascades based on observational data. We explore several large open areas that not only enrich our understanding of how cascades grow, but also challenge several underlying assumptions of the theoretical models that we have seen thus far.

Temporal Dynamics of Cascades

Where our prior models have assumed a largely sequential, time-independent process through which a cascade develops, temporal analyses of actual cascades reveal a much richer story about how cascades develop over time.

One primary observation is that cascades tend to be bursty, with a spike of activity happening within days of days of its introduction [196]. As such, prior work focusing on the temporal dynamics of cascades tends to assume that the temporal shape of a cascade consists primarily of a rising and falling pattern [4, 208]. This burstiness has typically been modeled as a direct result of external stimuli (e.g., breaking news [180]) or of periodic behavior (e.g., seasonality effects [16]). Nonetheless, the temporal patterns exhibited by cascades over significantly longer timescales has remained largely unexplored.

Contagion and Network Evolution

The models we have discussed thus far largely abstract the notion of what contagion is being transmitted, as well as what network a cascade is spreading on. Nonetheless, the nature of the contagion being transmitted, as well as the underlying network on which a cascade spreads can influence a cascade's growth.

In social media, work has shown how different topics have different transmission rates on Twitter [248]. Other work has studied how contagion can evolve over time as well – Memetracker demonstrated how phrases changed over time as different news media sources quoted each other [193]. More recently, drawing from work on the distribution of biological taxa [317], information copying on Facebook has been found to follow a Yule process in how information mutates over time [1].

In addition to work that experimentally demonstrated how network structure can alter the spread of a cascade [57], other work has shown the reverse, or how the cascade can itself alter the network structure [301]. Follow-up work has also described models

for how diffusion and network evolution can be jointly modeled [105].

Nonetheless, research on contagion and network evolution in this space has been relatively sparse, given the relative challenge of obtaining sufficient data to study such phenomena, in addition to technical challenges in identifying when contagion mutations or network changes occur.

External Influence

Rather than assuming that networks are self-contained, can we quantify external influence on a network? For example, what impact might news events reported in the mainstream media have on the topics discussed on social media?

The impact of external influence has been discussed in literature on social movements, where the media can influence the success or failure of social movements. For instance, spikes in protest participation is more likely attributable to media attention rather than diffusion through the social network of potential protesters [95].

More recently, one study on Twitter found that external influences to the network can be used to explain how information "jumps" from one part of the network to another, and estimated that about 30% of the information transmitted happens due to factors external to the network rather than network diffusion [226]. Another demonstrated how the instant messaging network in a large hedge fund was affected by price shocks [250]. External shocks have also considered in temporal cascade models [208].

2.2.3 Influence Maximization

One significant area of research interest arose from a graph theory perspective of cascades: given a certain budget for activating nodes in a graph, how can we pick the optimum set of nodes that results in the largest induced cascade? Practically, one might want to identify the "influencers" in a social network to target to maximize the spread of a viral marketing campaign. While many instances of this problem are

NP-hard, approximation algorithms (along with their corresponding approximation guarantees) have been developed [162, 66].

Relatedly, in outbreak detection (which can generalize some instances of influence maximization), the goal is to pick nodes that detect as many cascades that spread in a network as possible. To this end, greedy approximation algorithms such as CELF have been developed and applied to detecting cascades in blog and water networks [195].

Nonetheless, one study of Twitter suggests that simply identifying well-connected users may not be the most cost-effective mechanism for maximizing influence [23], suggesting that such models should additionally account for both variable acquisition costs, as well as differences in the strength of influence that nodes have on their neighbors.

2.2.4 Cascade Prediction

We now turn our attention to *predictive* models of cascades rather than *generative* models of cascades (as was the case with influence maximization). Rather than describing models of cascades from first principles, the work described here develops an understanding based on observation; Instead of directly answering the question of how cascades grow large by directly architecting them, it instead seeks to identify measurable properties associated with large cascades.

Nonetheless, research on predicting cascades or content popularity in general has seen varying levels of success. For example, predictions have been made at both the aggregate level (e.g., the total volume of new stories [311]) and the individual level (e.g., predicting if a user will retweet a given tweet [241]). In many of these studies, the prediction task is presented as a regression problem or binary classification problem with large bucket sizes. However, such problem formulations are biased towards studying extremely large or rare cascades, leading to criticism that cascades are only predictable after they grow large [299].

In fact, a recent line of work cautions against making such attempts, and even suggests that the eventual scope of a cascade may be inherently unpredictable. In the MusicLab experiment [257], participants used a music ranking service to rate songs. The goal was to see which songs would become most popular when the experiment was repeated many times. What the study found was that it was hard to predict which songs would get popular in each instance of the experiment. Other work has also made similar conclusions with regard to the predictability of large cascades in social networks [23, 220].

In this thesis, we develop methods for predicting cascade growth as a cascade continues to evolve over time, and conduct longitudinal studies to answer questions about recurrence.

2.2.5 Application Domains

Altogether, cascades has been studied in several different domains, with the following being a non-exhaustive list:

Technology Adoption. For instance, the adoption of hybrid seed corn by Iowan farmers [256] and the adoption of a novel drug among doctors [84] as described in the introduction.

Social Movements. In addition to the initial work on threshold models inspired by social movements [129], cascades can explain how the fall of the Berlin Wall started from just a handful of protesters in Leipzig [266].

Policy Adoption. Work has also examined the diffusion of anti-smoking policies in US cities[265], as well as liberal economic practices through the world [269].

Disease Epidemics. Examples include the use of compartmental models (e.g., the SIR model) to explain the spread of foot-and-mouth disease [106] and measles [44].

Other Social Phenomena. Beyond disease, one study suggests how obesity may spread

through a social network [78, 175], while another explores how loneliness can also be contagious [54].

Financial Markets. Herding behavior has been used to explain how investors, firms and analysts make decisions [92, 147].

Social Media. Beyond the numerous examples listed under previous headers, one line of work has involved primarily observational analyses of large data (e.g., on Flickr [61], Digg and Twitter [191] or Facebook [100]). Other work has also begun to look at the spread of rumors in social networks [110] or cascades of invitations to them [11].

2.3 Cascades and the Individual

In the previous section, the primary unit of analysis was the cascade; in this section, we now turn to studying individuals and how they influence and are influenced by other individuals and their environment as a whole. Thus, this section draws primarily on research in sociology and psychology to better understand the entities that drive cascades – human beings.

To begin, we discuss the broad topic of influence and how people look to others to decide what actions to take. Then, to ground this thesis's contributions on how cascades can propagate bad behavior, we present an overview of antisocial behavior, with a specific focus on antisocial behavior in online settings.

2.3.1 Influence

When an informational or behavioral contagion spreads from one person to another, it happens because influence is being exerted on the person who is just adopting the contagion.

First, we discuss two primary forms of influence – conformity, or how people align their actions with the observed behaviors of others or agreed-on social norms, and obedience, or how people are influenced by authority or status to act in specific ways, and in some cases, contrary to their prior beliefs. Next, we discuss several moderators of influence, including group size, gender, and culture. Last, we discuss several large-scale studies of influence that inspire the work in this thesis.

Conformity

Conformity is behavior that is in accordance with a larger group. The experiments described here underscore the strength of this desire to conform to social norms and to behave in accordance with the majority.

One of the most famous studies of conformity are the Asch experiments [14], which are themselves a modification of an earlier study by Muzafer Sheif [263]. In this study, groups of participants viewed a card with a line on it, followed by another with three lines of different lengths, and were then asked to state which of the three lines was the same as the previous. In each group, all but one participant were accomplices of the experimenter, and were instructed to always unanimously pick one of the options, but where this option was sometimes incorrect. What the study found was that people tended to conform one third of the time, with 74% of participants conforming at least once in twelve trials.

This experimental finding has been replicated with acoustic tones [213]. A more recent study that conducted MRI scans of participants suggests that conformance can even change one's perception, rather than it simply being a conscious decision to conform [40]. Similar experiments include the street corner experiment described previously, where crowds stationed at a street corner looking up led passers-by to also stop and look up [214].

Conformity in the form of social proof (or observations of others' behavior) can also produce more negative outcomes. One study demonstrated how participants were

more likely to litter if their environment was already littered, and even more so if they observed another person littering [80].

Obedience

Another type of influence is obedience, where people conform to the desires and expectations of authority or individuals with higher status. The experiments described here show how far people are willing to obey authority, even to the extent of causing others serious injury or distress.

The first of these experiments is the Milgram experiment on obedience to authority figures [215]. In this experiment, participants were told to administer electric shocks of increasing voltage to another participant who, unknown to the participant, was a confederate, and to whom no electric shocks were actually administered and was just acting. What the experiment found was that, at the prodding of the experimenter to keep going, a majority of participants ended up administering the maximum voltage possible, demonstrating the "extreme willingness of adults to go to almost any lengths on the command of an authority".

The second of these experiments is the Stanford Prison experiment [138]. In this experiment, students were randomly recruited as guards and prisoners for a simulated prison. Similar to the Milgram experiment, the guards here obeyed the orders of the experimenter and turned aggressive, enforcing authoritarian measures and subjecting prisoners to psychological torture. The prisoners on the other hand, became very stressed and hostile to the guards.

Moderators of Influence

While indicative, the results above do not suggest that anyone can be influenced to act in ways that contradict their morals or be induced to obey their superiors regardless of what they are told to do. In many of the the experiments above, a majority of

participants were reported to have unwillingly conformed; the Asch experiments also found that there was a group of individuals who always selected the correct answer regardless of what the majority said.

At a finer level, prior work has also identified several factors that moderate influence:

Group size. In general, increasing group size creates additional social pressure for individuals to conform. The Asch line experiments found that conformity peaked with four or five in a group [14]. Similarly, in the street corner experiment, the percent of passers-by that stopped to look up peaked with crowd sizes of about five [214].

Culture. Culture is a strong moderator of conformity and obedience. One study reported participants from Norway, a collectivistic society, exhibited greater conformity than participants from France, a more individualistic culture [213]. A meta-analysis of studies based on the Asch line experiments found that collectivist countries such as Japan showed higher levels of conformity than individualist countries such as the United States [48].

Other Factors. Beyond group size and culture, other studies have also identified several other predictors of influence. For instance, conformity tends to decreases with age [295]. Gender differences in social influence have also been identified, but these effects tend to be confounded by status and the effects of existing social structures [102].

Large-scale Experiments of Social Influence

Against the backdrop of these earlier experiments, a line of work has aimed to measure effects of social influence at large scale.

These include the MusicLab study [257] mentioned in the previous section, which demonstrated how social influence led to different songs becoming popular in different experimental "universes".

Another study on a Reddit-like platform where users could up-vote and down-vote content found that the initial vote on a post can strongly influence the type of votes that it receives in the future [220].

Facebook has also been a testbed for several experiments. One demonstrated how messages encouraging users to vote in the 2010 US congressional elections influenced the real-world voting behavior of not only the users that saw these messages, but their friends as well [47]. Another studying emotional contagion [176] demonstrated evidence that emotional states can be transferred from person to person by manipulating the amount of positive and negative content shown on the Facebook newsfeed, albeit with relatively small effect sizes (e.g., when positive posts were reduced, the percentage of positive words in people's status updates decreased by 0.1%).

All of these studies demonstrate the effects of social proof, where seeing the number of times others had downloaded a song, existing votes from other users, or messages about how others have already voted induced users to act in a similar fashion.

2.3.2 Antisocial Behavior

Thus far, we have described how people can be influenced to act in specific ways, and sometimes in ways that lead to negative behavior. In this subsection, we examine how such behavior manifests online.

Online antisocial behavior, which encompasses a broad range of negative activity (e.g., harassment [101], trolling [140], flaming [22], and (cyber)bullying [272]) is especially important to study because of its prevalence. It comprises a substantial fraction of user activity on many web sites [94, 112], with a large proportion of online users are victims of harassment [101]. This negative behavior increases anger and sadness [200] and threatens social and emotional development in adolescents [244].

In this subsection, after surveying the different types of antisocial behavior in online (and offline) contexts, we then examine their possible causes and mechanisms for their

spread.

Types of Antisocial Behavior

Negative or antisocial behavior as covering a broad spectrum of behavior that involve an individual inflicting some form of unpleasantness on other individuals.

Aggression. Aggression, or hostile behavior towards others, is one commonly studied form of negative behavior, and has been examined along several axes. It can be verbal (i.e., an insult), physical (i.e., a punch), or relational (i.e., manipulating a victim's social standing) [89]. It can also be reactive (i.e., in response to provocation) or proactive (i.e., used to achieve a goal) [88].

Bullying and Cyberbullying. One type of aggression is bullying, or the use of strength or influence to intimidate. Like aggression, bullying can be physical, verbal, or relational. Bullying tends to be caused by feelings of envy or resentment [103], with depression being a significant risk factor in adolescents for both bullies and victims [172].

Some research has also focused on how adolescents are bullied online, typically through the use of mobile phones [273]. One study found that one in four junior high students had been a victim of cyberbullying [199]. Suggesting that cyberbullying is not any less harmful than offline bullying, another has also linked cyberbullying to suicide ideation [146].

Trolling. But perhaps the most widespread and most widely studied of the many forms of online antisocial behavior is trolling, in part because definitions of trolling have varied widely across different work in terms of scope and intent.

Some on the earliest work on trolling started with analyses of Usenet newsgroups and discussion forums, and emphasized intent and deception as essential to what constituted trolling. Trolls were defined as those who "disrupt[ed] a group while remaining

undercover" [98], "lur[ed] others into pointless and time-consuming discussions" [143], or "[took] pleasure in upsetting others".

Examples of such trolls include "Ultimatego", who trolled other users in a wedding-related Usenet newsgroup by being condescending [98], "Macho Joe", a troll that attacked the Melrose Place newsgroup by posting homophobic content [22], and "Kent", a hostile male participant who attacked members of a feminist forum [143]. More recently, one study examined the activity of a small group of trolls on Twitter [284].

At the same time, other work examined the motivations that trolls had in trolling conversations. For instance, boredom (or mischief), attention-seeking, and revenge have been suggested as motivations for trolling [262], while coping mechanisms for trolling (e.g., to ignore the troll) have also been described [22].

More recently, definitions of trolling have substantially broadened, perhaps owing to increased interest, awareness, and broader mainstream use. Trolling has been defined as "engaging in negatively marked online behavior" [140], "not following the rules" [169], or where one "makes trouble" for a discussion forums' stakeholders [42].

Overall, trolling has been characterized through such unusual instances of negative behavior, suggesting that it is a primarily innate characteristic, a narrative that has also been common in the popular media [259]. In this thesis, by taking a broader view of trolling and drawing on work in influence and cascading behavior, we demonstrate how trolling can also be situational.

Cyberstalking. Closely related to cyberbullying is cyberstalking, which is generally differentiated by the repeated, directed harassment of one individual by another [275]. Studies have found how cyberstalking, like stalking, tends to be perpetrated by former intimate partners [6].

Griefing. Yet another line of work has focused on how online harassment happens in online games [75]. A study of Second Life, an online virtual world, found that 95% of its residents were the victims of griefing [86]. Another compared how different affordances of different video games affected opportunities for griefing [298].

Vandalism. In contrast to many of the other types of antisocial behavior, there has been a significant line of work developing computational models for identifying when vandalism happens, for example, on Wikipedia [242, 3] or on OpenStreetMap [229]).

Causes of Antisocial Behavior

Having identified several classes of antisocial behavior, we now turn to possible factors that lead to the appearance and propagation of such behavior.

Biology and Personality. Early work has linked aggression, bullying, and antisocial behavior to individual biological and personality traits. For instance, men tend to be quicker to aggression than females and more likely to express aggression physically [43]. Further, antisocial individuals tend to have lower arousal, and purposefully seek excessive stimulation [243]. Bullying has been linked to genetic influences [27]. One study has even demonstrated how several measures of psychopathic behavior are correlated with trolling [51].

Influence. However, influence and norms may have a role to play in the spread of antisocial behavior. As noted previously, negative social norms such as littering can propagate [80] (even despite their undesirability [305]), with negative emotions and behavior transferable from person to person [116, 176].

Environment. Prior literature on aggression suggests that the physical environment a person is in can also increase aggressive behavior.

Being exposed to pain or discomfort can also increase aggression: violence increases on days with higher temperatures [254], as does car honking [164]. Violent objects such as weapons can also elicit aggression. In an experiment where participants were made angry and left with either a gun or badminton racket, and then told to administer electric shocks to a peer, those left in the presence of a gun delivered more electric shocks [39].

The findings on influence and environment together generally point toward a "Broken Windows" hypothesis, which postulates that untended behavior can lead to the breakdown of a community [306].

Anonymity. Anonymity has also been identified as an important factor that contributes to a greater occurrence of antisocial behavior in online contexts. The relative anonymity of online interactions can result in a disinhibition effect [282], reducing accountability and resulting in people more likely to act antisocially (e.g., on 4chan [41]).

Negativity Bias. Last, negativity in general tends to persist longer because of negativity bias – negative events have a greater impact on individuals than positive events of the same intensity [35]. In other words, negative incidents are more likely to be remembered and likely to have an impact on one's future behavior.

Overall, while innate and situational factors can contribute to the spread of antisocial behavior, whether one type of factor dominates remains an open question. Experimental studies have been conducted in offline settings but their ecological validity in online settings remains unclear. On the other hand, as analysis of online antisocial behavior has been largely qualitative and survey-based, evidence for causal mechanisms remains sparse.

To conclude this section, work on cascades spans a wide range of disciplines and methods, from physics to psychology, and from mathematical models to laboratory experiments. While rich datasets of cascades have since become feasible to analyze, little work has attempted to synthesize observations in these datasets with existing theoretical work. This thesis contributes to this rich literature on cascading behavior by first addressing the predictability of cascades, even over long periods of time. Drawing both on theories of influence and cascading behavior, it contributes to the literature on online antisocial behavior by proposing causal mechanisms that lead to the propagation of such behavior.

Chapter 3

Cascade Growth

On many social networking web sites such as Facebook and Twitter, "re-sharing" or "re-posting" functionality allows users to share others' content with their own friends or followers. For example, on Facebook, users can "reshare" text, photos, videos, and links with one another; on Twitter, users can similarly "retweet" tweets. As content is reshared from user to user, large cascades of reshares can form.

As such, a growing body of research has focused on analyzing and characterizing such online cascades [136], including predicting their future behavior. By better predicting the future behavior of such cascades, one can then better forecast these cascades, for example, to estimate the reach of a social media campaign or to follow the spread of collective action. Nonetheless, a recent, parallel line of work has argued that the future trajectory of a cascade may be inherently unpredictable. For example, the same content can achieve very differing levels of popularity in separate independent settings [257], and small changes in the initial conditions can significantly influence a cascade's future trajectory [220]. And while several features suggestive of popularity have been found, the identification of cascades that eventually become large has typically been unreliable [23].

In this chapter, we develop a framework for addressing cascade prediction problems.

On a large sample of photo reshare cascades on Facebook, we find strong performance in predicting whether a cascade will continue to grow in the future. We find that the relative growth of a cascade becomes more predictable as we observe more of its reshares, that temporal and structural features are key predictors of cascade size, and that initially, breadth, rather than depth in a cascade is a better indicator of larger cascades. This prediction performance is robust in the sense that multiple distinct classes of features all achieve similar performance. We also discover that temporal features are predictive of a cascade's eventual shape. Observing independent cascades of the same content, we find that while these cascades differ greatly in size, we are still able to predict which ends up the largest.

This framework for cascade prediction takes a fundamentally different approach to predicting a cascade's future growth. Instead of asking if a cascade will eventually grow very large, we instead track a cascade over time, and ask, at successive stages of its growth, if it will continue growing. More generally, this framework also allows us to predict other aspects of a cascade's trajectory, such as its future "shape" in terms of its network structure. This chapter, together with the next, explores several open questions around the predictability of these cascades through several large-scale, longitudinal analyses of reshares on Facebook, which involve cascades that, in aggregate, comprise hundreds of millions of users and billions of reshares. Under what conditions do they grow large? Do they recur, or resurface again in the future? And perhaps, most importantly, can the future trajectory of a cascade be predicted?

3.1 Introduction

The sharing of content through social networks has become an important mechanism by which people discover and consume information online. In certain instances, a photo, link, or other piece of information may get *reshared* multiple times: a user shares the content with her set of friends, several of these friends share it with their respective sets of friends, and a *cascade* of resharing can develop, potentially reaching

a large number of people. Such cascades have been identified in settings including blogging [2, 132, 196], e-mail [122, 201], product recommendation [192], and social sites such as Facebook and Twitter [100, 179]. A growing body of research has focused on characterizing cascades in these domains, including their structural properties and their content.

In parallel to these investigations, there has been a recent line of work adding notes of caution to the study of cascades. These cautionary notes fall into two main genres: first, that large cascades are rare [119]; and second, that the eventual scope of a cascade may be an inherently unpredictable property [257, 299]. The first concern — that large cascades are rare — is a widespread property that has been observed quantitatively in many systems where information is shared. The second concern is arguably more striking, but also much harder to verify quantitatively: to what extent is the future trajectory of a cascade predictable; and which features, if any, are most useful for this prediction task?

Part of the challenge in approaching this prediction question is that the most direct ways of formulating it do not fully address the two concerns above. Specifically, if we are presented with a short initial portion of a cascade and asked to estimate its final size, then we are faced with a pathological prediction task, since almost all cascades are small. Alternately, if we radically overrepresent large cascades in our sample, we end up studying an artificial setting that does not resemble how cascades are encountered in practice. A set of recent initial studies have undertaken versions of cascade prediction despite these difficulties [182, 207, 241, 285], but to some extent they are inherent in these problem formulations.

These challenges reinforce the fact that finding a robust way to formulate the problem of cascade prediction remains an open problem. And because it is open, we are missing a way to obtain a deeper, more fundamental understanding of the predictability of cascades. How should we set up the question so that it becomes possible to address these issues directly, and engage more deeply with arguments about whether cascades might, in the end, be inherently unpredictable?

3.1.1 The present work: cascade growth prediction

In this work, we propose a new approach to the prediction of cascades, and show that it leads to strong and robust prediction results. We are motivated by a view of cascades as complex dynamic objects that pass through successive stages as they grow. Rather than thinking of a cascade as something whose final endpoint should be predicted from its initial conditions, we think of it as something that should be *tracked* over time, via a sequence of prediction problems in which we are constantly seeking to estimate the cascade's next stage from its current one.

What would it mean to predict the "next stage" of a cascade? If we think about all cascades that reach size k, there is a distribution of eventual sizes that these cascades will reach. Then the distribution of cascade sizes has a median value $f(k) \geq k$. This number $f(k)$ is thus the "typical" final size for cascades that reached size at least k. Hence, the most basic way to ask about a cascade's next stage of growth, given that it currently has size k, is to ask whether it reaches size $f(k)$.

We therefore propose the following *cascade growth prediction problem*: given a cascade that currently has size k, predict whether it grow beyond the median size $f(k)$. (As we show later, the prediction problem is equivalent to asking: given a cascade of size k, will the cascade double its size and reach at least $2k$ nodes?) This implicitly defines a family of prediction problems, one for each k. We can thus ask how cascade predictability behaves as we sweep over larger and larger values of k. (There are natural variants and generalizations in which we ask about reaching target sizes other than the median $f(k)$.) This problem formulation has a number of strong advantages over standard ways of trying to define cascade prediction. First, it leads to a prediction problem in which the classes are balanced, rather than highly unbalanced. Second, it allows us to ask for the first time how the predictability of a cascade varies over the range of its growth from small to large. Finally, it more closely approximates the real tasks that need to be solved in applications for managing viral content, where many evolving cascades are being monitored, and the question is which are likely to grow significantly as time moves forward.

For studying cascade growth prediction, it is important to work with a system in which the sharing and resharing of information is widespread, the complete trajectories of many cascades—both large and small—are observable, and the same piece of content shared separately by many people, so that we can begin to control for variation in content. For this purpose, we use a month of complete photo-resharing data from Facebook, which provides a rich ecosystem of shared content exhibiting all of these properties.

In this setting, we focus on several categories of questions:

(a) How high an accuracy can we achieve for cascade growth prediction? If we cannot improve on baseline guessing, then this would be evidence for the inherent unpredictably of cascades. But if we can significantly improve on this baseline, then there is a basis for non-trivial prediction. In the latter case, it also becomes important to understand the features that make prediction possible.

(b) Is growth prediction more tractable on small cascades or large ones? In other words, does the future behavior of a cascade become more or less predictable as the cascade unfolds?

(c) Beyond just the growth of a cascade, can we predict its "shape" — that is, its network structure?

3.1.2 Summary of results

Given the challenges in predicting cascades, we find surprisingly strong performance for the growth prediction problem. Moreover, the performance is robust in the sense that multiple distinct classes of features, including those based on time, graph structure, and properties of the individuals resharing, can achieve accuracies well above the baseline. Cascades whose initial reshares come quickly are more likely to grow significantly; and from a structural point of view, breadth rather than depth in the resharing tree is a better predictor of significant growth.

We investigate the performance of growth prediction as a function of the size of the cascade so far — when we want to predict the growth of a cascade of size k, how does our accuracy depend on k? It is not a priori clear whether accuracy should increase or decrease as a function of k, since for any value of k the challenge is to determine what the cascade will do in the future. Seeing more of the cascade (larger k) does not make the problem easier, as it also involves predicting "farther" into the future (i.e., whether the cascade will reach size at least $2k$). We find that accuracy increases with k, so that it is possible to achieve better performance on large cascades than small ones. The features that are most significant for prediction change with k as well, with properties of the content and the original author becoming less important, and temporal features remaining relatively stable.

We also consider a related question: how much of a cascade do we need to see in order to obtain good performance? Specifically, suppose we want to predict the growth of a cascade of size at least R, but we are only able to see the first $k < R$ nodes in the cascade. How does prediction performance depend on k, and in particular, is there a "sweet spot" where a relatively small value of k gives most of the performance benefits? We find in fact that there is no sweet spot: performance essentially climbs linearly in k, all the way up to $k = R$. Perhaps surprisingly, more information about the cascade continues to be useful even up to the full snapshot of size R.

In addition to growth, we also study how well we can predict the eventual "shape" of the cascade, using metrics for evaluating tree structures as a numerical measure of the shape. We obtain performance significantly above baseline for this task as well; and perhaps surprisingly, multiple classes of features including temporal ones perform well for this task, despite the fact that the quantity being predicted is a purely structural one.

One of the compelling arguments that originally brought the issue of inherent unpredictability onto the research agenda was a striking experiment by Salganik, Dodds, and Watts, in which they showed that the same piece of content could achieve very

different levels of popularity in separate independent settings [257]. Given the richness of our data, we can study a version of this issue here in which we can control for the content being shared by analyzing many cascades all arising from the sharing of the same photo. As in the experiment of Salganik et al., we find that independent resharings of the same photo can generate cascades of very different sizes. But we also show that this observation can be compatible with prediction: after observing small initial portions of these distinct cascades for the same photo, we are able to predict with strong performance which of the cascades will end up being the largest. In other words, our data shows wide variation in cascades for the same content, but also predictability despite this variation.

Overall, our goal is to set up a framework in which prediction questions for cascades can be carefully analyzed, and our results indicate that there is in fact a rich set of questions here, pointing to important distinctions between different types of features characterizing cascades, and between the essential properties of large and small cascades.

3.2 Related Work

Significant work has analyzed and cataloged properties of empirically observed information cascades, while others have considered theoretical models of cascade formation in networks; Most relevant to our analysis is work that focuses on predicting the future popularity of a given piece of content. These studies have proposed rich sets of features for prediction, which we discuss in section 3.3.3.

Much prior work aims to predict the *volume* of *aggregate* activity — the total number of up-votes on Digg stories [285], total hourly volume of news phrases [311], or total daily hashtag use [207]. At the other end of the spectrum, research has focused on *individual* user-level prediction tasks: whether a user will retweet a specific tweet [241] or share a specific URL [111]. Rather than attempt to predict aggregate popularity or individual behavior in the next time step, we instead look at whether an information

cascade grows over the median size (or doubles in size, as we later show).

Research on communities defined by user interests [18] or hashtag content [249] has also looked at a notion of growth, predicting whether a group will increase in size by a given amount. Nevertheless, these focused on groups of already non-trivial size, and their growth predicted without an explicit internal cascade topology, and without tracking predictability over different size classes.

Several papers focus on predictions after having observed a cascade for a given fixed time frame [182, 207, 290]. In contrast, rather than studying specific time slices, we continuously observe the cascade over its entire lifetime and attempt to understand how predictive performance varies as the cascade develops. Moreover, our methodology does not penalize slowly but persistently growing cascades. Thus, we predict the size and the structure after having observed a certain number of initial reshares.

Many studies consider the cascade prediction task as a regression problem [24, 182, 285, 290] or a binary classification problem with large bucket sizes [150, 154, 182]. The danger with these approaches is that they are biased towards studying extremely large but also extremely rare cascades, bypassing the whole issue about the general predictability of cascades. For example, research has specifically focused on content and users that create extremely large cascades, such as popular hashtags [148, 314] and very popular users [100, 134], which has led to criticism that cascades may only be predictable after they have already grown large [299]. While it is useful to understand the dynamics of extremely popular content, such content is also very rare. Thus, we rather seek to understand predictability along cascade's entire lifetime. We consider cascades that have as few as five reshares, and introduce a classification task which is not skewed towards very large cascades.

3.3 Predicting Cascade Growth

To examine the cascade growth prediction problem, we first define and motivate our experimental setup and the feature sets used, then report our prediction results with respect to different k.

3.3.1 Experimental setup

Mechanics of information passing on Facebook. We focus on content consisting of posts the author has designated as public, meaning that anyone on Facebook is eligible to view it, and we further restrict our attention to content in the form of photos, which comprise the majority of reshare cascades on Facebook [100]. Such posts are then distributed by Facebook's News Feed, typically at first to users who are either friends of the poster or who subscribe to their content, e.g. as followers. Each post is accompanied by a "share" link that allows friends and followers to "reshare" the post with her own friends and followers, thus expanding the set of users exposed to the content. This explicit sharing mechanism creates information cascades, starting with the root node (user or page) that originally created the content, and consisting of all subsequent reshares of that content.

Figure 3.1 illustrates the process with an example: a node v_0 posts a public photo, seen by v_0's friends and followers in their News Feeds. Friends v_1 and v_3 then share the photo with their own friends. This way the photo propagates over the edges of the Facebook network and creates an information cascade. We represent the cascade graph as $\hat{G}$, and the induced subgraph of all photo sharers, including all friendship or follow links between them as G'. Notice that some users (ex. v_5) are exposed via multiple sources (v_0, v_1, v_3, v_4).

An important issue for our understanding of reshare cascades is the following distinction: content can be produced by *users* — individual Facebook accounts whose primary audience consists of friends and any subscribers the individual has — and it

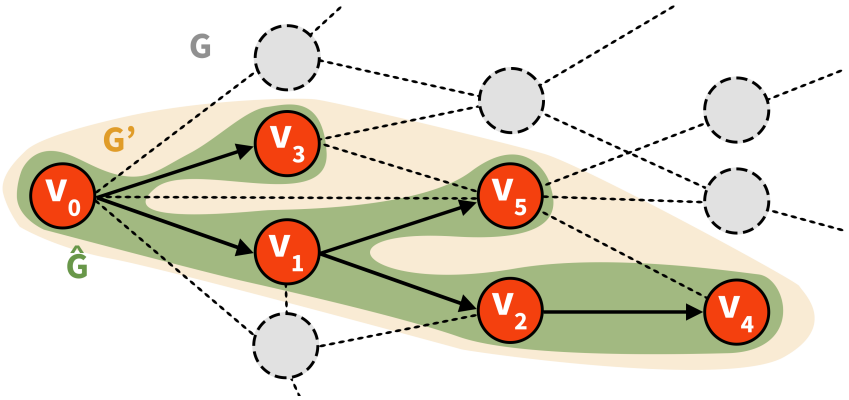


Figure 3.1: An information cascade represented by solid edges on a graph G, starting at v_0 ($\hat{G}$). Dashed lines indicate friendship edges; the edges between resharers make up the friend subgraph G'.

can also be produced by *pages*, which correspond to the Facebook accounts of companies, brands, celebrities, and other highly visible public entities. In the common parlance around cascades, reshared content originally produced by a user is often informally viewed as more "organic," developing a following in a more bottom-up way. In contrast, reshared content from pages is seen as more top-down, and generally broadcast via News Feed to a larger set of initial followers. A natural question, and a theme that will run through several analyses in the work, is to understand if these distinctions carry over to the properties we study here: do user-initiated cascades differ in their predictability and their underlying structure from page-initiated cascades?

Dataset description. We sampled our anonymized dataset from photos uploaded to Facebook in June 2013 and observed any reshares occurring within 28 days of initial upload. The dataset only includes photos posted publicly (viewable by anyone), and not deleted during the observation period. Further, we exclude photos with fewer than five reshares as is required by the prediction tasks described below. We constructed diffusion trees first by taking the explicit cascade, e.g. C clicking "share" on B's "share" of A's photo forms the cascade $A \to B \to C$. However, it is possible that user C clicked on user B's share, and then directly reshared from A. Since we want to know how the information actually flowed in the network, we reconstruct the path $A \to B \to C$ based on click, impression, and friend/follower data [100].

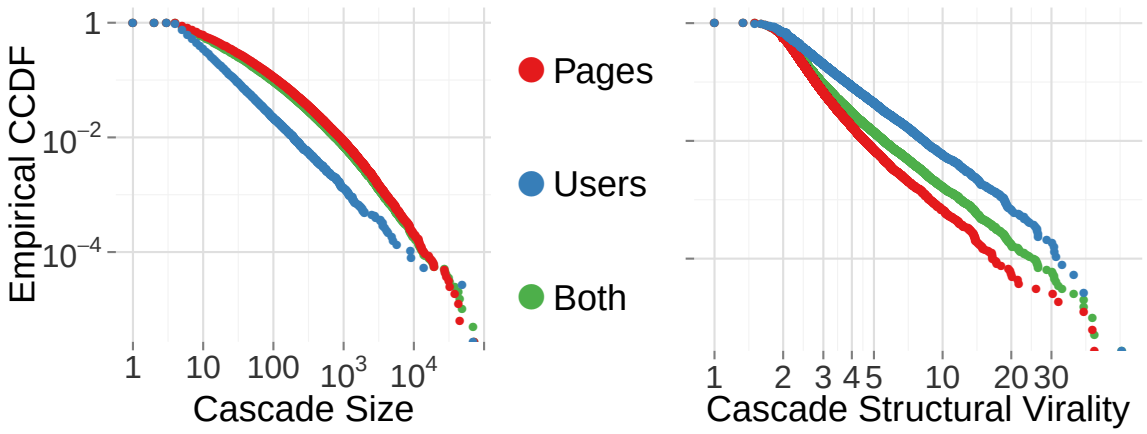
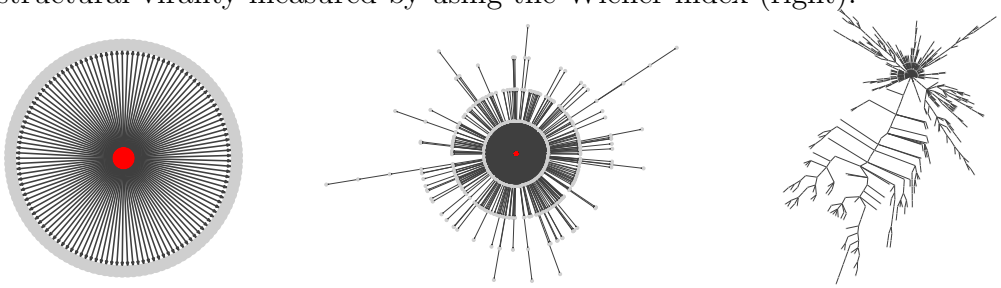


Figure 3.2: The complementary cumulative distribution (CCDF) of cascade size (left) and structural virality measured by using the Wiener index (right).



(a) $d = 1.98$ (b) $d = 2.47$ (c) $d = 14.4$

Figure 3.3: Cascades with a low Wiener index d resemble star graphs, while those with a high index appear more viral (the root is red).

Figure 3.2 begins to show how photos uploaded by pages generate cascades that differ from those uploaded by users. In our dataset, 81% of cascades are initiated by pages. Figure 3.2 shows the cascade size distribution for pages, users, and the two combined. Page cascades are typically larger than user cascades, e.g., 11% of page cascades reach at least 100 reshares, while only 2% of user cascades do, though both follow heavy tailed distributions. Fitting power-law curves to their tails, we observe power-law exponents of α equal to 2.2, 2.1, and 2.1 for user, page, and both, respectively ($x_{\min} = 10$, 2000, 2000).

In addition to cascade growth, we quantify the shape of a cascade using the Wiener

index, defined as the average distance between all pairs of nodes in a cascade. Recent work has proposed the Wiener index as a measure of the structural virality of a cascade [118]. Figure 3.3 shows examples of cascades with varying Wiener index values. Intuitively, a cascade with low structural virality has most of its distribution following from a small number of hub nodes, while a cascade with high virality will have many long paths. Figure 3.2 shows the distribution of cascade virality (as measured by Wiener index) in our dataset, which, as we saw with cascade size, follows a heavy-tailed distribution. While user cascades are typically smaller than page cascades in our dataset, they tend to have greater structural virality, supporting the intuition that the structure of user-initiated cascades is richer and deeper than that of page-initiated cascades.

3.3.2 Defining the cascade growth prediction problem

Our aim in this work is to study how well cascades can be predicted. Moreover, we are interested in understanding how various aspects of the prediction task affect the predictive performance.

There are several formulations of the task. If we were to define the task as a regression problem, predictions may be skewed towards large cascades, as cascade size follows a heavy-tailed distribution (Figure 3.2(right)). Similarly, if we define it as a classification problem of predicting whether a cascade reaches a specific size, we may end up with unbalanced classes, and an overrepresentation of large cascades. Also, if we simply observed a small initial portion of a cascade, and predict its future size, the problem is pathological as almost all cascades are small. And, if we only varied the initial period of observation, the task of predicting whether a cascade reaches a certain size gets easier as we observe more of it.

To remedy these issues, we define a classification task that does not suffer from these deficiencies. We consider a binary classification problem where we observe the first k reshares of a cascade and predict whether the eventual size of a cascade reaches the

median size of all the cascades with at least k reshares, $f(k)$. This allows us to study how cascade predictability varies with k. As exactly half the cascades reach a size greater than the median by definition, random guessing achieves accuracy of 50%.

Interestingly, the question of whether the cascade will reach $f(k)$ is equivalent to that of whether a cascade will double in size. This follows directly from the fact that cascade size distribution follows a power-law with exponent $\alpha \approx 2$. Consider a power-law distribution on the interval (x_{min}, ∞) with a power-law exponent $\alpha \approx 2$. Then the median $f(x)$ of this distribution is $2 \cdot x_{min}$, as demonstrated by the following calculation:

$$\int_{x_{\min}}^{f(x)} \frac{\alpha - 1}{x_{\min}} \left(\frac{x}{x_{\min}} \right)^{-\alpha} dx = \frac{1}{2} \Rightarrow f(x) = 2^{\frac{1}{\alpha - 1}} x_{\min} = 2x_{\min}$$

As we examine cascades of size greater than $k = x_{min}$, the median size of these cascades is thus $2 \cdot k$ from this derivation. In each of our prediction tasks, we observe that this is indeed true.

Methods used for learning. Our general methodology for the cascade prediction problem will be to represent a cascade by a set of features and then use machine learning classifiers to predict its future size. We used a variety of learning methods, including linear regression, naive Bayes, SVM, decision trees and random forests. However, we primarily report performance of the logistic regression classifier for ease of comparison. In many cases, the performance of most classifiers was similar, although non-linear classifiers such as random forests usually performed slightly better than linear classifiers such as logistic regression. In all cases, we performed 10-fold cross validation and report the classification accuracy, F1 score, and area under the ROC curve (AUC).

3.3.3 Factors driving cascade growth

We proceed by describing factors that contribute to the growth and spreading of cascades. We group these factors into five classes: properties of the content that is spreading, features of the original poster, features of the resharer, structural features of the cascade, and temporal characteristics of the cascade. Table 3.1 contains a detailed list of features.

Content features. The first natural factor contributing to the ability of the cascade to spread is the content itself [37]. On Twitter, tweet content and in particular, hashtags, are used to generate content features [207, 290], and identify topics affecting retweet likelihood [241]. LDA topic models have also been incorporated into these prediction tasks [150], and human raters employed to infer the interestingness of content [23, 241]. In our work, we relied on a linear SVM model, trained using image GIST descriptors and color histogram features, to assign likelihood scores of a photo being a closeup shot, taken indoors or outdoors, synthetically generated (e.g., screenshots or pure text vs. photographs), or contained food, a landmark, person, nature, water, or overlaid text (e.g., a meme). We also analyzed words in the caption accompanying an image for positive sentiment, negative sentiment, and sociality [154, 240].

Nevertheless, while content features affected the performance of structural and temporal features, we find that they are weak predictors of how widely disseminated a piece of content would become.

Original poster/resharer features. Some prior work focused on features of the root note in a cascade to predicting the cascade's evolution, finding that content from highly-connected individuals reaches larger audiences, and thus spreads further. Users with large follower counts on Twitter generated the largest retweet cascades [23]. Separately, features of an author of a tweet were shown to be more important than features of the tweet itself [241]. In many Twitter studies predicting cascade size or popularity, a user's number of followers ranks among the top, if not the most, important predictor of popularity [23, 207].

Other features of the root node have also been studied, such as the number of prior retweets of a user's posts [23, 150], and how many Twitter lists a user was included in [241]. The number of @-mentions of a Twitter user was used to predict whether, and how soon a tweet would be retweeted, how many users would directly retweet, and the depth a cascade would reach [314]. Still, [60] found that various measures of a user's popularity are not very correlated with his or her influence.

We capture the intuition behind these factors by defining demographic as well as network features of the original poster as well as the features of the users who reshared the content so far. We use Facebook's distinction of users (individuals) and pages (entities representing an interest) to further distinguish different origin types, in addition to the influence features mentioned above.

Structural features of the cascade. Networks provide the substrate through which information spreads, and thus their structure influences the path and reach of the cascade. As illustrated in Figure 3.1, we generate features from both the graph of the first k reshares ($\hat{G}$), as well as the induced friend subgraph of the first k resharers (G'). Whereas the reshare graph $\hat{G}$ describes the actual spread of a cascade, the friend subgraph G' provides information about the social ties between these initial resharers. The social graph G allows us to compute the potential reach of these reshares.

Previous work considered the network structure of the underlying graph in inferring the virality of content [300], with highly viral items spreading across communities. We use the density of the initial reshare cascade ($subgraph'_k$) and the proximity to the root node ($orig\_connections_k$, did_leave) as proxies for whether an item is spreading primarily within a community or across many. One can also look outside the network between resharers, and count the number of users reachable via all friendship and follow edges of the first k users ($border\_nodes_k$). This relates to total number of exposed users, and has been demonstrated to be an important feature in predicting Twitter hashtag popularity [207].

As we can trace information flow on Facebook exactly, we need not worry about independent entry points influencing a cascade [24, 226]; external influence instead

allows us to investigate multiple independent cascades arising from the same content (see Section 3.4.1).

Temporal features. Properties related to the "speed" of the cascade (e.g., $time_k$) were shown to be the most important features in predicting thread length on Facebook [19], and are a primary mechanism in predicting online content popularity [285]. Moreover, as the speed of diffusion changes over time, this may have a strong effect on the ability of the cascade to continue spreading through the network [314].

We characterize a number of temporal properties of cascade diffusion (see Table 3.1). In particular, we measure the change in the speed of reshares ($time''_{1..k}$), compare the differences between the speed in the first and second half of the measurement period ($time'_{1..k/2}$, $time'_{k/2..k}$), and quantify the number of users who were exposed to the cascade per time unit ($views'_{1..k-1,\ k}$).

Content Features	
$score_{food/nature/...}$	The probability of the photo having a specific feature (food, overlaid text, landmark, nature, etc.)
is_en	Whether the photo was posted by an English-speaking user or page
$has\_caption$	Whether the photo was posted with a caption
$liwc_{pos/neg/soc}$	Proportion of words in the caption that expressed positive or negative emotion, or sociality, if English
Root (Original Poster) Features	
$views_{0,\ k}$	Number of users who saw the original photo until the kth reshare was posted
$orig\_is\_page$	Whether the original poster is a page
$outdeg(v_0)$	Friend, subscriber or fan count of the original poster
age_0	Age of the original poster, if a user
$gender_0$	Gender of the original poster, if a user
fb_age_0	Time since the original poster registered on Facebook, if a user

$activity_0$	Average number of days the original poster was active in the past month, if a user
Resharer Features	
$views_{1..k-1,\ k}$	Number of users who saw the first $k-1$ reshares until the kth reshare was posted
$pages_k$	Number of pages responsible for the first k reshares, including the root, or $\sum_{i=0}^{k} \mathbb{1}\{v_i \text{ is a page}\}$
$friends_k^{avg/90p}$	Average or 90th percentile friend count of the first k resharers, or $\frac{1}{k}\sum_{i=1}^{k} outdeg_{friends}(v_i)\mathbb{1}\{v_i \text{ is a user}\}$
$fans_k^{avg/90p}$	Average or 90th percentile fan count of the first k resharers, or $\frac{1}{k}\sum_{i=1}^{k} outdeg(v_i)\mathbb{1}\{v_i \text{ is a page}\}$
$subscribers_k^{avg/90p}$	Average or 90th percentile subscriber count of the first k resharers, or $\frac{1}{k}\sum_{i=1}^{k} outdeg_{subscriber}(v_i)\mathbb{1}\{v_i \text{ is a user}\}$
$fb\_ages_k^{avg/90p}$	Average or 90th percentile time since the first k resharers registered on Facebook, or $\frac{1}{k}\sum_{i=1}^{k} fb\_age_i$
$activities_k^{avg/90p}$	Average number of days the first k resharers were active in July, or $\frac{1}{k}\sum_{i=1}^{k} activity_i$
$ages_k^{avg/90p}$	Average age of the first k resharers, or $\frac{1}{k}\sum_{i=1}^{k} age_i$
$female_k$	Number of female users among the first k resharers, or $\sum_{i=1}^{k} \mathbb{1}\{gender_i \text{ is female}\}$
Structural Features	
$outdeg(v_i)$	Connection count (sum of friend, subscriber and fan counts) of the ith resharer (or out-degree of v_i on $G = (V, E)$)
$outdeg(v_i')$	Out-degree of the ith reshare on the induced subgraph $G' = (V', E')$ of the first k resharers and the root
$outdeg(\hat{v}_i)$	Out-degree of the ith reshare on the reshare graph $\hat{G} = (\hat{V}, \hat{E})$ of the first k reshares
$orig\_connections_k$	Number of first k resharers who are friends with, or fans of the root, or $\lvert\{v_i \mid (v_0, v_i) \in E,\ 1 \le i \le k\}\rvert$

$border\_nodes_k$	Total number of users or pages reachable from the first k resharers and the root, or $\lvert\{v_i \mid (v_i, v_j) \in E,\ 0 \le i, j \le k\}\rvert$
$border\_edges_k$	Total number of first-degree connections of the first k resharers and the root, or $\lvert\{(v_i, v_j) \mid (v_i, v_j) \in E,\ 0 \le i, j \le k\}\rvert$
$subgraph'_k$	Number of edges on the induced subgraph of the first k resharers and the root, or $\lvert\{(v_i, v_j) \mid (v_i, v_j) \in E',\ 0 \le i, j \le k\}\rvert$
$depth'_k$	Change in tree depth of the first k reshares, or $\min_\beta \sum_{i=1}^{k} (depth_i - \beta i)^2$
$depths_k^{avg/90p}$	Average or 90th percentile tree depth of the first k reshares, or $\frac{1}{k}\sum_{i=1}^{k} depth_i$
did_leave	Whether any of the first k reshares are not first-degree connections of the root
Temporal Features	
$time_i$	Time elapsed between the original post and the ith reshare
$time'_{1..k/2}$	Average time between reshares, for the first $k/2$ reshares, or $\frac{1}{k/2-1}\sum_{i=1}^{k/2-1}(time_{i+1} - time_i)$
$time'_{k/2..k}$	Average time between reshares, for the last $k/2$ reshares, or $\frac{1}{k/2-1}\sum_{i=k/2}^{k-1}(time_{i+1} - time_i)$
$time''_{1..k}$	Change in the time between reshares of the first k reshares, or $\min_\beta \sum_{i=1}^{k-1}(time_{i+1} - time_i) - \beta i)^2$
$views'_{0,k}$	Number of users who saw the original photo, until the kth reshare was posted, per unit time, or $\frac{views_{0,\ k}}{time_k}$
$views'_{1..k-1,\ k}$	Number of users who saw the first $k-1$ reshares, until the kth reshare was posted, per unit time, or $\frac{views_{1..k-1,k}}{time_k}$

Table 3.1: List of features used for learning. We compute these features given the cascade until the kth reshare.

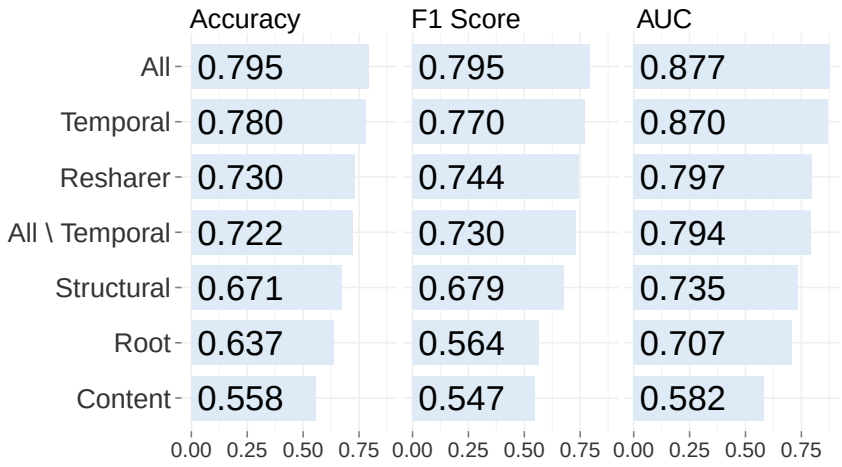


Figure 3.4: Using logistic regression, we are able to predict with near 80% accuracy whether the size of a cascade will reach the median (10) after observing the first k =5 reshares.

3.3.4 Predicting cascade growth

To illustrate the general performance of the features described in the previous section we consider a simple prediction task, where we observe the first 5 reshares of the cascade and want to predict whether it will reach the median cascade size (or equivalently, whether it will double and be reshared at least 10 times). For the experiment we use a set of N_c = 150,572 photos, where each photo was shared at least 5 times. The total number of reshares of these photos was N_r = 9,233,300.

Figure 3.4 shows logistic regression performance using all features from Table 3.1. For this task, random guessing would obtain a performance of 0.5, while our method achieves surprisingly strong performance: classification accuracy of 0.795 and AUC of 0.877. If we relax the task and instead of predicting above vs. below median size, we predict top vs. bottom quartile (top 25% vs. bottom 25%) the accuracy rises even further to 0.926, and the AUC to 0.976.

Overall, while each feature set is individually significantly better than predicting at random, it is the set of temporal features that outperforms all other individual feature sets, obtaining performance scores within 0.025 of those obtained when using

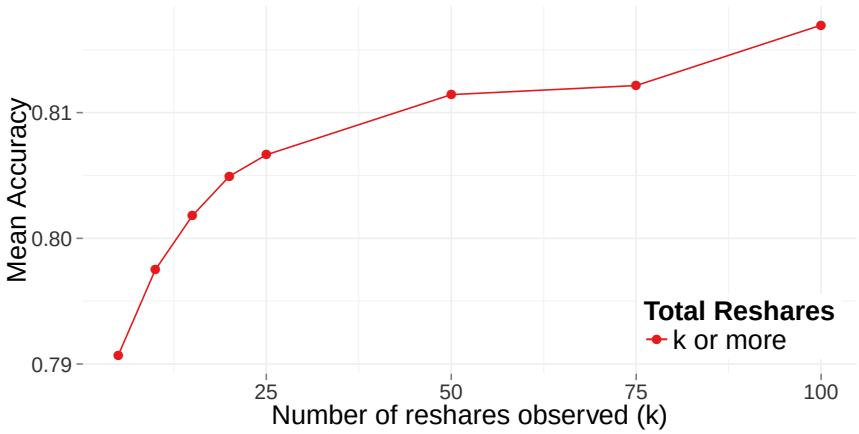


Figure 3.5: If we observe the first k reshares of a cascade, and want to predict whether the cascade will double in size, our prediction improves as we observe more of it.

all features. To understand if we could do well without temporal features, we trained a classifier which excluded them and were still able to obtain reasonable performance even without these features. This is especially useful when one knows through *whom* information was passed, but not *when* it was passed. The lack of reliance on any individual set of features demonstrates that the predictions are robust.

Studied individually, we also find that temporal features generally performed best, followed by structural features. The reshare rate in the second half ($time'_{k/2..k}$) was most predictive, attaining accuracy of 0.73. This was followed by the rate of user views of the original photo, $views'_{0,k}$, and the time elapsed between the original post and fifth reshare, $time_5$ (both 0.72). In fact, $time_{k+1}$ is always more accurate than $time_k$. The most accurate structural features were *did_leave* and $outdeg(v_0)$ (both 0.65). We examine individual feature importance in more detail later.

3.3.5 Predictability and the observation window of size k

It is also natural to ask whether cascades get more or less predictable as we observe more of the initial growth of a cascade. One may think that observing more of the cascade may allow us to extrapolate its future growth better; on the other hand,

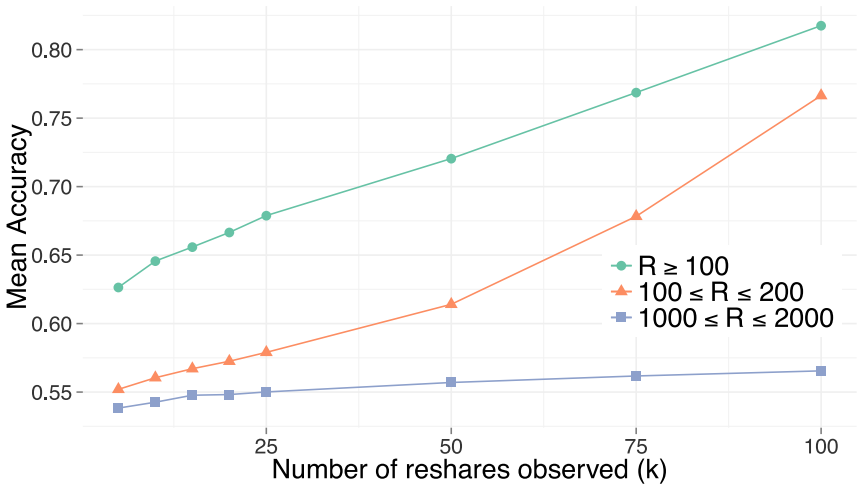


Figure 3.6: Knowing that a cascade obtains at least R reshares, prediction performance increases linearly with k, $k \leq R$. However, differentiating among cascades with large R also becomes more difficult.

additional observed reshares may also introduce noise and uncertainty in the future growth of the cascade. Note that the task does not get easier as we observe more of the cascade, as we are predicting whether the cascade will reach size $2k$ (or equivalently, the median) given that we have seen k reshares so far.

Figure 3.5 shows that the predictive performance of whether a cascade doubles in size increases as a function of the number of observed reshares k. In other words, it is easier to predict whether a cascade that has reached 25 reshares will get another 25, than to predict whether a cascade that has reached 5 reshares will obtain an additional 5. Thus, the prediction accuracy for larger cascades is above the already high accuracy for smaller values of k. The change in the F1 score and AUC also follow a very similar trend.

Overall, these results demonstrate that observing more of the cascade, while also predicting "farther" into the future, is easier than observing a cascade early in its life and predicting what it will do next (i.e., $k = 5$ vs. $k = 25$).

Fixing the minimum cascade size R. In the previous version of the task, cascades are required only to have at least k reshares. Thus, the set of cascades changes with k.

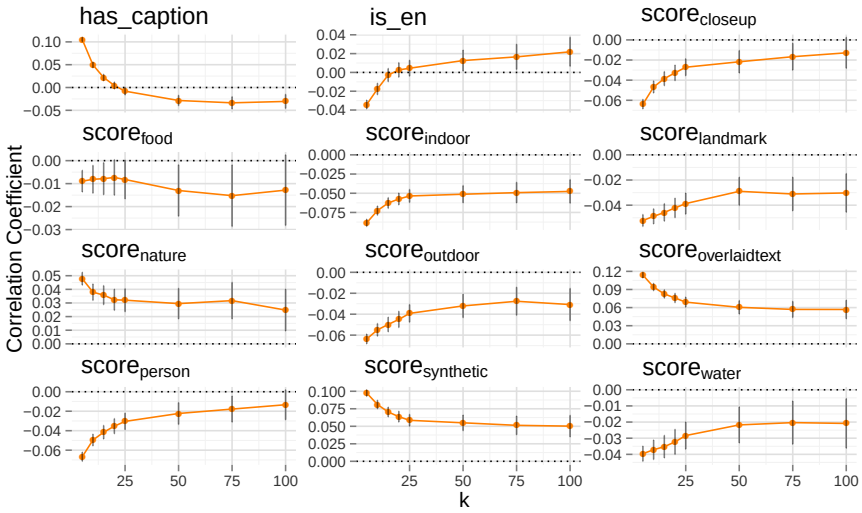


Figure 3.7: The importance of each feature varies as we observe more of a cascade, as shown by the change in correlation coefficients. This figure depicts these changes for content features.

Here, we examine a variation of this task, where we compose a dataset of cascades that have at least R reshares. We observe the first k ($k \leq R$) reshares of the cascade and aim to predict whether the cascade will grow over the median size (over all cascades of size $\geq R$). As we increase k, the task gets easier as we observe more of the cascade and the predicted quantity does not change.

With the task, we find that performance increases linearly with k up to R, or that there is no "sweet spot" or region of diminishing returns ($p < 0.05$ using a Harvey-Collier test). For example, the top-most line in Figure 3.6 shows that when each observed cascade has obtained 100 or more reshares, performance increases linearly as more of the cascade is observed. This demonstrates that more information is always better: the greater the number of observed reshares, the better the prediction.

However, Figure 3.6 also shows that larger cascades are less predictable than smaller cascades. For example, predicting whether cascades with 1,000 to 2,000 reshares grow large is significantly more difficult than predicting cascades of 100 to 200 reshares. This shows that once one knows that a cascade will grow to be large, knowing the characteristics of the very beginning of its spread is less useful for prediction.

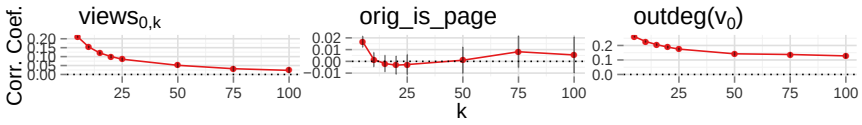


Figure 3.8: This figure depicts these changes in importance for root features.

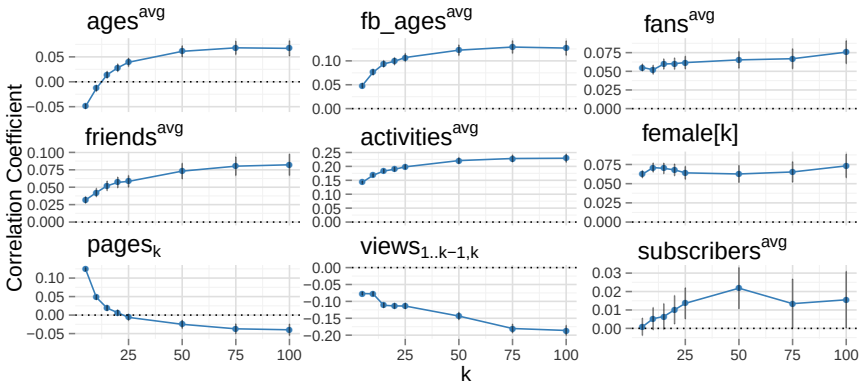


Figure 3.9: This figure depicts these changes in importance for resharer features.

3.3.6 Changes in feature importance

We now examine how feature importance changes as more and more of the cascade is observed. In this experiment, we compute the value of the feature after observing first k reshares and measure the correlation coefficient of the feature value with the log-transformed number of reshares (or cascade size).

Figures 3.7 to 3.11 show the results for the five feature types. We summarize the results by the following observations:

- *Correlations of averages increase with the number of observations.* As we obtain more examples, naturally averages get less noisy, and more predictive (e.g., $ages^{avg}$ and $friends^{avg}$).

- *The original post gets less important with increasing k.* After observing 100 reshares, it becomes less important that the original post was made by a page ($orig\_is\_page$), or that the original poster had many connections to other users ($outdeg(v_0)$).

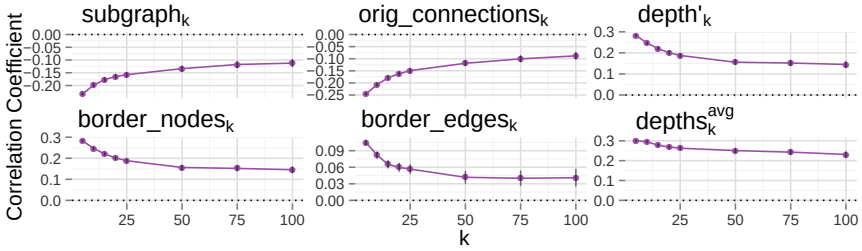


Figure 3.10: This figure depicts these changes in importance for structural features.

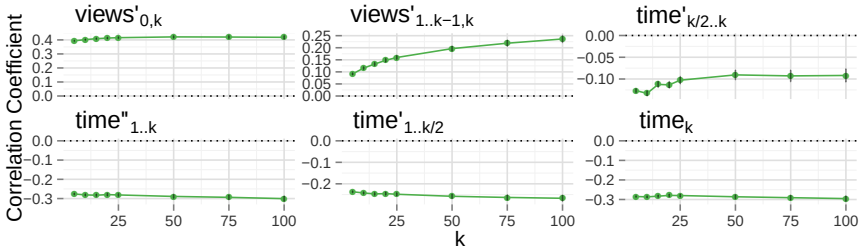


Figure 3.11: This figure depicts these changes in importance for temporal features.

- *Similarly, the actual content being reshared gets less important with increasing k.* Almost all content features tend to zero as k increases, except for *has_caption* and *is_en*. This can be explained by the fact that cascades of photos with captions have a unimodal distribution, and cascades started by English speakers have a bimodal distribution. Thus, these features become correlated in opposite directions.

- *Successful cascades get many views in a short amount of time, and achieve high conversion rates.* The number of users who have viewed reshares of a cascade is more negatively correlated with increasing k ($views_{1..k-1,k}$), suggesting that requiring "fewer tries" to achieve a given number of reshares is a positive indicator of its future success. On the other hand, while requiring fewer views is good, rapid exposure, or reaching many users within a short amount of time is also a positive predictor ($views'_{1..k-1,k}$).

- *Structural connectedness is important, but gets less important over time.* Nevertheless, reshare depth remains highly correlated: the deeper a cascade goes, the

more likely it is to be long-lasting, as even users "far away" from the original poster still find the content interesting.

- *The importance of timing features remains relatively stable.* While highly correlated, timing features remain remarkably stable in importance as k increases.

We note individual features' logistic regression coefficients empirically follow similar shapes, but have the downside of having interactions with one another. Using either the slope of the best-fit line of the cascade size against the normalized feature value, or individual feature performance also reveals similar trends. Further LIWC text content features (positive, negative, and social categories) consistently performed poorly, attaining performance no better than chance, with accuracy between 0.49 and 0.52.

3.3.7 Predicting Cascade Structure

Similar to predicting cascade size we can also attempt to predict the *structure* of the cascade. We now turn to examining how structural features of the cascade determine its evolution and spread.

3.3.8 User-started and page-started cascades

Earlier we discussed the notion of *structural virality* as a measure of how much the structure of a cascade is dominated by a few hub nodes, and we saw that user-initiated cascades have significantly higher structural virality than page-initiated cascades, reflecting their richer graph structure. It is natural to ask how these distinctions vary with the size of the cascade — are large user-initiated cascades more similar to page-initiated ones, e.g. are they driven by popular hub nodes?

Figure 3.12 shows that the opposite is the case — user and page-initiated cascades remain structurally distinct, with this distinction even increasing with cascade size.

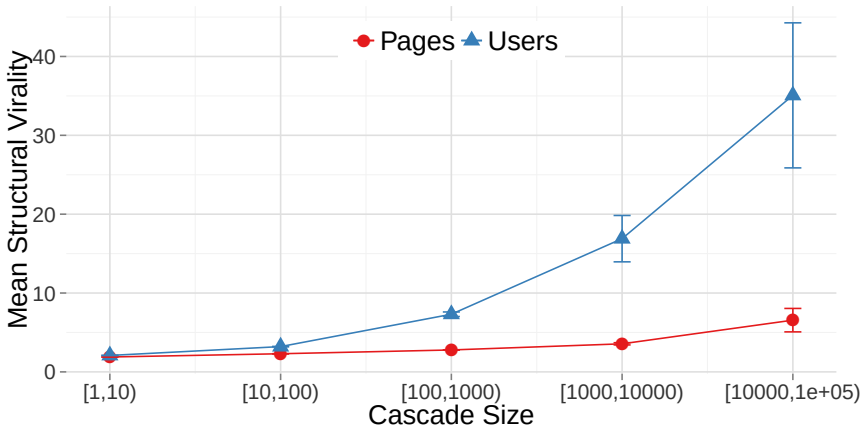


Figure 3.12: The mean structural virality (Wiener index) increases with cascade size, but is significantly higher for user cascades.

Moreover, this difference continues to hold even when controlling for the number of first-degree reshares (directly from the root), suggesting a certain robustness to their richer structure. Because of these structural differences, we handle user and page cascades separately in the analyses that follow.

These distinctions may also help explain a large difference in the predictability of user-initiated vs. page-initiated cascades. We observe that for page cascades accuracy exceeds 80%, while that for user cascades is slightly under 70%. (These results also hold for the F1 score and AUC, with a difference of about 0.1.) The fact that much more of the structure of a page-initiated cascade is typically carried by a small number of hub nodes may suggest why the prediction task is more tractable in this case.

3.3.9 The initial structure of a cascade influences its eventual size

To understand how structure bears on the future growth of the cascade, we examine how the configuration of the first three reshares (and the root) correlates with the cascade size. In particular, we measure the proportion of cascades starting from each configuration that reach the median size. We do this separately for two different

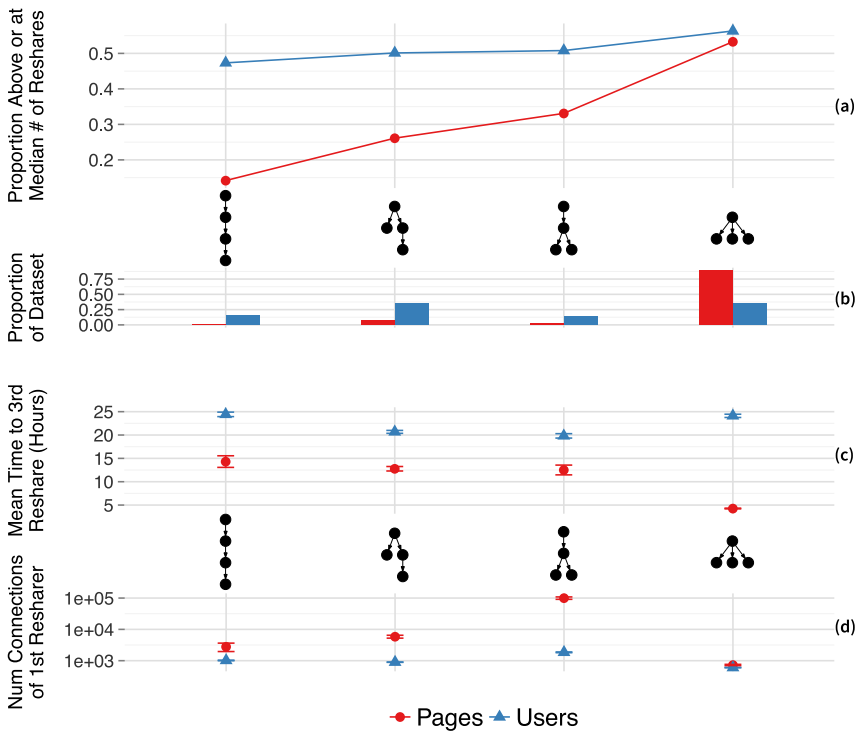


Figure 3.13: Shallow initial cascade structures are indicative of larger cascades. In contrast to page-started cascades, where the mean time to the 3rd reshare decreases with decreasing depth of the initial cascade, shallow cascades take a much longer time to form for user-started cascades. For these, the connections of the 1st resharer also significantly impacts the time to the 3rd resharer, especially when it receives two reshares before the original receives a second.

initial poster types: a user, and a page. We discard "celebrity" users who may large followings like the most popular pages. Figure 3.13a shows that as the initial cascade structure becomes shallower, the proportion of cascades that double in size increases. To examine why this would be the case, we also examined the time needed for the 3rd reshare to happen (Figure 3.13c). For pages, shallower cascades tend to happen more rapidly, consistent with being initiated by a popular page and achieving a large number of reshares directly from its fans. Interestingly, the configuration having the second and third reshares stemming from the first reshare correspond to having a first resharer with many connections, and indicating that the initial poster is less popular, be it a page or user (Figure 3.13d).

Curiously, for user-started cascades, the star configuration tends to grow into the largest cascades, but is also the slowest. It also tends to correspond to the first resharer having a low degree, both for page and user roots. One might speculate that this pattern is indicative of the item's appeal to less well-connected users, who also happen to be more likely to reshare. In fact, a median resharer has 35 fewer friends than someone who is active on the site nearly every day. Thus, an item's appeal, rather than the initial network structure, may drive the eventual cascade size in the long run.

3.3.10 Predicting cascade structure

The observations above naturally lead to the question of whether it is possible to predict future cascade structure. In particular, we aim to distinguish cascades that spread like a virus in a shallow forest fire-like pattern (Figure 3.3a) and cascades which spread in long, narrow string-like pattern (Figure 3.3c). As discussed earlier, this difference is related to the structural virality of a cascade and is quantified by the Wiener index. Here, we observe $k = 5$ reshares of a cascade and aim to predict whether the final cascade will have a Wiener index above or below the median. We obtain accuracy of 0.725 (F1 = 0.715, AUC = 0.796), while random guessing would, by construction, achieve accuracy of 0.5.

Temporal and structural features are most predictive of structure. For this task we expect structural features to be most important, while we expect temporal features not to be indicative of the cascade structure. However, when we train the model on individual classes of features we surprisingly find that both temporal and structural features are almost equally useful in predicting cascade structure: 0.622 vs. 0.620. Nevertheless, structural features remain individually more accurate (≈ 0.58) and highly correlated ($0.161 \leq |r| \leq 0.255$) with the Wiener index. Individually, one temporal feature, $views'_{1..k-1,k}$, is slightly more accurate (0.602) compared to the best-performing structural feature, $outdeg(\hat{v}_0)$ (0.600), but is significantly less correlated (0.041 vs. -0.255). The two classes of features nicely complement each other, since

when combined, accuracy increases to 0.72.

Cascade structure also becomes more predictable with increasing k. Like for cascade growth prediction, our prediction performance improves as we observe more of the cascade, with accuracy linearly increasing from 0.724 when k is 5 to 0.808 when k is 100. A linear relation also exists in the alternate task where we set the minimum cascade size R to be 100, varying k between 5 and 100.

Changes in feature importance. As we increase k, we find that the structural features become highly correlated with the Wiener index, suggesting that the initial shape of a cascade is a good indicator of its final structure. Rapidly growing cascades also result in final structures that are shallower—temporal features become more strongly correlated with the Wiener index as k increases. Unlike with cascade size, views were generally weakly correlated with structure, while content features had a weak, near-constant effect. Nonetheless, some of these features still provided reasonable performance in the prediction task.

User vs. page-started cascades. In predicting the shape of a cascade, we find that our overall prediction accuracy for pages is slightly higher (0.724) than for users (0.700). While using only structural features alone results in a higher prediction accuracy for users (0.643) than for pages (0.601), user and content features are significantly more predictive of cascade structure in the case of pages.

To sum up, we find that predicting the shape of a cascade is not as hard as one might fear. Nevertheless, predicting cascade size is still much easier than predicting cascade shape, though classifiers for either achieve non-trivial performance.

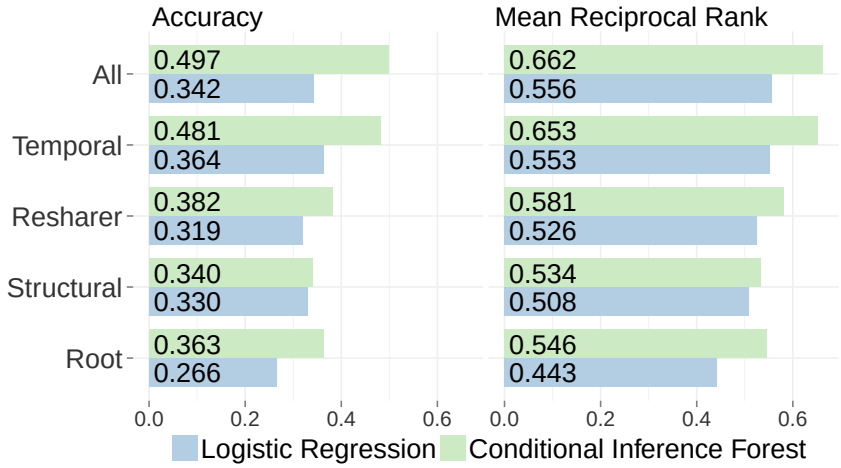


Figure 3.14: In predicting the largest cascade in clusters of 10 or more cascades of identical photos, we perform significantly above the baseline of 0.1.

3.4 Predictability and Content

3.4.1 Controlling for cascade content

In our analyses thus far, we examined cascades of uploads of different photos, and tried to account for content differences by including photo and caption features. However, temporal and structural features may still capture some of the difference in content. Thus, we now study how well we can predict cascade size if we control for the content of the photo itself. We consider *identical photos* uploaded to Facebook by different users and pages, which is not a rare occurrence. We used an image matching algorithm to identify copies of the same image and place their corresponding cascades into clusters (983 clusters, N_c = 38,073, N_r = 12,755,621). As one might expect, even the same photo uploaded at different times by different users can fare dramatically differently; a cluster typically consists of a few or even a single cascade with a large number of reshares, and many smaller cascades with few reshares. The average Gini coefficient, a measure of inequality, is 0.787 (σ = 0.104) within clusters. Thus, a natural task is to try to predict the largest cascade within a cluster. For every cluster we select 10 random cascades, placing the accuracy of random guessing at 10%.

As shown in Figure 3.14, in all cases we significantly outperform the baseline. Using a random forest model, we can identify the most popular cascade nearly half the time (accuracy 0.497); a mean reciprocal rank of 0.662 indicates that this cascade also appears in the top two predicted cascades almost all the time.

In terms of feature importance we notice that best results are obtained using temporal features, followed by resharer, root node, and structural features. Essentially, if one upload of the photo is initially spreading more rapidly than other uploads of the same photo, that cascade is also likely to grow to be the largest. This points to the importance of landing in the right part of the network at the right time, as the same photo tends to have widely and predictably varying outcomes when uploaded multiple times.

3.4.2 Feature importance in context

Some features may be more or less important for our prediction tasks in different contexts. Figure 3.15 shows how several features correlate with log-transformed cascade size when conditioned on one of four different variables, including a) source node type—user vs page, b) language—English versus Portuguese, the two most common languages of cascade root nodes in our dataset, c) whether text is overlaid on a photo—a common feature of recent Internet memes, and d) content category. We determine content category by matching entities in photo captions to Wikipedia articles, and in turn articles to seven higher-level categories: animal, entertainment, politics, religion, famous people (excluding religious and political figures), food, and health.

Figure 3.15 shows that the initial rate of exposure of the uploaded photo is generally more important for page cascades than for user cascades ($views'_{0,5}$). This is likely due to the higher variance in the distribution of the number of followers for a user versus a page. For page cascades in our sample, the median number of followers is 73,855 with a standard deviation of 675,203, while for users at the root of cascades

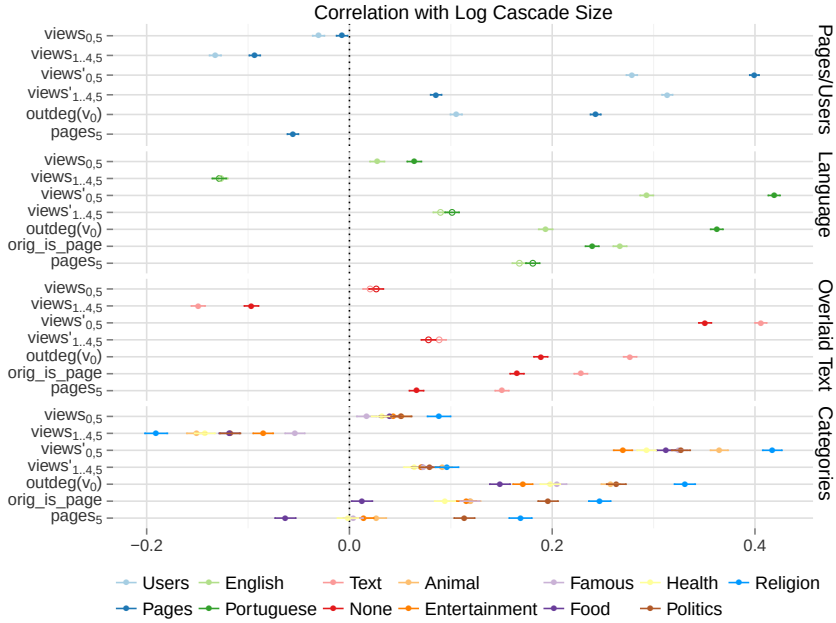


Figure 3.15: The initial exposure of the uploaded photo and initial reshares serve to differentiate datasets from one another, as can be seen by comparing the correlation coefficients of each feature with the log cascade size. Solid circles indicate significance at $p < 10^{-3}$, and lines through each circle indicate the 95% confidence interval.

the median number of friends and subscribers is 1,042 with a standard deviation of 26,482. Though rate of exposure to the original photo is more important for pages, we see that rate of exposure to the initial reshares ($views'_{1..4,5}$) is much more important for user cascades.

The number and rate of views also act to differentiate topical categories, with religion having the highest correlation between views and cascade size. Correlation for the rate of views of the uploaded photo is also higher for those with a Portuguese-speaking root node as opposed to an English one. The feature $outdeg(v_0)$ indicates the ability of the root to broadcast content, and we see this playing an important role for page cascades, Portuguese content, photos with text, and religious photos. This indicates that much of the success of these cascades is related to the root nodes being directly connected to large audiences.

In addition to the analysis of Figure 3.15, we also examined how the features correlate

with the structural virality of the final cascades. (Each of the reported correlation coefficient comparisons that follow are significant at $p < 10^{-3}$ using a Fisher transformation.) Photos relating to food differ significantly from all other categories in that features of the root, such as $outdeg(v_0)$, are less negatively correlated (>–0.18 vs. –0.11), and depth features, such as $depth_k^{avg}$, are less positively correlated (>0.18 vs. 0.11). This relationship also holds for English compared to Portuguese photos. While users with many friends or followers are more likely to generate cascades of larger size and greater structural virality, pages with many fans create cascades of larger size, although not necessarily greater virality (0.05 vs. –0.01). However, if the initial structure of a cascade is already deep, the final structure of the cascade is likely to have greater structural virality for both user and page-started cascades (>0.16). A user-started cascade whose initial reshares are viewed more quickly is also more likely to become viral than that for a page-started cascade (0.23 vs. 0.06).

3.5 Discussion and Conclusion

This work examines the problem of predicting the growth of cascades over social networks. Although predictive tasks of similar spirit have been considered in the past, we contribute a novel formulation of the problem which does not suffer from skew biases. Our formulation allows us to study predictability throughout the life of a cascade. We examine not only how the predictability changes as more and more of the cascade is observed (it improves), but also how predictable large cascades are if we only observe them initially (larger cascades are more difficult to predict). While some features, e.g., the average connection count of the first k resharers, have increasing predictive ability with increasing k, others weaken in importance, e.g., the connectivity of the root node. We find that the importance of features depends on properties of the original upload as well: the topics present in the caption, the language of the root node, as well as the content of the photo.

Despite the rich set of results we were able to obtain, there are some limitations

to this study. Most importantly, the study was conducted entirely with Facebook data and only with photos. Still, one advantage of this is the scale of the medium; hundreds of millions of photos are uploaded to Facebook every day, and photos, more than other content types, tend to dominate reshares. This also gives us high-fidelity traces of how the photo moves within Facebook's ecosystem, which allows us to precisely overlay the spreading cascade over the social network. Moreover, we are able to identify uploads of the same photo and track them individually. This eliminates the concern of shares being driven by an external entity and only appearing to be spreading over the network. Instead, external drivers benefit our study by creating independent 'experiments' where the same photo gets multiple chances to spread, helping us control for the role of content in some of our experiments. Another disadvantage of our setup is that diffusion within Facebook is driven by the mechanics of the site. The distinction between pages and users is specific to Facebook, as are the mechanisms by which users interact with content, e.g., liking and resharing. Despite these limitations, we believe the results give general insights which will be useful in other settings.

The present work only examines each cascade independently from others. Future work should examine interactions between cascades, both between different content competing for the same attention, and between the same content surfacing at different times and in different parts of the network. We found that when the same photo is uploaded at least 10 times, the largest cascade was twice as likely to be among the first 20% of uploads than the last 20%. Similarly, for photos uploaded 20 times, the largest cascade was 2.3 times as likely to be among the first 20% than the last. Figure 3.16 shows the friendship edges between users participating in different cascades of a single, specific photo. The high connectivity *between* different cascades demonstrates that users are likely being exposed to the same photo via different cascades, which could be a contributing factor in why earlier uploads of the same photo tend to generate larger cascade than later ones. Between-cascade dynamics like this should provide ample opportunities for further research.

Addressing questions like these will lead to a richer understanding of how information

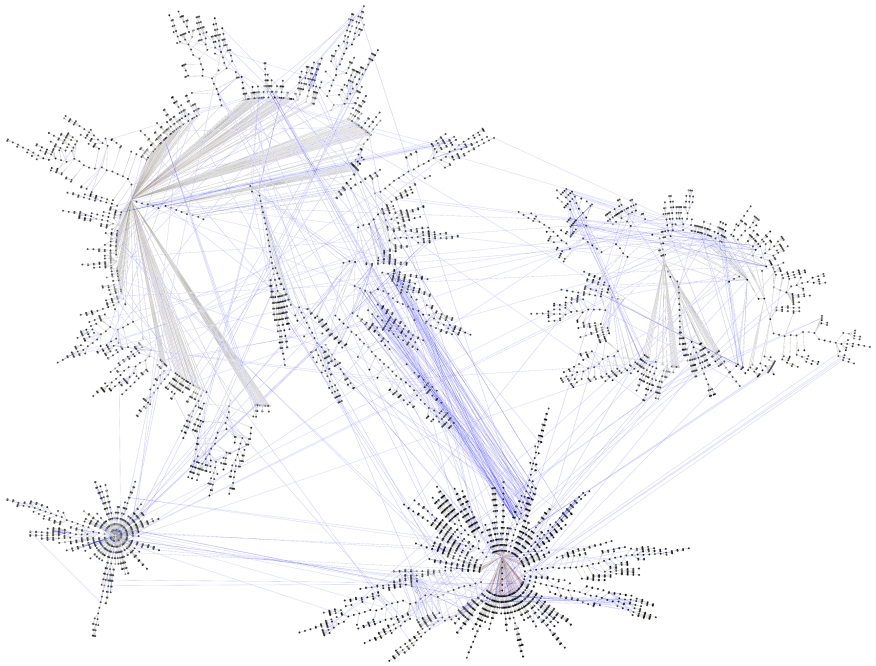


Figure 3.16: There is considerable overlap in friendship edges (blue) between four independent cascades of the same photo.

spreads online and pave the way towards better management of socially shared content and applications that can identify trending content in its early stages.

Chapter 4

Cascade Recurrence

In this chapter, we extend the line of work on cascade prediction that we embarked on in the previous chapter, where we asked how a cascade will behave in the future. Here, we focus on the long-term temporal dynamics of cascades, and in so doing, identify recurrence as a major phenomenon that can only be studied over long periods of time.

Existing work studying the temporal patterns of cascades has typically focused on the basic rising-and-falling pattern that characterizes the initial onset of a cascade [4, 34, 208, 312]. In this chapter, we perform a large-scale analysis of cascades on Facebook over significantly longer time scales, and find that a more complex picture emerges, in which many large cascades *recur*, exhibiting multiple bursts of popularity with periods of quiescence in between. Rather than cascades slowly decaying in activity over time, that they instead common exhibit more complex behavior in which they recur, or experience renewed bursts of popularity. We characterize recurrence by measuring the time elapsed between bursts, their overlap and proximity in the social network, and the diversity in the demographics of individuals participating in each peak. We discover that content virality, as revealed by its initial popularity, is a main driver of recurrence, with the availability of multiple copies of that content helping to spark new bursts. Still, beyond a certain popularity of content, the rate of recurrence drops as cascades start exhausting the population of interested individuals. We reproduce

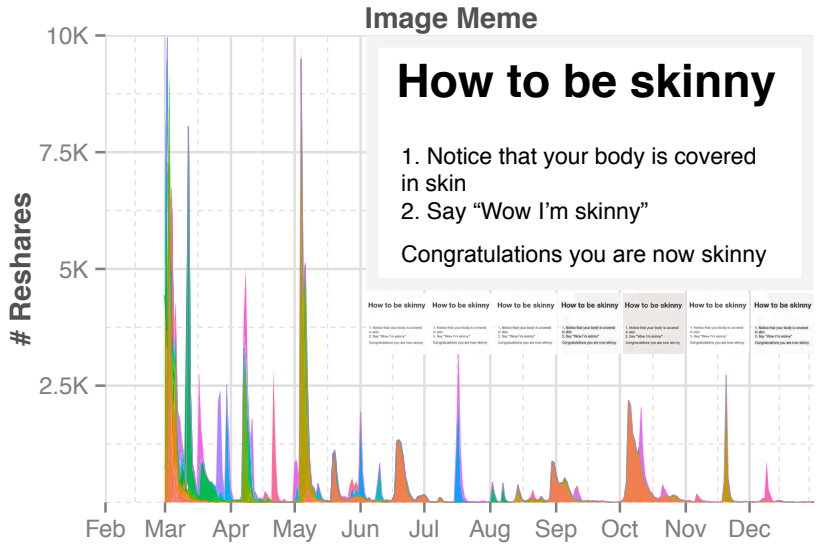
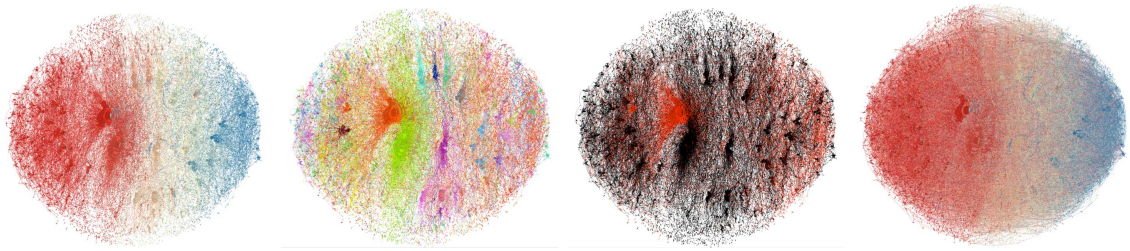


Figure 4.1: An example of a image meme that has recurred, or resurfaced in popularity multiple times, sometimes as a continuation of the same copy, and sometimes as a new copy of the same meme (example copies are shown as thumbnails). This recurrence appears as multiple peaks in the plot of reshares as a function of time.

these observed patterns in a simple model of content recurrence simulated on a real social network. Building on our work in the previous chapter on predicting a cascade's future growth, we also demonstrate strong performance in predicting whether it will recur in the future, by using only characteristics of that cascade's initial burst.

4.1 Introduction

In many online social networks, people share content in the form of photos, videos, and links with one another. As others reshare this content with their friends or followers in turn, cascades of resharing can develop [69]. Substantial previous work has studied the formation of such information cascades with the aim of characterizing and predicting their growth [25, 132, 314]. Cascades tend to be bursty, with a spike of activity occurring within a few days of the content's introduction into the network [196, 225]. This property forms the backdrop to a line of temporal analyses that focus



(a) (b) (c) (d)

Figure 4.2: (a) The diffusion cascade of the example meme from Figure 4.1 as it spreads over time, colored from red (early) to blue (late). Only reshares that prompted subsequent reshares are shown. (b) The cascade is made up of separately introduced copies of the same content; in this drawing of the cascade from (a), each copy is represented in a different color. (c) Sometimes, individual copies experience a resurgence in popularity; again we draw the cascade from (a), but now highlight a single resurgent copy in red with the spread of all other copies depicted in black. (d) A different network on the same set of users who took part in the cascade, showing friendship edges rather than reshare edges. These edges span reshares across copies and time, showing that multiple copies of the meme are not well-separated in the friendship network.

on the basic rising-and-falling pattern that characterizes the initial onset of a cascade [4, 34, 208, 312].

However, the temporal patterns exhibited by cascades over significantly longer time scales is largely unexplored. Do successful cascades display a long monotonic decline after their initial peak, or do they exhibit more complex behavior in which they can *recur*, experiencing renewed bursts of popularity long after their initial introduction? Anecdotally, many of us have experienced déjà vu when a friend shared content we had seen weeks or months ago, but it is not clear whether these are isolated occurrences or glimpses into a robust phenomenon. Resolving these basic distinctions in the long-time-scale behavior of cascades is crucial to understanding the longevity of content beyond its initial popularity, and points toward a more holistic view of how content spreads in a network.

The present work: Cascade recurrence. We perform a year-long large-scale analysis of cascades of public content on Facebook, measuring them over significantly longer

time scales than previously investigated. Our first main finding is that recurrence is widespread in the temporal dynamics of large cascades. Among large cascades appearing in 2014, over half come back in one or more subsequent bursts. While reshare activity does peak and then drop to very low or even zero levels relatively soon after introduction, the same content can recur after a short or extended lull.

The prevalence of recurrence prompts several questions about how and why content recurs. Is more broadly or narrowly appealing content more likely to recur? Does a larger initial burst indicate a greater likelihood of recurrence, or does it inhibit subsequent bursts by exposing and thus satiating many people in the initial wave? Do different bursts of the same content spread in different parts of the network? Is the second burst a continuation of the initial cascade, or a fresh re-introduction of the content into the network? Does the media type of the reshared content matter — for example, whether it is a photo or a video? Finally, how well can one combine such features to predict whether a piece of widely reshared content is likely to experience additional bursts in popularity later on?

We motivate our discussion with an example of content recurrence. Figure 4.1 shows an image meme that first became popular on Facebook at the end of February 2014, and it depicts how the number of reshares of that meme changed over time. Here, while an initial burst in resharing activity is followed by a gradual decrease, this meme recurred, experiencing multiple resurgences in popularity — first in mid-March, then several times over the next few months. Perhaps surprisingly, there is little to no resharing between consecutive bursts. Additionally, multiple near-identical *copies* of this image meme, represented in different colors, are shared in the network. This distinction between different copies of the same content will prove important in our later analyses: when a user reshares content through the reshare mechanism provided by the site, the content continues onward as the same copy; in contrast, when a user reposts or re-uploads the same content and thus shares it afresh, this is a new copy.

Figure 4.2 sketches the diffusion cascade of this meme, or its propagation over edges in the social network. As shown in (a), bursts in activity are connected through the

same large long-lived cascade and can be traced through the network, from the initial bursts in March (shown in red), to the smaller bursts nearer the end of 2014 (shown in blue). In (b), where the same network is now colored according to the copy of the image being reshared, different copies of the same content appear at different times, sometimes corresponding to when bursts occur, suggesting that recurrence sometimes occurs from the introduction of new copies. However, recurrence may also occur as a continuation of a previous copy: the copy highlighted in red in (c) experiences an initial burst in March, but then resurfaces in popularity later in the year. Further, we see in (d) that friendship ties exist between even the earliest and latest reshares — the meme appears to be diffusing rapidly, but also revisits parts of the network through which it had earlier diffused.

While the meme in our example recurred several times, are such memes the exception or the norm? And if such memes are in fact typical, what are the bases for such robust patterns of recurrence? To answer these questions, we use a dataset of reshare activity of publicly viewable photos and videos on Facebook in 2014.

Characterizing recurrence. First, we develop a simple definition of a *burst*, corresponding informally to a spike in the number of reshares over time, that we can use to quantify when recurrence occurs (via multiple observed bursts), and when it does not (a single burst). We show that a significant volume of popularly reshared content recurs (59% of image memes and 33% of videos), and that recurring bursts tend to take place over a month apart from each other. Recurrence is itself relatively bursty — rarely do we observe long sustained periods of resharing.

Studying the temporal patterns of recurrence, user characteristics of the resharing population, and the network structure of cascades, we find that the recurrence of a piece of content is moderated to a large extent by its *virality*, or broadness of appeal: cascades with initial bursts of activity that are larger, last longer, and have a more diverse population of resharers are more likely to recur. Nonetheless, it is not the cascades that start out the largest or most viral that recur, but those that are moderately appealing. Specifically, a moderate number of initial reshares, as well as

a moderate amount of homophily (or diversity) in the initial resharing population is correlated with higher rates of recurrence. This lies in contrast to more appealing (or popular) content, where one is likely to see a single large outbreak which results in a large single burst, as well as less appealing content, where one is likely to only see a single small outbreak and thus a smaller single burst. In the former case, we show evidence that a large initial burst inhibits subsequent recurrence by effectively "immunizing" a large proportion of the susceptible population.

While individual copies of content already recur in the network (18% for image memes and 30% for videos), the presence of multiple copies catalyzes recurrence, allowing that content to spread rapidly to different parts of the network, significantly boosting the rate of recurrence. To a smaller extent, the principle of homophily, suggesting that people are more likely to share content received from users similar to themselves, also plays a role in recurrence, with user similarity positively correlated with the rate of spreading.

Modeling recurrence. Motivated by the above picture of recurrence, and inspired by classic epidemiological models of diffusion [230] and disease recurrence [7, 155, 233], we present a simple model of cascading behavior that is primarily driven by content virality and the availability of multiple copies, and is able to reproduce the observed recurrence features. A simulation of this model, which introduces multiple copies of the same content into the network, can cause independent cascades that peak at different times and in aggregate are observed as recurring. As the virality of the content increases, the shape of a plot of overall reshares in the network over time transforms from a shorter independent single burst, to multiple bursts of differing sizes, to a single large burst of a longer duration. Replicating our previous findings, increasing virality increases recurrence, up to a point: once a meme has exposed a large part of the network, further recurrence is inhibited.

Predicting recurrence. Finally, we show how temporal, network, demographic, and multiple-copy features may be used to predict *whether* a cascade will recur, if the recurrence will be *smaller* or *larger* than the original burst, and *when* the recurrence

occurs. We demonstrate strong performance in predicting whether the same content will recur after observing its initial burst of popularity (ROC AUC=0.89 for image memes), as well as in predicting the relative size of the resulting burst (0.78). The time of recurrence, on the other hand, appears to be more unpredictable (0.58). Features relating to content virality and multiple copies perform best. Though multiple-copy features account for significant performance in predicting the recurrence of content, we obtain similarly strong performance (0.88) when predicting the recurrence of an individual copy of a piece of content.

Together, these results not only provide the first large-scale study of content recurrence in social media, but also begin to suggest some of the factors that underpin the process of recurrence.

4.2 Related Work

Significant prior work has studied information diffusion in online social media [25, 69, 225] — with respect to memes, work has demonstrated the effect of meme similarity [85] and competition for limited attention on subsequent popularity [302]. Most relevant is previous work that looked at the temporal dynamics of diffusion and developed epidemiological models of recurrence.

Among work that aims to predict the future popularity of online content [50, 185, 277], one relevant line of research has involved modeling the temporal patterns of the diffusion of information in social media [4, 208, 313] or using these patterns to predict future popularity or forecast trends [15, 33, 34, 69, 77, 135]. Beyond the initial burst of activity, we studied the long-term temporal dynamics of content on Facebook over a year.

In prior work, when multiple bursts are observed in a time series, they tend to be of a topic or hashtag rather than an individual piece of content, and are commonly attributed to external stimuli [132, 180, 193, 226] (e.g., news related to that topic).

While knowing about external events can help forecast the temporal pattern of the resulting spike [208], there has been little work in predicting if new spikes will appear in the future lacking such knowledge. In particular, rumor recurrence is bursty, with or without external stimuli, and sometimes with embellishments and other mutations [1, 184, 110], but there is little understanding of this phenomenon. Patterns of human activity can also explain periodicity in popularity [16, 131, 196], but the vast majority of recurrence we observe in this work is aperiodic. While external stimuli explains some instances of recurrence, we discover other factors that influence recurrence. In contrast to most work that has observed multiple bursts in topics, we observed recurrence even at the level of an individual copy.

Finally, substantial work has studied how bursts in streams or time series can be detected [166, 171, 236]. In this work, we adopted a simple definition of burstiness, parameterizing peaks and bursts relative to the mean activity observed.

Recurrence has also been studied in the context of epidemiology, though primarily from a modeling perspective. Many base their analysis on SIR models [230], simulating recurrence through introducing dormant periods [155], seasonality effects [7], or changes in contagion fitness [117], which may be periodic [233]. More recently, some work studied content popularity using these models, while accounting for user login dynamics and content aging [59]. The structure of the network can also cause periodicity in epidemics [183, 294]. Many focus on modeling specific types of recurrence (e.g., historical disease epidemics [7]). In contrast, many recurrences we observe are aperiodic, and findings on synthetic networks may not easily generalize. Inspired by this line of work, we adapted an SIR model assuming multiple points of infection on a real social network, and show that key characteristics of recurrence we observed can be reproduced.

4.3 Technical Preliminaries

Studying cascade recurrence requires both sufficiently rich data that accurately measures activity throughout a network over long periods of time, as well as a robust definition of what recurrence is.

4.3.1 Dataset Description

In this work, we use over a year of sharing data from Facebook. All data was de-identified and analyzed in aggregate. Facebook presents a particularly rich ecosystem of users and pages (entities that can represent organizations or brands) sharing a large amount of content over long periods of time.

Reliably measuring the spread of content in a network over time is challenging because multiple copies of the same content may exist at any time. As we will later show, the presence of multiple copies in a cascade is an important catalyst for recurrence. On Facebook, users and pages may introduce a new copy of the same content by re-posting or re-uploading it; resharing an existing copy instead creates an attribution back to that same copy. Content may be reintroduced, instead of reshared, for various reasons — multiple users may have independently discovered the same content, or downloaded and then re-uploaded an image.

To construct a dataset of popularly shared content, we initially selected a seed set of reshared content uploaded to Facebook in March 2014. We selected the top 200,000 most reshared images, which were publicly viewable, counting only reshares within the 180 days since the image was uploaded,

Next, we tried to identify other copies of content that exist in this seed set. Beyond exact copies of the same image, many near-identical images, which have slightly different dimensions or introduce compression artifacts or borders, also exist (as seen in Figure 4.1). As such, a binary k-means algorithm [125] was used to identify clusters of near-identical images to which each of these candidates belonged, including images

beyond the original set. For each cluster, we then obtained all reshares of images in that cluster that were made in 2014. To verify the quality of the clustering, we manually examined the top 100 most-reshared copies in each of 100 randomly sampled clusters. In 94 clusters, all 100 copies were near-identical. The remaining clusters mainly comprised the same image overlaid with different text.

This sample of resharing activity in 2014 that we use consists of 395,240,736 users and pages that made 5,167,835,292 reshares of 105,198,380 images. These images were aggregated into 76,301 clusters. Repeating the process above for videos shared on Facebook, we obtain a sample comprising 323,361,625 users and pages that made 2,187,047,135 reshares of 6,748,622 videos, aggregated into 156,145 clusters. Images, videos, users, and pages that were deleted were excluded from analysis. On average, each image cluster is made up of 1379 copies of the same content. Video clusters were smaller, with 43 copies in each cluster on average.

As we only measured reshares for a year, we may only be observing part of a cascade's spread if it began prior to 2014. Thus, we also considered subsets of each dataset containing only clusters that began in 2014. We identified these subsets by additionally measuring reshares of content in the three months prior to 2014 (October to December 2013) and excluding clusters where activity was observed during this period.

Though we mainly analyze recurrence at the cluster level, we also investigate the recurrence of individual copies by studying the top 100,000 individually most reshared copies in each dataset.

	# Clusters	**Copies/Cluster**	**Prop. Recurrence**
Image Memes	51,415 (76,793)	523 (1378)	0.40 (0.59)
Videos	149,253 (156,145)	13 (43)	0.30 (0.33)
	# Peaks	**Days Observed**	**Days betw. 1st/2nd Burst**
Image Memes	2.3 (4.6)	202 (280)	31 (32)

Videos	1.6 (2.0)	170 (182)	47 (44)

Table 4.1: Recurrence occurs in a large proportion of popular image memes and videos shared on Facebook. We note in parentheses statistics computed on all cascades, as opposed to the cascades that began in 2014 whose initial spread we can observe.

4.3.2 Defining Recurrence

In this work, we define recurrence relative to *peaks* and *bursts* in popularity over time. In practice, almost all popular content on Facebook experiences at least one peak in popularity. If content peaks in popularity more than once, we say that it *recurs*.

To identify these peaks, and thus whether a cascade recurs, we measure the number of reshares of content over time. Figure 4.4 shows several examples of recurring memes. Empirically, reshare activity is varied across different content but is generally bursty, with long periods of inactivity between peaks. As recurrence occurs over a long amount of time, we discretize time into days.

Intuitively, a recurrence occurs when a *peak* is observed in the time series. Not only should these peaks be relative outliers on a timeline, but they should also last for a significant amount of time. Further, we should be able to tell these peaks apart from each other. Motivated by this intuition, suppose we observe a meme for t days. Let $r_i,\ i \in \{1, 2, ..., t\}$ be the number of reshares observed on day i. We parameterize recurrence using four variables — h_0, m, and w place constraints on identified peaks, and v places a constraint on the "valley" between peaks. Specifically, the height h of each peak must be at least h_0 and at least m times the mean reshares per day $\bar{r}$ (Figure 4.3). Additionally, a peak day must be a local maximum within $\pm w$ days. Finally, between any two adjacent peaks p_i and p_{i+1}, the number of reshares must drop below $v \cdot \min\{r_{p_i}, r_{p_{i+1}}\}$. We call the area around the peak a *burst* (b_0, b_1 respectively for p_0, p_1 in Figure 4.3), whose duration or width w, is defined as the sum of the number of days the number of reshares is increasing before p_i and falling after p_i,

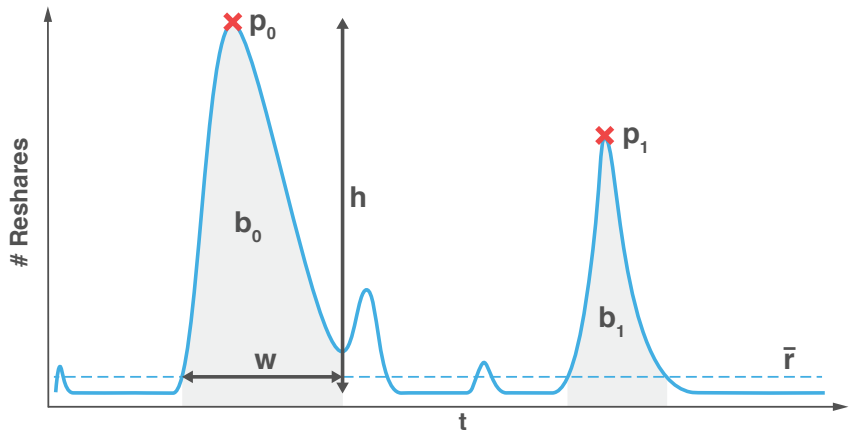


Figure 4.3: Recurrence occurs when we observe multiple peaks (p_0, p_1, red crosses) in the number of reshares over time. Bursts (b_0, b_1) capture the activity around each peak.

while remaining above $\bar{r}$. There is a one-to-one correspondence between peaks and bursts.

In practice, we set h_0=10, m=2, w=7, and v=0.5 so that each burst is relatively well-defined. The red crosses in Figure 4.4 show the identified peaks under this regime. While this definition does not strictly minimize activity between bursts, empirically, activity does drop significantly (and in many cases, falls to zero) in between bursts. Stricter definitions that reduce the number of identified peaks (e.g., requiring a well-defined "valley floor" between two peaks, or increasing h_0 or m) also resulted in qualitatively similar findings. The approach we take is fairly rudimentary; future work may involve developing more specific definitions of recurrence which take into account the shape of resulting bursts.

4.4 Characterizing Recurrence

We first introduce recurrence at a high level, showing that it is both common and bursty, with the same content sometimes resurfacing multiple times. We then discuss four important classes of observations that we later draw on to model and predict

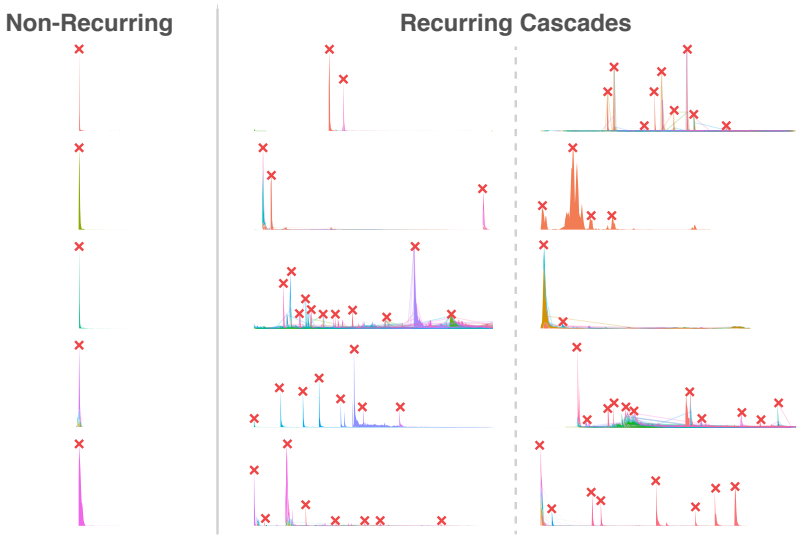
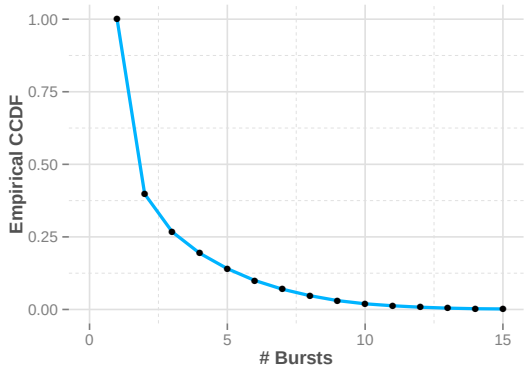


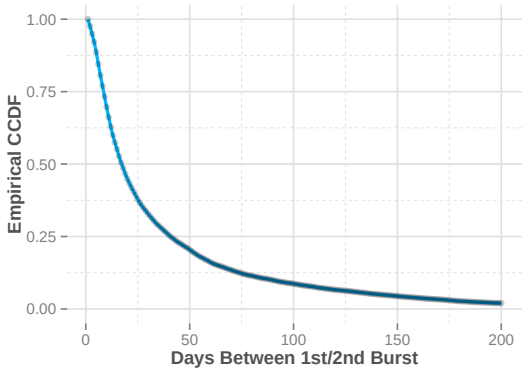
Figure 4.4: Examples of time series of recurring and non-recurring cascades over a year, colored by copy. Identified peaks are marked with red crosses; the number of reshares is normalized per cascade.

recurrence:

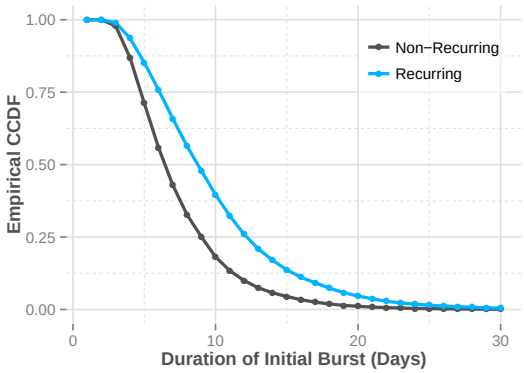
- Temporal patterns: cascades with longer initial bursts, but a moderate number of reshares, are more likely to recur.
- Sharer characteristics: recurring and non-recurring cascades differ in demographic makeup, and moderate diversity in the initial sharing population encourages recurrence. Further, changes in homophily in the network affect the speed at which content spreads, and hence burstiness.
- Network structure: bursts in a cascade occur in different, but nonetheless connected parts of the network. Also, large initial bursts tend to exhaust the supply of susceptible users, potentially accounting for why moderate, but not high cascade volume or diversity results in greater recurrence.
- Catalysts of recurrence: the availability of multiple copies in the network may catalyze recurrence. Still, neither does the presence of multiple copies suggest that recurrence is entirely an externally-driven phenomenon, nor is it a necessary



(a) Number of Peaks



(b) Days Between Bursts

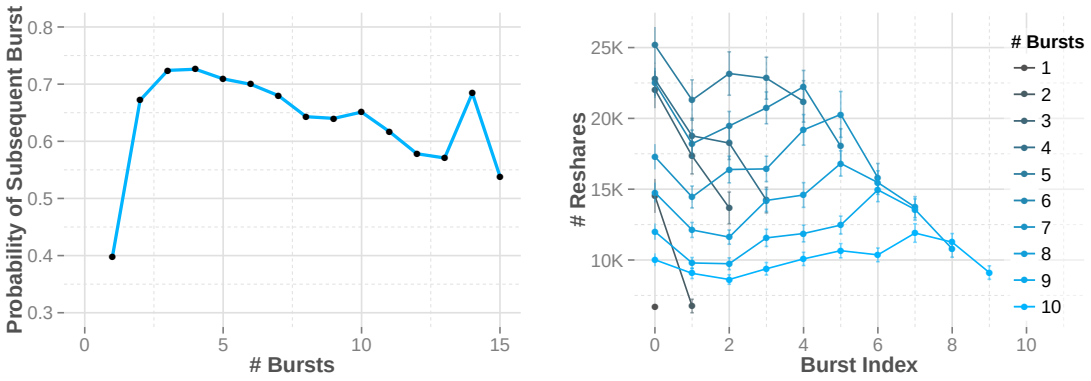


(c) Cascade Duration

Figure 4.5: (a) 40% of cascades that began in 2014 came back, and (b) over 30% of recurring cascades only resurfaced after a month or more. (c) Further, the initial burst of a recurring cascade tends to last longer than that of a non-recurring cascade.

condition for recurrence.

In the remainder of this work, we report results primarily on image memes, and note any salient differences with videos. All differences reported are significant at $p<10^{-10}$ using a t-test unless otherwise noted.



(a) Probability of Subsequent Bursts (b) # Reshares in Each Burst

Figure 4.6: (a) The probability of subsequent recurrences increases after the initial recurrence. (b) Cascades that recur less tend to have bursts that diminish in size over time, while those that recur more tend to have a stable burst size.

4.4.1 Recurrence is common

Once introduced on Facebook, popular content continues spreading for a long time. On average, the maximum time between reshares of the same content is 280 days. But rather than being shared at a constant rate (among popularly reshared content, less than 1% of memes have no discernible peak), resharing tends to be bursty, with bursts typically separated by substantial periods of relative inactivity. A mean of 32 days separates the initial and subsequent bursts for image memes (Figure 4.5b).

Previously, we defined recurrence as observing multiple peaks in the number of reshares observed over time, and non-recurrence as observing only a single peak. Over these long periods of time, 59% of popular image memes recur. In fact, a significant proportion of these cascades experience resurgences in popularity (Figure 4.5a), and may even have experienced bursts prior to our observation window. If we limit the sample to the set of image memes which *began* spreading in 2014, 40% of these memes recur (Table 4.1).

4.4.2 Temporal patterns

Cascades with larger initial bursts of activity that last longer are more likely to recur, suggesting that more viral, or appealing cascades are more likely to recur. However, it is not the most popular cascades that recur the most, but those that are only moderately popular — while recurrence initially increases with the size of the initial peak, it subsequently decreases.

Recurring cascades have larger, longer-lived initial bursts. The initial burst of a cascade is already indicative of recurrence. Recurring cascades start out larger (15,547) and initially last longer (9.3 days) than non-recurring cascades (6128 reshares, 6.9 days), spending more time "building up" (Figure 4.5c) and "winding down". The greater initial popularity of recurring cascades suggests that more viral cascades are more likely to recur, but is this the case?

Recurring content is moderately popular. Plotting the total number of reshares in the initial burst against the subsequent number of bursts observed, rather than the number of reshares monotonically increasing or decreasing the rate of recurrence, we observe a striking interior maximum at approximately 10^5 reshares for both image memes and videos (Figure 4.7a). Neither the initially best-performing (or most viral), nor poorest-performing (or least viral) cascades tend to resurface. In the former case, a single large burst tends to dominate with smaller bursts after; in the latter case, a small number of small bursts is typically observed.

They keep coming back! While most of our analyses focus on the initial burst and subsequent recurrence, several general trends arise as more recurrence is observed:

- Once a cascade has recurred, it is more likely to resurface again. The probability of recurrence jumps from 0.40 initially, to 0.60 for subsequent recurrences before gradually decreasing (Figure 4.6a). This observation parallels prior work showing that the prior popularity of Youtube videos predicts their future popularity [49]. In fact, for 26% of all image meme cascades, we observe resharing activity on the first and last day of our observation period. These image memes

may be "evergreen", tending to continuously recur.

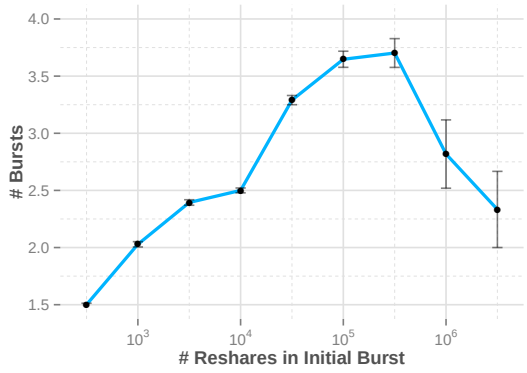
- For cascades that recur less, subsequent bursts tend to be smaller; for cascades that recur more, subsequent bursts are more similar in size (Figure 4.6b), suggesting that they depend less on external factors (e.g., breaking news) to spread.
- Subsequent recurrences are briefer than their predecessors. Burst duration monotonically decreases from a mean of 7.6 days for the first burst to 6.3 for the tenth.
- On average, the lull between recurrences is substantial, with bursts happening an 28 to 32 days apart for image memes, and 30 to 44 days apart for videos. Again, these long periods between bursts suggest that recurrence can only be observed over substantial periods of time.

4.4.3 Sharer characteristics

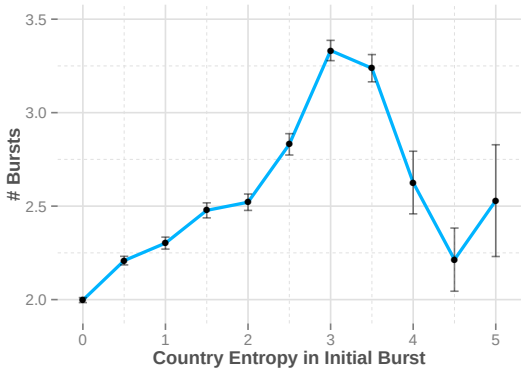
People who participate in recurring cascades differ significantly from those who participate in non-recurring cascades. While a diverse user population encourages recurrence, moderately diverse cascades recur the most. Homophily, the concept that similar people are likely to share the same content, also affects how quickly content spreads, suggesting that it modulates recurrence.

Demographics vary with recurrence. For recurring cascades, the average age of people participating in the initial burst is lower (40 vs. 42), but the proportion of women is higher (65% vs. 58%). The latter observation corroborates previous work that showed a correlation with eventual cascade size [69].

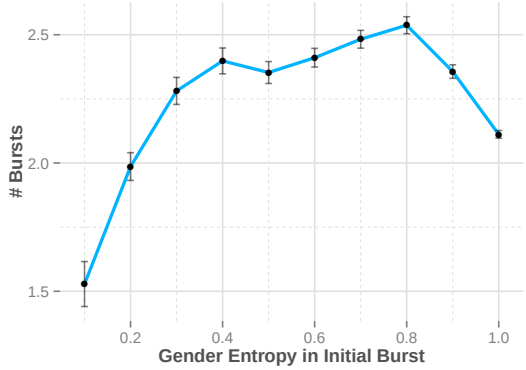
Demographics also change across bursts. In the case of image meme cascades, the mean age changes by 2.7 years, and the proportion of women by 6.1 percentage points (in absolute terms). The same content may become popular in different parts of the world at different times, resulting in recurrence: 13% of the time, the majority of people in the initial two bursts come from different countries.



(a) # Bursts vs. # Reshares



(b) # Bursts vs. Country-Entropy



(c) # Bursts vs. Gender-Entropy

Figure 4.7: (a) A moderate number of reshares results in more recurrence. (b), (c) Similarly, recurrence is more likely when the entropy of the distribution of users across countries, as well as gender, is moderate.

Diversity encourages recurrence. We now turn our attention to the diversity (or homophily) of people who take part in a cascade. We quantify homophily in the network by measuring the entropy of the distribution of demographic characteristics. A low entropy in the distribution of countries users are from (or country-entropy) corresponds to high homophily, suggesting that a majority of sharers belong to a small number of countries. On the other hand, a high country-entropy suggests that the countries sharers belong to are more diverse and distributed more evenly.

It is not *a priori* clear whether homophily encourages or inhibits recurrence. Homophily within a community, meaning that connected users are receptive to sharing the same content, may help a cascade gain the initial traction it needs to spread, but may also result in the content getting "trapped" in a local part of the network. In contrast, diversity in the users sharing that content suggests it has wider appeal and might come back, but may also result in only a single burst if the initial spread overwhelms the network.

We find that diversity in the country distribution is predictive of recurrence. Controlling for the duration (w), peak height (h), and the number of reshares in the initial bursts of recurring and non-recurring cascades [252], a Wilcoxon Signed-rank test shows that a higher country-entropy is indicative of recurrence ($W>10^8$, $p<10^{-10}$, effect size r=0.19). Thus, if the initial burst of a cascade occurs in more countries, it is more likely to recur. Higher gender-entropy (i.e., greater gender-balance) also predicts recurrence, but its effect is weaker ($W>10^8$, $p<10^{-2}$, r=0.02). The effect of age is inconsistent across image memes and videos.

Recurring content is moderately diverse. Again, it is not the most diverse populations that bring about recurrence: a moderate country-entropy of approximately 3.0 in the initial burst of a cascade results in the most recurrence (Figure 4.7b). An interior maximum can also be observed with respect to the gender-entropy of the initial burst (Figure 4.7c). These results, combined with the previous observation of a similar interior maximum with respect to the initial number of reshares, suggests that the virality of content plays a significant role in recurrence.

Cascades spread quickly in pockets of homophily. The virality of a cascade and homophily in the network are closely related, and perhaps represent two perspectives on the spread of content. Greater virality enables content to appeal to a larger population; more homophily suggests that receptive users are closer in the network. In fact, homophily in the network modulates the speed of resharing (and thus bursts in a cascade). If we measure the average country-entropy of a sliding window of 100 reshares ordered in time and the time elapsed, and then compute the average correlation, we

find a slight positive correlation between the two (0.08), suggesting that homophily among those sharing results in faster resharing, and hence burstiness. Gender-entropy and age-entropy are also positively, but more weakly correlated with burstiness (0.06 and 0.04 respectively). One potential cause of this is that pages, with their substantial and homophilous followings, are driving resharing. The more pages that share content, the more homophilous the set of potentially reachable people, and thus the quicker content is reshared. However, as the proportion of reshares attributable to pages increases, the entropy in these demographic characteristics instead increases, an effect opposite to what we observe.

4.4.4 Network structure

The initial bursts of recurring cascades tend to be better connected. Further, successful recurrence tends to occur in different, but not disconnected parts of the network. Considering the people potentially exposed in each burst beyond the resharers, large initial cascades may exhaust the population of susceptible people in the network, a fact that will subsequently become important in explaining the mechanism of recurrence.

Bursts of recurring cascades are internally more connected. More people and pages share in the initial burst of recurring cascades than non-recurring cascades (15,050 users and 59 pages, and 5855 users and 24 pages respectively). To measure connectivity within a burst, we used the induced subgraph G_0 of the Facebook network made up of the people and pages resharing in the initial burst. The subgraph includes two kinds of edges: friend edges between people, and follow edges between people and the pages they like. On this subgraph, people in the initial burst of recurring cascades have an average of 3.4 connections to other people and pages in the same burst, relative to 3.1 connections in non-recurring cascades, suggesting that the initial bursts of recurring cascades are slightly more connected.

Subsequent bursts happen in different parts of the network. Bursts of a cascade are

separate in time, but may overlap in terms of sharers, or be connected via friend and follow edges.

To start, the sharer overlap between bursts is small. For recurring cascades, an average of 15,050 people and 59 pages make up the initial burst, and 8892 people and 28 pages the subsequent burst. Comparing people and pages across these bursts, Jaccard similarities of 0.02 (i.e., 2% of people share the same content in both bursts) and 0.03 mean that bursts have very little direct overlap.

We also find evidence of community structure within bursts by considering whether the second burst is proximate in the network to the first. Even if individuals in the second burst do not repost content a second time, they may not be far removed in the social network from someone in the first burst. Here, we instead consider the induced subgraph G_{0+1} of the individuals and pages who reshared in the initial two bursts for each cascade. If these bursts correspond to communities within the network, we would expect more edges within bursts than between them. An average of 17,457 friend and 17,242 follower edges exist in the first burst; 10,094 friend and 6,310 follower edges exist in the second. Note that these communities are very sparse. A person within a burst has a connection to an average of 3.2 friends or pages within the same burst, indicating that memes tend to diffuse out through the network rather than stay within a narrow community. Still, across bursts, we observe an average of 8,273 friend and 4,755 follower edges, resulting in an average of 1.4 connections to friends and pages in a different burst, indicating that these bursts are somewhat separated.

Large initial bursts exhaust the supply of susceptible people. As noted above, the second burst in a cascade has fewer sharers. Intuitively, people may tire of content they have seen before, but is this the case? Studying overlap a third way, we look at the susceptible populations of the initial and subsequent bursts of a cascade. We approximate these populations by considering people who could have been exposed through their connections (friend and follow edges) to those who shared content in these two bursts.

On average, 6.4 million unique individuals are potentially exposed in the initial burst,

with recurring cascades, having a greater number of initial resharers, having greater reach (8.2 million vs. 4.0 million for non-recurring cascades). For recurring cascades, the potential reach of the second burst is smaller, but still sizable at 6.8 million. Comparing the sets of individuals exposed in the first two bursts of recurring cascades, we obtain a Jaccard similarity of 0.15, indicating this second burst is mostly reaching a different set of people, but where some people will have seen the same content twice, creating a sense of déjà vu for many.

In particular, 28% of people who are potentially exposed in the second peak would have also been exposed in the first. This high overlap could be a contributing factor to the second peak being smaller. Further, the proportion of exposed individuals in the second peak is positively correlated (ρ=0.39) with the size of the first peak, meaning that the larger the initial peak, the more likely that those exposed in the second peak have previously seen the content. In other words, in the case of large initial bursts, subsequent bursts are likely reaching a similar part of the network.

4.4.5 Catalyzing recurrence

As shown in Figures 4.1 and 4.2, cascades are made up of reshares of multiple copies of the same content, and the presence of these copies can help catalyze recurrence. Still, neither are copies the only cause of recurrence (recurrence is substantial even with a single copy), nor must they be independently or externally introduced (many later copies are attributable to previously seen copies).

Cascades whose reshares are divided across multiple copies tend to recur. Recurring cascades are made up of more copies than non-recurring cascades (2277 vs. 93). Reshares are also more spread out across multiple copies in the former case (841 and 3445 reshares per copy for recurring and non-recurring cascades respectively), suggesting recurrence may be characterized by multiple smaller outbreaks. The most reshared copy accounts for 72% of reshares in the initial burst for recurring cascades, and 93% for non-recurring cascades. Altogether, the substantial differences here

suggest the strong predictive power of these characteristics.

The appearance of new copies correlates with recurrence. Further, the introduction of new copies and the number of reshares over time is significantly correlated (Pearson's r=0.66), suggesting that the appearance of new copies causes bursts, and thus recurrence. On a related note, prior work showed that reposting content helps make it popular [277].

Copies are not the only cause of recurrence. Nonetheless, not all copies burst (only 6% are reshared at least 10 times on any single day), and not all bursts are caused by new copies, as we will later show. And while correlations between the number of copies and other characteristics such as duration and country-entropy also exist, when we control for the number of copies in the initial bursts of recurring and non-recurring cascades [252], all previously observed differences in the temporal, sharer, and network characteristics of these cascades still hold ($W>10^8$, $p<10^{-10}$, mean effect size r=0.08). Comparing recurring and non-recurring cascades with similar numbers of copies in their initial bursts, the initial bursts of recurring cascades are still larger, longer-lived, and more diverse. In all, this suggests that recurrence is not simply caused by distinct copies of the same content spreading through the network, but is a result of a more complex phenomenon which we explain in Section 4.5.

A majority of copies are internal to the network. Still, where do these copies come from, and are they internal or external to the network? By using the network to identify friends and pages who may have previously shared a different copy of some content, we can attribute 75% of newly uploaded copies to previously seen copies in the network (this approach roughly estimates content-copying that occurs within Facebook, as users who share a new copy may not have seen a friend's shared copy). This suggests a nuanced approach to studying recurrence — external sources may drive some of the introduction of new copies to a social system, but a large proportion of activity, which we can study, occurs within the network.

Pages may also catalyze recurrence. Pages are responsible for a large proportion of

highly-reshared copies (over 70% of reshares are attributable to page-created copies in the second burst of recurring cascades). In recurring cascades, pages tend to re-upload, rather than reshare content, doing so 50% of the time, as opposed to 2% for users. Further, the most popular copy in the second burst is likely to have been created by a page (70%). Given the relatively higher degree of pages, which tend to have tens of thousands of followers, as opposed to users who typically only have hundreds of friends, pages may spark recurrence by posting a new copy of the same content, rapidly exposing a number of followers to it.

Individual copies recur too! Recurrence of the individually most popular copies in our datasets, while lower than when copies are studied in clusters, is still substantial (18%). These individual copies last a significant amount of time (261 days), with bursts further apart (41 days). Like cascades of multiple copies, the initial bursts of recurring individual-copy cascades are larger and longer-lived than those of non-recurring cascades, with later bursts occurring in different parts of the network. Recurrence of the same copy can also be observed within clusters — 22% of the time, the most reshared copy in a burst was also most reshared in a previous burst.

4.5 Modeling Recurrence

Tying our observations together, we present an overall picture of the mechanisms of recurrence, then suggest a model of recurrence which we evaluate through simulations on a real social network.

4.5.1 Why do cascades recur?

Our findings as a whole suggest a model of recurrence where virality is a primary factor, and where the availability of multiple copies can help spark recurrence.

Virality plays a primary role in recurrence. Virality, or broadness of appeal, affects

recurrence: cascades with initial bursts that are larger, last longer, and are more demographically diverse are more likely to recur. Specifically, *moderately* popular and diverse cascades are *most* likely to recur. While recurrence typically occurs in different parts of the network, the larger the initial burst of a cascade, the larger the proportion of the potentially exposed population in the subsequent burst that was already previously exposed. This observation, coupled with the fact that users tend not to reshare the same content multiple times, suggests that large initial bursts inactivate a significant portion of the network, inhibiting a cascade's future spread. Our subsequent simulations show more clearly that this may indeed happen as the initial burst grows large.

Multiple copies in the network help spark recurrence. Bursts in a cascade are separated by relatively long periods of inactivity. By studying the availability of multiple copies of the same content, we find that these copies can act as catalysts for recurrence in different parts of the network. Indeed, multiple introductions of the same content correlate with recurrence. However, while more copies initially increases the chance of recurrence, they are not the only cause of it; recurring and non-recurring cascades with similar numbers of copies differ significantly in virality. Moreover, multiple copies do not explain the substantial recurrence of individual copies. To a lesser extent, we also discover that homophily in the network affects the speed of the spread of a cascade in a network.

Together, moderate content virality and the presence of multiple copies results in recurrence. While the likelihood of recurrence does increase with the number of copies (or potential "sparks"), we can still observe an interior maximum in how recurrence varies with the number of reshares after fixing the number of copies, where a moderate number of reshares results in the most recurrence.

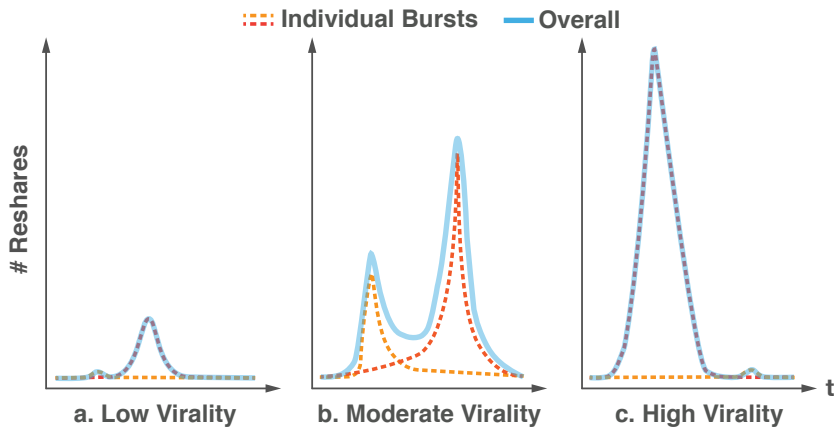


Figure 4.8: When virality is low, only a small number of attempts at infection succeed. When virality is moderate, more attempts succeed, which aggregate into observable recurrence. When virality is high, rather than a large number of bursts aggregating to form a single large peak, the first successful burst infects a large portion of the network, making it difficult for other copies to spread.

4.5.2 A simple model of recurrence

Motivated by these findings, we suggest a simple model of cascading behavior where recurrence depends on content virality:

- If the virality of a cascade is low, it may only appeal to a small group of people, and is thus unable to spread far in the network. Thus, a single, small peak results, with many attempts to propagate in the network failing (Figure 4.8a).
- As virality increases, the cascade is able to spread substantially further in the network, and may occasionally even jump to other local communities in the network, spreading faster within them. As several bursts occur in the network, they may be observed as recurring in aggregate (Figure 4.8b).
- However, as virality increases beyond some threshold, any individual burst is likely to spread through a large portion of the susceptible population, inhibiting the transmission of subsequent copies (Figure 4.8c). This last point lies in contrast to the trivial hypothesis that more independent copies leads to more independent bursts that aggregate to form a single large burst, which does not

appear to be the case, as most reshares in initial bursts can be attributed to a single copy.

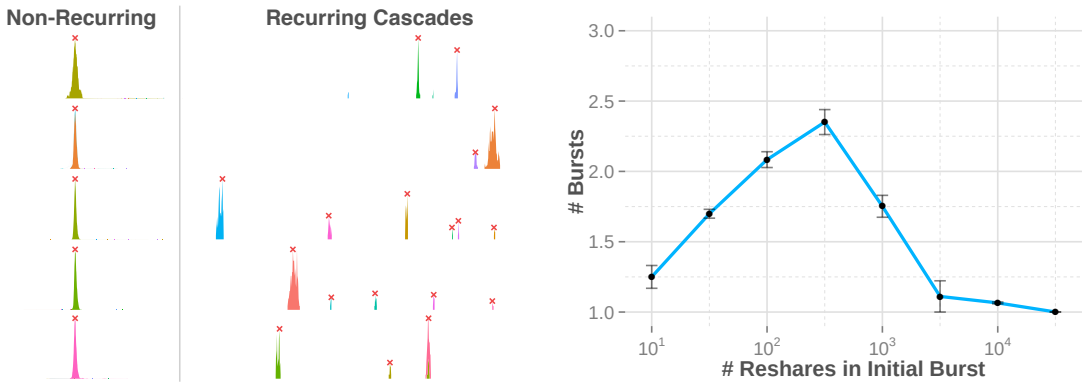
4.5.3 Simulating recurrence

To see if such a model of recurrence can reproduce characteristics of recurrence observed in the data, we now simulate recurrence on a real social network. Our observations and model suggest the use of an SIR model, where nodes in a network are initially susceptible (S) to a contagion, and then may become infected (I) when exposed. Infected nodes subsequently recover (R) and become resistant to the contagion. These models have been used to study the spread of disease [20] and information [33, 59, 230] in a network.

Setup. Our simulation thus consists of an SIR model with multiple outbreaks introduced at different times, and with resistant nodes reinfectable at a lower rate. We parameterize our model as follows: For a given contagion c, its virality, or equivalently, the susceptibility of every node in the network, is p_0^c. In other words, if exposed to the contagion, the probability that the node will be infected is p_0^c. Infected nodes attempt to infect all neighbors in the subsequent time step, and then become resistant. As users sometimes share the same content multiple times, resistant nodes have a constant lower probability $p_1^c < p_0^c$ of being re-infected. The introduction of each copy of a contagion is normally distributed in time ($N(\mu, \sigma)$).

Here, we make a simplifying assumption that independent copies of the same content are introduced into the network at different points in time. Following the intuition that more connected entities (e.g., pages) are likely to start outbreaks, the target nodes to infect are sampled, with replacement, proportional to the node's degree. m copies are introduced in total.

We simulate this model for 1000 discrete time steps with μ=500, σ=250, and m=50, varying p_0^c between 5×10^{-4} and 10^{-3} and where p_1^c=$0.5 \cdot p_1^c$. We run our simulation on the network of a country with approximately 1.4 million nodes and 160 million



(a) Simulated Time Series (b) # Peaks vs. Reshares, Simulated

Figure 4.9: (a) By varying content virality, a model of recurrence that assumes independent introductions of copies of the same content can simulate recurrence. (b) It also replicates the observation that a moderate number of reshares results in more recurrence.

friendship edges, repeating the simulation 5000 times. We measure the total number of infections (or reshares) in each time step, and identify bursts as defined in Section 4.3.2.

Results. Within certain ranges of virality ($6\times10^{-4} \leq p_0^c < 8\times10^{-4}$), we can consistently reproduce recurring cascades. Figure 4.9a shows several examples of the time series of these simulations. In aggregate, we can obtain a distribution of number of peaks similar in shape to Figure 4.5. Plotting the number of peaks against number of reshares in the initial burst (or alternatively, p_0^c), we observe an interior maximum — a moderate amount of virality results in the most recurrence (Figure 4.9b), replicating our previous findings.

When the virality of the contagion is high ($p_0^c \geq 8\times10^{-4}$), a large fraction of the highly connected portion of the graph becomes infected by a single copy in the initial burst, suppressing subsequent bursts as many nodes are now resistant. To show this happening, we consider for each simulation, in addition to our original model, an alternate-universe setting where the resistances of nodes are reset following the initial burst. We can then measure how much the initial burst inhibited the second by observing

the likelihood of a second burst in the alternate case, as well as the overlap of nodes infected in the second burst with nodes in the initial burst. A significant difference in the total number of peaks when virality is high (1.0 vs. 2.0, t=92, $p<10^{-10}$), but not when virality is low ($p_0^c \leq 7\times10^{-4}$, n.s.) suggests that the supply of susceptible nodes is indeed being used up in the former case, but not the latter. A significant positive correlation of initial peak size with the size of the overlap of the second peak in the alternate setting (0.76) further supports this hypothesis and our prior observations.

Likewise, the connectivity of graph deteriorates significantly after a large initial burst ($p_0^c \geq 8\times10^{-4}$). Here, we measure the algebraic connectivity [107] of the graph if all the nodes involved in the initial burst are removed, and compare this to a baseline that removes the same number of nodes at random. Connectivity is significantly lower in the former case (579 vs. 1065, $t>17$, $p<10^{-10}$), especially in comparison to the graph's initial connectivity (1105).

These results together suggest that under such a model of recurrence, a large initial burst does indeed inhibit subsequent bursts, as we previously hypothesized (Figure 4.8c). Also in support of our prior observations, increasing the number of introduced copies m monotonically increases recurrence.

Limitations and alternatives. Importantly, our model assumes that recurrence is sparked primarily by independent copies introduced to the network. However, the reality of recurrence is subtler: individual copies recur significantly in the network, and homophily may also moderate recurrence. Allowing virality to vary with time [135] or having nodes wait according to a power-law distribution [87, 201] may also reproduce recurrence with only a single copy. Decision-based queuing processes [29] may also help model the long periods of inactivity between bursts.

4.6 Predicting Recurrence

Is it possible to predict if a cascade will resurface in the future? Observing just the initial burst of a cascade, we use features related to the temporality, network structure, user demographics, and presence of multiple copies to determine *a*) *whether* recurrence occurs, *b*) if the recurrence will be relatively *smaller or larger*, and *c*) *when* the recurrence occurs. Overall, we find that cascades with longer initial bursts that consist of multiple small outbreaks tend to recur, supporting the hypothesis that content virality and multiple copies play a significant role in recurrence. Nonetheless, we obtain similarly strong performance predicting recurrence for individual copies of content. Predicting recurrence may enable us to better forecast content longevity in a network.

4.6.1 Factors driving recurrence

Based on our observations, we develop several features that help predict recurrence, and group them into four categories:

Temporal features (7). Initially longer-lived bursts are suggestive of recurrence, motivating the importance of the *number of days before* and *after* the peak is reached, as well as the *number of reshares before* and *after*, and the *height of the initial peak*. The average *gradient of the initial burst before* and *after* the peak further characterize the shape of the initial burst.

Demographic features (5). The differences in user characteristics and diversity we previously observed suggest the importance of *age*, *gender*, as well as the *entropy in the distribution of age*, *gender* and *country* of the initial burst.

Network features (6). Recurring cascades appear to be more connected in their initial bursts, having more *friendship* and *follower edges*, in addition to having a larger potentially *exposed population*. The *number of users*, *pages*, and *proportion of pages*

in the initial burst also vary.

Multiple-copy features (8). The availability of multiple copies plays a significant role in recurrence, motivating the use of the *number of copies observed in the initial peak*, the *entropy in the distribution of reshares of each copy*, the *mean reshares per copy*, and the *proportion of reshares attributable to the most popular copy*. Pages also play a role in recurrence, suggesting that the *proportion of copies created by pages*, the *proportion of all reshares made by pages* or *attributable to page-created copies*, and *whether the most popular copy was created by a page* are useful features.

AUC on Feature Sets			
Temporal	0.74	0.76	0.55
+ Demographic	0.78 (0.63)	0.76 (0.58)	0.56 (0.52)
+ Network	0.81 (0.72)	0.77 (0.66)	0.57 (0.53)
+ Multiple-Copy	0.89 (0.82)	0.78 (0.70)	0.58 (0.54)

Table 4.2: We obtain strong performance in predicting *whether* recurrence occurs and if the subsequent burst will be *smaller or larger*, but not in predicting *when* recurrence occurs. Individual feature set performance is in parentheses. The column headers correspond to these three prediction tasks, and are described subsequently in this section.

4.6.2 Does it recur?

Prediction task. We formulate our prediction task as a binary classification problem: given only the initial burst of a cascade, we aim to predict if a second burst will be observed (i.e., if the cascade will recur). We use a balanced dataset of recurring and non-recurring cascades (N=40,912 for image memes, 89,368 for videos) so that guessing results in a baseline accuracy of 0.5. Given the non-linear relation of several features to recurrence (e.g., that a moderate number of reshares results in the most recurrence), we use a random forest classifier. In all cases, we perform 10-fold cross-validation and report the classification accuracy, F1 score, and area under the ROC

curve (AUC).

Results. Overall, we find strong performance in predicting recurrence (Accuracy=0.82, F1=0.81, AUC=0.89). A logistic regression classifier results in slightly worse performance (AUC=0.78). Table 4.2 shows how performance improves as features are added to the model, as well as individual feature set performance. While multiple-copy features perform best, temporal and network features, and to a lesser extent demographic features, also individually exhibit robust performance, suggesting that each significantly contributes to recurrence. In the absence of strong multiple-copy features (fewer copies of any one video exist), we obtain worse performance in predicting the recurrence of videos (Acc=0.69, F1=0.66, AUC=0.76), with temporal features instead performing best.

For image meme cascades, the most predictive features of recurrence relate to cascades having multiple small outbreaks (fewer reshares per copy (0.78) and a higher entropy in the distribution of reshares across copies (0.72)), and longer initial bursts (more days before (0.63) and after (0.63) the peak). These features remain important for video cascades. Mirroring the dual importance of multiple-copy and temporal features, just the number of reshares per copy and the average gradient of the initial burst after its peak alone achieve strong performance (0.81). Though the initial burst of a recurring cascade is on average significantly larger, size-related features are weaker signals of recurrence ($\leq$0.59).

4.6.3 Will the recurrence be smaller/larger?

Prediction task. Assuming that we know that a cascade will recur, how much smaller or larger will the second burst be? Knowing the relative size of the next recurrence can differentiate bursty cascades that are rising or falling in popularity. Given the initial burst of a cascade, we aim to predict if the relative size of the second burst, or the ratio of the size of the second burst to that of the first, is above or below the median (0.28). As the median evenly divides the dataset, we again have a balanced

binary classification task with a random guessing baseline accuracy of 0.5.

Results. We also find strong performance in predicting the relative size of the subsequent burst (Acc=0.72, F1=0.69, AUC=0.78 for image memes, AUC=0.85 for videos). Temporal features here outperform all other feature sets, with the most predictive features relating to the cascade having a long initial burst.

4.6.4 When does it recur?

Prediction task. If a cascade will recur, when will we observe the next burst? With a cascade's initial burst, can we predict if the duration between bursts will be greater than the median (14 days)?

Results. We find that the timing of recurrence is far less predictable (Acc=0.56, F1=0.51, AUC=0.58 for image memes, AUC=0.60 for videos). Nevertheless, longer initial bursts are most indicative of recurrence happening earlier.

4.6.5 Predicting recurrence for individual copies

Given the correlation of the appearance of multiple copies with bursts, multiple-copy features perform strongest in predicting recurrence. But what if we want to predict recurrence of a single instance of some content, where multiple copies do not exist by definition? Surprisingly, we obtain similarly strong performance in predicting the recurrence of individual copies (N=28,454, Acc=0.80, F1=0.79, AUC=0.88 for image memes, AUC=0.82 for videos). Network features are strongest (AUC=0.84), with fewer edges between users and pages (0.68) in the initial peak the most predictive of recurrence. As individual copies have a single point of origin, fewer edges between pages and users and more edges between users (0.61) suggests that the burst may have resulted more from users sharing content from other users than high-degree pages sharing that content with their followers. This observation, together with the

fact that longer initial bursts continue to be strongly predictive of recurrence (>0.65), suggests the continued significance of virality with respect to individual copies.

The relative size of the subsequent burst is similarly predictable for individual copies (0.83 for image memes, 0.84 for videos), but interestingly, the time of recurrence is more predictable (0.68 and 0.63 respectively), which may be because any recurrence must be a continuation of the initial copy, as opposed to possibly being sparked by a new, less related copy.

4.7 Discussion and Conclusion

Our results start to shed light on the mechanism of content recurrence — studying a large dataset of popularly reshared content, we find that recurrence is common, and that content can come back not just once, but several times. Strikingly, content may nearly cease to circulate for days, weeks or even months, prior to experiencing another surge in popularity. Such a phenomenon may seem highly unpredictable, but we find trends in how recurring cascades behave, and can predict whether content will come back. The virality, or appeal of a cascade plays a role in recurrence: cascades whose initial bursts are long-lasting, moderately popular, and moderately diverse are most likely to recur. The presence of multiple copies of the same content sparks recurrence, though homophily in the network may also influence recurrence.

One limitation of our work is that we only analyze content within a single network. Though most copies of the same content were made within the network, a minority appeared without a prior path. Analyzing the transfer of content between different social networks may reveal different mechanisms of recurrence. Separately, while the appearance of multiple copies correlates with recurrence, this does not hold in the case of individual-copy recurrence. Understanding recurrence in the absence of multiple copies (e.g., through studying homophily in more detail) remains future work.

Based on our observations, we presented a simple model that exhibits some features

of recurrence (e.g., pronounced bursts with little activity in-between, and an internal maximum in the number of bursts as a function of the number of reshares). Future work could extend such models to account for homophily and community structure in the network.

While the temporal shape, network structure, and user attributes are already highly predictive of resharing behavior, other factors may improve prediction accuracy further: sentimentality or humor may make content evergreen, while content tied to current events may have an expiration date. Seasonality effects may also cause periodic recurrence: we did observe an instance of a daylight-savings image meme which appeared, as expected, exactly at the two points during the year when people needed to adjust their clocks. Also, other types of content may exhibit different properties of recurrence (e.g., link sharing may be more externally driven); the interactions of users with shared content (e.g., comments) may also reveal the reasons why some content came back; the societal context of memes, as well as their interactions (or competition) with other content, may also reveal more insight into their popularity [274]. Perhaps most suggestive that much remains to be studied is that while we can predict if recurrence will happens, it remains a significant challenge to predict *when* recurrence will happen.

Chapter 5

The Causes of Antisocial Behavior

In the previous two chapters on cascade growth and recurrence, we focused primarily on macro-level characteristics of a cascade (e.g., its overall size and network structure). In this chapter and the next, we now turn to studying the micro-level characteristics of cascades, and focus on understanding the behavior of the individuals that make up cascades. Motivated by the prevalence of antisocial behavior online [101], we focus our analyses on understanding its spread. We examine how such negative behavior is transmitted from person to person, and in particular, focus on trolling in online discussions.

In this chapter, we identify the primary causes of trolling behavior in online discussions, and thus the underlying factors that enable cascades of such behavior to spread in online discussion communities. In contrast to prior work which suggests that trolls are a vocal and antisocial minority, we demonstrate that ordinary people can engage in trolling behavior as well. We propose two primary trigger mechanisms: the individual's mood, and the surrounding context of a discussion (e.g., exposure to prior trolling behavior). Through an experiment simulating an online discussion, we find that both negative mood and seeing troll posts by others significantly increases the probability of a user trolling, and together double this probability. To support and extend these results, we study how these same mechanisms play out in the wild via a

data-driven, longitudinal analysis of a large online news discussion community. This analysis reveals temporal mood effects, and explores long range patterns of repeated exposure to trolling. A predictive model of trolling behavior shows that mood and discussion context together can explain trolling behavior better than an individual's history of trolling. These results combine to suggest that ordinary people can, under the right circumstances, behave like trolls.

A Brief Note on Definitions

In this thesis, we take a relatively broad view of antisocial behavior, and define trolling as behavior that occurs outside of community norms. Given that our analysis takes place primarily on discussion platforms, this definition is also based on the community guidelines present on several online discussion forums [83, 133, 97], where undesirable behavior includes no name-calling, personal attacks, profanity, threat, hate speech, and offensive material. Under this definition, regardless of intent, trolling happens as long as people are seen to behave like trolls. Nonetheless, we see future work in identifying particular types of trolling, which we explore in Chapter 7.

We operationalize these definitions in two ways depending on the study design. In our controlled experiments, posts are labeled as trolling or not by humans (either expert raters or multiple crowd workers) based on these community guidelines; In our observational studies, we use moderator labels to identify instances of trolling. In the online communities we study, moderators can either delete posts or ban users for violating community guidelines, and we use these as proxies for trolling behavior.

5.1 Introduction

As online discussions become increasingly part of our daily interactions [94], antisocial behavior such as trolling [140, 159], harassment, and bullying [272] is a growing concern. Not only does antisocial behavior result in significant emotional distress

[5, 200, 244], but it can also lead to offline harassment and threats of violence [304]. Further, such behavior comprises a substantial fraction of user activity on many web sites [73, 94, 112] – 40% of internet users were victims of online harassment [101]; on CNN.com, over one in five comments are removed by moderators for violating community guidelines. What causes this prevalence of antisocial behavior online?

In this work, we focus on the causes of *trolling behavior* in discussion communities, defined in the literature as behavior that falls outside acceptable bounds defined by those communities [42, 83, 140]. Prior work argues that trolls are born and not made: those engaging in trolling behavior have unique personality traits [51] and motivations [22, 143, 262]. However, other research suggests that people can be influenced by their environment to act aggressively [79, 156]. As such, is trolling caused by particularly antisocial individuals or by ordinary people? Is trolling behavior innate, or is it situational? Likewise, what are the conditions that affect a person's likelihood of engaging in such behavior? And if people can be influenced to troll, can trolling spread from person to person in a community? By understanding what causes trolling and how it spreads in communities, we can design more robust social systems that can guard against such undesirable behavior.

This work reports a field experiment and observational analysis of trolling behavior in a popular news discussion community. The former allows us to tease apart the causal mechanisms that affect a user's likelihood of engaging in such behavior. The latter lets us replicate and explore finer grained aspects of these mechanisms as they occur in the wild. Specifically, we focus on two possible causes of trolling behavior: a user's mood, and the surrounding discussion context (e.g., seeing others' troll posts before posting).

Online experiment. We studied the effects of participants' prior mood and the context of a discussion on their likelihood to leave troll-like comments. Negative mood increased the probability of a user subsequently trolling in an online news comment section, as did the presence of prior troll posts written by other users. These factors combined to double participants' baseline rates of engaging in trolling behavior.

Large-scale data analysis. We augment these results with an analysis of over 16 million posts on *CNN.com*, a large online news site where users can discuss published news articles. One out of four posts flagged for abuse are authored by users with no prior record of such posts, suggesting that many undesirable posts can be attributed to ordinary users. Supporting our experimental findings, we show that a user's propensity to troll rises and falls in parallel with known population-level mood shifts throughout the day [121], and exhibits cross-discussion persistence and temporal decay patterns, suggesting that negative mood from bad events linger [156, 163]. Our data analysis also recovers the effect of exposure to prior troll posts in the discussion, and further reveals how the strength of this effect depends on the volume and ordering of these posts.

Drawing on this evidence, we develop a logistic regression model that accurately (AUC=0.78) predicts whether an individual will troll in a given post. This model also lets us evaluate the relative importance of mood and discussion context, and contrast it with prior literature's assumption of trolling being innate. The model reinforces our experimental findings – rather than trolling behavior being mostly intrinsic, such behavior can be mainly explained by the discussion's context (i.e., if prior posts in the discussion were flagged), as well as the user's mood as revealed through their recent posting history (i.e., if their last posts in other discussions were flagged).

Thus, not only can negative mood and the surrounding discussion context prompt ordinary users to engage in trolling behavior, but such behavior can also spread from person to person in discussions and persist across them to spread further in the community. Our findings suggest that trolling, like laughter, can be contagious, and that ordinary people, given the right conditions, can act like trolls. In summary, we:

- present an experiment that shows that both negative mood and discussion context increases the likelihood of trolling,
- validate these findings with a large-scale analysis of a large online discussion community, and

- use these insights to develop a predictive model that suggests that trolling may be more situational than innate.

5.2 Related Work

To begin, we review literature on antisocial behavior (e.g., aggression and trolling) and influence (e.g., contagion and cascading behavior), and identify open questions about how trolling spreads in a community.

5.2.1 Antisocial behavior in online discussions

Antisocial behavior online can be seen as an extension of similar behavior offline, and includes acts of aggression, harassment, and bullying [5, 159]. Online antisocial behavior increases anger and sadness [200], and threatens social and emotional development in adolescents [244]. In fact, the pain of verbal or social aggression may also linger longer than that of physical aggression [67].

Antisocial behavior can be commonly observed in online public discussions, whether on news websites or on social media. Methods of combating such behavior include comment ranking [151], moderation [187, 237], early troll identification [62, 73], and interface redesigns that encourage civility [177, 178]. Several sites have even resorted to completely disabling comments [108]. Nonetheless, on the majority of popular web sites which continue to allow discussions, antisocial behavior continues to be prevalent [73, 94, 112]. In particular, a rich vein of work has focused on understanding trolling on these discussion platforms [98, 140], for example discussing the possible causes of malicious comments [189].

A troll has been defined in multiple ways in previous literature – as a person who initially pretends to be a legitimate participant but later attempts to disrupt the community [98], as someone who "intentionally disrupts online communities" [259],

or "takes pleasure in upsetting others" [169], or more broadly as a person engaging in "negatively marked online behavior" [140] or that "makes trouble" for a discussion forums' stakeholders [42]. In this work, similar to the latter studies, we adopt a definition of trolling that includes flaming, griefing, swearing, or personal attacks, including behavior outside the acceptable bounds defined by several community guidelines for discussion forums [83, 97, 133].[1] In our experiment, we code posts manually for trolling behavior. In our longitudinal data analysis, we use posts that were flagged for unacceptable behavior as a proxy for trolling behavior.

Who engages in trolling behavior? One popular recurring narrative in the media suggests that trolling behavior comes from *trolls*: a small number of particularly sociopathic individuals [246, 259]. Several studies on trolling have focused on a small number of individuals [22, 42, 143, 262]; other work shows that there may be predisposing personality (e.g., sadism [51]) and biological traits (e.g., low baseline arousal [243]) to aggression and trolling. That is, trolls are born, not made.

Even so, the prevalence of antisocial behavior online suggests that these trolls, being relatively uncommon, are not responsible for all instances of trolling. Could ordinary individuals also engage in trolling behavior, even if temporarily? People are less inhibited in their online interactions [282]. The relative anonymity afforded by many platforms also deindividualizes and reduces accountability [319], decreasing comment quality [167]. This disinhibition effect suggests that people, in online settings, can be more easily influenced to act antisocially. Thus, rather than assume that only trolls engage in trolling behavior, we ask: *RQ: Can situational factors trigger trolling behavior?*

[1] In contrast to cyberbullying, defined as behavior that is repeated, intended to harm, and targeted at specific individuals [272], this definition of trolling encompasses a broader set of behaviors that may be one-off, unintentional, or untargeted.

5.2.2 Causes of antisocial behavior

Previous work has suggested several motivations for engaging in antisocial behavior: out of boredom [292], for fun [262], or to vent [189]. Still, this work has been largely qualitative and non-causal, and whether these motivations apply to the general population remains largely unknown. Out of this broad literature, we identify two possible trigger mechanisms of trolling – mood and discussion context – and try to establish their effects using both a controlled experiment and a large-scale longitudinal analysis.

Mood. Bad moods may play a role in how a person later acts. Negative mood correlates with reduced satisfaction with life [261], impairs self-regulation [190], and leads to less favorable impressions of others [109]. Similarly, exposure to unrelated aversive events (e.g., higher temperatures [254] or secondhand smoke [156]) increases aggression towards others. An interview study found that people thought that malicious comments by others resulted from "anger and feelings of inferiority" [189].

Nonetheless, negative moods elicit greater attention to detail and higher logical consistency [260], which suggests that people in a bad mood may provide more thoughtful commentary. Prior work is also mixed on how affect influences prejudice and stereotyping. Both positive [45] and negative affect [130] can increase stereotyping, and thus trigger trolling [143]. Still, we expect the negative effects of negative mood in social contexts to outweigh these other factors.

Circumstances that influence mood may also modify the rate of trolling. For instance, mood changes with the time of day or day of week [121]. As negative mood rises at the start of the week, and late at night, trolling may vary similarly. "Time-outs" or allowing for a period of calming down [163] can also reduce aggression – users who wait longer to post after a bout of trolling may also be less susceptible to future trolling. Thus, we may be able to observe how mood affects trolling, directly through experimentation, and indirectly through observing factors that influence mood:

H1: Negative mood increases a user's likelihood of trolling.

Discussion context. A discussion's context may also affect what people contribute. The discussion starter influences the direction of the rest of the discussion [139]. Qualitative analyses suggest that people think online commenters follow suit in posting positive (or negative) comments [189]. More generally, standards of behavior (i.e., social norms) are inferred from the immediate environment [64, 79, 215]. Closer to our work is an experiment that demonstrated that less thoughtful posts led to less thoughtful responses [281]. We extend this work by studying amplified states of antisocial behavior (i.e., trolling) in both experimental and observational settings.

On the other hand, users may not necessarily react to trolling with more trolling. An experiment that manipulated the initial votes an article received found that initial downvotes tended to be corrected by the community [220]. Some users respond to trolling with sympathy or understanding [22], or apologies or joking [188]. Still, such responses are rarer [22].

Another aspect of a discussion's context is the subject of discussion. In the case of discussions on news sites, the topic of an article can affect the amount of abusive comments posted [112]. Overall, we expect that previous troll posts, regardless of who wrote them, are likely to result in more subsequent trolling, and that the topic of discussion also plays a role:

H2: The discussion context (e.g., prior troll posts by other users) affects a user's likelihood of trolling.

5.2.3 Influence and antisocial behavior

That people can be influenced by environmental factors suggests that trolling could be contagious – a single user's outburst might lead to multiple users participating in a flame war. Prior work on social influence [28] has demonstrated multiple examples of herding behavior, or that people are likely to take similar actions to previous others [80, 214, 319]. Similarly, emotions and behavior can be transferred from person to person [31, 57, 116, 176, 303]. More relevant is work showing that getting downvoted

leads people to downvote others more and post content that gets further downvoted in the future [72].

These studies generally point toward a "Broken Windows" hypothesis, which postulates that untended behavior can lead to the breakdown of a community [306]. As an unfixed broken window may create a perception of unruliness, comments made in poor taste may invite worse comments. If antisocial behavior becomes the norm, this can lead a community to further perpetuate it despite its undesirability [305].

Further evidence for the impact of antisocial behavior stems from research on negativity bias – that negative traits or events tend to dominate positive ones. Negative entities are more contagious than positive ones [255], and bad impressions are quicker to form and more resistant to disconfirmation [35]. Thus, we expect antisocial behavior is particularly likely to be influential, and likely to persist. Altogether, we hypothesize:

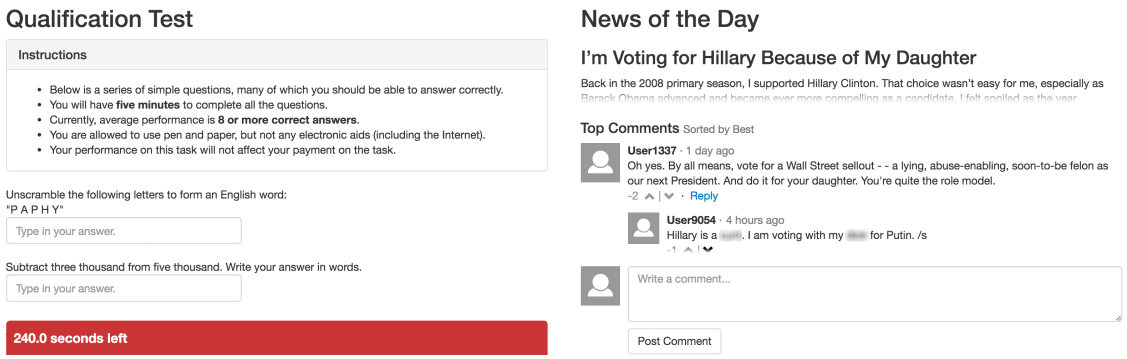
H3: Trolling behavior can spread from user to user.

We test H1 and H2 using a controlled experiment, then verify and extend our results with an analysis of discussions on CNN.com. We test H3 by studying the evolution of discussions on CNN.com, finally developing an overall model for how trolling might spread from person to person.

5.3 Experiment: Mood and Discussion Context

To establish the effects of mood and discussion context, we deployed an experiment designed to replicate a typical online discussion of a news article.

Specifically, we measured the effect of mood and discussion context on the quality of the resulting discussion across two factors: a) PosMood or NegMood: participants were either exposed to an unrelated positive or negative prior stimulus (which in turn affected their prevailing mood), and b) PosContext or NegContext: the initial



(a) (b)

Figure 5.1: To understand how a person's mood and discussion's context (i.e., prior troll posts) affected the quality of a discussion, we conducted an experiment that varied (a) how difficult a quiz, given prior to participation in the discussion, was, as well as (b) whether the initial posts in a discussion were troll posts or not.

posts in the discussion thread were either benign (or not troll-like), or troll-like. Thus, this was a two-by-two between-subjects design, with participants assigned in a round robin to each of the four conditions.

We evaluated discussion quality using two measures: a) trolling behavior, or whether participants wrote more troll-like posts, and b) affect, or how positive or negative the resulting discussion was, as measured using sentiment analysis.

If negative mood (NegMood) or troll posts (NegContext) affects the probability of trolling, we would expect these conditions to reduce discussion quality.

5.3.1 Experimental Setup

The experiment consisted of two main parts – a quiz, followed by a discussion – and was conducted on Amazon Mechanical Turk (AMT). Past work has also recruited workers to participate in experiments with online discussions [210]. Participants were restricted to residing in the US, only allowed to complete the experiment once, and compensated $2.00, for an hourly rate of $8.00. To avoid demand characteristics, participants were not told of the experiment's purpose prior, and were only instructed

to complete a quiz, and then participate in an online discussion. After the experiment, participants were debriefed and told of its purpose (i.e., to measure the impact of mood and trolling in discussions). The experimental protocol was reviewed and conducted under IRB Protocol #32738.

*Quiz (*PosMood *or* NegMood*).* The goal of the quiz was to see if participants' mood prior to participating in a discussion had an effect on subsequent trolling. Research on mood commonly involves giving people negative feedback on tasks that they perform in laboratory experiments regardless of their actual performance [173, 310, 127]. Adapting this to the context of AMT, where workers care about their performance on tasks and qualifications (which are necessary to perform many higher-paying tasks), participants were instructed to complete an experimental test qualification that was being considered for future use on AMT. They were told that their performance on the quiz would have no bearing on their payment at the end of the experiment.

The quiz consisted of 15 open-ended questions, and included logic, math, and word problems (e.g., word scrambles) (Figure 5.1a). In both conditions, participants were given five minutes to complete the quiz, after which all input fields were disabled and participants forced to move on. In both the PosMood and NegMood conditions, the composition and order of the types of questions remained the same. However, the NegMood condition was made up of questions that were substantially harder to answer within the time limit: for example, unscramble "DEANYON" (NegMood) vs. "PAPHY" (PosMood). At the end of the quiz, participants' answers were automatically scored, and their final score displayed to them. They were told whether they performed better, at, or worse than the "average", which was fixed at eight correct questions. Thus, participants were expected to perform well in the PosMood condition and receive positive feedback, and expected to perform poorly in the NegMood condition and receive negative feedback, being told that they were performing poorly, both absolutely and relatively to other users. While users in the PosMood condition can still perform poorly, and users in the NegMood condition perform well, this only reduces the differences later observed.

To measure participants' mood following the quiz, and acting as a manipulation check, participants then completed 65 Likert-scale questions on how they were feeling based on the Profile of Mood States (POMS) questionnaire [211], which quantifies mood on six axes such as anger and fatigue.

*Discussion (*PosContext *or* NegContext*).* Participants were then instructed to take part in an online discussion, and told that we were testing a comment ranking algorithm. Here, we showed participants an interface similar to what they might see on a news site — a short article, followed by a comments section. Users could leave comments, reply to others' comments, or upvote and downvote comments (Figure 5.1b). Participants were required to leave at least one comment, and told that their comments may be seen by other participants. Each participant was randomly assigned a username (e.g., User1234) when they commented. In this experiment, we showed participants an abridged version of an article arguing that women should vote for Hillary Clinton instead of Bernie Sanders in the Democratic primaries leading up to the 2016 US presidential election [157]. In the NegContext condition, the first three comments were troll posts, e.g.,:

> *Oh yes. By all means, vote for a Wall Street sellout – a lying, abuse-enabling, soon-to-be felon as our next President. And do it for your daughter. You're quite the role model.*

In the PosContext, they were more innocuous:

> *I'm a woman, and I don't think you should vote for a woman just because she is a woman. Vote for her because you believe she deserves it.*

These comments were abridged from real comments posted by users in comments in the original article, as well as other online discussion forums discussing the issue (e.g., Reddit).

To ensure that the effects we observed were not path-dependent (i.e., if a discussion breaks down by chance because of a single user), we created eight separate "universes" for each condition [257], for a total of 32 universes. Each universe was seeded with

the same comments, but were otherwise entirely independent. Participants were randomized between universes within each condition. Participants assigned to the same universe could see and respond to other participants who had commented prior, but not interact with participants from other universes.

Measuring discussion quality. We evaluated discussion quality in two ways: if subsequent posts written exhibited trolling behavior, or if they contained more negative affect. To evaluate whether a post was a troll post or not, two experts (including one of the authors) independently labeled posts as being troll or non-troll posts, blind to the experimental conditions, with disagreements resolved through discussion. Both experts reviewed CNN.com's community guidelines [83] for commenting – posts that were offensive, irrelevant, or designed to elicit an angry response, whether intentional or not, were labeled as trolling. To measure the negative affect of a post, we used LIWC [240] (Vader [152] gives similar results).

5.3.2 Results

667 participants (40% female, mean age 34.2, 54% Democrat, 25% Moderate, 21% Republican) completed the experiment, with an average of 21 participants in each universe. In aggregate, these workers contributed 791 posts (with an average of 37.8 words written per post) and 1392 votes.

Manipulation checks. First we sought to verify that the quiz did affect participants' mood. On average, participants in the PosMood condition obtained 11.2 out of 15 questions correct, performing above the stated "average" score of 8. In contrast, participants in the NegMood condition answered only an average of 1.9 questions correctly, performing significantly worse ($t(594)$=63.2, p<0.001 using an unequal variances t-test), and below the stated "average". Correspondingly, the post-quiz POMS questionnaire confirmed that participants in the NegMood condition experienced higher mood disturbance on all axes, with higher anger, confusion, depression, fatigue, and tension scores, and a lower vigor score ($t(534)$>7.0, p<0.001). Total mood

	Proportion of Troll Posts		Negative Affect (LIWC)	
	PosMood	NegMood	PosMood	NegMood
PosContext	35%	49%	1.1%	1.4%
NegContext	47%	**68%**	2.3%	**2.9%**

Table 5.1: The proportion of user-written posts that were labeled as trolling (and proportion of words with negative affect) was lowest in the (PosMood, PosContext) condition, and highest, and almost double, in the (NegMood, NegContext) condition (highlighted in bold).

disturbance, where higher scores correspond to more negative mood, was 12.2 for participants in the PosMood condition (comparable to a baseline level of disturbance measured among athletes [286]), and 40.8 in the NegMood condition. Thus, the quiz put participants into a more negative mood.

Verifying that the initial posts in the NegContext condition were perceived as being more troll-like than those in the PosContext condition, we found that the initial posts in the NegContext condition were less likely to be upvoted (36% vs. 90% upvoted for PosContext, $t(507)=15.7$, $p<0.001$).

Negative mood and negative context increase trolling behavior. Table 5.1 shows how the proportion of troll posts and negative affect (measured as the proportion of negative words) differ in each condition. The proportion of troll posts was highest in the (NegMood, NegContext) condition with 68% troll posts, drops in both the (NegMood, PosContext) and (PosMood, NegContext) conditions with 47% and 49% each, and is lowest in the (PosMood, PosContext) condition with 35%. For negative affect, we observe similar differences.

Fitting a mixed effects logistic regression model, with the two conditions as fixed effects, an interaction between the two conditions, user as a random effect, and whether a contributed post was trolling or not as the outcome variable, we do observe a significant effect of both NegMood and NegContext ($p<0.05$) (Table 5.2). These results confirm both H1 and H2, that negative mood and the discussion context (i.e.,

Fixed Effects	Coef.	SE	z
(Intercept)	-0.70^{***}	0.17	-4.23
NegMood	0.64^{**}	0.24	2.66
NegContext	0.52^{*}	0.23	2.38
NegMood × NegContext	0.41	0.33	1.23
Random Effects	Var.	SE	
User	0.41	0.64	

Table 5.2: A mixed effects logistic regression reveals a significant effect of both NegMood and NegContext on troll posts (*: $p<0.05$, **: $p<0.01$, ***: $p<0.001$). In other words, both negative mood and the presence of initial troll posts increases the probability of trolling.

prior troll posts) increase a user's likelihood of trolling. Negative mood increases the odds of trolling by 89%, and the presence of prior troll posts increases the odds by 68%. A mixed model using MCMC revealed similar effects ($p<0.05$), and controlling for universe, gender, age, or political affiliation also gave similar results. Further, the effect of a post's position in the discussion on trolling was not significant, suggesting that trolling tends to persist in the discussion.

With the proportion of words with negative affect as the outcome variable, we observed a significant effect of NegContext ($p<0.05$), but not of NegMood – such measures may not accurately capture types of trolling such as sarcasm or off-topic posting. There was no significant effect of either factor on positive affect.

Examples of troll posts. Contributed troll posts comprised a relatively wide range of antisocial behavior: from outright swearing ("What a dumb c***") and personal attacks ("You're and idiot and one of the things that's wrong with this country.") to veiled insults ("Hillary isn't half the man Bernie is!!! lol"), sarcasm ("You sound very white, and very male. Must be nice."), and off-topic statements ("I think Ted Cruz has a very good chance of becoming president."). In contrast, non-troll posts tended to be more measured, regardless of whether they agreed with the article ("Honestly I agree too. I think too many people vote for someone who they identify with rather

than someone who would be most qualified.").

Other results. We observed trends in the data. Both conditions reduced the number of words written relative to the control condition: 44 words written in the (PosMood, PosContext) vs. 29 words written in the (NegMood, NegContext) condition. Also, the percentage of upvotes on posts written by other users (i.e., excluding the initial seed posts) was lower: 79% in the (PosMood, PosContext) condition vs. 75% in the (NegMood, NegContext) condition. While suggestive, neither effect was significant.

Discussion. Why did NegContext and NegMood increase the rate of trolling? Drawing on prior research explaining the mechanism of contagion [303], participants may have an initial negative reaction to reading the article, but are unlikely to bluntly externalize them because of self-control or environmental cues. NegContext provides evidence that others had similar reactions, making it more acceptable to also express them. NegMood further accentuates any perceived negativity from reading the article and reduces self-inhibition [190], making participants more likely to act out.

Limitations. In this experiment, like prior work [210, 281], we recruited participants to participate in an online discussion, and required each to post at least one comment. While this enables us isolate both mood and discussion context (which is difficult to control for in a live Reddit discussion for example) and further allows us to debrief participants afterwards, payment may alter the incentives to participate in the discussion. Users also were commenting pseudonymously via randomly generated usernames, which may reduce overall comment quality [167]. Different initial posts may also elicit different subsequent posts. While our analyses did not reveal significant effects of demographic factors, future work could further examine their impact on trolling. For example, men may be more susceptible to trolling as they tend to be more aggressive [38]. Anecdotally, several users who identified as Republican trolled the discussion with irrelevant mentions of Donald Trump (e.g., "I'm a White man and I'm definitely voting for Donald Trump!!!"). Understanding the effects of different

types of trolling (e.g., swearing vs. sarcasm) and user motivations for such trolling (e.g., just to rile others up) also remains future work. Last, different articles may be trolled to different extents [112], so we examine the effect of article topic in our subsequent analyses.

Overall, we find that both mood and discussion context significantly affect a user's likelihood of engaging in trolling behavior. For such effects to be observable, a substantial proportion of the population must have been susceptible to trolling, rather than only a small fraction of atypical users – suggesting that trolling can be generally induced. But do these results generalize to real-world online discussions? In the subsequent sections, we verify and extend our results with an analysis of CNN.com, a large online news discussion community. After describing this dataset, we study how trolling behavior tracks known daily mood patterns, and how mood persists across multiple discussions. We again find that the initial posts of discussions have a significant effect on subsequent posts, and study the impact of the volume and ordering of multiple troll posts on subsequent trolling. Extending our analysis of discussion context to include the accompanying article's topic, we find that it too mediates trolling behavior.

5.4 Data: Introduction

CNN.com is a popular American news website where editors and journalists write articles on a variety of topics (e.g., politics and technology), which users can then discuss. In addition to writing and replying to posts, users can *up-* and *down-vote*, as well as *flag* posts (typically for abuse or violations of the community guidelines [83]). Moderators can also *delete* posts or even *ban* users, in keeping with these guidelines. Disqus, a commenting platform that hosted these discussions on CNN.com, provided us with a complete trace of user activity from December 2012 to August 2013, consisting of 865,248 users (20,197 banned), 16,470 discussions, and 16,500,603 posts, of which 571,662 (3.5%) were flagged and 3,801,774 (23%) were deleted. Out of all

flagged posts, 26% were made by users with no prior record of flagging in previous discussions; also, out of all users with flagged posts who authored at least ten posts, 40% had less than 3.5% of their posts flagged (the baseline probability of a random post being flagged on CNN). These observations suggest that ordinary users are responsible for a significant amount of trolling behavior, and that many may have just been having a bad day.

In studying behavior on CNN.com, we consider two main units of analysis: a) a *discussion*, or all the posts that follow a given news article, and b) a *sub-discussion*, or a top-level post and any replies to that post. We make this distinction as discussions may reach thousands of posts, making it likely that users may post in a discussion without reading any previous responses. In contrast, a sub-discussion necessarily involves replying to a previous post, and would allow us to better study the effects of people reading and responding to each other.

In our subsequent analyses, we filter banned users (of which many tend to be clearly identifiable trolls [73]), as well as any users who had all of their posts deleted, as we are primarily interested in studying the effects of mood and discussion context on the general population.

We use flagged posts (posts that CNN.com users marked for violating community guidelines) as our primary measure of trolling behavior. In contrast, moderator deletions are typically incomplete: moderators miss some legitimate troll behavior and tend to delete entire discussions as opposed to individual posts. Likewise, written negative affect misses sarcasm and other trolling behaviors that do not involve common negative words, and downvoting may simply indicate disagreement. To validate this approach, two experts (including one of the authors) labeled 500 posts (250 flagged) sampled at random, blind to whether each post was flagged, using the same criteria for trolling as for the experiment. Comparing the expert labels with post flags from the dataset, we obtained a precision of 0.66 and recall of 0.94, suggesting that while some troll posts remain unflagged, almost all flagged posts are troll posts. In other words, while instances of trolling behavior go unnoticed (or are ignored), when a post

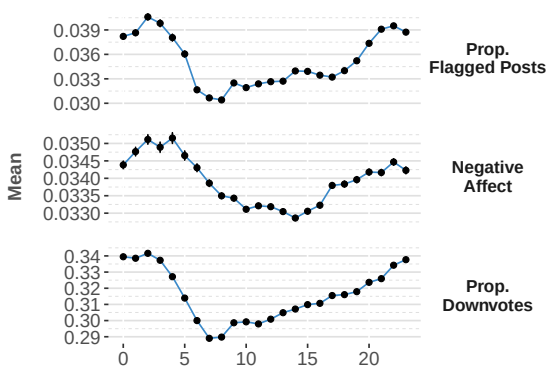
is flagged, it is highly likely that trolling behavior did occur. So, we use flagged posts as a primary estimate of trolling behavior in our analyses, complementing our analysis with other signals such as negative affect and downvotes. These signals are correlated: flagged posts are more likely than non-flagged posts to have greater negative affect (3.7% vs. 3.4% of words, Cohen's d=0.06, t=40, p<0.001), be downvoted (58% vs. 30% of votes, d=0.76, t=531, p<0.001), or be deleted by a moderator (79% vs. 21% of posts, d=1.4, t=1050, p<0.001).

5.5 Data: Understanding Mood

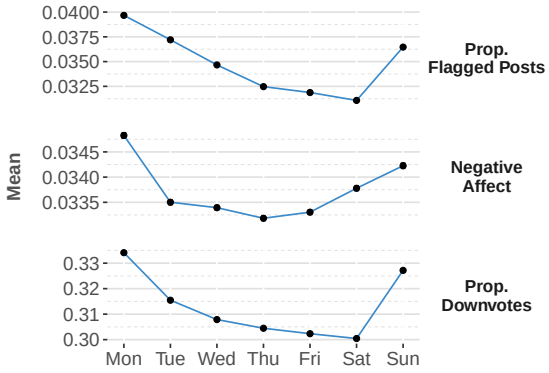
In the earlier experiment, we showed that bad mood increases the probability of trolling. In this work, using large-scale and longitudinal observational data, we verify and expand on this result. While we cannot measure mood directly, we can study its known correlates. Seasonality influences mood [121], so we study how trolling behavior also changes with the time of day or day of week. Aggression can linger beyond an initial unpleasant event [156], thus we also study how trolling behavior persists as a user participates in multiple discussions.

5.5.1 Happy in the day, sad at night

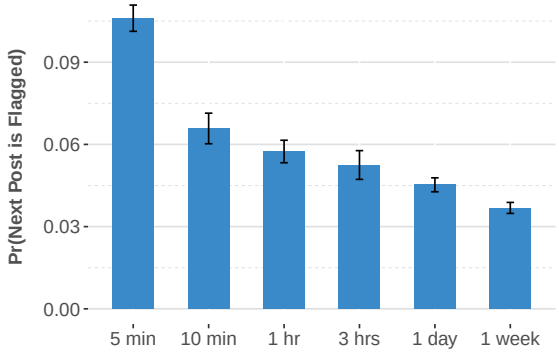
Prior work that studied changes in linguistic affect on Twitter demonstrated that mood changes with the time of the day, and with the day of the week – positive affect peaks in the morning, and during weekends [121]. If mood changes with time, could trolling be similarly affected? Are people more likely to troll later in the day, and on weekdays? To evaluate the impact of the time of day or day of week on mood and trolling behavior, we track several measures that may indicate troll-like behavior: a) the proportion of flagged posts (or posts reported by other users as being abusive), b) negative affect, and c) the proportion of downvotes on posts (or the average fraction of downvotes on posts that received at least one vote).



(a) Time of Day (EDT)



(b) Day of Week



(c) Time Since Last Post

Figure 5.2: Like negative mood, indicators of trolling peak (a) late at night, and (b) early in the work week, supporting a relation between mood and trolling. Further, (c) the shorter the time between a user's subsequent posts in unrelated discussions, where the first post is flagged, the more likely the second will also be flagged, suggesting that negative mood may persist for some time.

Figures 5.2a and 5.2b show how each of these measures changes with the time of day and day of week, respectively, across all posts. Our findings corroborate prior work – the proportion of flagged posts, negative affect, and the proportion of downvotes are all lowest in the morning, and highest in the evening, aligning with when mood is worst [121]. These measures also peak on Monday (the start of the work week in the US).

Still, trolls may simply wake up later than normal users, or post on different days. To understand how the time of day and day of week affect the same user, we compare

these measures for the same user in two different time periods: from 6 am to 12 pm and from 11 pm to 5 am, and on two different days: Monday and Friday (i.e., early or late in the work week). A paired t-test reveals a small, but significant increase in negative behavior between 11 pm and 5 am (flagged posts: 4.1% vs. 4.3%, d=0.01, $t(106300)$=2.79, p<0.01; negative affect: 3.3% vs. 3.4%, $t(106220)$=3.44, d=0.01, p<0.01; downvotes: 20.6% vs. 21.4%, d=0.02, $t(26390)$=2.46, p<0.05). Posts made on Monday also show more negative behavior than posts made on Friday ($d\geq$0.02, t>2.5, p<0.05). While these effects may also be influenced by the type of news that gets posted at specific times or days, limiting our analysis to just news articles categorized as "US" or "World", the two largest sections, we continue to observe similar results.

Thus, even without direct user mood measurements, patterns of trolling behavior correspond predictably with mood.

5.5.2 Anger begets more anger

Negative mood can persist beyond the events that brought about those feelings [161]. If trolling is dependent on mood, we may be able to observe the aftermath of user outbursts, where negative mood might spill over from prior discussions into subsequent, unrelated ones, just as our experiment showed that negative mood that resulted from doing poorly on a quiz affected later commenting in a discussion. Further, we may also differentiate the effects that stem from actively engaging in negative behavior in the past, versus simply being exposed to negative behavior. Correspondingly, we ask two questions, and answer them in turn. First, a) if a user wrote a troll post in a prior discussion, how does that affect their probability of trolling in a subsequent, unrelated discussion? At the same time, we might also observe indirect effects of trolling: b) if a user participated in a discussion where trolling occurred, but did not engage in trolling behavior themselves, how does that affect their probability of trolling in a subsequent, unrelated discussion?

To answer the former, for a given discussion, we sample two users at random, where one had a post which was flagged, and where one had a post which was not flagged. We ensure that these two users made at least one post prior to participating in the discussion, and match users on the total number of posts they wrote prior to the discussion. As we are interested in these effects on ordinary users, we also ensure that neither of these users have had any of their posts flagged in the past. We then compare the likelihood of each user's next post in a new discussion also being flagged. We find that users who had a post flagged in a prior discussion were twice as likely to troll in their next post in a different discussion (4.6% vs. 2.1%, d=0.14, $t(4641)$=6.8, p<0.001) (Figure 5.3a). We obtain similar results even when requiring these users to also have no prior deleted posts or longer histories (e.g., if they have written at least five posts prior to the discussion).

Next, we examine the indirect effect of participating in a "bad" discussion, even when the user does not directly engage in trolling behavior. We again sample two users from the same discussion, but where each user participated in a different sub-discussion: one sub-discussion had at least one other post by another user flagged, and the other sub-discussion had no flagged posts. Again, we match users on the number of posts they wrote in the past, and ensure that these users have no prior flagged posts (including in the sampled discussions). We then compare the likelihood of each user's next post in a new discussion being flagged. Here, we also find that users who participated in a prior discussion with at least one flagged post were significantly more likely to subsequently author a post in an new discussion that would be flagged (Figure 5.3b). However, this effect is significantly weaker (2.2% vs. 1.7%, d=0.04, $t(7321)$=2.7, p<0.01).

Thus, both trolling in a past discussion, as well as participating in a discussion where trolling occurred, can affect whether a user trolls in the future discussion. These results suggest that negative mood can persist and transmit trolling norms and behavior across multiple discussions, where there is no similar context to draw on. As none of the users we analyzed had prior flagged posts, this effect is unlikely to arise simply because some users were just trolls in general.

5.5.3 Time heals all wounds

One typical anger management strategy is to use a "time-out" to calm down [163]. Thus, could we minimize negative mood carrying over to new discussions by having users wait longer before making new posts? Assuming that a user is in a negative mood (as indicated by writing a post that is flagged), the time elapsed until the user's next post may correlate with the likelihood of subsequent trolling. In other words, we might expect that the longer time the time between posts, the greater the temporal distance from the origin of the negative mood, and hence the lower the likelihood of trolling.

Figure 5.2c shows how the probability of a user's next post being flagged changes with the time since that user's last post, assuming that the previous post was flagged. So as not to confuse the effects of the initial post's discussion context, we ensure that the user's next post is made *in a new discussion with different other users*. The probability of being flagged is high when the time between these two subsequent posts is short (five minutes or less), suggesting that a user might still be in a negative mood persisting from the initial post. As more time passes, even just ten minutes, the probability of being flagged gradually decreases. Nonetheless, users with better impulse control may wait longer before posting again if they are angry, and isolating this effect would be future work. Our findings here lend credence to the rate-limiting of posts that some forums have introduced [17].

5.6 Data: Understanding Discussion Context

From our experiment, we identified mood and discussion context as influencing trolling. The previous section verified and extended our results on mood; in this section, we do the same for discussion context. In particular, we show that posts are more likely to be flagged if others' prior posts were also flagged. Further, the number and ordering of flagged posts in a discussion affects the probability of subsequent trolling, as does

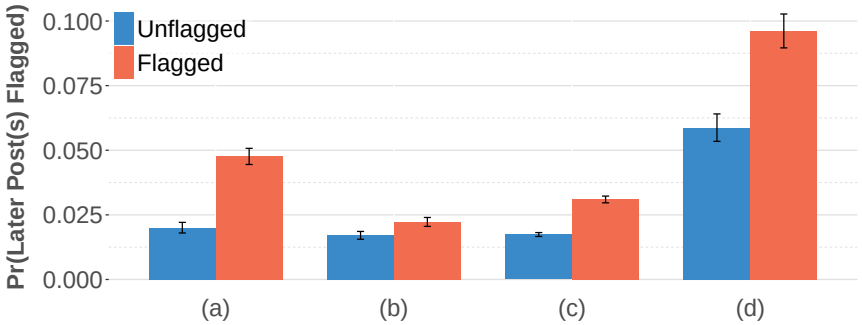


Figure 5.3: Suggesting that negative mood may persist across discussions, users with no prior history of flagged posts, who either (a) make a post in a prior unrelated discussion that is flagged, or (b) simply participates in a sub-discussion in a prior discussion with at least one flagged post, without themselves being flagged, are more likely to be subsequently flagged in the next discussion they participate in. Demonstrating the effect of discussion context, (c) discussions that begin with a flagged post are more likely to have a greater proportion of flagged posts by other users later on, as do (d) sub-discussions that begin with a flagged post.

the topic of the discussion.

5.6.1 "FirST!!1"

How strongly do the initial posts to a discussion affect the likelihood of subsequent posts to troll? To measure the effect of the initial posts on subsequent discussions, we first identified discussions of at least 20 posts, separating them into those with their first post flagged and those without their first post flagged. We then used propensity score matching to create matched pairs of discussions where the topic of the article, the day of week the article was posted, and the total number of posts are controlled for [252]. Thus, we end up with pairs of discussions on the same topic, started on the same day of the week, and with similar popularity, but where one discussion had its first post flagged, while the other did not. We then compare the probability of the subsequent posts in the discussion being flagged. As we were interested in the impact of the initial post on other ordinary users, we excluded any posts written by the user who made the initial post, posts by users who replied (directly or indirectly) to that

post, and posts by users with prior flagged or deleted posts in previous discussions.

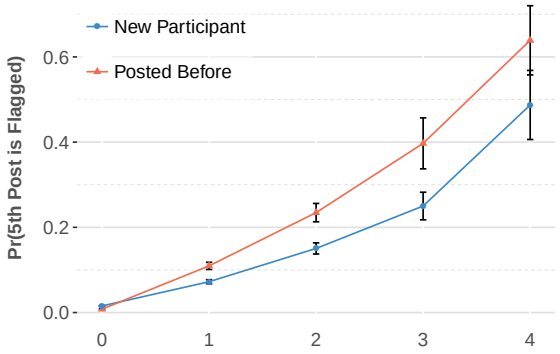
After an initial flagged post, we find that subsequent posts by other users were more likely to be flagged, than if the initial post was not flagged (3.1% vs. 1.7%, d=0.32, $t(1545)$=9.1, p<0.001) (Figure 5.3c). This difference remains significant even when only considering posts made in the second half of a discussion (2.1% vs. 1.3%, d=0.19, $t(1545)$=5.4, p<0.001). Comparing discussions where the first three posts were all flagged to those where none of these posts were flagged (similar to NegContext vs. PosContext in our experiment), the gap widens (7.1% vs. 1.7%, d=0.61, $t(113)$=4.6, p<0.001).

Nonetheless, as these different discussions were on different articles, some articles, even within the same topic, may have been more inflammatory, increasing the overall rate of flagging. To control for the article being discussed, we also look at sub-discussions (a top-level post and all of its replies) within the same discussion. Sub-discussions tend to be closer to actual conversations between users as each subsequent post is an explicit reply to another post in the chain, as opposed to considering the discussion as a whole where users can simply leave a comment without reading or responding to anyone else. From each discussion we select two sub-discussions at random, where one sub-discussion's top-level post was flagged, and where the other's was not, and only considered posts not written by the users who started these sub-discussions. Again, we find that sub-discussions whose top-level posts were flagged were significantly more likely to result in more flagging later in that sub-discussion (9.6% vs. 5.9%, d=0.16, $t(501)$=3.9, p<0.001) (Figure 5.3d).

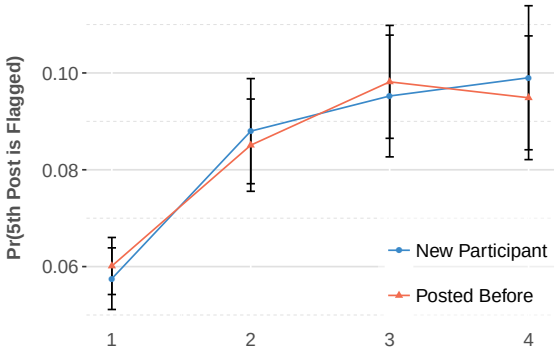
Altogether, these results suggest that the initial posts in a discussion set a strong, lasting precedent for later trolling.

5.6.2 From bad to worse: sequences of trolling

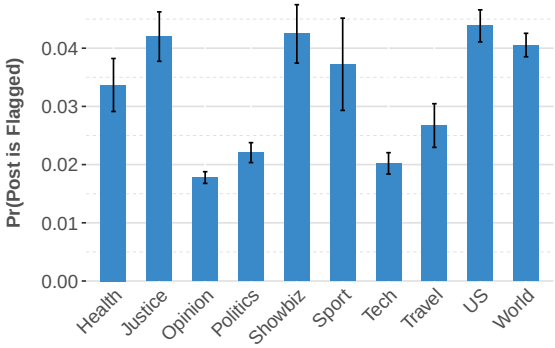
By analyzing the volume and ordering of troll posts in a discussion, we can better understand how discussion context and trolling behavior interact. Here, we study



(a) # of Prior Flagged Posts



(b) Position of Flagged Post



(c) Flagged Posts by Topic

Figure 5.4: In discussions with at least five posts, (a) the probability that a post is flagged monotonically increases with the number of prior flagged posts in the discussion. (b) If only one of the first four posts was flagged, the fifth post is more likely to be flagged if that flagged post is closer in position. (c) The topic of a discussion also influences the probability of a post being flagged.

sub-discussions at least five posts in length, and separately consider posts written by users new to the sub-discussion and posts written by users who have posted before in the sub-discussion to control for the impact of having already participated in the discussion.

Do more troll posts increase the likelihood of future troll posts? Figure 5.4a shows that as the number of flagged posts among the first four posts increases, the probability that the fifth post is also flagged increases monotonically. With no prior flagged posts, the chance of the fifth post by a new user to the sub-discussion being flagged

is just 2%; with one other flagged post, this jumps to 7%; with four flagged posts, the odds of the fifth post also being flagged are almost one to one (49%). These pairwise differences are all significant with a Holm correction ($\chi^2(1)>7.6$, $p<0.01$). We observe similar trends for users new to the sub-discussion, as well as users that had posted previously, with the latter group of users more likely to be subsequently flagged.

Further, does a troll post made later in a discussion, and closer to where a user's post will show up, have a greater impact than a troll post made earlier on? Here, we look at discussions of at least five posts where there was exactly one flagged post among the first four, and where that flagged post was not written by the fifth post's author. In Figure 5.4b, the closer in position the flagged post is to the fifth post, the more likely that post is to be flagged. For both groups of users, the fifth post in a discussion is more likely to be flagged if the fourth post was flagged, as opposed to the first ($\chi^2(1)>6.9$, $p<0.01$).

Beyond the presence of troll posts, their conspicuousness in discussions substantially affects if new discussants troll as well. These findings, together with our previous results showing how simply participating in a previous discussion having a flagged post raises the likelihood of future trolling behavior, support H3: that trolling behavior spreads from user to user.

5.6.3 Hot-button issues push users' buttons?

How does the subject of a discussion affect the rate of trolling? Controversial topics (e.g., gender, GMOs, race, religion, or war) may divide a community [197], and thus lead to more trolling. Figure 5.4c shows the average rate of flagged posts of articles belonging to different sections of CNN.com.

Post flagging is more frequent in the health, justice, showbiz, sport, US, and world sections (near 4%), and less frequent in the opinion, politics, tech, and travel sections (near 2%). Flagging may be more common in the health, justice, US, and world

sections because these sections tend to cover controversial issues: a linear regression unigram model using the titles of articles to predict the proportion of flagged posts revealed that "Nidal" and "Hasan" (the perpetrator of the 2009 Fort Hood shooting) were among the most predictive words in the justice section. For the showbiz and sport sections, inter-group conflict may have a strong effect (e.g., fans of opposing teams) [264]. Though political issues in the US may appear polarizing, the politics section has one of the lowest rates of post flagging, similar to tech. Still, a deeper analysis of the interplay of these factors (e.g., personal values, group membership, and topic) with trolling remains future work.

The relatively large variation here suggests that the topic of a discussion influences the baseline rate of trolling, where hot-button topics spark more troll posts.

5.6.4 Summary

Through experimentation and data analysis, we find that situational factors such as mood and discussion context can induce trolling behavior, answering our main research question (RQ). Bad mood induces trolling, and trolling, like mood, varies with time of day and day of week; bad mood may also persist across discussions, but its effect diminishes with time. Prior troll posts in a discussion increase the likelihood of future troll posts (with an additive effect the more troll posts there are), as do more controversial topics of discussion.

5.7 A Model of How Trolling Spreads

Thus far, our investigation sought to understand whether ordinary users engage in trolling behavior. In contrast, prior work suggested that trolling is largely driven by a small population of trolls (i.e., by intrinsic characteristics such as personality), and our evidence suggests complementary hypotheses – that mood and discussion context

also affect trolling behavior. In this section, we construct a combined predictive model to understand the relative strengths of each explanation.

We model each explanation through features in the CNN.com dataset. First, the impact of mood on trolling behavior can be modeled indirectly using *seasonality*, as expressed through time of day and day of week; and a user's *recent posting history* (outside of the current discussion), in terms of the time elapsed since the last post and whether the user's previous post was flagged. Second, the effect of discussion context can be modeled using the *previous posts* that precede a user's in a discussion (whether any of the previous five posts in the discussion were flagged, and if they were written by the same user); and the *topic* of discussion (e.g., politics). Third, to evaluate if trolling may be innate, we use a user's *User ID* to learn each user's base propensity to troll, and the user's *overall history* of prior trolling (the total number and proportion of flagged posts accumulated).

Our prediction task is to guess whether a user will write a post that will get flagged, given features relating to the discussion or user. We sampled posts from discussions at random (N=116,026), and balance the set of users whose posts are later flagged and users whose posts are not flagged, so that random guessing results in 50% accuracy. To understand trolling behavior across all users, this analysis was not restricted to users who did not have their posts previously flagged. We use a logistic regression classifier, one-hot encoding features (e.g., time of day) as appropriate. A random forest classifier gives empirically similar results.

Our results suggest that trolling is better explained as situational (i.e., a result of the user's environment) than as innate (i.e., an inherent trait). Table 5.3 describes performance on this prediction task for different sets of features. Features relating to discussion context perform best (AUC=0.74), hinting that context alone is sufficient in predicting trolling behavior; the individually most predictive feature was whether the previous post in the discussion was flagged. Discussion topic was somewhat informative (0.58), with the most predictive feature being if the post was in the opinion section. In the experiment, mood produced a stronger effect than discussion context.

Feature Set	AUC
Mood	
Seasonality (31)	0.53
Recent User History (4)	0.60
Discussion Context	
Previous Posts (15)	0.74
Article Topic (13)	0.58
User-specific	
Overall User History (2)	0.66
User ID (45895)	0.66
Combined	
Previous Posts + Recent User History (19)	0.77
All Features	0.78

Table 5.3: In predicting trolling in a discussion, features relating to the discussion's context are most informative, followed by user-specific and mood features. This suggests that while some users are inherently more likely to troll, the context of a discussion plays a greater role in whether trolling actually occurs. The number of binary features is in parentheses.

However, here we cannot measure mood directly, so its feature sets (seasonality and recent user history) were weaker (0.60 and 0.53 respectively). Most predictive was if the user's last post in a different discussion was flagged, and if the post was written on Friday. Modeling each user's probability of trolling individually, or by measuring all flagged posts over their lifetime was moderately predictive (0.66 in either case). Further, user features do not improve performance beyond the using just the discussion context and a user's recent history. Combining previous posts with recent history (0.77) resulted in performance nearly as good as including all features (0.78). We continue to observe strong performance when restricting our analysis only to posts by users new to a discussion (0.75), or to users with no prior record of reported or deleted posts (0.70). In the latter case, it is difficult to detect trolling behavior without discussion context features ($<$0.56).

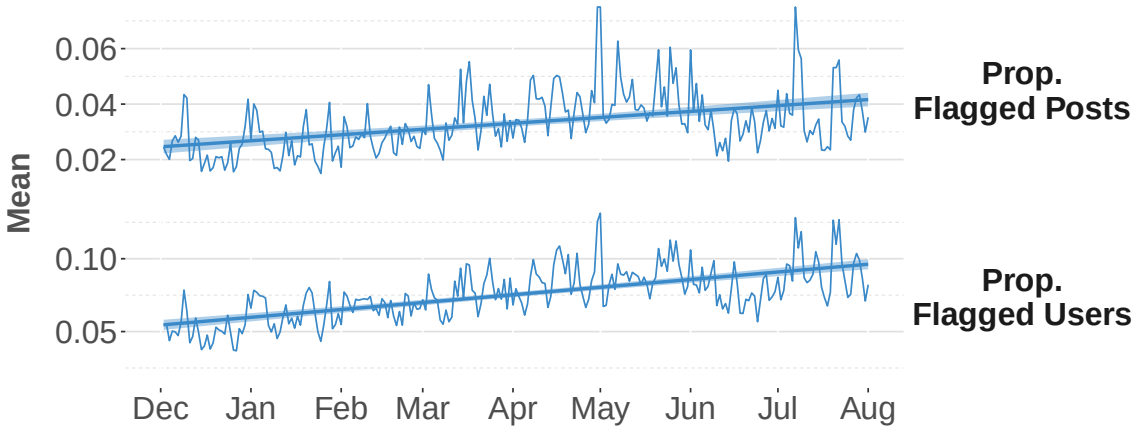


Figure 5.5: On CNN.com, the proportion of flagged posts, as well as users with flagged posts, is increasing over time, suggesting that trolling behavior can spread and be reinforced.

Overall, we find that the context in which a post is made is a strong predictor of a user later trolling, beyond their intrinsic propensity to troll. A user's recent posting history is also predictive, suggesting that mood carries over from previous discussions, and that past trolling predicts future trolling.

5.8 Discussion and Conclusion

While prior work suggests that some users may be born trolls and innately more likely to troll others, our results show that ordinary users will also troll when mood and discussion context prompt such behavior.

5.8.1 The spread of negativity

If trolling behavior can be induced, and can carry over from previous discussions, could such behavior cascade and lead to the community worsening overall over time? Figure 5.5 shows that on CNN.com, the proportion of flagged posts and proportion of users with flagged posts are rising over time. These upward trends suggest that

trolling behavior is becoming more common, and that a growing fraction of users are engaging in such behavior. Comparing posts made in the first half and second half of the CNN.com dataset, the proportion of flagged posts and proportion of users with flagged posts increased (0.03 vs. 0.04 and 0.09 vs. 0.12, $p<0.001$). There may be several explanations for this (e.g., that users joining later are more susceptible to trolling), but our findings, together with prior work showing that negative norms can be reinforced [305] and that downvoted users go on to downvote others [72], suggest that negative behavior can persist in and permeate a community when left unchecked.

5.8.2 Designing better discussion platforms

The continuing endurance of the idea that trolling is innate may be explained using the fundamental attribution error [253]: people tend to attribute a person's behavior to their internal characteristics rather than external factors – for example, interpreting snarky remarks as resulting from general mean-spiritedness (i.e., their disposition), rather than a bad day (i.e., the situation that may have led to such behavior). This line of reasoning may lead communities to incorrectly conclude that trolling is caused by people who are unquestionably trolls, and that trolling can be eradicated by banning these users. However, not only are some banned users likely to be ordinary users just having a bad day, but such an approach also does little to curb such situational trolling, which many ordinary users may be susceptible to. How might we design discussion platforms that minimize the spread of trolling behavior?

Inferring mood through recent posting behavior (e.g., if a user just participated in a heated debate) or other behavioral traces such as keystroke movements [174], and selectively enforcing measures such as post rate-limiting [17] may discourage users from posting in the heat of the moment. Allowing users to retract recently posted comments may help minimize regret [297]. Alternatively, reducing other sources of user frustration (e.g., poor interface design or slow loading times [56]) may further temper aggression.

Altering the context of a discussion (e.g., by hiding troll comments and prioritizing constructive ones) may increase the perception of civility, making users less likely to follow suit in trolling. To this end, one solution is to rank comments using user feedback, typically by allowing users to up- and downvote content, which reduces the likelihood of subsequent users encountering downvoted content. But though this approach is scalable, downvoting can cause users to post worse comments, perpetuating a negative feedback loop [72]. Selectively exposing feedback, where positive signals are public and negative signals are hidden, may enable context to be altered without adversely affecting user behavior. Community norms can also influence a discussion's context: reminders of ethical standards or past moral actions (e.g., if users had to sign a "no trolling" pledge before joining a community) can also increase future moral behavior [209, 227].

5.8.3 Limitations and future work

Though our results do suggest the overall effect of mood on trolling behavior, a more nuanced understanding of this relation should require improved signals of mood (e.g., by using behavioral traces as described earlier). Models of discussions that account for the reply structure [19], changes in sentiment [296], and the flow of ideas [231, 318] may provide deeper insight into the effect of context on trolling behavior.

Different trolling strategies may also vary in prevalence and severity (e.g., undirected swearing vs. targeted harassment and bullying). Understanding the effects of specific types of trolling may also allow us to design measures better targeted to the specific behaviors that may be more pertinent to deal with. The presence of social cues may also mediate the effect of these factors: while many online communities allow their users to use pseudonyms, reducing anonymity (e.g., through the addition of voice communication [91] or real name policies [76]) can reduce bad behavior such as swearing, but may also reduce the overall likelihood of participation [76]. Finally, differentiating the impact of a troll post and the intent of its author (e.g., did its writer intend to hurt others, or were they just expressing a different viewpoint? [189]) may

help separate undesirable individuals from those who just need help communicating their ideas appropriately.

Future work could also distinguish different types of users who end up trolling. Prior work that studied users banned from communities found two distinct groups – users whose posts were consistently deleted by moderators, and those whose posts only started to get deleted just before they were banned [73]. Our findings suggest that the former type of trolling may have been innate (i.e., the user was constantly trolling), while the latter type of trolling may have been situational (i.e., the user was involved in a heated discussion).

5.8.4 Conclusion

Trolling stems from both innate and situational factors – where prior work has discussed the former, this work focuses on the latter, and reveals that both mood and discussion context affect trolling behavior. This suggests the importance of different design affordances to manage either type of trolling. Rather than banning all users who troll and violate community norms, also considering measures that mitigate the situational factors that lead to trolling may better reflect the reality of how trolling occurs.

Chapter 6

The Spread of Antisocial Behavior

In Chapter 5, we showed how trolling is influenced by both a person's mood and the context of a discussion. In this chapter, we now analyze how social feedback mechanisms (e.g., voting) on many online platforms allow negative behavior to propagate, potentially amplifying the effects of negative behavior. Many social media systems rely on such feedback mechanisms for personalization, ranking, and content filtering. However, when users evaluate content contributed by fellow users (e.g., by liking a post or voting on a comment), these evaluations create complex social feedback effects.

Here, we investigate how votes on a piece of content affect its author's future behavior. By studying four large comment-based news communities, we find that negative feedback leads to significant behavioral changes that are detrimental to the community. Not only do authors of negatively-evaluated content contribute more, but also their future posts are of lower quality, and are perceived by the community as such. Moreover, these authors are more likely to subsequently evaluate their fellow users negatively, percolating these effects through the community. In contrast, positive feedback does not carry similar effects, and neither encourages rewarded authors to write more, nor improves the quality of their posts. Interestingly, the authors that receive no feedback are most likely to leave a community. Furthermore, a structural

analysis of the voter network reveals that evaluations polarize the community the most when positive and negative votes are equally split.

In summary, looking beyond trolling, we show, perhaps unfortunately, how the design of online discussions can also propagate negative behavior and exacerbate its spread. We find that downvoting, a common mechanism in many online communities, exacerbates negative behavior. People not only write worse after being downvoted, but they are also perceived worse by the community-at-large independent of what they write.

The studies covered in these two chapters suggest how negative behavior can lead to other users also behaving negatively as well, and demonstrate how antisocial behavior may spread through a network if left unchecked. They also show how multi-methods analyses can reveal new insights about human behavior. In the previous chapter, we combined a controlled online experiment with large data analysis of a real online community, which enables us to establish causality, as well as demonstrate both the replicability and ecological validity of these findings. In this chapter, we adopt quasi-experimental approach and develop a measure of post quality that is validated by crowdsourced labels, and establish findings that generalize across multiple communities.

6.1 Introduction

The ability of users to rate content and provide feedback is a defining characteristic of today's social media systems. These ratings enable the discovery of high-quality and trending content, as well as personalized content ranking, filtering, and recommendation. However, when these ratings apply to content generated by fellow users—helpfulness ratings of product reviews, likes on Facebook posts, or up-votes on news comments or forum posts—evaluations also become a mean of social interaction. This can create social feedback loops that affect the behavior of the author whose content was evaluated, as well as the entire community.

Online rating and evaluation systems have been extensively researched in the past. The focus has primarily been on predicting user-generated ratings [168, 114, 235, 9] and on understanding their effects at the community-level [90, 8, 220, 270]. However, little attention has been dedicated to studying the effects that ratings have on the behavior of the author whose content is being evaluated.

Ideally, feedback would lead users to behave in ways that benefit the community. Indeed, if positive ratings act as "reward" stimuli and negative ratings act as "punishment" stimuli, the *operant conditioning* framework from behavioral psychology [271] predicts that community feedback should guide authors to generate better content in the future, and that punished authors will contribute less than rewarded authors. However, despite being one of the fundamental frameworks in behavioral psychology, there is limited empirical evidence of operant conditioning effects on humans[1] [30]. Moreover, it remains unclear whether community feedback in complex online social systems brings the intuitive beneficial effects predicted by this theory.

In this work we develop a methodology for quantifying and comparing the effects of rewards and punishments on multiple facets of the author's future behavior in the community, and relate these effects to the broader theoretical framework of *operant conditioning*. In particular, we seek to understand whether community feedback regulates the quality and quantity of a user's future contributions in a way that benefits the community.

By applying our methodology to four large online news communities for which we have complete article commenting and comment voting data (about 140 million votes on 42 million comments), we discover that community feedback does *not* appear to drive the behavior of users in a direction that is beneficial to the community, as predicted by the operant conditioning framework. Instead, we find that community feedback is likely to perpetuate undesired behavior. In particular, punished authors

[1] The framework was developed and tested mainly through experiments on animal behavior (e.g., rats and pigeons); the lack of human experimentation can be attributed to methodological and ethical issues, especially with regards to punishment stimuli (e.g. electric shocks).

actually write worse[2] in subsequent posts, while rewarded authors do not improve significantly.

In spite of this detrimental effect on content *quality*, it is conceivable that community feedback still helps regulate *quantity* by selectively discouraging contributions from punished authors and encouraging rewarded authors to contribute more. Surprisingly, we find that negative feedback actually leads to more (and more frequent) future contributions than positive feedback does.[3] Taken together, our findings suggest that the content evaluation mechanisms currently implemented in social media systems have effects contrary to the interest of the community.

To further understand differences in social mechanisms causing these behavior changes, we conducted a structural analysis of the voter network around popular posts. We discover that not only does positive and negative feedback tend to come from communities of users, but that the voting network is most polarized when votes are split equally between up- and down-votes.

These observations underscore the asymmetry between the effects of positive and negative feedback: the detrimental impact of punishments is much more noticeable than the beneficial impact of rewards. This asymmetry echoes the *negativity effect* studied extensively in social psychology literature: negative events have a greater impact on individuals than positive events of the same intensity [158, 35].

To summarize our contributions, in this work we

- validate through a crowdsourcing experiment that the proportion of up-votes is a robust metric for measuring and aggregating community feedback,
- introduce a framework based on propensity score matching for quantifying the effects of community feedback on a user's post quality,

[2]One important subtlety here is that the observed quality of a post (i.e., the proportion of up-votes) is not entirely a direct consequence of the actual textual quality of the post, but is also affected by community bias effects. We account for this through experiments specifically designed to disentangle these two factors.

[3]We note that these observations cannot simply be attributed to flame wars, as they spread over a much larger time scale.

- discover that effects of community evaluations are generally detrimental to the community, contradicting the intuition brought up by the operant conditioning theory, and
- reveal an important asymmetry between the mechanisms underlying negative and positive feedback.

Our results lead to a better understanding of how users react to peer evaluations, and point to ways in which online rating mechanisms can be improved to better serve individuals, as well as entire communities.

6.2 Related Work

Our contributions come in the context of an extensive literature examining social media voting systems. One major research direction is concerned with predicting the helpfulness ratings of product reviews starting from textual and social factors [114, 203, 235, 289, 205, 221] and understanding the underlying social dynamics [65, 90, 307, 270]. The mechanisms driving user voting behavior and the related community effects have been studied in other contexts, such as Q&A sites [8], Wikipedia [52, 194, 9], YouTube [268], social news aggregation sites [187, 186, 220] and online multiplayer games [267]. Our work adds an important dimension to this general line of research, by providing a framework for analyzing the effects votes have on the author of the evaluated content.

The setting considered in this work, that of comments on news sites and blogs, has also been used to study other social phenomena such as controversy [68], political polarization [238, 26], and community formation [123, 126]. News commenting systems have also been analyzed from a community design perspective [217, 115, 94], including a particular focus on understanding what types of articles are likely to attract a large volume of user comments [288, 315]. In contrast, our analysis focuses on the effects of voting on the behavior of the author whose content is being evaluated.

Community	# Threads	# Posts	# Votes (Prop. Upvotes)
CNN	200,576	26,552,104	58,088,478 (0.82)
IGN	682,870	7,967,414	40,302,961 (0.84)
Breitbart	376,526	4,376,369	18,559,688 (0.94)
allkpop	35,620	3,901,487	20,306,076 (0.95)

Table 6.1: Summary statistics of the four communities analyzed in this study.

Our findings here reveal that negative feedback does not lead to a decrease of undesired user behavior, but rather attenuates it. Given the difficulty of moderating undesired user behavior, it is worth pointing out that anti-social behavior in social media systems is a growing concern [145], as emphasized by work on review spamming [202, 222, 234], trolling [262], social deviance [267] and online harassment [316].

6.3 Data: Measuring Encouragement

We aim to develop a methodology for studying the subtle effects of community-provided feedback on the behavior of content authors in realistic large-scale settings. To this end, we start by describing a longitudinal dataset where millions of users explicitly evaluate each others' content. Following that, we discuss a crowdsourcing experiment that helps establish a robust aggregate measure of community feedback.

6.3.1 Dataset description

We investigate four online news communities: *CNN.com* (general news), *Breitbart.com* (political news), *IGN.com* (computer gaming), and *Allkpop.com* (Korean entertainment), selected based on diversity and their large size. Common to all these sites is that community members post comments on (news) articles, where each comment

Community	# Registered Users	Prop. Up-votes	
		Q_1	Q_3
CNN	1,111,755	0.73	1.00
IGN	289,576	0.69	1.00
Breitbart	214,129	0.96	1.00
allkpop	198,922	0.84	1.00

Table 6.2: Additional summary statistics of these four communities. The lower (Q_1) and upper (Q_3) quartiles for the proportion of up-votes only takes into account posts with at least ten votes.

can then be up- or down-voted by other users. We refer to a comment as a *post* and to all posts relating to the same article as a *thread*.

From the commenting service provider, we obtained complete timestamped trace of user activity from March 2012 to August 2013.[4] We restrict our analysis to users who joined a given community after March 2012, so that we are able to track users' behavior from their "birth" onwards. As shown in Tables 6.1 and 6.2, the data includes 1.2 million threads with 42 million comments, and 140 million votes from 1.8 million different users. In all communities around 50% of posts receive at least one vote, and 10% receive at least 10 votes.

6.3.2 Measures of Community Feedback

Given a post with with some number of up- and down-votes we next require a measure that aggregates the post's votes into a single number, and that corresponds to the magnitude of reward/punishment received by the author of the post. However, it is not a priori clear how to design such a measure. For example, consider a post that received P up-votes and N down-votes. How can we combine P and N into a single number that best reflects the overall evaluation of the community? There are several natural candidates for such a measure: the total number of up-votes (P) received by

[4]This is prior to an interface change that hides down-votes.

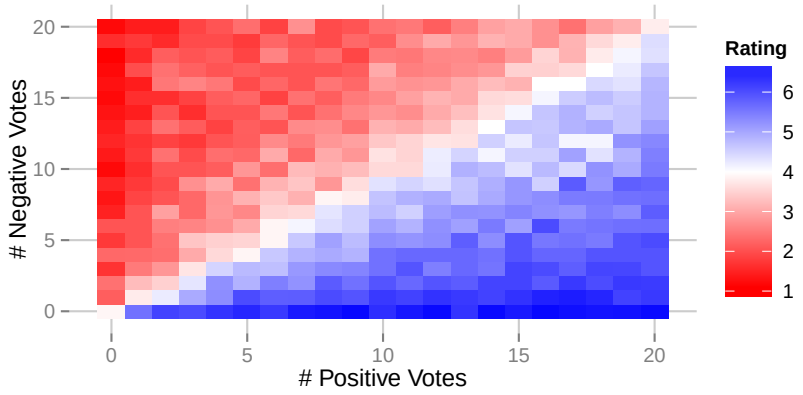


Figure 6.1: People perceive votes received as proportions, rather than as absolute numbers. Higher ratings correspond to more positive perceptions.

a post, the proportion of up-votes ($P/(P+N)$), or the difference in number of up-/down-votes ($P-N$). However, each of these measures has a particular drawback: P does not consider the number of down-votes (e.g., 10+/0- vs. 10+/20-); $P/(P+N)$ does not differentiate between absolute numbers of votes received (e.g., 4+/1- vs. 40+/10-); and, $P-N$ does not consider the effect relative to the total number of posts (e.g., 5+/0- vs. 50+/45-).

To understand what a person's "utility function" for votes is, we conducted an Amazon Mechanical Turk experiment that asked users how they would perceive receiving a given number of up- and down-votes. On a seven-point Likert scale, workers rated how they would feel about receiving a certain number of up- and down-votes on a comment that they made. The number of up- and down-votes was varied between 0 and 20, and each worker responded to 10 randomly-selected pairs of up-votes and down-votes. We then took the average response as the mean rating for each pair. 66 workers labeled 4,302 pairs in total, with each pair obtaining at least 9 independent evaluations.

We find that the proportion of up-votes ($P/(P+N)$) is a very good measure of how positively a user perceives a certain number of up-votes and down-votes. In Figure 6.1, we notice a strong "diagonal" effect, suggesting that increasing the total

number of votes received, while maintaining the proportion of up-votes constant, does not significantly alter a user's perception. In Table 6.3 we evaluate how different measures correlate with human ratings, and find that $P/(P+N)$ explains almost all the variance and achieves the highest R^2 of 0.92. While other more complex measures could result in slightly higher R^2, we subsequently use $p = P/(P+N)$ because it is both intuitive and fairly robust.

Thus, for the rest of this section, we use the *proportion of up-votes* (denoted as p) as the measure of the overall feedback of the community. We consider a post to be *positively evaluated* if the proportion of up-votes received is in the upper quartile Q_3 (75^{th} percentile) of all posts, and *negatively evaluated* if the fraction is instead in the lower quartile Q_1 (25^{th} percentile). This lets us account for differences in community voting norms: in some communities, a post may be perceived as bad, even with a high fraction of up-votes (e.g. Breitbart). As the proportion of up-votes is skewed in most communities, at the 75th percentile all votes already tend to be up-votes (i.e, feedback is 100% positive). Further, in order to obtain sufficient precision of community feedback, we require that these posts have at least ten votes.

Unless specified otherwise, all reported observations are consistent across all four communities we studied. For brevity, the figures that follow are reported only for CNN, with error bars indicating 95% confidence intervals.

Measure	$\mathbf{R^2}$	**F-Statistic**	**p-value**
P	0.410	$F(439) = 306.1$	$< 10^{-16}$
$P - N$	0.879	$F(438) = 1603$	$< 10^{-16}$
$P/(P+N)$	**0.920**	$F(438) = 5012$	$< 10^{-16}$

Table 6.3: The proportion of up-votes $p = P/(P+N)$ best captures a person's perception of up-voting and down-voting, according to a crowdsourcing experiment.

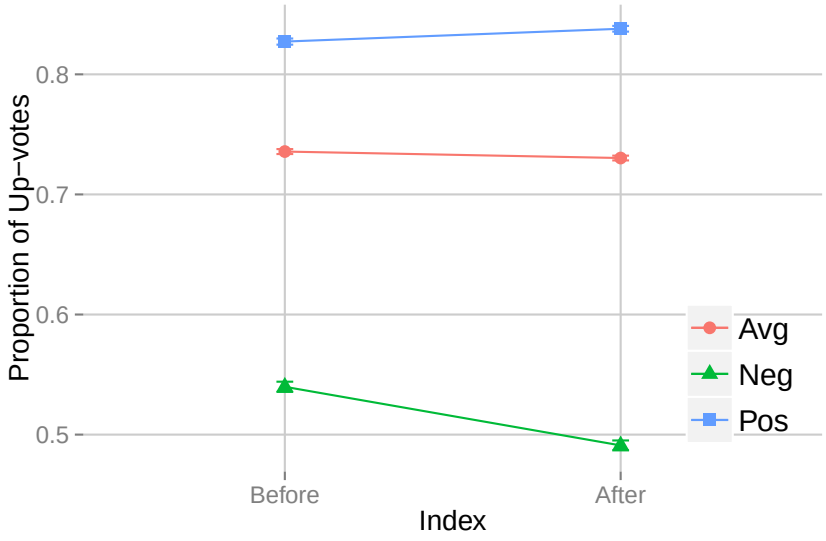


Figure 6.2: Proportion of up-votes before/after a user receives a positive ("Pos"), negative ("Neg") or neutral ("Avg") evaluation. After a positive evaluation, future evaluations of an author's posts do not differ significantly from before. However, after a negative evaluation, an author receives worse evaluations than before.

6.4 Post Quality

The *operant conditioning framework* posits that an individual's behavior is guided by the consequences of its past behavior [271]. In our setting, this would predict that community feedback would lead users to produce better content. Specifically, we expect that users punished via negative feedback would either improve the quality of their posts, or contribute less. Similarly, users rewarded by receiving positive feedback would write higher quality posts, and contribute more often.

In this section, we focus on understanding the effects of positive and negative feedback on the the quality of one's posts. We start by simply measuring the *post quality* as the *proportion of up-votes* a given post received, denoted by p. Figure 6.2 plots the proportion of up-votes p as a function of time for users who received a positive, negative or neutral evaluation. We compare the proportion of up-votes of posts written before receiving the positive/negative evaluation with that of posts written

after the evaluation. Interestingly, there is no significant difference for positively evaluated users (i.e., there is no significant difference between p before and after the evaluation event).

In the case of a negative evaluation however, punishment leads to worse community feedback in the future. More precisely, the difference in the proportion of up-votes received by a user before/after the feedback event is statistically significant at $p < 0.05$. This means that negative feedback seems to have exactly the opposite effect than predicted by the operant conditioning framework [271]. Rather than feedback leading to better posts, Figure 6.2 suggests that punished users actually get worse, not better, after receiving a negative evaluation.

6.4.1 Textual vs. Community Effects

One important subtlety is that the observed proportion of up-votes is not entirely a direct consequence of the actual textual quality of the post, but could also be due to a community's biased perception of a user. In particular, the drop in the proportion of up-votes received by users after negative evaluations observed in Figure 6.2 could be explained by two non-mutually exclusive phenomena: (1) after negative evaluations, the user writes posts that are of lower quality than before (*textual quality*), or (2) users that are known to produce low quality posts automatically receive lower evaluations in the future, regardless of the actual textual quality of the post (*community bias*).

We disentangle these two effects through a methodology inspired by propensity matching, a statistical technique used to support causality claims in observational studies [252].

First, we build a machine learning model that predicts a post's quality (q) by training a binomial regression model using *only* textual features extracted from the post's content, i.e. q is the *predicted* proportion of a post's up-votes. This way we are able to model the relationship between the content of a post and the post's *quality*. This model was trained on half the posts in the community, and used to predict q for the

other half (mean $R = 0.22$).

We validate this model using human-labeled text quality scores obtained for a sample of posts ($n = 171$). Using Crowdflower, a crowdsourcing platform, workers were asked to label posts as either "good" (defined as something that a user would want to read, or that contributes to the discussion), or "bad" (the opposite). They were only shown the text of individual posts, and no information about the post's author. Ten workers independently labeled each post, and these labels were aggregated into a "quality score" q', the proportion of "good" labels. We find that the correlation of q' with q ($R^2 = 0.25$) is more than double of that with p ($R^2 = 0.12$), suggesting that q is a reasonable approximation of text quality. The low correlation of q' with p also suggests that a community effect influences the value of p.

Since the model was trained to predict the proportion of a post's fraction of up-votes p, but only encodes text features (bigrams), the predicted proportion of up-votes q corresponds to the *quality* of the post's text. In other words, when we compare changes in q, these can be attributed to changes in the text, rather than to how a community perceives the user.[5] Thus, this model allows us to assess the textual quality of the post q, while the difference between the predicted and true proportion of up-votes $(p - q)$ allows us to quantify community bias.

Using the textual regression model, we match pairs of users (A, B) that contributed posts of similar quality, but that received very different evaluations: A's post was positively evaluated, while B's post was negatively evaluated. This experimental design can be interpreted as selecting pairs of users that appear indistinguishable before the "treatment" (i.e., evaluation) event, but where one was punished while the other rewarded. The goal then is to measure the effect of the treatment on the users' future behavior. As the two users "looked the same" before the treatment, any change in their future behavior can be attributed to the effect of the treatment (i.e., the act of receiving a positive or a negative evaluation).

[5]Even though the predicted proportion of up-votes q can be biased by user and community effects, this bias affects all posts equally (since the model is only trained on textual features). In fact, we find the model error, $p - q$, to be uniformly distributed across all values of p.

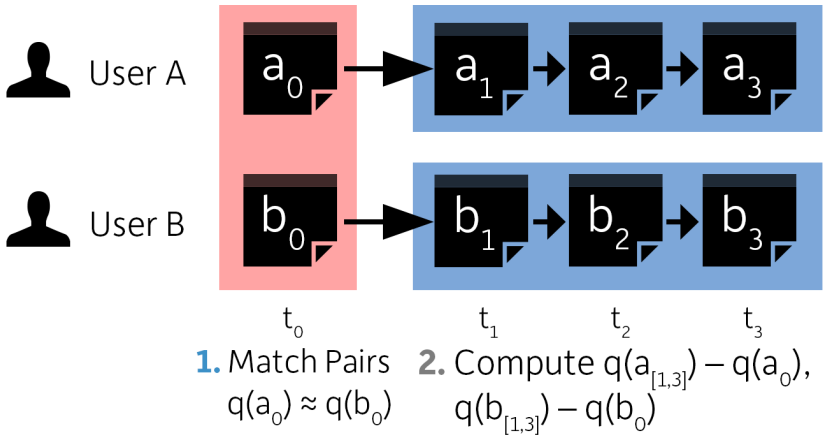


Figure 6.3: To measure the effects of positive and negative evaluations on post quality, we match pairs of posts of similar textual quality $q(a_0) \approx q(b_0)$ written by two users A and B with similar post histories, where A's post a_0 received a positive evaluation, and B's post b_0 received a negative evaluation. We then compute the change in quality in the subsequent three posts: $q(a_{[1,3]}) - q(a_0)$ and $q(b_{[1,3]}) - q(b_0)$.

	Before matching		**After matching**	
	Positive (A)	Negative (B)	Positive (A)	Negative (B)
	$n = 72463$	$n = 39788$	$n = 35640$	$n = 35640$
Textual quality $q(a_0)$ / $q(b_0)$	0.885	0.810	0.828	0.828
Number of words	42.1	34.5	29.8	30.0
Number of past posts	507	735	596	607
Prop. positive votes on past posts	0.833	0.650	0.669	0.668

Table 6.4: To obtain pairs of positively and negatively evaluated users that were as similar as possible, we matched these user pairs on post quality and the user's past behavior. On the CNN dataset, the mean values of these statistics were significantly closer after matching. Similar results were also obtained for other communities.

Figure 6.3 summarizes our experimental setting. Here, A's post a_0 received a positive evaluation and B's post b_0 received a negative evaluation, and we ensure that these posts are of the same textual quality, $|q(a_0) - q(b_0)| \leq 10^{-4}$. We further control for the number of words written by the user, as well as for the user's past behavior: both the

number of posts written before the evaluation was received, and the mean proportion of up-votes votes received on posts in the past (Table 6.4). To establish the effect of reward and punishment we then examine the next three posts of A ($a_{[1,3]}$) and the next three posts of B ($b_{[1,3]}$).

It is safe to assume that when a user contributes the first post a_1 after being punished or rewarded, the feedback on her previous post a_0 has already been received. Depending on the community, in roughly 70 to 80% of the cases the feedback on post a_0 at time of a_1's posting is within 1% of a_0's final feedback $p(a_0)$.

How feedback affects a user's post quality. To understand whether evaluations result in a change of text quality, we compare the post quality for users A and B before and after they receive a punishment or a reward. Importantly, we do not compare the actual fraction p of up-votes received by the posts, but rather the fraction q as predicted by the text-only regression model.

By design, both A and B write posts of similar quality $q(a_0) \approx q(b_0)$ at time $t = 0$. We then compute the quality of the three posts following $t = 0$ as the average predicted fraction of up-votes $q(a_{[1,3]})$ of posts a_1, a_2, a_3. Finally, we compare the post quality before/after the treatment event, by computing the difference $\Delta_a = q(a_{[1,3]}) - q(a_0)$ for the rewarded user A. Similarly, we compute $\Delta_b = q(b_{[1,3]}) - q(b_0)$ for the punished user B.

Now, if the positive (respectively negative) feedback has no effect and the post quality does not change, then the difference Δ_a (respectively Δ_b) should be close to zero. However, if subsequent post quality changes, then this quantity should be different from zero. Moreover, the sign of Δ_a (respectively Δ_b) gives us the direction of change: a positive value means that the post quality of positively (respectively negatively) evaluated users improves, while a negative value means that post quality drops after the evaluation.

Using a Mann-Whitney's U test, we find that across all communities, the quality of text significantly changes after the evaluation. In particular, we find that the post

quality significantly *drops* after a negative evaluation ($\Delta_b < 0$ at significance level $p < 0.05$ and effect size $r > 0.06$). This effect is similar both within and across threads (average $r = 0.19, 0.18$ respectively). While the effect of negative feedback is consistent across all communities, the effect of positive feedback is inconsistent and not significant.

These results are interesting as they establish the effect of reward and punishment on the quality of a user's future posts. Surprisingly, our findings are in a sense exactly the opposite than what we would expect under the operant conditioning framework. Rather than evaluations increasing the user's post quality and steering the community towards higher quality discussions, we find that negative evaluations actually decrease post quality, with no clear trend for positive evaluations having an effect either way.

How feedback affects community perception. We also aim to quantify whether evaluations changes the community's perception of the evaluated user (community bias). That is, do users that generally contribute good posts "undeservedly" receive more up-votes for posts that may actually not be that good? And similarly, do users that tend to contribute bad posts receive more down-votes even for posts that are in fact good?

To measure the community perception effect we use the experimental setup already illustrated in Figure 6.3. As before, we first match users (A, B) on the predicted fraction of up-votes q; we then measure the residual difference between the *true* and the *predicted* fraction of up-votes $p(a_{[1:3]}) - q(a_{[1:3]})$ after user A's treatment (analogously for user B). Systematic non-zero residual differences are suggestive of community bias effects, i.e., posts get evaluated differently from how they should be based solely on their textual quality. Specifically, if the community evaluates a user's posts higher than expected then the residual difference is positive, and if a user's posts are evaluated lower than expected then the residual difference is negative.

Across all communities, posts written by a user after receiving negative evaluations are perceived worse than the text-only model prediction, and this discrepancy is much larger than the one observed after positive evaluations ($p < 10^{-16}, r > 0.03$). This

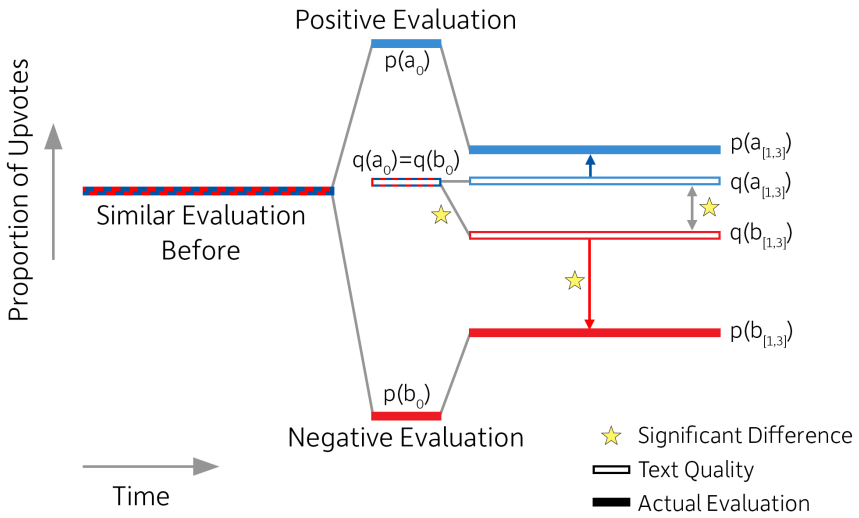


Figure 6.4: The effect of evaluations on user behavior. We observe that both community perception and text quality is significantly worse after a negative evaluation than after a positive evaluation (in spite of the initial post and user matching). Significant differences are indicated with stars, and the scale of effects have been edited for visibility.

effect is also stronger within threads (average $r = 0.57$) than across threads (average $r = 0.13$). For instance, after a negative evaluation on CNN, posts written by the punished author in the same thread are evaluated on average 48 percentage points lower than expected by just considering their text.

Note that the true magnitude of these effects could be smaller than reported, as using a different set of textual features could result in a more accurate classifier, and hence smaller residuals. Nevertheless, the experiment on post quality presented earlier does not suffer from these potential classifier deficiencies.

Summary. Figure 6.4 summarizes our observations regarding the effects of community feedback on the textual and perceived quality of the author's future posts. We plot the textual quality and proportion of up-votes before and after the evaluation event ("treatment"). Before the evaluation the textual quality of the posts of two users A and B is indistinguishable, i.e., $q(a_0) \approx q(b_0)$. However, after the evaluation event, the textual quality of the posts of the positively evaluated user A remains at the same

level, i.e., $q(a_0) \approx q(a_{[1:3]})$, while the quality of the posts of the negatively evaluated user B drops significantly, i.e., $q(b_0) > q(b_{[1:3]})$. We conclude that community feedback does not improve the quality of discussions, as predicted by the operand conditioning theory. Instead, punished authors actually write worse in subsequent posts, while rewarded authors do not improve significantly.

We also find suggestive evidence of community bias effects, creating a discrepancy between the perceived quality of a user's posts and their textual quality. This perception bias appears to mostly affect negatively evaluated users: the perceived quality of their subsequent posts $p(b_{[1:3]})$ is much lower than their textual quality $q(b_{[1:3]})$, as illustrated in Figure 6.4. Perhaps surprisingly, we find that community perception is an important factor in determining the proportion of up-votes a post receives.

Overall, we notice an important asymmetry between the effects of positive and negative feedback: the detrimental effects of punishments are much more noticeable than the beneficial effects of rewards. This asymmetry echoes the *negativity effect* studied extensively in the social psychology literature [158, 35].

6.5 User Activity

Despite the detrimental effect of community feedback on an author's content quality, community feedback could still have a beneficial effect by selectively regulating *quantity*, i.e., discouraging contributions from punished authors and encouraging rewarded authors to contribute more.

To establish whether this is indeed the case we again use a methodology based on propensity score matching, where our variable of interest is now posting frequency. As before, we pair users that wrote posts of the same textual quality (according to the textual regression model), ensuring that one post was positively evaluated, and the other negatively evaluated. We further control for the variable of interest by considering matching pairs of users that had the same posting frequency before the

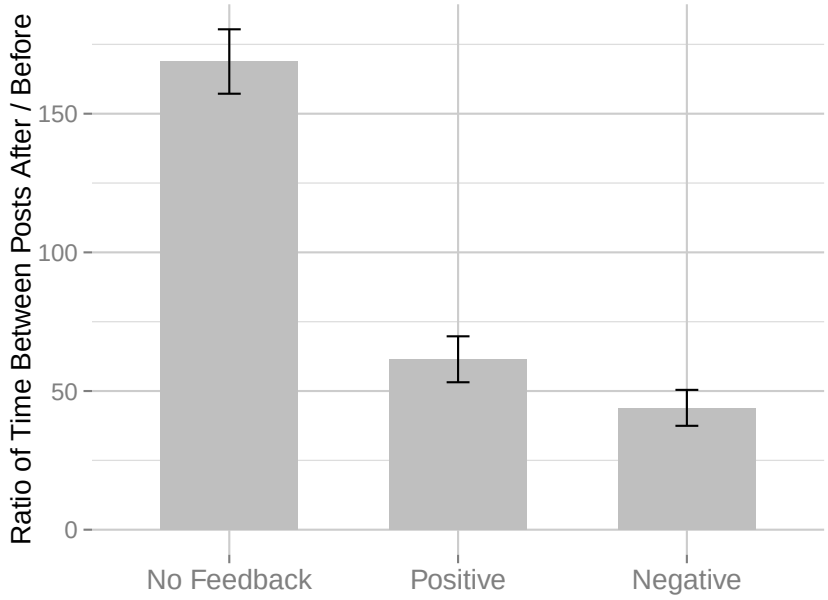


Figure 6.5: Negative evaluations increase posting frequency more than positive evaluations; in contrast, users that received no feedback slow down. (Values below 100 correspond to an increase in posting frequency after the evaluation; a lower value corresponds to a larger increase.)

evaluation. This methodology allows us to compare the effect of positive and negative feedback on the author's future posting frequency.

Figure 6.5 plots the ratio between inverse posting frequency after the treatment and inverse posting frequency before the treatment, where inverse frequency is measured as the average time between posts in a window of three posts after/before the treatment. Contrary to what operant conditioning would predict, we find that negative evaluations encourage users to post more frequently. Comparing the change in frequency of the punished users with that of the rewarded users, we also see that negative evaluations have a greater effect than positive evaluations ($p < 10^{-15}, r > 0.18$). Moreover, when we examine the users who received no feedback on their posts, we find that they actually slow down. In particular, users who received no feedback write about 15% less frequently, while those who received positive feedback write 20% more frequently than before, and those who received negative feedback write 30% more frequently

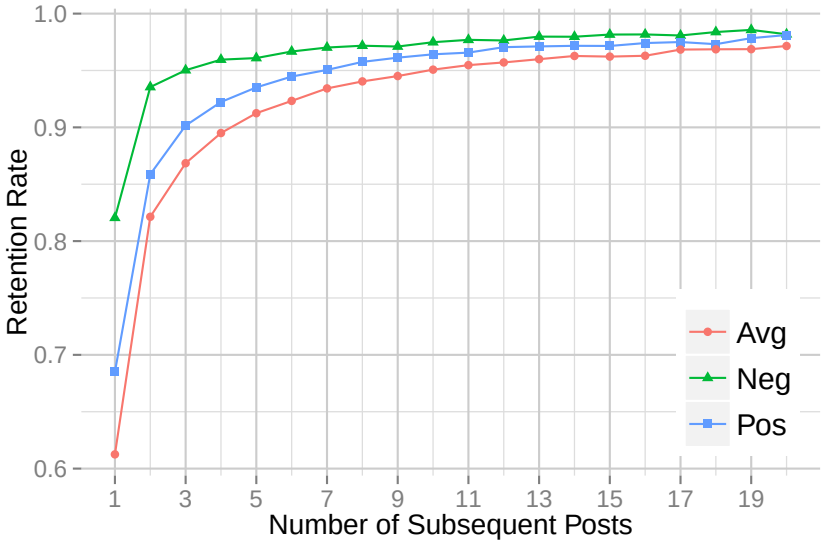


Figure 6.6: Rewarded users ("Pos") are likely to leave the community sooner than punished users ("Neg"). Average users ("Avg") are most likely to leave. For a given number of subsequent posts x the retention rate is calculated as the fraction of users that posted at least x more posts.

than before. These effects are also statistically significant, and consistent across all four communities.

The same general trend is true when considering the impact of evaluations on user retention (Figure 6.6): punished users ("Neg") are more likely than rewarded users ("Pos") to stay in the community and contribute more posts ($\chi^2 > 6.8, p < 0.01$); also both types or users are less likely to leave the community than the control group ("Avg"). Note, however, that the nature of this experiment does not allow one to control for the value of interest (retention rate) before the evaluation.

The fact that both types of evaluations encourage users to post more frequently suggests that providing negative feedback to "bad" users might not be a good strategy for combating undesired behavior in a community. Given that users who receive no feedback post less frequently, a potentially effective strategy could be to ignore undesired behavior and provide no feedback at all.

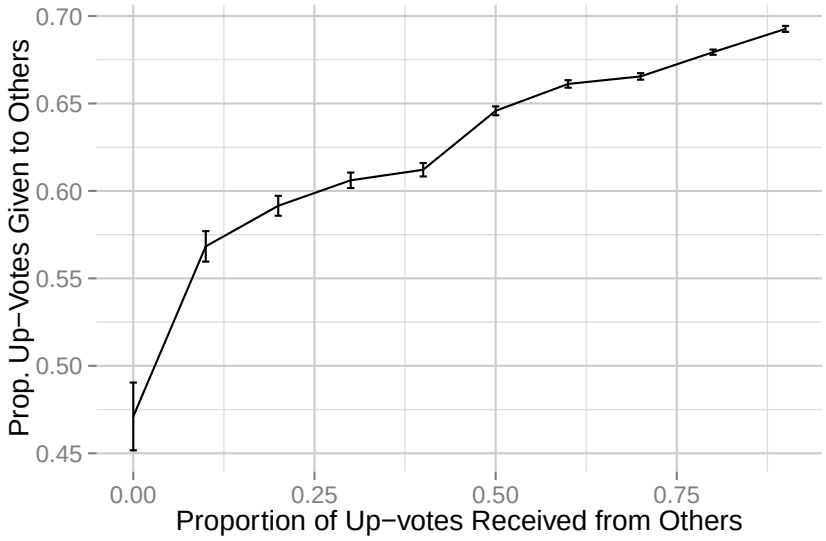


Figure 6.7: Users seem to engage in "tit-for-tat" — the more up-votes a user receives, the more likely she is to give up-votes to content written by others.

6.6 Voting Behavior

Our findings so far suggest that negative feedback worsens the quality of future interactions in the community as punished users post more frequently. As we will discuss next, these detrimental effects are exacerbated by the changes in the voting behavior of evaluated users.

Tit-for-tat. As users receive feedback, both their posting and voting behavior is affected. When comparing the fraction of up-votes received by a user with the fraction of up-votes given by a user, we find a strong linear correlation (Figure 6.7). This suggests that user behavior is largely "tit-for-tat". If a user is negatively/positively evaluated, she in turn will negatively/positively evaluate others. However, we also note an interesting deviation from the general trend. In particular, very negatively evaluated people actually respond in a positive direction: the proportion of up-votes they give is *higher* than the proportion of up-votes they receive. On the other hand, users receiving many up-votes appear to be more "critical", as they evaluate others more negatively. For example, people receiving a fraction of up-votes of 75% tend to

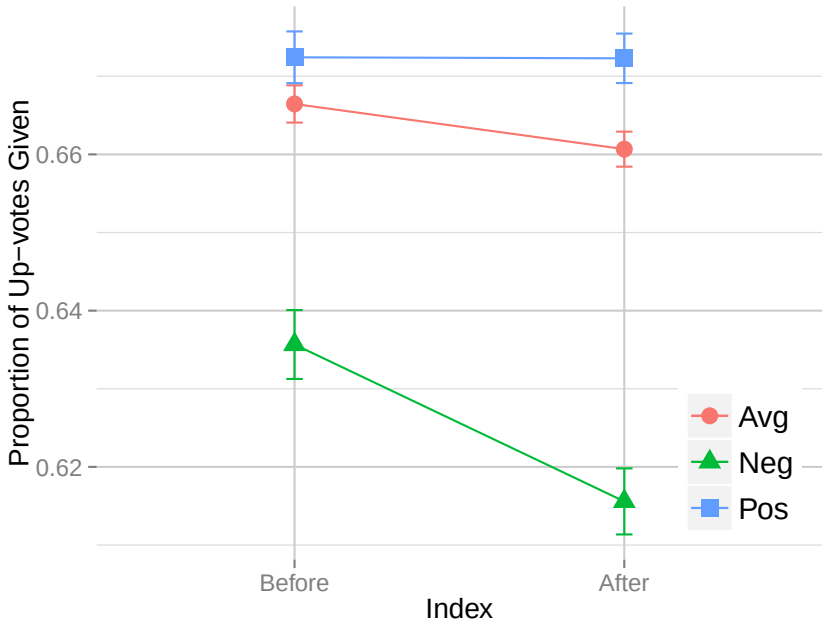


Figure 6.8: If a user is evaluated negatively, then she also tends to vote on others more negatively in the week following the evaluation than in the week before the evaluation. However, we observe no statistically significant effect on users who receive a positive evaluation.

give up-votes only 67% of the time.

Nevertheless, this overall perspective does not directly distinguish between the effects of positive and negative evaluations on voting behavior. To achieve that, Figure 6.8 compares the change in voting behavior following a positive or negative evaluation. We find that *negatively-evaluated users are more likely to down-vote others* in the week following an evaluation, than in the week before it ($p < 10^{-13}, r > 0.23$). In contrast, we observe no significant effect for the positively evaluated users.

Overall, punished users not only change their posting behavior, but also their voting behavior by becoming more likely to evaluate their fellow users negatively. Such behavior can percolate the detrimental effects of negative feedback through the community.

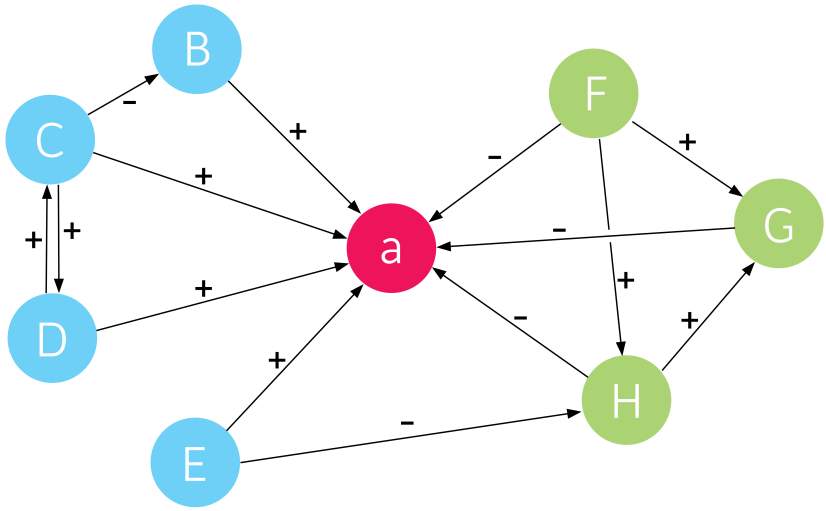
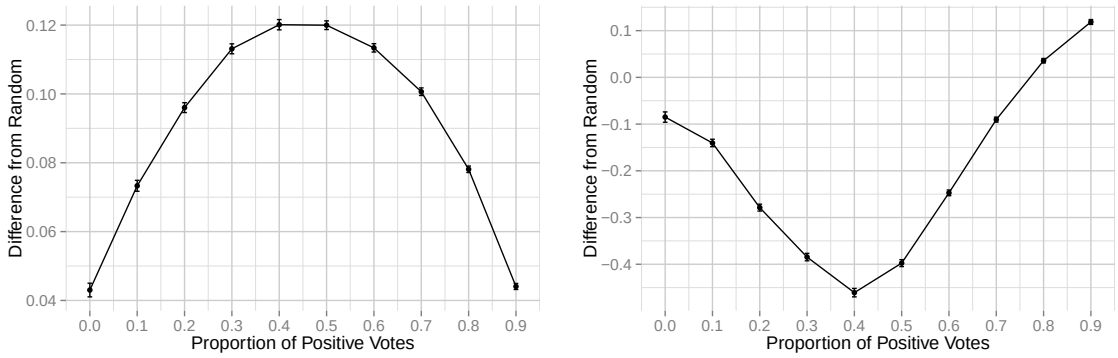


Figure 6.9: An example voting network G around a post a. Users B to H up-vote (+) or down-vote (-) a, and may also have voted on each other. The graph induced on the four users who up-voted a (B, C, D, E) forms G_+ (blue nodes), and that induced on the three users who down-voted a (F, G, H) forms G_- (green nodes).

6.7 Organization of Voting Networks

Having observed the effects community feedback has on user behavior, we now turn our attention to structural signatures of positive and negative feedback. In particular, we aim at studying the structure of the social network around a post and understanding a) when do evaluations most polarize this social network, and b) whether positive/negative feedback comes from independent people or from tight groups.

Experimental setup. We define a social networks around each post, a *voting network*, as illustrated in Figure 6.9. For a given post a, we generate a graph $G = (V, E)$, with V being the set of users who voted on a. An edge (B, C) exists between voters B and C if B voted on C in the 30 days prior to when the post a was created. Edges are signed: positive for up-votes, negative for down-votes. We examine voting networks for posts which obtained at least ten votes, and have at least one up-vote and one down-vote.



(a) Balance (b) Edges across the camps

Figure 6.10: (a) The difference between the observed fraction of balanced triangles and that obtained when edge signs are shuffled at random. The peak at 50% suggests that when votes are split evenly, these voters belong to different groups. (b) The difference of the observed number of edges between the up-voters and negative voters, versus those in the case of random rewiring. The lowest point occurs when the votes are split evenly.

When is the voting network most polarized? And, to what degree do coalitions or factions form in a post's voting network? As our networks are signed, we apply structural balance theory [55], and examine the fraction of balanced triads in our network. A triangle is balanced if it contains three positive edges (a set of three "friends"), or two negative edges and a positive edge (a pair of "friends" with a common "enemy"). The more balanced the network, the stronger the separation of the network into coalitions — nodes inside the coalition up-vote each other, and down-vote the rest of the network.

Figure 6.10a plots the fraction of balanced triangles in a post's voting network as a function of the proportion of up-votes that post receives (normalized by randomly shuffling the edge signs in the original voting network). When votes on a post are split about evenly between up- and down-votes, the network is most balanced. This means that when votes are split evenly, the coalitions in the network are most pronounced and thus the network most polarized. This observation holds in all four studied communities.

We also compare the number of edges between the up-voters and down-voters (i.e., edges crossing between G_+ and G_-) to that of a randomly rewired network (a random network with the same degree distribution as the original). Figure 6.10b plots the normalized number edges between the two camps as a function of the proportion of up-votes the post received. We make two interesting observations. First, the number of edges crossing the positive and negative camp is lowest when votes are split about evenly. Thus, when votes are split evenly not only is the network most balanced, but also the number of edges crossing the camps is smallest. Second, the number of edges is always below what occurs in randomly rewired networks (i.e., the Z-scores are negative). This suggests that the camp of up-voters and the camp of down-voters are generally not voting on each other. These effects are qualitatively similar in all four communities.

Where does feedback come from? Having observed the formation of coalitions, we are next interested in their relative size. Is feedback generally given by isolated individuals, or by tight groups of like-minded users? We find interesting differences between communities. In general-interest news sites like CNN, up-votes on positively-evaluated posts are likely to come from multiple groups — the size of the largest connected component decreases as the proportion of up-votes increases. In other words, negative voters on a post are likely to have voted on each other. However, on special-interest web sites like Breitbart, IGN, and Allkpop, the size of the largest connected component also peaks when votes are almost all positive. Thus, up-voters in these communities are also likely to have voted on each other suggesting that they come from tight groups.

6.8 Discussion and Conclusion

Rating, voting and other feedback mechanisms are heavily used in today's social media systems, allowing users to express opinions about the content they are consuming. In this work, we contribute to the understanding of how feedback mechanisms are used

in online systems, and how they affect the underlying communities. We start from the observation that when users evaluate content contributed by a fellow user (e.g., by liking a post or voting on a comment) they also implicitly evaluate the author of that content, and that this can lead to complex social effects.

In contrast to previous work, we analyze effects of feedback at the user level, and validate our results on four large, diverse comment-based news communities. We find that negative feedback leads to significant changes in the author's behavior, which are much more salient than the effects of positive feedback. These effects are detrimental to the community: authors of negatively evaluated content are encouraged to post more, and their future posts are also of lower quality. Moreover, these punished authors are more likely to later evaluate their fellow users negatively, percolating these undesired effects through the community.

We relate our empirical findings to the operand conditioning theory from behavioral psychology, which explains the underlying mechanisms behind reinforcement learning, and find that the observed behaviors deviate significantly from what the theory predicts. There are several potential factors that could explain this deviation. Feedback in online settings is potentially very different from that in controlled laboratory settings. For example, receiving down-votes is likely a much less severe punishment than receiving electric shocks. Also, feedback effects might be stronger if a user trusts the authority providing feedback, e.g., site administrators down-voting author's posts could have a greater influence on the author's behavior than peer users doing the same.

Crucial to the arguments made in this work is the ability of the machine learning regression model to estimate the textual quality of a post. Estimating text quality is a very hard machine learning problem, and although we validate our model of text quality by comparing its output with human labels, the goodness of fit we obtain can be further improved. Improving the model could allow for finer-grained analysis and reveal even subtler relations between community feedback and post quality.

Localized back-and-forth arguments between people (i.e. flame wars) could also potentially affect our results. However, we discard these as being the sole explanation

since the observed behavioral changes carry on across different threads, in contrast to flame wars which are usually contained within threads. Moreover, we find that the set of users providing feedback changes drastically across different posts of the same user. This suggests that users do not usually create "enemies" that continue to follow them across threads and down-vote any posts they write. Future work is needed to understand the scale and effects of such behavior.

There are many interesting directions for future research. While we focused only on one type of feedback — votes coming from peer users — there are several other types that would be interesting to consider, such as feedback provided through textual comments. Comparing voting communities that support both up- and down-votes with those that only allow for up-votes (or likes) may also reveal subtle differences in user behavior. Another open question is how the relative authority of the feedback provider affects author's response and change in behavior. Further, building machine learning models that could identify which types of users improve after receiving feedback and which types worsen could allow for targeted intervention. Also, we have mostly ignored the content of the discussion, as well as the context in which the post appears. Performing deeper linguistic analysis, and understanding the role of the context may reveal more complex interactions that occur in online communities.

More broadly, online feedback also relates to the key sociological issues of norm enforcement and socialization [239, 81, inter alia], i.e. what role does peer feedback play in directing users towards the behavior that the community expects, and how a user's reaction to feedback can be interpreted as her desire to conform to (or depart from) such community norms. For example, our results suggest that negative feedback could be a potential trigger of deviant behavior (i.e., behavior that goes against established norms, such as trolling in online communities). Here, surveys and controlled experiments can complement our existing data-driven methodology and shed light on these issues.

Chapter 7

Discussion

In all, this thesis represents our initial attempt at better understanding cascading behavior in online social networks. Chapters 3 and 4 described how we might go about predicting the behavior of cascades, while chapters 5 and 6 demonstrated how antisocial behavior can spread from person to person.

Nevertheless, in interpreting these findings, it is worthwhile to consider their limitations, as well as their implications for both future work and system design.

7.1 Limitations

The findings presented in this thesis are limited in terms of a) data (what was measured), b) definitions (how concepts were defined), c) causality (how results generalize), d) methodology (how a particular research approach limits what can be discovered), and e) specificity (how applicable these findings are to individuals).

7.1.1 Data

The results of the work described are limited in that they only apply to the systems that we evaluated. The studies on cascading behavior are limited to analyses of Facebook, and to the sharing of photos and videos. While the studies of antisocial behavior were validated across multiple different web sites, the form of interaction was nonetheless similar – of public discussions surrounding news articles.

Further, a majority of these analyses (a notable exception being the work studying cascade recurrence) involve data on English-speaking users in the United States. While this lets us interpret the content being shared or discussed, and ensures that the people interacting have at least some common cultural source to draw on, a better understanding of different geographies and cultures would be valuable to study. A more global view may allow us to better identify the original source of a cascade, and better account for external shocks to individual networks.

At the same time, while data allows us to study the interactions of many people at scale, there are limits to what can be passively measured. For example, intent and motivation remains relatively opaque, and can typically only be inferred. In the case of a post identified as trolling, was the poster trolling intentionally or legitimately angry? What if they were paid to post from a particular point of view? Similarly, was a user paid to share content with their friends or followers? To this end, techniques for detecting manipulative content (e.g., [293]) could be valuable.

7.1.2 Definitions

One significant challenge in this work is in defining the phenomena that we study. Though we test recurrence in a few different settings with similar results, other parameterizations or definitions may result in different findings. For example, though we do not require a subsequent cascade burst to be at least as large as the initial burst (i.e., if recurrence means something coming back with at least the same level

of popularity it initially enjoyed), that would result in a different set of cascades to analyze.

Similarly, we adopted several different definitions of what antisocial behavior is throughout this work. While we saw value in using relatively broader definitions that line up with what constitutes unacceptable behavior in typical discussion forums, this meant that we were not studying trolls with respect to their historically narrower definition that lies closer to intentional deception. Differentiating between people who are simply behaving badly and those who are intentionally causing harm to others is extremely valuable in identifying people who may be likely to improve over time, and is a significant avenue for potential work.

7.1.3 Causality

Across the multiple studies presented, there remain challenges in identifying the causal factors that lead to both virality and antisocial behavior.

The work presented on cascading behavior is primarily observational, and its goal was to predict the future trajectory of cascades using features related to its prior spread. Still, this should be differentiated from identifying what actually causes cascades to go viral.

To this end, to understand the causes of antisocial behavior, we combined longitudinal data analysis with an experiment. The former, while large-scale and reflective of real behavior, is observational and cannot prove causality. The latter, though being able to demonstrate causality, was done in a laboratory setting and can have issues with ecological validity. By demonstrating similar findings in both data and in experiments, we can be more confident about whether our findings are in fact valid.

Still, there are situations where experimentation is infeasible (e.g., where participants cannot be exposed to multiple conditions, when there are adverse effects, or when effects can only be observed over very long periods of time). Here, we use causal

inference methods (e.g., propensity score matching [252] or more recently, coarsened exact matching [153]) to control for the factors that are likely to correlate with the outcome. Controlling for all possible covariates though, remains exceedingly difficult. To this end, future work should also look to replicate of the findings described here in different settings and without the same experimental biases [251].

7.1.4 Methodology

The quantitative, largely data-driven nature of this research contrasts especially with prior work on antisocial behavior which has been primarily qualitative. Nonetheless, there is significant value in using qualitative methods such as interviews and surveys to complement data analyses. By combining survey results with behavioral data, one can better explain changes in tie strength over time [53], or understand motivation on Pinterest [74]. Further, the challenge in identifying intent could be resolved with an approach that also involves directly interacting with, rather than simply observing participants.

Qualitative work also remains a rich source of hypotheses, and can provide better context for understanding exactly what sets people off. Such work may also help us better understand any differences between community policies and the prevailing culture if they exist, better understand perceptions of negative behavior and inform future definitions, and evaluate the impact of interventions that mitigate negative behavior.

7.1.5 Specificity

Last, when we say that cascades can be predicted and antisocial behavior can be identified before it happens, we note that this does not necessarily mean that we understand the behavior of any individual user. While large cascades generally tend not to recur, there do exist large, viral, and recurring outliers; while an average

user's likelihood of trolling is strongly influenced by their mood or the presence of other troll posts, this does not preclude the existence of users who will never troll no matter what, or those that will regardless. Understanding how we might predict the future behavior of a particular individual user well remains future work.

7.2 Implications

How should we design the social systems of tomorrow? This thesis, through analyzing how social systems today enable the spread of information and behavior, provides a starting point from which we can begin to envision how future social systems should function.

7.2.1 Most behavior is situational

Rather than outcomes (e.g., widespread popularity or trolling) being either unpredictable or inevitable, a majority of action taken online is can be seen situational because of the inherent social nature of our interactions, where our behavior is influenced by others' behavior. Knowing the quality of a piece of content alone is insufficient to predict whether it will go viral, and 4chan-esque trolls cannot independently cause discussions to break down. Instead, people observe how others react and act accordingly. Something that gets shared widely will be shared even more in the future; if bad behavior appears to be the norm, people will be less inhibited in how they interact.

And because behavior cascades, it remains important to understand and monitor the growth of these cascades, especially at in the beginning stages. The initial conditions of any given system can have a sizable impact on what happens in the future. In chapters chapter 3 and chapter 4, we showed how the initial spread of a cascade makes its future spread predictable. In chapters chapter 5 and chapter 6, we showed

how widespread antisocial behavior can arise from a small number of individuals behaving badly.

As such, the reason that cascades appear unpredictable is because they are highly situational. On one hand, as the initial behavior of a cascade strongly affects the trajectory of its subsequent growth, it is difficult to predict how a cascade will fare before it spreads. On the other hand, if we are able to observe even a small fraction of a cascade's initial growth, we then have the knowledge to make reliable predictions about its future growth.

There is also perhaps a lesson to be drawn from the fact that a majority of behavior is situational. Prior psychology studies have found that people are more likely to attribute patterns of behavior to inherent characteristics rather than situational factors (i.e., that they make a fundamental attribution error) [141, 287]. As such, when one encounters some popularly-shared content or a troll-like user, one should be reasonably skeptical of that content's intrinsic quality, and forgiving of that user since they just may be having a bad day.

7.2.2 Encouraging prosocial behavior

Several implications for the design of social systems arise from our findings on how behavior, and negative behavior in particular, cascades.

Designing for Nth-Order Effects. First, our results on the negative effects of downvoting underscore the difficulty of designing social systems that do not have unintended negative consequences. While being able to vote on content can help with content ranking, it also impacts producers of that content negatively. Users, through indirectly evaluating other users, can cause the average quality of content to decrease over time. In other words, rather than simply thinking about features as changing an individual user's experience, system designers should also consider how they also impact interactions between users (or "second-order" effects).

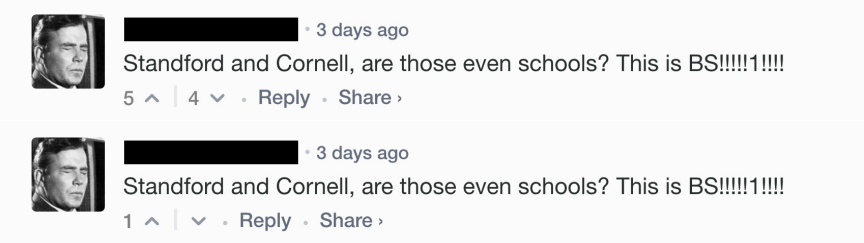


Figure 7.1: A previous version of Disqus (above), a popular comments plugin, displayed both upvotes and downvotes separately. Today (below), only a single score is shown.

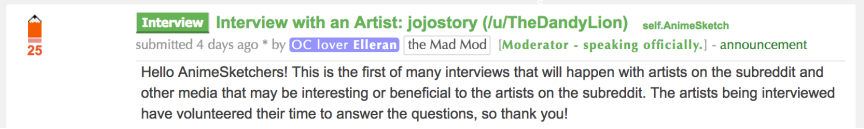


Figure 7.2: On Reddit, certain subreddits such as /r/animesketch restrict users to only being able to upvote posts, rather than being able to both upvote and downvote them.

The value of predictive modeling. In this work, we developed several predictive models for identifying users likely to be banned by a community in the future, as well as posts likely to be flagged or deleted by moderators. Nonetheless, as these predictions are not perfect, we cannot rely entirely on automated methods to identify undesirable content. Further, the development of improved models is complicated by the fact that definitions of acceptable content differ across communities and individuals. One potential approach then, is to adopt hybrid systems that combine both machine and human intelligence [71]. Content production today far outpaces the rate at which human moderators can moderate content, so predictive algorithms may act as an initial filter to flag potentially abusive content that human moderators can then focus on.

Validation of existing approaches. At the same time, our results also lend credence to several changes in interface design that have happened in recent years. One such change has been the reduction of the availability of negative evaluation mechanisms. For instance, Disqus has since stopped showing the number of downvotes that a post

got, and require users to log in to downvote content (Figure 7.1). On Reddit, several subreddits have also removed downvoting altogether (Figure 7.2). Our findings suggesting that waiting at least several minutes before participating in a new discussion reduces negative spillover effects validate the post rate-limiting implemented by several discussion platforms, including Discourse [17] and Reddit.

7.2.3 Managing the spread of content

Our work on information cascades also informs how social systems can manage the spread of content, and to either encourage or discourage it.

On one hand, knowledge of the factors that lead to content going viral can provide improved tools for tracking content of interest over time, which is potentially useful for tracking the effectiveness of viral advertisements, or to quantify the stickiness of news stories.

On the other hand, we may also begin to think about how we may manage the spread of undesirable content (e.g., false rumors). As above, our results suggest that it is easiest to stop the spread of such content in its earliest stages when social influence plays less of a role in its spread.

7.2.4 Ethics, experimentation, and influence

Finally, as this line of work focuses on studies human behavior, it also necessitates a discussion on the ethics of to what extent experimentation can take place.

While experiments can be seen as a type of A/B test, which many social platforms conduct daily and in line with their terms of service, some experiments may result in negative outcomes (e.g., the Facebook contagion study [176]). Additional oversight is especially important in cases where experiments can potentially result in participant harm (e.g., if one were to manipulate the presence of trolls on an actual web site).

Social platforms also have extraordinary power to shape behavior, through influencing the type of content that users are shown and the mechanisms (e.g., voting or even the introduction of badges [10]) through which users interact with these platforms and each other. One challenge worth considering here is of balancing conformity (everyone should behave in a certain way) and diversity (everyone can act however they want without repercussions).

Chapter 8

Conclusion

Cascades are an important way to understand the spread of content and virality in social networks, as well as the spread of negative behavior in online discussions. Rather than the web being just about making information access convenient, it is increasingly about enabling people to interact with one another. These interpersonal interactions transmit information and behavior, and can lead to complex feedback effects that can have wide-ranging and long-lasting impact.

Information cascades are unpredictable insofar as we assume that we know nothing about the starting conditions or the structure of the social network that a cascade is developing in. However, even knowing just a little about a cascade's starting conditions can make its growth predictable. Not only is it possible to reliable predict a cascade's future trajectory as it continues to develop, but it is also possible to predict if it will recur in the future after it dies down.

While some bad behavior online can be attributed to sociopathic individuals, ordinary people can behave like trolls under the right circumstances. In online discussions, ordinary individuals can be influenced by troll posts or downvotes, which leads them behave worse in the future – even an otherwise level-headed individual posting a single troll post in the heat of the moment can lead to negative behavior cascading

through a network if left unchecked.

8.1 Summary of Contributions

The work in this dissertation aimed to understand how information and behavior spreads in social networks, including in online discussions. Broadly, this thesis makes contributions to network science, online communities, and social computing:

- Network science. Continuously observing the spread of information cascades in social networks allow us to predict their future behavior, including their growth and recurrence.
- Online communities. A significant fraction of negative behavior in online communities is situational rather than innate. Negative behavior can spread from person to person, and the design of feedback mechanisms can exacerbate this spread.
- Social computing. Combining large-scale longitudinal data analysis with experimentation via crowdsourcing makes it possible to both demonstrate ecological validity and establish causality.

8.2 Recent Developments

Since the publication of the work described in this thesis, the broader research community has built on and extended this work significantly.

Some work has adopted the methodologies we introduced for cascade prediction (e.g., in predicting scientific impact [99] or board growth on Pinterest [204]), while others have started to explore the bounds on cascade predictability [149].

A growing line of work has also started to explore other quantitative approaches to

understanding antisocial behavior (e.g., [258]), in identifying specific types of behavior (e.g., sockpuppets [181] or ad hominem attacks [308]), or in developing scalable methods for predicting abusive behavior (e.g., by using "bags of communities" [63]).

Finally, the release of new datasets (e.g., the Reddit [93] and Wikipedia [309] comment datasets) presents new opportunities for analyzing other aspects of antisocial behavior that occur on different parts of the social web.

8.3 Future Work

Here, we outline potential future work that can be undertaken to both broaden and deepen research on cascading and antisocial behavior. These future directions reflect the inherent interdisciplinarity of work that aims to understand properties of human behavior. We see work in this space making the greatest impact through building on foundational theories from the social sciences and combining both qualitative (e.g., fieldwork and user studies) and quantitative (e.g., experiments and data analysis) methods.

8.3.1 Recipes for successful cascades

What makes a cascade successful? While we identified classes of factors that can help predict the future trajectory of a cascade, we see significant value in studying cascades in the context of the larger sharing ecosystem.

Understanding the role of contagion and network evolution

To begin, natural selection and evolution may provide clues to how a contagion may change over time. Prior work has suggested how individual memes and conventions may mutate or evolve similar to DNA [1]. Similarly, we may be able to trace the

propagation or evolution of general traits or mechanisms that inherently favor replication. It may also be valuable to identify and predict the relative success of different variants of the same contagion ahead of time.

The spread of contagion may also change the structure of the network. For instance, uncertainty can cause a network to "turtle up" and favor strong-tie interaction [250]. Thus, could contagion lead to a network become more or less favorable to other contagion spreading in the future?

Quantifying interaction effects

The spread of a contagion is not only influenced by other contagion, but also by the environment that it spreads in. For example, being infected with measles increases one's susceptibility to other diseases in the future [216]; on Twitter, shared URLs may either cooperate or compete with one another [224]. Similarly, how do properties of memes (e.g., on Facebook) help or hinder one another? How do recommender systems influence the propagation of information on social platforms? Could we encourage or hinder the spread of a meme by simultaneously introducing another (e.g., to mitigate the spread of misinformation)?

Tracking cascades at planetary scale

Contagion can also spread across multiple social networks. While prior work has tried to quantify the impact of external influence or shocks on the network [226], can we also develop methods for tracking the spread of contagion from one network to another? Rather than only tracking phrases on blogs and the news media [193], can we track contagion at web scale?

Relating influence, homophily, and causality

At the same time, it is important to distinguish influence from homophily [13, 78, 206] – how much is a contagion spreading due to people directly influencing each others' actions as opposed to people simply sharing similar backgrounds and interests? Separating out the various effects that contribute to cascading behavior and evaluating their statistical significance remains important to future studies that attempt to draw conclusions about human behavior. To this end, causal models of cascading behavior can support even more reliable predictions of future behavior.

Studying mechanisms of transmission

Most research, including the present work, assumes that the sharing of content from person-to-person is instantaneous. While this may be true in many cases (i.e., clicking "reshare" or "retweet"), many cascades involve more complex transmission mechanisms. For example, cascades may be limited in who can share them (e.g., content celebrating motherhood), or require specific tasks to be completed as a prerequisite (e.g., pouring a bucket of ice water over one's head). A better understanding of how these mechanisms influence the resulting cascades can paint a richer picture of how cascades become successful.

Investigating alternative approaches to cascade prediction

Finally, future work can also explore alternative models for cascade prediction. To this end, some work has begun to look at deep learning methods (e.g., [198]) for cascade prediction, or even network-agnostic techniques (e.g., [279]). Other work has also begun to take into account the multimodal nature of information shared online [144].

8.3.2 Supporting prosocial discourse

The findings described in this dissertation appear to paint a grim picture about the future of online discussions. One might even posit that antisocial behavior is inevitable, given how easily a single negative event (e.g., a troll post or downvote) can have a ripple effect throughout the community. However, we can only combat antisocial behavior by understanding how people end up behaving negatively. While the present work focuses on general findings about antisocial behavior, future work could explore the multifaceted nature of antisocial behavior and also begin to design systems that mitigate such behavior.

Redefining antisocial behavior

This work adopted relatively broad definitions of antisocial behavior and trolling. Nonetheless, trolling can vary depending on the community (e.g., CNN.com vs. 4chan), by the intent of the troll (e.g., genuine disagreement vs. purposeful deception), and by the type of trolling conducted (e.g., swearing vs. stalking, and only once or repeatedly). By better identifying the specific type of trolling that takes place, we can also better differentiate users who are situationally trolling and those that are not.

Understanding other causes of antisocial behavior

There is also value in understanding other factors that contribute to negative behavior in online settings. As polarization leads people to disagree more and adopt more extreme viewpoints [283], how may it factor into the likelihood of trolling a conversation? On a separate note, Dunbar's number suggests a cognitive limit on the maintenance of stable social relationships [124], suggesting that there may be a corresponding limit on the size of group discussions, beyond which they become difficult to sustain.

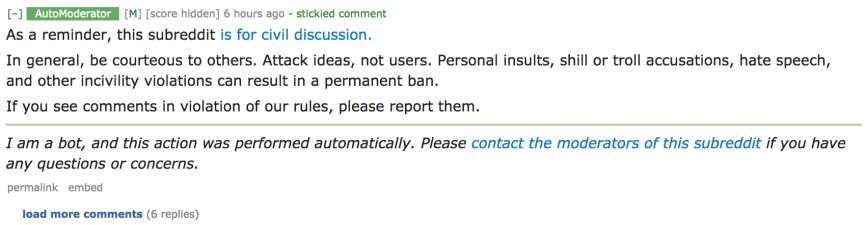


Figure 8.1: On Reddit's /r/politics, every discussion is preceded by a notice reminding users to be civil. However, this pushes relevant discussion further down on the website, potentially reducing user engagement? Further, what percentage of users learn to ignore the "banner" post?

Apart from the presence of troll posts in discussion, other aspects of the context of an interaction can also influence one's likelihood of behaving badly. This can include the visual design of a user interface [281]. For instance, visual complexity and colorfulness affect users' first impressions of web sites [245]. Further, the mechanism of communication can both discourage and amplify negative behavior – adding audio communication can decrease aversive behavior [91], but can have a larger impact on victims if used for harassment [273].

Some antisocial behavior is inescapably the product of individuals or groups with malicious intent. Identifying such types of antisocial behavior (e.g., sockpuppets [181] or autonomous bots [293]) remains valuable, so that we do not erroneously attribute the causes of these instances of such behavior to situational factors.

Mitigating interventions

Another significant area of future work is in identifying effective online interventions. Many such interventions exist, including hellbanning, or banning users from an online community without them realizing they have been banned [46]. Nonetheless, the efficacy of such interventions remain unmeasured.

Even simple changes such as adding reminders can be effective. Prior work showed how reminders of ethical and moral standards lead people to behave more morally [209]; a recent experiment on Reddit demonstrated how simply reminding people to

fact-check posts halved the Reddit scores of tabloid news submissions [228]. The challenge here is managing the usability tradeoff – showing users lots of reminders to behave well may work in the short-term, but may lead to reduced engagement or even banner blindness [36].

Bots may potentially be used to manage the flow of conversations. Twitter users reduced their use of racist language after being told off by bot [223], suggesting that conversational agents can help mitigate negative behavior. For example, a bot may step in to intervene if a discussion becomes heated or goes off-topic, verify statements made by other users, or even help steer conversations towards constructive goals.

Interventions that also account for the user's physical environment and disposition may also help. These may include factors such as the weather (e.g., higher temperatures and air pollution increase aggression [254]), physiological signals (e.g., heart rate and skin conductance can predict stress [142]), or even typing speed. Together, these may help identify periods when users may be more likely to troll in the heat of the moment.

Reflection can also improve on the quality of public deliberation [137], with some work framing reflection around comparing pros and cons [177]. Future work could further draw on conflict resolution techniques (e.g., [218]) such as joint problem solving to resolve arguments in online settings.

Social translucence may also be a useful concept in exploring such interventions [104]. For instance, WikiDashboard helped improve accountability on Wikipedia articles though making editors' activity more salient to other users [280]. Reputation systems (e.g., on eBay or StackOverflow) can also increase translucency by providing indicators of past activity and can help foster trust between users [96]. Other than these examples, what are appropriate online correlates of subtle changes in body language when we engage in face-to-face conversations? How can we best balance translucency and privacy (e.g., to prevent the misuse of anonymity)?

Developing tools for prototyping social interactions

While the systems we use today are overwhelmingly social, the testing and evaluation of such systems is still primarily concerned with understanding how individual users interact with an interface. By enabling more effective multi-user prototyping, designers may be able to better foresee challenges that arise when multiple users interact with one another.

For example, future work could use crowdsourcing to enable on-demand multi-user testing. Adversarial crowdworkers may also inform the creation of algorithms that can automatically generate adversarial users for a given social system. Future work may also involve creating a browsable pattern library for social mechanisms (e.g., voting) that describes their benefits and drawbacks, like what has been done for user interfaces in general [232].

Building better models for predicting bad behavior

Future work could further improve on the predictive models of bad behavior presented in this thesis. Recent work shows how sockpuppets can be differentiated from ordinary users [181], or how abusive behavior can be identified without training examples by drawing on data from multiple communities [63]. Identifying both polarization and controversy [113] in discussions may also be helpful.

One challenge of developing such prediction models is managing the tradeoff between false positives (incorrectly identifying ordinary users as trolls) and false negatives (incorrectly identifying trolls as ordinary users) – the former being significantly more costly. To this end, work that combination both improved predictions but allows for human intervention when necessary may end up being most effective in the near future.

8.3.3 Bridging the online and offline worlds

A final interesting area of exploration is in better characterizing the online-offline interface. How does behavior in the online world (e.g., on Facebook) compare with that in the offline world (e.g., in face-to-face interactions)?

Comparing online and offline cascades

With better sensing in the offline world, we might be able to better characterize how information sharing via word-of-mouth differs from sharing behavior on social networks. Similarly, it would be valuable to contrast collective action offline with collective action online. This can also inform the design of novel social systems – Catalyst [70], a platform for helping to organize events with sufficient critical mass, was inspired by the threshold theories of collective action [129].

Quantifying online and offline bias

Research in this space would also be able to better characterize differences in how people present themselves in online settings and offline settings. It would further be able to quantify bias of information online with actual public opinion. Some initial work here has tried to quantify novelty bias [212], or how representative topics on the web are relative to their true incidence in the general population.

8.4 Looking Ahead

While the web presents an opportunity to study collective behavior at unprecedented scale, we still lack the tools to properly harness this opportunity. Challenges remain in deriving causal theories from observational data, translating small-scale laboratory studies to large-scale online experiments, or even in acquiring the necessary breadth

of knowledge to conduct research in this growing area. While this thesis attempts to address a few of these challenges, it remains an initial attempt.

Future work should not only work towards a better understanding of human behavior, but also consider on how this understanding can improve existing social systems and the interactions that they support. The environment a person interacts in strongly influences their behavior, and thus the resulting observations made and any theories developed in relation to this particular context. However, these theories may also suggest how these environments could be altered to achieve better outcomes. Though such an approach is infeasible in the physical world, the more easily-malleable online world allows us to consider research approaches where the environment can be altered, perhaps even dynamically, depending on the behaviors being observed.

To this end, we see future research involving a virtuous cycle, where the design of social systems is not only analyzed, but where analyses also inform changes to their design, and ultimately results in both healthier online communities and richer theories of human behavior.

Bibliography

[1] Lada A Adamic, Thomas M Lento, Eytan Adar, and Pauline C Ng. Information evolution in social networks. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, 2016.

[2] Eytan Adar, Li Zhang, Lada A Adamic, and Rajan M Lukose. Implicit structure and the dynamics of blogspace. In *Workshop on the Weblogging Ecosystem*, 2004.

[3] B Thomas Adler, Luca De Alfaro, Santiago M Mola-Velasco, Paolo Rosso, and Andrew G West. Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In *Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing*, 2011.

[4] Mohamed Ahmed, Stella Spagna, Felipe Huici, and Saverio Niccolini. A peek into the future: Predicting the evolution of popularity in user generated content. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, 2013.

[5] Yavuz Akbulut, Yusuf Levent Sahin, and Bahadir Eristi. Cyberbullying victimization among turkish online social utility members. *Journal of Educational Technology and Society*, 2010.

[6] Eileen M Alexy, Ann W Burgess, Timothy Baker, and Shirley A Smoyak. Perceptions of cyberstalking among college students. *Brief Treatment and Crisis Intervention*, 2005.

[7] Sonia Altizer, Andrew Dobson, Parviez Hosseini, Peter Hudson, Mercedes Pascual, and Pejman Rohani. Seasonality and the dynamics of infectious diseases. *Ecology Letters*, 2006.

[8] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. Discovering Value from Community Activity on Focused Question Answering Sites: A Case Study of Stack Overflow. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2012.

[9] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. Effects of user similarity in social media. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, 2012.

[10] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. Steering user behavior with badges. In *Proceedings of the International World Wide Web Conference*, 2013.

[11] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, Jure Leskovec, and Mitul Tiwari. Global diffusion via cascading invitations: Structure, growth, and homophily. *Proceedings of the International World Wide Web Conference*, 2015.

[12] Lisa R Anderson and Charles A Holt. Information cascades in the laboratory. *The American Economic Review*, 1997.

[13] Sinan Aral, Lev Muchnik, and Arun Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 2009.

[14] Solomon E Asch and H Guetzkow. Effects of group pressure upon the modification and distortion of judgments. *Groups, Leadership, and Men*, 1951.

[15] Sitaram Asur, Bernardo Huberman, et al. Predicting the future with social media. *Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology*, 2010.

[16] Sitaram Asur, Bernardo Huberman, Gabor Szabo, and Chunyan Wang. Trends in social media: Persistence and decay. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2011.

[17] Jeff Atwood. Why is there a topic reply limit for new users? `https://meta.discourse.org/t/why-is-there-a-topic-reply-limit-for-new-users/11513`, 2013.

[18] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks: Membership, growth, and evolution. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2006.

[19] Lars Backstrom, Jon Kleinberg, Lillian Lee, and Cristian Danescu-Niculescu-Mizil. Characterizing and curating conversation threads: Expansion, focus, volume, re-entry. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, 2013.

[20] Norman TJ Bailey et al. *The mathematical theory of infectious diseases and its applications*. 1975.

[21] Andy Baio. 72 hours of #gamergate. `https://medium.com/message/72-hours-of-gamergate-e00513f7cf5d`, 2014.

[22] Paul Baker. Moral panic and alternative identity construction in Usenet. *Journal of Computer-Mediated Communication*, 2001.

[23] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, 2011.

[24] Eytan Bakshy, Brian Karrer, and Lada A Adamic. Social influence and the diffusion of user-created content. In *Proceedings of the ACM Conference on Economics and Computation*, 2009.

[25] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. In *Proceedings of the International World Wide Web Conference*, 2012.

[26] Ramnath Balasubramanyan, William W Cohen, Doug Pierce, and David P Redlawsk. Modeling Polarizing Topics: When Do Different Political Communities Respond Differently to the Same News? In *Proceedings of the International AAAI Conference on Web and Social Media*, 2012.

[27] Harriet A Ball, Louise Arseneault, Alan Taylor, Barbara Maughan, Avshalom Caspi, and Terrie E Moffitt. Genetic and environmental influences on victims, bullies and bully-victims in childhood. *Journal of Child Psychology and Psychiatry*, 2008.

[28] Abhijit V Banerjee. A simple model of herd behavior. *The Quarterly Journal of Economics*, 1992.

[29] Albert-Laszlo Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 2005.

[30] A Baron, M Perone, and M Galizio. Analyzing the reinforcement process at the human level: can application and behavioristic interpretation replace laboratory research? *Behavioral Analysis*, 1991.

[31] Sigal G Barsade. The ripple effect: Emotional contagion and its influence on group behavior. *Administrative Science Quarterly*, 2002.

[32] Frank M Bass. A new product growth for model consumer durables. *Management Science*, 1969.

[33] Christian Bauckhage, Fabian Hadiji, and Kristian Kersting. How viral are viral videos? In *Proceedings of the International AAAI Conference on Web and Social Media*, 2015.

[34] Christian Bauckhage, Kristian Kersting, and Fabian Hadiji. Mathematical models of fads explain the temporal dynamics of internet memes. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2013.

[35] Roy F Baumeister, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D Vohs. Bad is stronger than good. *Review of General Psychology*, 2001.

[36] Jan P Benway and David M Lane. Banner blindness: Web searchers often miss obvious links. *Itg Newsletter*, 1998.

[37] Jonah Berger and Katherine L. Milkman. What makes online content viral. *J. Marketing Research*, 2012.

[38] Leonard Berkowitz. *Aggression: Its causes, consequences, and control.* 1993.

[39] Leonard Berkowitz and Anthony LePage. Weapons as aggression-eliciting stimuli. *Journal of Personality and Social Psychology*, 1967.

[40] Gregory S Berns, Jonathan Chappelow, Caroline F Zink, Giuseppe Pagnoni, Megan E Martin-Skurski, and Jim Richards. Neurobiological correlates of social conformity and independence during mental rotation. *Biological Psychiatry*, 2005.

[41] Michael S Bernstein, Andrés Monroy-Hernández, Drew Harry, Paul André, Katrina Panovich, and Gregory G Vargas. 4chan and/b: An analysis of anonymity and ephemerality in a large online community. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2011.

[42] Amy Binns. Don't feed the trolls! managing troublemakers in magazines' online communities. *Journalism Practice*, 2012.

[43] Kaj Bjorkqvist, Karin Osterman, and Kirsti MJ Lagerspetz. Sex differences in covert aggression among adults. *Aggressive Behavior*, 1994.

[44] Ottar N Bjørnstad, Bärbel F Finkenstädt, and Bryan T Grenfell. Dynamics of measles epidemics: estimating scaling of transmission rates using a time series sir model. *Ecological Monographs*, 2002.

[45] Galen V Bodenhausen, Geoffrey P Kramer, and Karin Süsser. Happiness and stereotypic thinking in social judgment. *Journal of Personality and Social Psychology*, 1994.

[46] Dieter Bohn. One of twitters new anti-abuse measures is the oldest trick in the forum moderation book. `https://www.theverge.com/2017/2/16/14635030/twitter-shadow-ban-moderation`, 2017.

[47] Robert M Bond, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 2012.

[48] Rod Bond and Peter B Smith. Culture and conformity: A meta-analysis of studies using asch's (1952b, 1956) line judgment task. *Psychological Bulletin*, 1996.

[49] Youmna Borghol, Sebastien Ardon, Niklas Carlsson, Derek Eager, and Anirban Mahanti. The untold story of the clones: content-agnostic factors that impact Youtube video popularity. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2012.

[50] Youmna Borghol, Siddharth Mitra, Sebastien Ardon, Niklas Carlsson, Derek Eager, and Anirban Mahanti. Characterizing and modelling popularity of user-generated videos. *Performance Evaluation*, 2011.

[51] Erin E Buckels, Paul D Trapnell, and Delroy L Paulhus. Trolls just want to have fun. *Personality and Individual Differences*, 2014.

[52] Moira Burke and Robert Kraut. Mopping up: modeling wikipedia promotion decisions. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*, 2008.

[53] Moira Burke and Robert E Kraut. Growing closer on facebook: changes in tie strength through social network site use. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2014.

[54] John T Cacioppo, James H Fowler, and Nicholas A Christakis. Alone in the crowd: the structure and spread of loneliness in a large social network. *Journal of Personality and Social Psychology*, 2009.

[55] Dorwin Cartwright and Frank Harary. Structural balance: a generalization of heider's theory. *Psychological Review*, 1956.

[56] Irina Ceaparu, Jonathan Lazar, Katie Bessiere, John Robinson, and Ben Shneiderman. Determining causes and severity of end-user frustration. *International Journal of Human-Computer Interaction*, 2004.

[57] Damon Centola. The spread of behavior in an online social network experiment. *Science*, 2010.

[58] Damon Centola and Michael Macy. Complex contagions and the weakness of long ties 1. *American Journal of Sociology*, 2007.

[59] Meeyoung Cha, Fabrício Benevenuto, Yong-Yeol Ahn, and Krishna P Gummadi. Delayed information cascades in Flickr: Measurement, analysis, and modeling. *Computer Networks*, 2012.

[60] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and P Krishna Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2010.

[61] Meeyoung Cha, Alan Mislove, Ben Adams, and Krishna P Gummadi. Characterizing social cascades in flickr. In *Proceedings of the First Workshop on Online Social Networks*, 2008.

[62] Stevie Chancellor, Zhiyuan Jerry Lin, and Munmun De Choudhury. this post will just get taken down: Characterizing removed pro-eating disorder social media content. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2016.

[63] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. The bag of communities: Identifying abusive behavior online with preexisting internet data. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2017.

[64] Daphne Chang, Erin L Krupka, Eytan Adar, and Alessandro Acquisti. Engineering information disclosure: Norm shaping designs. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2016.

[65] Pei-Yu Chen, Samita Dhanasobhon, and Michael D Smith. All Reviews are Not Created Equal: The Disaggregate Impact of Reviews and Reviewers at Amazon.Com. *SSRN eLibrary*, 2008.

[66] Wei Chen, Yajun Wang, and Siyu Yang. Efficient influence maximization in social networks. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2009.

[67] Zhansheng Chen, Kipling D Williams, Julie Fitness, and Nicola C Newton. When hurt will not heal: Exploring the capacity to relive social and physical pain. *Psychological Science*, 2008.

[68] Zoeh Chen and Jonah Berger. When, Why, and How Controversy Causes Conversation. *J. of Consumer Research*, 2013.

[69] Justin Cheng, Lada A Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. Can cascades be predicted? In *Proceedings of the International World Wide Web Conference*, 2014.

[70] Justin Cheng and Michael Bernstein. Catalyst: triggering collective action with thresholds. In *Proceedings of the ACM Conference on Computer-Supported*

Cooperative Work and Social Computing, 2014.

[71] Justin Cheng and Michael S Bernstein. Flock: Hybrid crowd-machine learning classifiers. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*, 2015.

[72] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. How community feedback shapes user behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2014.

[73] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Antisocial behavior in online discussion communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2015.

[74] Justin Cheng, Caroline Lo, and Jure Leskovec. Predicting intent using activity logs: How goal specificity and temporal range affect user behavior. In *Proceedings of the International World Wide Web Conference Companion*, 2017.

[75] Thomas Chesney, Iain Coyne, Brian Logan, and Neil Madden. Griefing in virtual worlds: causes, casualties and coping strategies. *Information Systems Journal*, 2009.

[76] Daegon Cho and Alessandro Acquisti. The more social cues, the less trolling? an empirical study of online commenting behavior. In *Proceedings of the Workshop on the Economics of Information Security*, 2013.

[77] Hyunyoung Choi and Hal Varian. Predicting the present with Google Trends. *Economic Record*, 2012.

[78] Nicholas A Christakis and James H Fowler. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 2007.

[79] Robert B Cialdini and Noah J Goldstein. Social influence: Compliance and conformity. *Annual Review of Psychology*, 2004.

[80] Robert B Cialdini, Raymond R Reno, and Carl A Kallgren. A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, 1990.

[81] Robert B Cialdini and Melanie R Trost. *Social influence: Social norms, conformity and compliance*. 1998.

[82] Catherine Cloutier. Facebook: 1.2 million #icebucketchallenge videos posted. `https://www.bostonglobe.com/business/2014/08/15/facebook-million-icebucketchallenge-videos-posted/24D8bnxFlrMce5BRTixAEM/story.html`, 2014.

[83] CNN. Community Guidelines. `http://www.cnn.com/terms/comment_policy.html`, n.d.

[84] James Coleman, Elihu Katz, and Herbert Menzel. The diffusion of an innovation among physicians. *Sociometry*, 1957.

[85] Michele Coscia. Average is boring: How similarity kills a meme's success. *Scientific Reports*, 2014.

[86] Iain Coyne, Thomas Chesney, Brian Logan, and Neil Madden. Griefing in a virtual community: An exploratory survey of second life residents. *Zeitschrift für Psychologie/Journal of psychology*, 2009.

[87] Riley Crane and Didier Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 2008.

[88] Nicki R Crick and Kenneth A Dodge. Social information-processing mechanisms in reactive and proactive aggression. *Child Development*, 1996.

[89] Nicki R Crick and Jennifer K Grotpeter. Relational aggression, gender, and social-psychological adjustment. *Child Development*, 1995.

[90] Cristian Danescu-Niculescu-Mizil, Gueorgi Kossinets, Jon Kleinberg, and Lillian Lee. How opinions are received by online communities: A case study on Amazon.com helpfulness votes. In *Proceedings of the International World Wide Web Conference*, 2009.

[91] John P Davis, Shelly Farnham, and Carlos Jensen. Decreasing online 'bad' behavior. In *Proceedings of the ACM Conference on Human Factors in Computing Systems Extended Abstracts*, 2002.

[92] Andrea Devenow and Ivo Welch. Rational herding in financial economics. *European Economic Review*, 1996.

[93] dewarim. Updated reddit comment dataset as torrents. `https://www.reddit.com/r/datasets/comments/65o7py/updated_reddit_comment_dataset_as_torrents/`, 2017.

[94] Nicholas Diakopoulos and Mor Naaman. Towards quality discourse in online news comments. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*, 2011.

[95] Mario Diani and Doug McAdam. *Social movements and networks: Relational approaches to collective action*. 2003.

[96] Andreas Dieberger, Paul Dourish, Kristina Höök, Paul Resnick, and Alan Wexelblat. Social navigation: Techniques for building more usable systems. *interactions*, 2000.

[97] Discourse. This is a civilized place for public discussion. `https://meta.discourse.org/guidelines`, n.d.

[98] Judith S Donath. Identity and deception in the virtual community. *Communities in Cyberspace*, 1999.

[99] Yuxiao Dong, Reid A Johnson, and Nitesh V Chawla. Will this paper increase

your h-index?: Scientific impact prediction. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, 2015.

[100] P Alex Dow, Lada A Adamic, and Adrien Friggeri. The anatomy of large facebook cascades. *Proceedings of the International AAAI Conference on Web and Social Media*, 2013.

[101] Maeve Duggan. Online harassment. 2014.

[102] Alice H Eagly. Gender and social influence: A social psychological analysis. *American Psychologist*, 1983.

[103] Stale Einarsen, Helge Hoel, and Cary Cooper. *Bullying and emotional abuse in the workplace: International perspectives in research and practice.* 2003.

[104] Thomas Erickson and Wendy A Kellogg. Social translucence: an approach to designing systems that support social processes. *ACM TOCHI*, 2000.

[105] Mehrdad Farajtabar, Yichen Wang, Manuel Gomez Rodriguez, Shuang Li, Hongyuan Zha, and Le Song. Coevolve: A joint point process model for information diffusion and network co-evolution. In *Proceedings of the Conference on Neural Information Processing Systems*, 2015.

[106] Neil M Ferguson, Christl A Donnelly, and Roy M Anderson. The foot-and-mouth epidemic in great britain: pattern of spread and impact of interventions. *Science*, 2001.

[107] Miroslav Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 1973.

[108] Klint Finley. A brief history of the end of the comments. `https://www.wired.com/2015/10/brief-history-of-the-demise-of-the-comments-timeline/`, 2015.

[109] Joseph P Forgas and Gordon H Bower. Mood effects on person-perception judgments. *Journal of Personality and Social Psychology*, 1987.

[110] Adrien Friggeri, Lada A Adamic, Dean Eckles, and Justin Cheng. Rumor cascades. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2014.

[111] Wojciech Galuba, Karl Aberer, Dipanjan Chakraborty, Zoran Despotovic, and Wolfgang Kellerer. Outtweeting the twitterers-predicting information cascades in microblogs. In *Proceedings of the Workshop on Online Social Networks*, 2010.

[112] Becky Gardiner, Mahana Mansfield, Ian Anderson, Josh Holder, Daan Louter, and Monica Ulmanu. The dark side of guardian comments. 2016.

[113] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Quantifying controversy in social media. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, 2016.

[114] Anindya Ghose and Panagiotis G Ipeirotis. Designing Novel Review Ranking Systems: Predicting the Usefulness and Impact of Reviews. In *Proceedings of the ACM Conference on Economics and Computation*, 2007.

[115] Eric Gilbert, Tony Bergstrom, and Karrie Karahalios. Blogs are Echo Chambers: Blogs are Echo Chambers. In *Proceedings of the Hawaii International Conference on System Sciences*, 2009.

[116] Francesca Gino, Shahar Ayal, and Dan Ariely. Contagion and differentiation in unethical behavior: the effect of one bad apple on the barrel. *Psychological Science*, 2009.

[117] Michelle Girvan, Duncan S Callaway, Mark EJ Newman, and Steven H Strogatz. Simple model of epidemics with pathogen mutation. *Physical Review E*, 2002.

[118] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J Watts. The structural virality of online diffusion. *Management Science*, 2015.

[119] Sharad Goel, Duncan J Watts, and Daniel G Goldstein. The structure of online diffusion networks. In *Proceedings of the ACM Conference on Economics and Computation*, 2012.

[120] Jacob Goldenberg, Barak Libai, and Eitan Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 2001.

[121] Scott A Golder and Michael W Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 2011.

[122] Benjamin Golub and Matthew O Jackson. Using selection bias to explain the observed structure of internet diffusions. *Proceedings of the National Academy of Sciences*, 2010.

[123] Vicenç Gómez, Andreas Kaltenbrunner, and Vicente López. Statistical analysis of the social network and discussion threads in Slashdot. In *Proceedings of the International World Wide Web Conference*, 2008.

[124] Bruno Gonçalves, Nicola Perra, and Alessandro Vespignani. Modeling users' activity on twitter networks: Validation of dunbar's number. *PloS one*, 2011.

[125] Yunchao Gong and Svetlana Lazebnik. Iterative quantization: A procrustean approach to learning binary codes. *CVPR*, 2011.

[126] Sandra Gonzalez-Bailon, Andreas Kaltenbrunner, and Rafael E Banchs. The structure of political discussion networks: A model for the analysis of online deliberation. *Journal of Information Technology*, 2010.

[127] Roberto González-Ibáñez and Chirag Shah. Investigating positive and negative affects in collaborative information seeking: A pilot study report. In *Proceedings of the Annual Meeting of the Association for Information Science and Technology*, 2012.

[128] Jeffrey Gottfried and Elisa Shearer. News use across social media platforms 2016. *Pew Research Center*, 2016.

[129] Mark Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 1978.

[130] Jeff Greenberg, Linda Simon, Tom Pyszczynski, Sheldon Solomon, and Dan Chatel. Terror management and tolerance: Does mortality salience always intensify negative reactions to others who threaten one's worldview? *Journal of Personality and Social Psychology*, 1992.

[131] Nir Grinberg, Mor Naaman, Blake Shaw, and Gilad Lotan. Extracting diurnal patterns of real world activity from social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2013.

[132] Daniel Gruhl, Ramanathan Guha, David Liben-Nowell, and Andrew Tomkins. Information diffusion through blogspace. In *Proceedings of the International World Wide Web Conference*, 2004.

[133] The Guardian. Community Standards and Participation Guidelines. `https://www.theguardian.com/community-standards`, n.d.

[134] Marco Guerini, Jacopo Staiano, and Davide Albanese. Exploring image virality in google plus. *Proceedings of the IEEE International Conference on Social Computing and Networking*, 2013.

[135] Adrien Guille and Hakim Hacid. A predictive model for the temporal dynamics of information diffusion in online social networks. In *Proceedings of the International World Wide Web Conference Companion*, 2012.

[136] Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel A Zighed. Information diffusion in online social networks: A survey. *ACM SIGMOD Record*, 2013.

[137] Jürgen Habermas. *The structural transformation of the public sphere: An inquiry into a category of bourgeois society.* 1991.

[138] Craig Haney, Curtis Banks, and Philip Zimbardo. Interpersonal dynamics in a simulated prison. Technical report, 1972.

[139] Noriko Hara, Curtis Jay Bonk, and Charoula Angeli. Content analysis of online discussion in an applied educational psychology course. *Insturctional Science*, 2000.

[140] Claire Hardaker. Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *Journal of Politeness Research*, 2010.

[141] Gilbert Harman. Moral philosophy meets social psychology: Virtue ethics and the fundamental attribution error. In *Proceedings of the Aristotelian Society*, 1999.

[142] Jennifer A Healey and Rosalind W Picard. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems*, 2005.

[143] Susan Herring, Kirk Job-Sluder, Rebecca Scheckler, and Sasha Barab. Searching for safety online: Managing "trolling" in a feminist forum. *The Information Society*, 2002.

[144] Jack Hessel, Lillian Lee, and David Mimno. Cats and captions vs. creators and the clock: Comparing multimodal content to context in predicting relative popularity. In *Proceedings of the International World Wide Web Conference*, 2017.

[145] P Heymann, G Koutrika, and H Garcia-Molina. Fighting Spam on Social Web Sites: A Survey of Approaches and Future Challenges. *IEEE Internet Computing*, 2007.

[146] Sameer Hinduja and Justin W Patchin. Bullying, cyberbullying, and suicide. *Archives of Suicide Research*, 2010.

[147] David Hirshleifer and Siew Hong Teoh. Herd behaviour and cascading in capital markets: A review and synthesis. *European Financial Management*, 2003.

[148] Tuan-Anh Hoang and Ee-Peng Lim. Virality and susceptibility in information diffusions. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2012.

[149] Jake M Hofman, Amit Sharma, and Duncan J Watts. Prediction and explanation in social systems. *Science*, 2017.

[150] Liangjie Hong, Ovidiu Dan, and Brian D. Davison. Predicting popular messages in twitter. In *Proceedings of the International World Wide Web Conference Companion*, 2011.

[151] Chiao-Fang Hsu, Elham Khabiri, and James Caverlee. Ranking comments on the social web. In *Proceedings of the International Conference on Computational Science and Engineering*, 2009.

[152] Clayton J Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2014.

[153] Stefano M Iacus, Gary King, Giuseppe Porro, and Jonathan N Katz. Causal inference without balance checking: Coarsened exact matching. *Political analysis*, 2012.

[154] Maximilian Jenders, Gjergji Kasneci, and Felix Naumann. Analyzing and predicting viral tweets. In *Proceedings of the International World Wide Web Conference Companion*, 2013.

[155] Anders Johansen. A simple model of recurrent epidemics. *Journal of Theoretical Biology*, 1996.

[156] John W Jones and G Anne Bogat. Air pollution and human aggression. *Psychological Reports*, 1978.

[157] Laura June. I'm voting for Hillary because of my daughter. `https://www.thecut.com/2016/02/im-voting-for-hillary-because-of-my-daughter.html`, 2016.

[158] David E Kanouse and L Reid Hanson. *Negativity in Evaluations*. 1972.

[159] Joseph M Kayany. Contexts of uninhibited online behavior: Flaming in social newsgroups on usenet. *Journal of the Association for Information Science and Technology*, 1998.

[160] Matt J Keeling and Ken TD Eames. Networks and epidemic models. *Journal of the Royal Society Interface*, 2005.

[161] Dacher Keltner, Phoebe C Ellsworth, and Kari Edwards. Beyond simple pessimism: effects of sadness and anger on social perception. *Journal of Personality and Social Psychology*, 1993.

[162] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2003.

[163] Philip C Kendall, W Robert Nay, and John Jeffers. Timeout duration and contrast effects: A systematic evaluation of a successive treatments design. *Behavior Therapy*, 1975.

[164] Douglas T Kenrick and Steven W MacFarlane. Ambient temperature and horn honking: A field study of the heat/aggression relationship. *Environment and Behavior*, 1986.

[165] William O Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 1927.

[166] Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. In *Proceedings of the International Conference on Very Large*

Databases, 2004.

[167] Peter G Kilner and Christopher M Hoadley. Anonymity options and professional participation in an online community of practice. In *Proceedings of the Conference on Computer support for Collaborative Learning*, 2005.

[168] Soo-Min Kim, Patrick Pantel, Tim Chklovski, and Marco Pennacchiotti. Automatically assessing review helpfulness. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, 2006.

[169] Ben Kirman, Conor Lineham, and Shaun Lawson. Exploring mischief and mayhem in social computing or: how we learned to stop worrying and love the trolls. In *Proceedings of the ACM Conference on Human Factors in Computing Systems Extended Abstracts*, 2012.

[170] Jon Kleinberg. Cascading behavior in networks: Algorithmic and economic issues. *Algorithmic Game Theory*.

[171] Jon Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 2003.

[172] Anat Brunstein Klomek, Frank Marrocco, Marjorie Kleinman, Irvin S Schonfeld, and Madelyn S Gould. Bullying, depression, and suicidality in adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry*, 2007.

[173] Silvia Knobloch-Westerwick and Scott Alter. Mood adjustment to social situations through mass media use: How men ruminate and women dissipate angry moods. *Human Communication Research*, 2006.

[174] Agata Kołakowska. A review of emotion recognition methods based on keystroke dynamics and mouse movements. In *Proceedings of the International Conference on Human System Interaction*, 2013.

[175] Gina Kolata. Catching obesity from friends may not be so easy. *New York Times. Retrieved from http://www.nytimes.com*, 2011.

[176] Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 2014.

[177] Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. Supporting reflective public thought with considerit. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*, 2012.

[178] Travis Kriplean, Michael Toomim, Jonathan Morgan, Alan Borning, and Andrew Ko. Is this what you meant? promoting listening on the web with reflect. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2012.

[179] Ravi Kumar, Mohammad Mahdian, and Mary McGlohon. Dynamics of conversations. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2010.

[180] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. On the bursty evolution of blogspace. In *Proceedings of the International World Wide Web Conference*, 2005.

[181] Srijan Kumar, Justin Cheng, Jure Leskovec, and VS Subrahmanian. An army of me: Sockpuppets in online discussion communities. In *Proceedings of the International World Wide Web Conference*, 2017.

[182] Andrey Kupavskii, Liudmila Ostroumova, Alexey Umnov, Svyatoslav Usachev, Pavel Serdyukov, Gleb Gusev, and Andrey Kustarev. Prediction of retweet cascade size over time. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, 2012.

[183] Marcelo Kuperman and Guillermo Abramson. Small world effect in an epidemiological model. *Physical Review Letters*, 2001.

[184] Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. Prominent features of rumor propagation in online social media. In *Proceedings of the IEEE International Conference on Data Mining*, 2013.

[185] Himabindu Lakkaraju, Julian J McAuley, and Jure Leskovec. What's in a name? understanding the interplay between titles, content, and communities in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2013.

[186] Cliff Lampe and Erik Johnston. Follow the (Slash) Dot: Effects of Feedback on New Members in an Online Community. In *Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work*, 2005.

[187] Cliff Lampe and Paul Resnick. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2004.

[188] Hangwoo Lee. Behavioral strategies for dealing with flaming in an online forum. *The Sociological Quarterly*, 2005.

[189] So-Hyun Lee and Hee-Woong Kim. Why people post benevolent and malicious comments online. *Communications of the ACM*, 2015.

[190] Karen Pezza Leith and Roy F Baumeister. Why do bad moods increase self-defeating behavior? emotion, risk tasking, and self-regulation. In *Journal of Personality and Social Psychology*, 1996.

[191] Kristina Lerman and Rumi Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2010.

[192] Jure Leskovec, Lada A Adamic, and Bernardo A Huberman. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 2007.

[193] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2009.

[194] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Governance in Social Media: A case study of the Wikipedia promotion process. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2010.

[195] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. Cost-effective outbreak detection in networks. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2007.

[196] Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie S Glance, and Matthew Hurst. Patterns of cascading behavior in large blog graphs. *SDM*, 2007.

[197] Jerome M Levine and Gardner Murphy. The learning and forgetting of controversial material. *Journal of Abnormal Psychology*, 1943.

[198] Cheng Li, Jiaqi Ma, Xiaoxiao Guo, and Qiaozhu Mei. Deepcas: an end-to-end predictor of information cascades. In *Proceedings of the International World Wide Web Conference*, 2017.

[199] Qing Li. Cyberbullying in schools: A research of gender differences. *School Psychology International*, 2006.

[200] Tanya Beran Qing Li. Cyber-harassment: A study of a new method for an old behavior. In *Journal of Educational Computing Research*, 2005.

[201] David Liben-Nowell and Jon Kleinberg. Tracing information flow on a global scale using internet chain-letter data. *Proceedings of the National Academy of Sciences*, 2008.

[202] Ee-Peng Lim, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady Wirawan Lauw. Detecting product review spammers using rating behaviors. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, 2010.

[203] Jingjing Liu, Yunbao Cao, Chin-Yew Lin, Yalou Huang, and Ming Zhou. Low-quality product review detection in opinion summarization. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing and Computational Natural Language Learning*, 2007.

[204] Caroline Lo, Justin Cheng, and Jure Leskovec. Understanding online collection growth over time: A case study of pinterest. In *Proceedings of the International World Wide Web Conference Companion*, 2017.

[205] Yue Lu, Panayiotis Tsaparas, Alexandros Ntoulas, and Livia Polanyi. Exploiting social context for review quality prediction. In *Proceedings of the International World Wide Web Conference*, 2010.

[206] Russell Lyons. The spread of evidence-poor medicine via flawed social-network analysis. *Statistics, Politics, and Policy*, 2011.

[207] Zongyang Ma, Aixin Sun, and Gao Cong. On predicting the popularity of newly emerging hashtags in twitter. *Journal of the American Society for Information Science and Technology*, 2013.

[208] Yasuko Matsubara, Yasushi Sakurai, B Aditya Prakash, Lei Li, and Christos Faloutsos. Rise and fall patterns of information diffusion: model and implications. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2012.

[209] Nina Mazar, On Amir, and Dan Ariely. The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, 2008.

[210] Brian James McInnis, Elizabeth Lindley Murnane, Dmitry Epstein, Dan Cosley,

and Gilly Leshed. One and done: Factors affecting one-time contributors to ad-hoc online communities. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*, 2016.

[211] Douglas M McNair. *Manual profile of mood states*. 1971.

[212] Danaë Metaxa-Kakavouli, Gili Rusak, Jaime Teevan, and Michael S Bernstein. The web is flat: The inflation of uncommon experiences online. In *Proceedings of the ACM Conference on Human Factors in Computing Systems Extended Abstracts*, 2016.

[213] Stanley Milgram. Which nations conform most? `https://www.scientificamerican.com/article/milgram-nationality-conformity/`, 2011.

[214] Stanley Milgram, Leonard Bickman, and Lawrence Berkowitz. Note on the drawing power of crowds of different size. *Journal of Personality and Social Psychology*, 1969.

[215] Stanley Milgram and Christian Gudehus. Obedience to authority, 1978.

[216] Michael J Mina, C Jessica E Metcalf, Rik L de Swart, ADME Osterhaus, and Bryan T Grenfell. Long-term measles-induced immunomodulation increases overall childhood infectious disease mortality. *Science*, 2015.

[217] Gilad Mishne and Natalie Glance. Leave a reply: An analysis of weblog comments. In *WWE*, 2006.

[218] Jakki Mohr and Robert Spekman. Characteristics of partnership success: partnership attributes, communication behavior, and conflict resolution techniques. *Strategic Management Journal*, 1994.

[219] Stephen Morris. Contagion. *The Review of Economic Studies*, 2000.

[220] Lev Muchnik, Sinan Aral, and Sean J Taylor. Social influence bias: A randomized experiment. *Science*, 2013.

[221] Susan M Mudambi and David Schuff. What makes a helpful online review? A study of consumer reviews on Amazon.com. *MIS Quarterly*, 2010.

[222] A Mukherjee, B Liu, and N Glance. Spotting Fake Reviewer Groups in Consumer Reviews. In *Proceedings of the International World Wide Web Conference*, 2012.

[223] Kevin Munger. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, 2016.

[224] Seth A Myers and Jure Leskovec. Clash of the contagions: Cooperation and competition in information diffusion. In *Proceedings of the IEEE International Conference on Data Mining*, 2012.

[225] Seth A Myers and Jure Leskovec. The bursty dynamics of the twitter information network. In *Proceedings of the International World Wide Web Conference*, 2014.

[226] Seth A Myers, Chenguang Zhu, and Jure Leskovec. Information diffusion and external influence in networks. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2012.

[227] J. Nathan Matias. Posting rules in online discussions prevents problems & increases participation. `https://civilservant.io/moderation_experiment_r_science_rule_posting.html`, 2016.

[228] J. Nathan Matias. Persuading algorithms with an ai nudge. `https://civilservant.io/persuading_ais_preserving_liberties_r_worldnews.html`, 2017.

[229] Pascal Neis, Marcus Goetz, and Alexander Zipf. Towards automatic vandalism detection in openstreetmap. *ISPRS International Journal of Geo-Information*, 2012.

[230] Mark EJ Newman. Spread of epidemic disease on networks. *Physical Review E*, 2002.

[231] Vlad Niculae and Cristian Danescu-Niculescu-Mizil. Conversational markers of constructive discussions. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016.

[232] Erik G Nilsson. Design patterns for user interface for mobile applications. *Advances in Engineering Software*, 2009.

[233] Lars Folke Olsen, Gregory L Truty, and William Morris Schaffer. Oscillations and chaos in epidemics: a nonlinear dynamic study of six childhood diseases in Copenhagen, Denmark. *Theoretical Population Biology*, 1988.

[234] Myle Ott, Claire Cardie, and Jeff Hancock. Estimating the Prevalence of Deception in Online Review Communities. In *Proceedings of the International World Wide Web Conference*, 2012.

[235] Jahna Otterbacher. 'Helpfulness' in online communities: a measure of message quality. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2009.

[236] G Palshikar et al. Simple algorithms for peak detection in time-series. *ICAD-ABAI*, 2009.

[237] Deokgun Park, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. Supporting comment moderators in identifying high quality online news comments. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2016.

[238] Souneil Park, Minsam Ko, Jungwoo Kim, Ying Liu, and Junehwa Song. The politics of comments: predicting political orientation of news stories with commenters' sentiment patterns. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*, 2011.

[239] Talcott Parsons. *The social system*. 1951.

[240] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. 2001.

[241] Sasa Petrovic, Miles Osborne, and Victor Lavrenko. RT to win! predicting message propagation in twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2011.

[242] Martin Potthast, Benno Stein, and Robert Gerling. Automatic vandalism detection in wikipedia. In *European Conference on Information Retrieval*, 2008.

[243] Adrian Raine. Annotation: The role of prefrontal deficits, low autonomic arousal, and early health factors in the development of antisocial and aggressive behavior in children. *Journal of Child Psychology and Psychiatry*, 2002.

[244] Juliana Raskauskas and Ann D Stoltz. Involvement in traditional and electronic bullying among adolescents. *Developmental Psychology*, 2007.

[245] Katharina Reinecke, Tom Yeh, Luke Miratrix, Rahmatri Mardiko, Yuechen Zhao, Jenny Liu, and Krzysztof Z Gajos. Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2013.

[246] Emmett Rensin. Confessions of a former internet troll. 2014.

[247] Everett M Rogers. *Diffusion of innovations*. 2010.

[248] Daniel M Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the International World Wide Web Conference*, 2011.

[249] Daniel M Romero, Chenhao Tan, and Johan Ugander. On the interplay between social and topical structure. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2013.

[250] Daniel M Romero, Brian Uzzi, and Jon Kleinberg. Social networks under stress. In *Proceedings of the International World Wide Web Conference*, 2016.

[251] Paul R Rosenbaum. Replicating effects and biases. *The American Statistician*, 2001.

[252] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 1983.

[253] Lee Ross. The intuitive psychologist and his shortcomings: Distortions in the attribution process. *Advances in Experimental Social Psychology*, 1977.

[254] James Rotton and James Frey. Air pollution, weather, and violent crimes: concomitant time-series analysis of archival data. *Journal of Personality and Social Psychology*, 1985.

[255] Paul Rozin and Edward B Royzman. Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, 2001.

[256] Bryce Ryan and Neal C Gross. The diffusion of hybrid seed corn in two iowa communities. *Rural Sociology*, 1943.

[257] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 2006.

[258] Mattia Samory and Enoch Peserico. Sizing up the troll: A quantitative characterization of moderator-identified trolling in an online forum. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2017.

[259] Mattathias Schwartz. The trolls among us. *NY Times Magazine*, 2008.

[260] Norbert Schwarz and Herbert Bless. Happy and mindless, but sad and smart? the impact of affective states on analytic reasoning. *Emotion and Social Judgments*, 1991.

[261] Norbert Schwarz and Gerald L Clore. Mood, misattribution, and judgments of well-being: informative and directive functions of affective states. *Journal of Personality and Social Psychology*, 1983.

[262] Pnina Shachaf and Noriko Hara. Beyond vandalism: Wikipedia trolls. *Journal of Information Science*, 2010.

[263] Muzafer Sherif. The psychology of social norms. 1936.

[264] Muzafer Sherif, Oliver J Harvey, B Jack White, William R Hood, Carolyn W Sherif, et al. *Intergroup conflict and cooperation: The Robbers Cave experiment*. 1961.

[265] Charles R Shipan and Craig Volden. The mechanisms of policy diffusion. *American Journal of Political Science*, 2008.

[266] Clay Shirky. *Here comes everybody: The power of organizing without organizations*. 2008.

[267] Kenneth Shores, Yilin He, Kristina L Swanenburg, Robert Kraut, and John Riedl. The Identification of Deviance and its Impact on Retention in a Multiplayer Game. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*, 2014.

[268] Stefan Siersdorfer, Sergiu Chelaru, Wolfgang Nejdl, and Jose San Pedro. How useful are your comments?: Analyzing and predicting YouTube comments and comment ratings. In *Proceedings of the International World Wide Web Conference*, 2010.

[269] Beth A Simmons and Zachary Elkins. The globalization of liberalization: Policy diffusion in the international political economy. *American Political Science*

Review, 2004.

[270] Ruben Sipos, Arpita Ghosh, and Thorsten Joachims. Was This Review Helpful to You? It Depends! Context and Voting Patterns in Online Content. In *Proceedings of the International World Wide Web Conference*, 2014.

[271] B F Skinner. *The behavior of organisms: an experimental analysis*. 1938.

[272] Robert Slonje, Peter K Smith, and Ann FriséN. The nature of cyberbullying, and strategies for prevention. *Computers in Human Behavior*, 2013.

[273] Peter K Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell, and Neil Tippett. Cyberbullying: Its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry*, 2008.

[274] Brian H Spitzberg. Toward a model of meme diffusion (M3D). *Communication Theory*, 2014.

[275] Brian H Spitzberg and Gregory Hoobler. Cyberstalking and the technologies of interpersonal terrorism. *New Media & Society*, 2002.

[276] James B Stewart. Facebook has 50 minutes of your time each day. it wants more. `https://www.nytimes.com/2016/05/06/business/facebook-bends-the-rules-of-audience-engagement-to-its-advantage.html`, 2016.

[277] Greg Stoddard. Popularity dynamics and intrinsic quality in reddit and hacker news. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2015.

[278] David Strang and Sarah A Soule. Diffusion in organizations and social movements: From hybrid corn to poison pills. *Annual Review of Sociology*, 1998.

[279] Karthik Subbian, B Aditya Prakash, and Lada Adamic. Detecting large reshare cascades in social networks. In *Proceedings of the International World Wide Web Conference*, 2017.

[280] Bongwon Suh, Ed H Chi, Aniket Kittur, and Bryan A Pendleton. Lifting the veil: improving accountability and social transparency in wikipedia with wikidashboard. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2008.

[281] Abhay Sukumaran, Stephanie Vezich, Melanie McHugh, and Clifford Nass. Normative influences on thoughtful online participation. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2011.

[282] John Suler. The online disinhibition effect. *Cyberpsychology and Behavior*, 2004.

[283] Cass R Sunstein. The law of group polarization. *Journal of Political Philosophy*, 2002.

[284] John Synnott, Andria Coulias, and Maria Ioannou. Online trolling: The case of Madeleine McCann. *Computers in Human Behavior*, 2017.

[285] Gabor Szabo and Bernardo A Huberman. Predicting the popularity of online content. *Communications of the ACM*, 2010.

[286] Peter C Terry and Andrew M Lane. Normative values for the profile of mood states for use with athletic samples. *Journal of Applied Sport Psychology*, 2000.

[287] Philip E Tetlock. Accountability: A social check on the fundamental attribution error. *Social Psychology Quarterly*, 1985.

[288] Manos Tsagkias, Wouter Weerkamp, and Maarten de Rijke. Predicting the volume of comments on online news stories. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, 2009.

[289] Oren Tsur and Ari Rappoport. Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2009.

[290] Oren Tsur and Ari Rappoport. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, 2012.

[291] Johan Ugander, Lars Backstrom, Cameron Marlow, and Jon Kleinberg. Structural diversity in social contagion. *Proceedings of the National Academy of Sciences*, 2012.

[292] Kris Varjas, Jasmaine Talley, Joel Meyers, Leandra Parris, and Hayley Cutts. High school students perceptions of motivations for cyberbullying: An exploratory study. *The Western Journal of Emergency Medicine*, 2010.

[293] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2017.

[294] J Verdasca, MM Telo Da Gama, A Nunes, NR Bernardino, JM Pacheco, and MC Gomes. Recurrent epidemics in small world networks. *Journal of Theoretical Biology*, 2005.

[295] Michael B Walker and Maria G Andrade. Conformity in the Asch task as a function of age. *The Journal of Social Psychology*, 1996.

[296] Lu Wang and Claire Cardie. A Piece of My Mind: A Sentiment Analysis Approach for Online Dispute Detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2014.

[297] Yang Wang, Gregory Norcie, Saranga Komanduri, Alessandro Acquisti, Pedro Giovanni Leon, and Lorrie Faith Cranor. I regretted the minute i pressed share: A qualitative study of regrets on facebook. In *Proceedings of the Symposium on Usable Privacy and Security*, 2011.

[298] Dorothy E. Warner and Mike Raiter. Social context in massively-multiplayer online games (mmogs): Ethical questions in shared space. *International Review*

of Information Ethics, 2005.

[299] Duncan J. Watts. *Everything is obvious: How common sense fails us.* 2011.

[300] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. Virality prediction and community structure in social networks. *Scientific Reports*, 2013.

[301] Lilian Weng, Jacob Ratkiewicz, Nicola Perra, Bruno Gonçalves, Carlos Castillo, Francesco Bonchi, Rossano Schifanella, Filippo Menczer, and Alessandro Flammini. The role of information diffusion in the evolution of social networks. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2013.

[302] Lillian Weng, Alessandro Flammini, Alessandro Vespignani, and Fillipo Menczer. Competition among memes in a world with limited attention. *Scientific Reports*, 2012.

[303] Ladd Wheeler. Toward a theory of behavioral contagion. *Psychological Review*, 1966.

[304] David Wiener. Negligent publication of statements posted on electronic bulletin boards: Is there any liability left after Zeran? *Santa Clara Law Review*, 1998.

[305] Robb Willer, Ko Kuwabara, and Michael W Macy. The false enforcement of unpopular norms. *American Journal of Sociology*, 2009.

[306] James Q Wilson and George L Kelling. Broken Windows. *Atlantic Monthly*, 1982.

[307] Fang Wu and Bernardo A. Huberman. Opinion formation under costly expression. *ACM Transactions on Intelligent Systems and Technology*, 2010.

[308] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the International World Wide Web Conference*, 2017.

[309] Taraborelli Dario Thain Nithum Dixon Lucas Wulczyn, Ellery. Algorithms and insults: Scaling up our understanding of harassment on wikipedia. `https://blog.wikimedia.org/2017/02/07/scaling-understanding-of-harassment/`, 2017.

[310] Carrie L Wyland and Joseph P Forgas. On bad mood and white bears: The effects of mood state on ability to suppress unwanted thoughts. *Cognition and Emotion*, 2007.

[311] J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. In *Proceedings of the IEEE International Conference on Data Mining*, 2010.

[312] Jaewon Yang and Jure Leskovec. Modeling information diffusion in implicit networks. In *Proceedings of the IEEE International Conference on Data Mining*, 2010.

[313] Jaewon Yang and Jure Leskovec. Patterns of temporal variation in online media. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, 2011.

[314] Jiang Yang and Scott Counts. Predicting the speed, scale, and range of information diffusion in twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 2010.

[315] Tae Yano and Noah A. Smith. What's worthy of comment? Content and comment volume in political blogs. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2010.

[316] Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, and Lynne Edwards. Detection of harassment on Web 2.0. In *Proceedings of the Content Analysis for the WEB 2.0 Workshop at WWW*, 2009.

[317] George Udny Yule. A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS. *Philosophical Transactions of the Royal Society B*, 1925.

[318] Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. Conversational flow in Oxford-style debates. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016.

[319] Philip G Zimbardo. The human choice: Individuation, reason, and order versus deindividuation, impulse, and chaos. *Nebraska Symposium on Motivation*, 1969.