UNDERSTANDING INNOVATIONS AND CONVENTIONS AND THEIR DIFFUSION PROCESS IN ONLINE SOCIAL MEDIA

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Rahmtin Rotabi

December 2017

UNDERSTANDING INNOVATIONS AND CONVENTIONS AND THEIR DIFFUSION PROCESS IN ONLINE SOCIAL MEDIA

Rahmtin Rotabi, Ph.D.

Cornell University 2017

This thesis investigates innovations, trends, conventions and practices in online social media. Tackling these problems will give more insight into how their users use these online platforms with the hope that the results can be generalized to the offline world. Every major step in human history was accompanied by an innovation, from the time that mankind invented and mastered the production of fire, to the invention of the World Wide Web. The societal process of adopting innovations has been a case that has fascinated many researchers throughout the past century. Prior to the existence of online social networks, economists and sociologists were able to study these phenomena on small groups of people through microeconomics and microsociology. However, the data gathered from these online communities help us to take one step further, initiating studies on macroeconomic and macrosociologal problemsin addition to the previous two areas. Work in this thesis sheds light on the properties of both innovators and laggards, the expansion and adaptation of innovation, competition among innovations with the same purpose, and the eventual crowding out of competitor innovations in the target society. Lastly, we look at the bigger picture by studying the entire diffusion process as a whole, abstracting out a great deal of details. This offers a view on why every single idea, content, product, etc., fails to go viral.

BIOGRAPHICAL SKETCH

Rahmtin Rotabi grew up in Milton Keynes and Tehran and graduated from Allameh Tabatabaee high school in 2009. Due to his father's profession in software engineering, he was interested in computers since his childhood. At first he played in the gaming team that made it to the International competition in Korea but later on he changed his use of computers. In 2008, he won the national gold medal in Olympiads in Informatics in Iran. Given his enthusiasm towards computer programming and algorithms, he decided to pursue his undergraduate studies in Computer Engineering in Sharif University of Technology (SUT). His undergraduate thesis was on the theoretical side of computer science and focused on graph theory and provided a tight upper bound on a labeling problem called Signed Star Domination Number (SSDN). In 2013, he enrolled as a Ph.D. student in the computer science department at Cornell University in Ithaca, New York. During the summers of 2014, 2015, 2016 and 2017 he spent time as a researcher and software engineer intern at Google and Twitter in California, working on large scale problems in social media. After graduating from Cornell he will be joining Google in Mountain View, California.

Dedicated to my parents and stepmother

ACKNOWLEDGEMENTS

First things first - I would like to thank my thesis advisor Jon Kleinberg. He decided to work with me when I started as a confused graduate student and supported me through all good and bad days, both academically and emotionally in the past four and half years. He has taught me a great deal in computer science, how to do research, and most importantly how to be a better person with his calm, polite and logical behavior. Working with him has been one of the biggest privileges in my life. I cannot emphasize how appreciative I am for his mentorship, friendship and guidance.

I would also like to thank my mentor and friend at Twitter, Aneesh Sharma , who guided me through a new line of research and has always been sharing his wisdom on next steps in both research projects and my career. Aneesh hosted me for two summers at Twitter and changed my view on which research problems to approach.

I am also grateful of my collaborator Cristian Danescu-Niculescu-Mizil that has played a role in most of the research done in this thesis. His enthusiasm and thirst for finding answers to questions has been a strong motivation for me. I have learned from him to *fight* for what I believe. This PhD would have been much harder without valuable discussions with him and his guidance.

I am thankful of Krishna Kamath, Paul Ginsparg, Jack Hessel, Chenhao Tan, Julian McAuley, Dan Jurafsky and Peter Lepage who offered their generous help with datasets, discussions, advice and comments.

I would like to thank my committee members Robert Kleinberg for suggesting Jon as a potential advisor, Lillian Lee for her mentorship in the first semester I started working on projects in analyzing online social media and Arpita Ghosh who encouraged me with her kind words throughout the program.

Beyond my amazing academic collaborators and mentors, this thesis would have not been possible without the support and friendship of a few people.

My housemate for the past two years and friend for the past three years Amir Montazeri, who has been like family to me. It's hard to think of a time that I did not cook and have dinner with him while being housemates. My friend Jonathan DiLorenzo, who was there to go out with when I was not doing research. He has helped me adapt to the American culture with his awesome friendship.

My partners during my stay in Ithaca, Tina Ahmadi and Meghan Flyke who endlessly supported me, tolerated me and helped me become a better and happier person.

I am thankful of my friends Daniel Freund, Steffen Smolka, Sam Hopkins, Tobias Schnabel, Vikram Rao, Sarah Tan, Pooya Jalaly, Rad Niazadeh, Sameen Kiayei, Chenhao Tan, Mark Reitblatt, Eoin O'Mahony, Josh Moore, George Berry, Aurya Javeed, Kara Karpman, Dylan Foster, Jack Hessel, Vlad Niculae, Yolanda Lin, Molly Feldman, Geoff Pleis, Gilburt Chiang, Aria Rezaei, Khadijeh Sheikhan, Behrouz Rabiee, Pouya Samangouei, Isabel Kloumann, Maithra Raghu, Rediet Abebe, and Manish Raghavan who throughout these years have spent time with me at conferences, bars, gyms, running trails, restaurants, social cooking gatherings and etc., and made my life have a much higher quality.

Lastly, I want to thank my stepmother Roya, parents Hamid and Mitra and my younger brother Armin. All I have achieved in my life would have been far from attainable without their support, constant encouragement and inspiration. As I have always said my stepmother has done far more than a child could ever ask for. My parents although separated, created an amazing setup to spend time with both of them in a healthy and loving environment.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The study of diffusion of innovation dates back to 1903, when Gabriel Tarde [26] defined the innovation-decision process. The process had 5 steps; knowledge, forming an attitude, a decision to adopt or reject, implementation and use, and confirmation of the decision. Once the innovation occurs, it has the capability to spread to other people. The five step process starts out very slow. At some point, other users start adopting it, and the innovation gains momentum until it reaches critical mass and experiences its surge which results into it becoming prevalent. At this point, the process has reached its saturation, and from there on, the diffusion process slows down. No amount of effort will help the innovation spread any faster.
He showed that this process can be demonstrated by an S-curve shape shown in figure 1.1, where the x-axis is time and the y-axis is the fraction of people who adopted the innovation.

This work inspires a huge body of work, such as the famous book *Crossing the Chasm* [62] that has the Probability Density Function (PDF) of the S-shape curve on its cover or the early celebrated work of Ryan and Gross in 1943 [77]. In this work, Ryan and Gross studied the adoption of genetically modified seed corns by the farmers located in Iowa. They verified that there is indeed the S-shaped pattern in the behavior of farmers 1.2 and showed that there is delay between the time the farmers hear about these new seeds and when they start using it. In the beginning, very few farmers, the early adopters, dared to try this new type of seed. As time passes and the others see the outcome of farms that used these new seeds, they start using them as well.

A few decades later, in 1962, Rogers published a book entirely based on previous studies on diffusion of innovation in fields such as anthropology, early sociology, rural

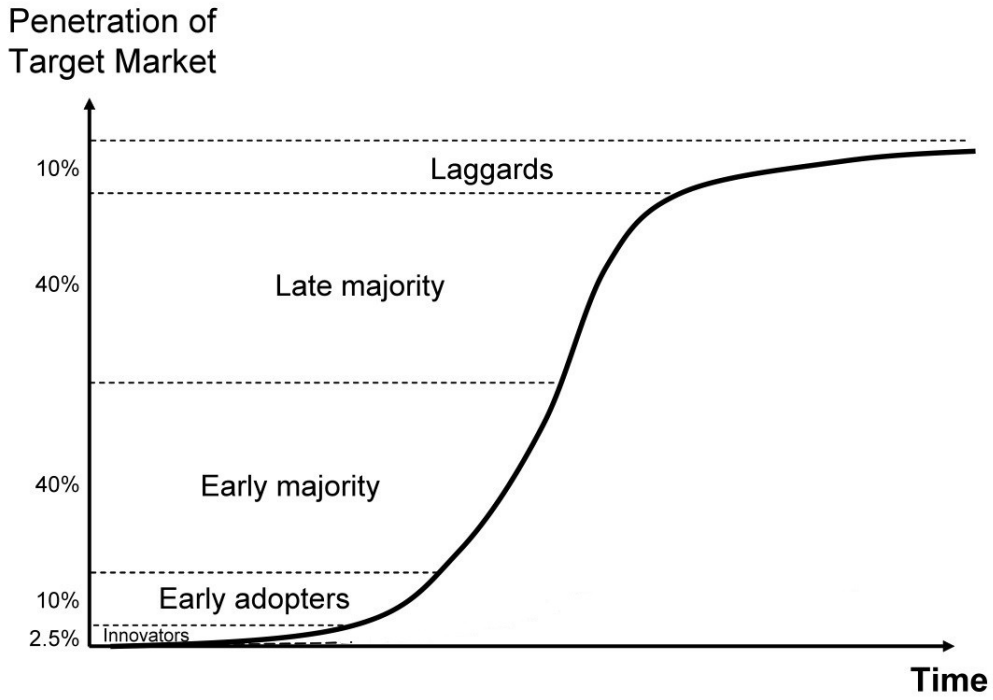


Figure 1.1: S-shape curve of the diffusion process

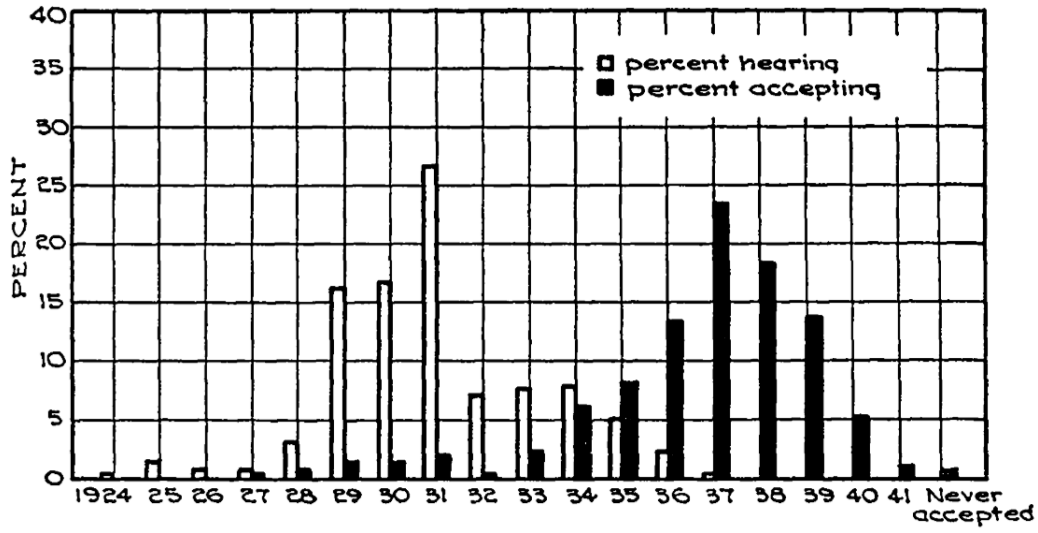


Figure 1.2: The time farmers hear about new seed corns versus the time they start using it

sociology, education, industrial sociology, and medical sociology.

More recently with the birth of the internet and common usage of online social media, online platforms play a huge role in our lives. An important aspect of the everyday experience on large on-line platforms is the emergence and spread of new activities and behaviors, including resharing of content, participation in new topics, and adoption of new features. These activities are described by various terms — as *trends* in the topic detection and social media literatures, and *innovations* by sociologists working on the diffusion of new behaviors [69]. An active line of recent research has used rich Web datasets to study the properties of such trends in on-line settings, and how they develop over time (e.g. [4, 39, 55, 52, 29, 38, 7, 10, 88]). The analyses performed in this style have extensively investigated the temporal aspects of trends, including patterns that accompany bursts of on-line activity [45, 48, 21, 89], and the network dynamics of their spread at both local levels [10, 52] and global levels [55, 29, 38]. In recent years, the online domain has provided a powerful setting in which to study this process. All of these online platforms have given a chance to revisit all the classic sociological problems using the massive data they create, such as, how do two innovation competitors compete to win the market when innovations of same purpose emerge in the ecosystem, or how do new words and language emerge in a community and how they spread [25, 30]?

In addition to observing the birth of these innovations and trends in these rich collections of online social media, the immense amount of data makes the process of observing the spread of new ideas and innovations through social networks much easier. A growing line of research has discovered principles for both the local mechanisms and global properties involved in the spread of pieces of information, such as messages, quotes, links, news stories, and photos [4, 39, 54, 55, 53, 3, 20, 37, 12], the diffusion of new products through viral marketing [52], and the cascading recruitment to on-line groups

[10, 6].

Another central question along this line is understanding how the theories of diffusion can address the process of competition and selection among conventions; when there are multiple possible behaviors and a group must choose among them, can we characterize how this selection process takes place, and how the latent interaction between competing options unfolds? The rise of new idioms and terminology [25]; technical standards in engineering and technological domains [9]; themes in political rhetoric [34]; and styles in artistic and other subjective domains [78] are all cases where we can pose such questions. It is important to note that the set of issues surrounding such conventions is far from monolithic — in particular, in cases with high costs to miscoordination, one tends to see a single convention crowd out all the others almost completely, while in cases where the convention has lower coordination effects and poses lower normative constraints, one typically finds extensive coexistence of conventions, with one convention dominating and others persisting in parts of the population [90].

However this is not all, nowadays given all these online social media and the easy mechanisms for resharing (as easy as one click), there are more high-level questions in this domain to ask. One could focus on the user experience on these network in response to these information diffusions. Users on modern social and information networks play dual roles as content producers and consumers; content they produce is seen by their friends or followers, and content they see (or consume) on the network is produced by users they are friends with or following. In addition to producing their own content, these networks also provide users with low-friction content-producing mechanisms. Users can switch from being consumers to content producers with a single click, as they can share or retweet content that they want to communicate to their followers. In some cases, consumption activity (such as “liking”) is akin to content production from the user in

terms of what their followers and friends observe.

Having a low barrier for content production is clearly important in activating the information-sharing aspects of social and information networks, but some of these mechanisms could be viewed as existing in tension with a basic contract of these networks. A key premise of a social or information networks is that users opt in to connect to friends or users that they are explicitly interested in hearing from. But, in the presence of sharing mechanisms, cascades originate on the network and hence a user can often see content from users they did not opt in to see content from. It is conceivable that cascades could overwhelm a user's homepage (news feed, timeline, ...) , rendering the network significantly less useful to the user. Therefore, one important question related to cascades asks whether they are beneficial to the health of the social network.

1.1 Contributions

Chapter 2 presents work on finding language innovation and trends in online social media and uses them to investigate properties of different users in these different stages of these innovations spreading in the community. Finding these vocabulary that are introduced to the language is a challenging task for several reasons. First, there is no clear definition for language innovations. Second, even if a definition is set, there might be words that existed before but find a new meaning (homonyms) after an event that happens in the community. Third, typos might fool the algorithm into counting them as innovations. In this chapter, we propose a method for detecting bursts of usage of different words which is inspired by an algorithm in [45]. Using these bursts will help in providing an algorithm, robust to challenges listed above. Previously, most algorithms require manual tuning of the parameters for each dataset independently, which leads to human biases. The outcome of running this algorithm on multiple datasets, such as pop-

ular subreddits, academic settings, review websites and online retailers verifies multiple findings in this research area. This chapter is a collaboration with Jon Kleinberg and appeared in proceedings of International Conference on Web and Social Media (ICWSM) 2016.

The work presented in chapter 3 is a follow-up work to Chapter 2. After detecting language innovations and the properties of the early and late adopters, a natural question to ask is: how do these innovations spread in networks? Studying diffusion of innovations requires a tag that enables us to track the spreading of the item and connections among people or users that use that innovation. This occurs in many previous studies such as the inheritance of DNA, spread of viruses, tweets, Facebook posts and online memes. However, once the context is changed to language, the problem becomes intractable. Keeping track of the communication among all people and how words spread in a community is simply impossible, even without dealing with homonyms and synonyms. In chapter 3, we present a way to get one step closer to answering this question. By relaxing the requirement of natural languages to programming languages the problem becomes much simpler while keeping the context. The benefit of using programming languages is that each command has a single deterministic meaning, which already removes two of the biggest stated challenges in the problem. This chapter is joint work with Jon Kleinberg and Cristian Danescu-Niculescu-Mizil and appeared in proceedings of International Conference on Web and Social Media (ICWSM) 2017.

In chapter 4, we continue working on the spread of innovations in the context of programming languages. Now that the problem of tracking these innovations is solved, the next natural step is to ask how these innovations interact with each other. More precisely, we ask: which one of the two language innovations that serve the same purpose (synonyms), will end up being the more popular one and winning over the crowd? But

we don't stop at this question and investigate the outcome of the scenario when two users of the language collaborate and disagree on terms. The second question that we undertake in this chapter is: which person has more influence in a collaboration? There have been general models proposed for diffusion of innovations in the literature, however, the underlying connections were known, and these models do not try to capture the properties of nodes (people) in the network. The work done in this chapter addresses the two problems mentioned and focuses on one-on-one interactions, giving us more insight on the spread of these innovations. This chapter is joint work with Jon Kleinberg and Cristian Danescu-Niculescu-Mizil and appeared in proceedings of World Wide Web (WWW) 2017.

Finally, in chapter 5, we look at diffusion on social media from the bird's-eye view and how it impacts the social medium as a whole. This work started as a collaboration with Twitter, to understand how content did not overflow the network when the retweet mechanism was introduced. There has been a lot of work on cascades and diffusion on social media, however, the work shown in this chapter takes a different approach to this problem. Instead of looking at these cascades and how they spread, we look at the effects of these cascades on their audience. In other words, cascade trees are usually known to have the producer of the content as the root and the content flowing outwards, but in this work, the root of the tree is the consumer of the content and the content is flowing inwards. This problem has not been studied before in academia, since it requires the private logs logged by the online platform. In addition to the data mining section, showing how far content travels in these networks we propose a theoretical model to exhibit why adding a resharing feature to the system does not make the content overflow the network. This chapter is joint work with Jon Kleinberg, Aneesh Sharma and Krishna Kamath; and appeared in proceedings of World Wide Web (WWW) 2017 [75].

CHAPTER 2

STATUS GRADIANT OF TRENDS IN SOCIAL MEDIA

An active line of research has studied the detection and representation of trends in social media content. There is still relatively little understanding, however, of methods to characterize the early adopters of these trends: who picks up on these trends at different points in time, and what is their role in the system? We develop a framework for analyzing the population of users who participate in trending topics over the course of these topics' lifecycles. Central to our analysis is the notion of a *status gradient*, describing how users of different activity levels adopt a trend at different points in time. Across multiple datasets, we find that this methodology reveals key differences in the nature of the early adopters in different domains.

2.1 Introduction

Online social media have become immersed in our daily lives and all the users activities provide us researchers with a huge body of data to study. An issue that has received less exploration using these types of datasets is the set of distinguishing characteristics of the participants themselves — those who take part in a trend in an on-line domain. This has long been a central question for sociologists working in diffusion more broadly: who adopts new behaviors, and how do early adopters differ from later ones [69]?

Key question: Who adopts new behaviors, and when do they adopt them?. When empirical studies of trends and innovations in off-line domains seek to characterize the adopters of new behaviors, the following crucial dichotomy emerges: is the trend proceeding from the "outside in," starting with peripheral or marginal members of the community and being adopted by the high-status central members; or is the innovation

proceeding from the "inside out," starting with the elites and moving to the periphery [1, 14, 22, 23, 65]?

Note that this question can be framed at either a broader population level or a more detailed network structural level. We pursue the broader population-level framing here, in which it is relevant to any distinction between elite and more peripheral members of a community, and not necessarily tied to measures based on network structure.

There are compelling arguments for the role of both the elites and the periphery in the progress of a trend. Some of the foundational work on adopter characteristics established that early adopters have significantly higher socioeconomic status in aggregate than arbitrary members of the population [27, 68]; elites also play a crucial role — as likely *opinion leaders* —in the two-step flow theory of media influence [42]. On the other hand, a parallel line of work has argued for the important role of peripheral members of the community in the emergence of innovations; Simmel's notion of "the stranger" who brings ideas from outside the mainstream captures this notion [79], as does the theory of *change agents* [60, 86] and the power of individuals who span *structural holes*, often from the periphery of a group [18, 47].

This question of how a trend flows through a population — whether from high-status individual to lower-status ones, or vice versa — is a deep issue at the heart of diffusion processes. It is therefore natural to ask how it is reflected in the adoption of trends in on-line settings. The interesting fact, however, is that there is no existing general framework or family of measures that can be applied to user activity in an on-line domain to characterize trends according to whether they are proceeding from elites outward or peripheral members inward. In contrast to the extensive definitions and measures that have been developed to characterize temporal and network properties of on-line diffusion, this progress of adoption along dimensions of status is an issue that to a large extent has

remained computationally unformulated.

Contribution: Formulating the status gradient of a trend. We define a formalism that we term the *status gradient*, which aims to take a first step toward characterizing how the adopter population of a trend changes over time with respect to their status in the community. Our goal in defining the status gradient is that it should be easy to adapt to data from different domains, and it should admit a natural interpretation for comparing the behavior of trends across these domains.

We start from the premise that the computation of a status gradient for a trend should produce a time series showing how the status of adopters in the community evolves over the life cycle of the trend. To make this concrete, we need (i) a way of assessing the status of community members, and (ii) a way of identifying trends.

While our methods can adapt to any way of defining (i) and (ii), for purposes of this work we operationalize them in a simple, concrete way as follows. Since our focus in this work is on settings where the output of the community is textual, we will think abstractly of each user as producing a sequence of posts, and the candidate trends as corresponding to words in these posts. (The adaptation to more complex definitions of status and trends would fit naturally within our framework as well.)

- We will use the activity level of each user as a simplified proxy for their status: users who produce more content are in general more visible and more actively engaged in the community, and hence we can take this activity as a simple form of status.[1] The current activity level of a user at a time t is the total number of posts they have produced up until t, and their final activity level is the total number of

[1] In the datasets with a non-trivial presence of high-activity spammers, we employ heuristics to remove such users, so that this pathological form of high activity is kept out of our analysis.

posts they have produced overall.

- We use a burst-detection approach for identifying trending words in posts [45]; thus, for a given trending word w, we have a time β_w when it entered its *burst state* of elevated activity. When thinking about a trending word w, we will generally work with "relative time" in which β_w corresponds to time 0.

We could try to define the status gradient simply in terms of the average activity level (our proxy for status) of the users who adopt a trend at each point in time. But this would miss a crucial point: high-activity users are already overrepresented in trends simply because they are overrepresented in *all* of a site's activities. This is, in a sense, a consequence of what it means to be high-activity. And this subtlety is arguably part of the reason why a useful definition of something like the status gradient has been elusive.

Our approach takes this issue into account. We provide precise definitions in the following section, but roughly speaking we say that the status gradient for a trending word w is a function f_w of time, where $f_w(t)$ measures the extent to which high-activity users are overrepresented or underrepresented in the use of w, *relative* to the baseline distribution of activity levels in the use of random words. The point is that since high-activity users are expected to be heavily represented in usage of both w and of "typical" words, the status gradient is really emerging from the difference between these two.

2.2 Overview of Approach and Summary of Results

We apply our method to a range of on-line datasets, including Amazon reviews from several large product categories [59], Reddit posts and comments from several active sub-communities [82], posts from two beer-reviewing communities [25], and paper titles

from DBLP and Arxiv.

We begin with a self-contained description of the status gradient we compute, before discussing the detailed implementation and results in subsequent sections. Recall that for purposes of our exposition here, we have an on-line community containing posts by users; each user's activity level is the number of posts he or she has produced; and a trending word w is a word that appears in a subset of the posts and has a burst starting at a time β_w.

Perhaps the simplest attempt to define a status gradient would be via the following function of time. First, abusing terminology slightly, we define the activity level *of a post* to be the activity level of the post's author. Now, let $P_w(t)$ be the set of all posts at time $t + \beta_w$ containing w, and let $g_w(t)$ be the median activity level of the posts in $P_w(t)$.

Such a function g_w would allow us to determine whether the median activity level of users of the trending word w is increasing or decreasing with time, but it would not allow us to make statements about whether this median activity level at a given time $t + \beta_w$ is high or low viewed as an isolated quantity in itself. To make this latter kind of statement, we need a baseline for comparison, and that could be provided most simply by comparing $g_w(t)$ to the median activity level g^* of the set of *all* posts in the community.

The quantity g^* has an important meaning: half of all posts are written by users of activity level above g^*, and half are written by users of activity level below g^*. Thus if $g_w(t) < g^*$, it means that the users of activity level at most $g_w(t)$ are producing half the occurrences of w at time $t + \beta_w$, but globally are producing less than half the posts in the community overall. In other words, the trending word w at time t is being over-produced by low-activity users and underproduced by high-activity users; it is being adopted mainly by the periphery of the community. The opposite holds true if $g_w(t) > g^*$.

Note how this comparison to g^* allows us to make absolute statements about the activity level of users of w at time $t + \beta_w$ without reference to the activity at other times.

This then suggests how to define the status gradient function $f_w(t)$ that we actually use, as a normalized version of $g_w(t)$. To do so, we first define the distribution of activity levels $H : [0, \infty) \to [0, 1]$ so that $H(x)$ is the fraction of *all* posts whose activity level is at most x. We then define

$$f_w(t) = H(g_w(t)).$$

This is the natural general formulation of our observations in the previous paragraph: the users of activity level at most $g_w(t)$ are producing half the occurrences of w at time $t+\beta_w$, but globally are producing an $f_w(t)$ fraction of the posts in the community overall. When $f_w(t)$ is small (and in particular below $1/2$), it means that half the occurrences of w at time $t + \beta_w$ are being produced by a relatively small slice of low-activity users, so the trend is being adopted mainly by the periphery; and again, the opposite holds when $f_w(t)$ is large.

Our proposal, then, is to consider $f_w(t)$ as a function of time. Its relation to 0.5 conveys whether the trend is being overproduced by high-activity or low-activity members of the community, and because it is monotonic in the more basic function $g_w(t)$, its dynamics over time show how this effect changes over the life cycle of the trend w.

Summary of Results. We find recurring patterns in the status gradients that reflect aspects of the underlying domains. First, for essentially all the datasets, the status gradient indicates that high-activity users are overrepresented in their adoption of trends (even relative to their high base rate of activity), suggesting their role in the development of trends.

We find interesting behavior in the status gradient right around time 0, the point

at which the burst characterizing the trend begins. At time 0, the status gradient for most of the sites we study exhibits a sharp drop, reflecting an influx of lower-activity users as the trend first becomes prominent. This is a natural dynamic; however, it is not the whole picture. Rather, for datasets where we can identify a distinction between *consumers* of content (the users creating posts on the site) and *producers* of content (the entities generating the primary material that is the subject of the posts), we generally find a sharp drop in the activity level of consumers at time 0, but *not* in the activity level of producers. Indeed, for some of our largest datasets, the activity levels of the two populations move inversely at time 0, with the activity level of consumers falling as the activity level of producers rises. This suggests a structure that is natural in retrospect but difficult to discern without the status gradient: in aggregate, the onset of a burst is characterized by producers of rising status moving in to provide content to consumers of falling status.

We now provide more details about the methods and the datasets where we evaluate them, followed by the results we obtain.

2.3 Data Description

Throughout this chapter, we will study multiple on-line communities gathered from different sources. The study uses the three biggest communities on Amazon.com, several of the largest sub-reddits from reddit.com, two large beer-reviewing communities that have been the subject of prior research, and the set of all papers on DBLP and Arxiv (using only the title of each).

- **Amazon.com**, in addition to allowing users to purchase items, hosts a rich set of reviews; these are the textual posts that we use as a source of trends. We take

all the reviews written before December 2013 for the top 3 departments: TV and Movies, Music, and Books [59].

- **Reddit** is one of the most active community-driven platforms, allowing users to post questions, ideas and comments. Reddit is organized into thousands of categories called sub-reddits; we study 5 of the biggest text-based sub-reddits. Our Reddit data includes all the Reddit posts and comments posted before January 2014 [82]. Reddit contains a lot of content generated by robots and spammers; heuristics were used to remove this content from the dataset.
- The two on-line beer communities **Beer Advocate** and **Rate Beer**, include reviews of beers from 2001 to 2011. Users on these two platforms describe a beer using a mixture of well-known and newly-adopted adjectives [25].
- **DBLP** is a website with bibliographic data for published papers in the computer science community. For this study we only use the title of the publications.
- **Arxiv** is a repository of on-line preprints of scientific papers in physics, mathematics, computer science, and an expanding set of other scientific fields. As with DBLP, we use the titles of the papers uploaded on Arxiv for our analysis, restricting to papers before November 2015. We study both the set of all Arxiv papers (denoted *Arxiv All*), as well as subsets corresponding to well-defined sub-fields. Two that we focus on in particular are the set of all statistics and computer science categories, denoted *Arxiv stat-cs*, and astrophysics — denoted *Arxiv astro-ph* — as an instance of a large sub-category of physics. In this study we only use papers that use *author* and *title* for including their title and their names.

More specific details about these datasets can be found in Table 2.1.

Dataset	Authors	Documents
Amazon Music	971,186	11,726,645
Amazon Movies and TV	846,915	14,391,833
Amazon Books	1,715,479	23,625,228
Reddit music	969,895	5,873,797
Reddit movies	930,893	1,0541,409
Reddit books	392,000	2,575,104
Reddit worldnews	1,196,638	16,091,492
Reddit gaming	1,811,850	33,868,254
Rate Beer	29,265	2,854,842
Beer Advocate	343,285	2,908,790
DBLP	1,510,698	2,781,522
Arxiv astro-ph	83,983	167,580
Arxiv stats-cs	63,128	71,131
Arxiv All	326,102	717,425

Table 2.1: Number of authors and documents in the studied datasets.

2.4 Details of Methods

In this section we describe our method for finding trends and then how we use these to compute the status gradient. We run this method for each of these datasets separately so we can compare the communities with each other. In each of these communities, users produce textual content, and so for unity of terminology we will refer to the textual output in any of these domains (in the form of posts, comments, reviews, and publication titles) as a set of *documents*; similarly, we will refer to the producers of any of this content (posters, commenters, reviewers, researchers) as the *authors*. For Amazon, Reddit, and the beer communities we use an approach that is essentially identical across all the domains; the DBLP and Arxiv datasets have a structure that necessitates some slight differences that we will describe below.

2.4.1 Finding Trends

As discussed above, the trends we analyze are associated with *word bursts* — words that increase in usage in a well-defined way. We compute word bursts using an underlying probabilistic automaton as a generative model, following [45]. These word bursts form the set of trends on which we then base the computation of status gradients.

For each dataset (among Amazon, Reddit, and the beer communities), and for each word w in the dataset, let α_w denote the fraction of documents in which it appears. We define a two-state automaton that we imagine to probabilistically generate the presence or absence of the word w in each document. In its "low state" q_0 the automaton generates the word with probability α_w, and in its "high state" q_1 it generates the word with probability $c_D\alpha_w$ for a constant $c_D > 1$ that is uniform for the given dataset D. Finally, it transitions between the two states with probability p. (In what follows we use $p = 0.1$, but other values give similar results.)

Now, for each word w, let $f_{w,1}, f_{w,2}, \ldots, f_{w,n}$ be a sequence in which $f_{w,i}$ denotes the fraction of documents in week i that contain w. We compute the state sequence $S_{w,1}, S_{w,2}, \ldots, S_{w,n}$ (with each $S_{w,i} \in \{q_0, q_1\}$) that maximizes the likelihood of observing the fractions $f_{w,1}, f_{w,2}, \ldots, f_{w,n}$ when the automaton starts in q_0. Intuitively, this provides us with a sequence of "low rate" and "high rate" time intervals that conform as well as possible to the observed frequency of usage, taking into account (via the transition probability p) that we do not expect extremely rapid transitions back and forth between low and high rates. Moreover, words that are used very frequently throughout the duration of the dataset will tend to produce state sequences that stay in q_0, since it is difficult for them to rise above their already high rate of usage.

A *burst* is then a maximal sequence of states that are all equal to q_1, and the be-

ginning of this sequence corresponds to a point in time at which w can be viewed as "trending." The *weight* of the burst is the difference in log-probabilities between the state sequence that uses q_1 for the interval of the burst and the sequence that stays in q_0.

To avoid certain pathologies in the trends we analyze, we put in a number of heuristic filters; for completeness we describe these here. First, since a word might produce several disjoint time intervals in the automaton's high state, we focus only on the interval with highest weight. For simplicity of phrasing, we refer to this as *the* burst for the word (Other choices, such as focusing on the first or longest interval, produce similar results). Next, we take a number of steps to make sure we are studying bursty words that have enough overall occurrences, and that exist for more than a narrow window of time. The quantity c_D defined above controls how much higher the rate of q_1 is relative to q_0; too high a value of c_D tends to produce short, extremely high bursts that may have very few occurrences of the word. We therefore choose the maximum c_D subject to the property that the median number of occurrences of words that enter the burst state is at least 5000. Further, we only consider word bursts of at least eight weeks in length, and only for words that were used at least once every three months for a year extending in either direction from the start of the burst.

With these steps in place, we take the top 500 bursty words sorted by the weight of their burst interval, and we use these as the trending words for building the status gradient. With our heuristics in place, each of these words occurred at least 200 times. For illustrative purposes, a list of top 5 words for each dataset is shown in Table 2.2.

Dataset	Words
Amazon Music	anger, metallica, coldplay, limp, kanye
Amazon Movies and TV	lohan, lindsay, sorcerers, towers, gladiator
Amazon Books	kindle, vinci, bush, phoenix, potter
Reddit Music	daft, skrillex, hipster, radiohead, arcade
Reddit movies	batman, bane, superman, bond, django
Reddit books	hunger, nook, borders, gatsby, twilight
Reddit worldnews	israel, hamas, isis, gaza, crimea
Reddit gaming	gta, skyrim, portal, diablo, halo
Rate Beer	cigar, tropical, winter, kernel, farmstead
Beer Advocate	finger, tulip, pine, funk, roast
DBLP	parallel, cloud, social, database, objectoriented
Arxiv astro-ph	chandra, spitzer, asca, kepler, xmmnewton
Arxiv stats-cs	deep, channels, neural, capacity, convolutional
Arxiv All	learning, chandra, xray, spitzer, bayesian

Table 2.2: The top 5 words that our algorithm finds using the burst detection method. Words in parenthesis are stop words that got removed by the algorithm.

2.4.2 Computing the Status Gradient

Now we describe the computation of the status gradient. This follows the overview from earlier in the chapter, with one main change. In the earlier overview, we described a computation that used only the the documents containing a single bursty word w. This, however, leads to status gradients (as functions of time) that are quite noisy. Instead, we compute a single, smoother aggregate status gradient over all the bursty words in the dataset.

Essentially, we can do this simply by merging all the time-stamped documents containing any of the bursty words, including each document with a multiplicity corresponding to the number of bursty words it contains, and shifting the time-stamp on each instance of a document with bursty word w to be relative to the start of the burst for w. Specifically, each of the bursty words w selected above has a time β_w at which its burst interval begins. For each document containing w, produced at time T, we define

its *relative time* to be $T - \beta_w$ — i.e. time is shifted so that the start of the burst is at time 0. (Time is measured in integer numbers of weeks for all of our datasets except DBLP and Arxiv, where it is measured in integer numbers of years and months, respectively.)

Now we take all the documents and we bucket them into groups that all have the same relative time: for each document produced at time T containing a bursty word w, we place it in the bucket associated with its relative time $T - \beta_w$.[2] From here, the computation proceeds as in the overview earlier in the chapter: for each relative time t, we consider the median activity level $g(t)$ of all documents in the bucket associated with t. This function $g(t)$ plays the role of the single-word function $g_w(t)$ from the overview, and the computation continues from there.

Final and Current Activity Levels. The computation of the status gradient involves the activity levels of users, and there are two natural ways to define this quantity, each leading to qualitatively different sets of questions. The first is the *final activity level*: defining each user's activity level to be the lifetime number of documents they produced. Under this interpretation, an author will have the same activity whenever we see them in the data, since it corresponds to their cumulative activity. This was implicitly the notion of activity level that was used to describe the status gradient computation earlier.

An alternate, also meaningful, way to define an author's activity level is to define it instantaneously at any time t to be the number of documents the author has produced up to time t. This reflects the author's involvement with the community at the time he or she produced the document, but it does not show his or her eventual activity in the community.

[2] If a document contains multiple bursty words, we place it in multiple buckets. Also, to reduce noise, in a post-processing step we combine adjacent buckets if they both have fewer than a threshold number of documents θ, and we continue this combining process iteratively from earlier to later buckets until all buckets have at least θ documents. In our analysis we use $\theta = 1500$.

Performing the analysis in terms of the final activity level is straightforward. For the analysis in terms of the current activity level, we need to be careful about a subtlety. If we directly adapt the method described so far, we run into the problem that users' current activity levels are increasing with time, resulting in status gradient plots that increase monotonically for a superficial reason. To handle this issue, we compare documents containing bursty words with documents which were written at approximately the same time. Document d written at time t_d that has a bursty word will be compared to documents written in the same week as d. We say that the *fractional rank* of document d is the fraction of documents written in the same week t_d whose authors have a smaller current activity level than the author of d. Note that the fractional rank is independent of the trending word; it depends only on the week. Now that each document has a score that eliminates the underlying monotone increase, we can go back to the relative time domain and use the same method that we employed for the final activity level, but using the fractional rank instead of the final activity level. Note that in this computation we thus have an extra level of indirection — once for finding the fractional rank and a second time for computing the status gradient function.

As it turns out, the analyses using final and current activity levels give very similar results; due to this similarity, we focus here on the computation and results for the final activity level.

Bigrams. Thus far we have performed all the analysis using trends that consist of single words (unigrams). But we can perform a strictly analogous computation in which the trends are comprised of bursty two-word sequences (bigrams), after stop-word removal. Essentially all aspects of the computation remain the same. The top 5 bigrams that the algorithm finds are shown in Table 2.3.

Dataset	Bigrams
Amazon Music	st-anger, green-day, limp-bizkit, 50-cent, x-y
Amazon Movies and TV	mean-Girls, rings-trilogy, lindsay-lohan, matrix-reloaded, two-towers
Amazon Books	da-Vinci, john-kerry, harry-potter, twilight-book, fellowship-(of the)-ring
Reddit Music	daft-punk, get-lucky, chance-rapper, mumford-(&)-sons, arctic-monkeys
Reddit movies	pacific-rim, iron-man, man-(of)-steel, guardians-(of the)-galaxy, dark-knight
Reddit books	hunger-games, shades-(of)-grey, gone-girl, great-gatsby, skin-game
Reddit worldnews	north-korea, chemical-weapons, human-shields, iron-dome, civilian-casualties
Reddit gaming	gta-v, last-(of)-us, mass-effect, bioshock-infinite, wii-u
Rate Beer	cigar-city, black-ipa, belgian-yeast, cask-handpump, hop-front
Beer Advocate	lacing-s, finger-head, moderate-carbonation, poured-tulip, head-aroma

Table 2.3: The top 5 bigrams that our algorithm finds using the burst detection method. Words in parenthesis are stop words that got removed by the algorithm.

The results for bigrams in all datasets are very similar to those for unigrams, and so in what follows we focus on the results for unigrams.

DBLP and Arxiv. Compared to other datasets that we use in this study, DBLP and Arxiv have a different structure in ways that are useful to highlight. We will point out two main differences.

First, documents on DBLP/Arxiv generally only arrive in yearly/monthly increments, rather than daily or weekly increments in the other datasets, and so we perform our analyses by placing documents into buckets corresponding to years/month rather than weeks. In our heuristics for burst detection on DBLP, we require a minimum burst length of 3 years (in place of the previous requirement of 8 weeks). We found it was not necessary to use any additional minimum-length filters.

The second and more dramatic structural difference from the other datasets is that a given document will generally have multiple authors. To deal with this issue, we adopt the following simple approach: We define the current and final activity level of a document as the highest current and final activity level, respectively, among all its authors.[3] Note, however, that a document still contributes to the activity level of all its authors.

We observe that the bursty words identified for these datasets appear in at least 70 documents each instead of the minimum 200 we saw for the other datasets. We scaled down other parameters accordingly, and did not compute bursty bigrams for DBLP and Arxiv.

2.5 Results

Now that we have a method for computing status gradients, we combine the curves $f_w(t)$ over the top bursty words in each dataset, as described above, aligning each bursty word so that time 0 is the start of its burst, β_w. In the underlying definition of the status gradient, we focus here on the final activity level of users; the results for current user activity are very similar.

2.5.1 Dynamics of Activity Levels

The panels of Figure 2.1 show the aggregate status gradient curves for the three Amazon categories, four of the sub-reddits, and one of the beer communities. (Results for the other sub-reddits and beer communities are similar.)

[3]The results for taking the median experience instead of the maximum for each paper leads to similar results.

The plots in Figure 2.1 exhibit two key commonalities.

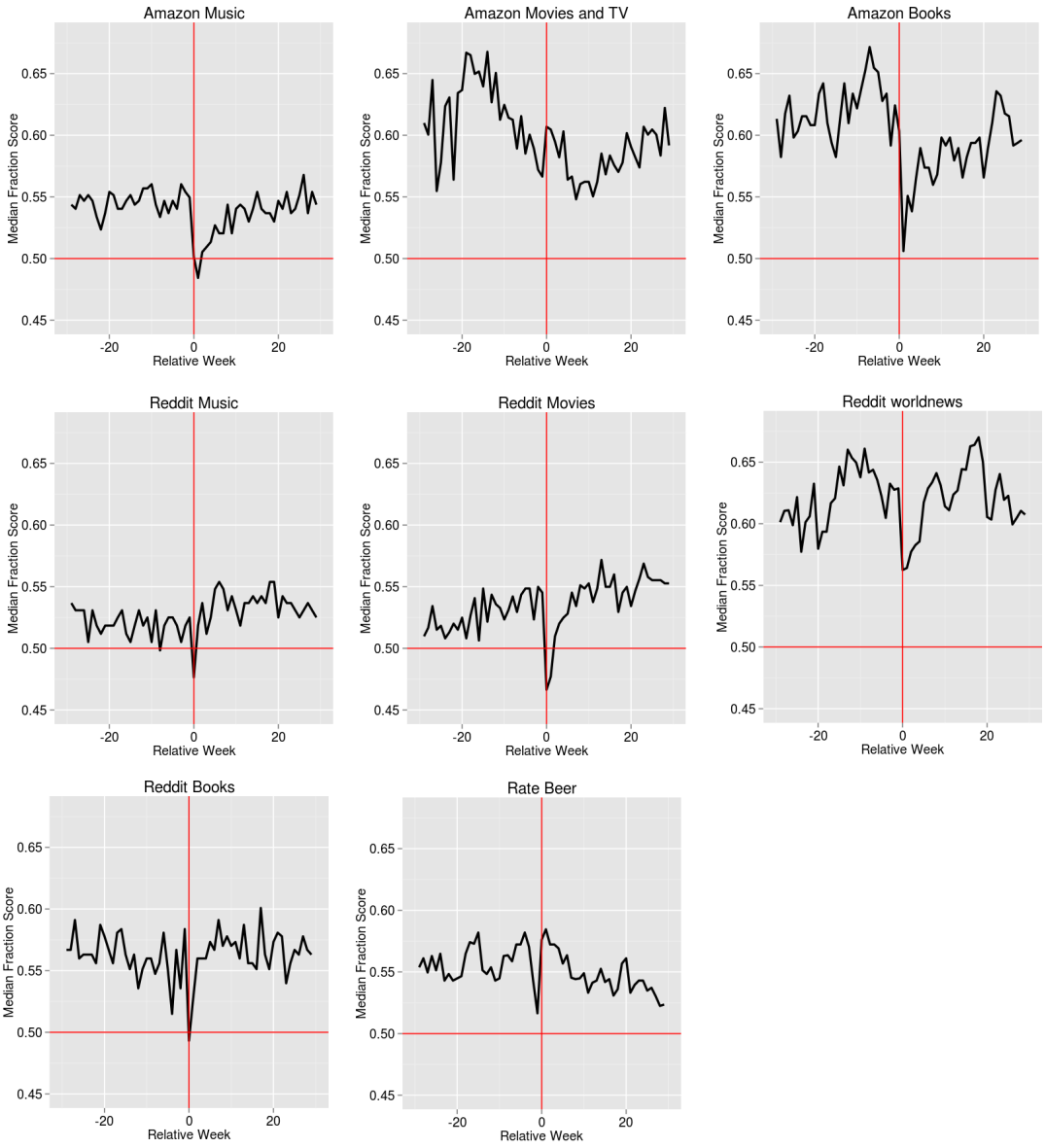


Figure 2.1: The status gradients for datasets from Amazon, Reddit, and an on-line beer community, based on the final activity level of users and a ranked set of 500 bursty words for each dataset.

- First, they lie almost entirely above the line $y = 0.5$. Recalling the definition of the status gradient, this means that high-activity individuals are using bursty words at a rate *greater* than what their overall activity level would suggest. That is, even

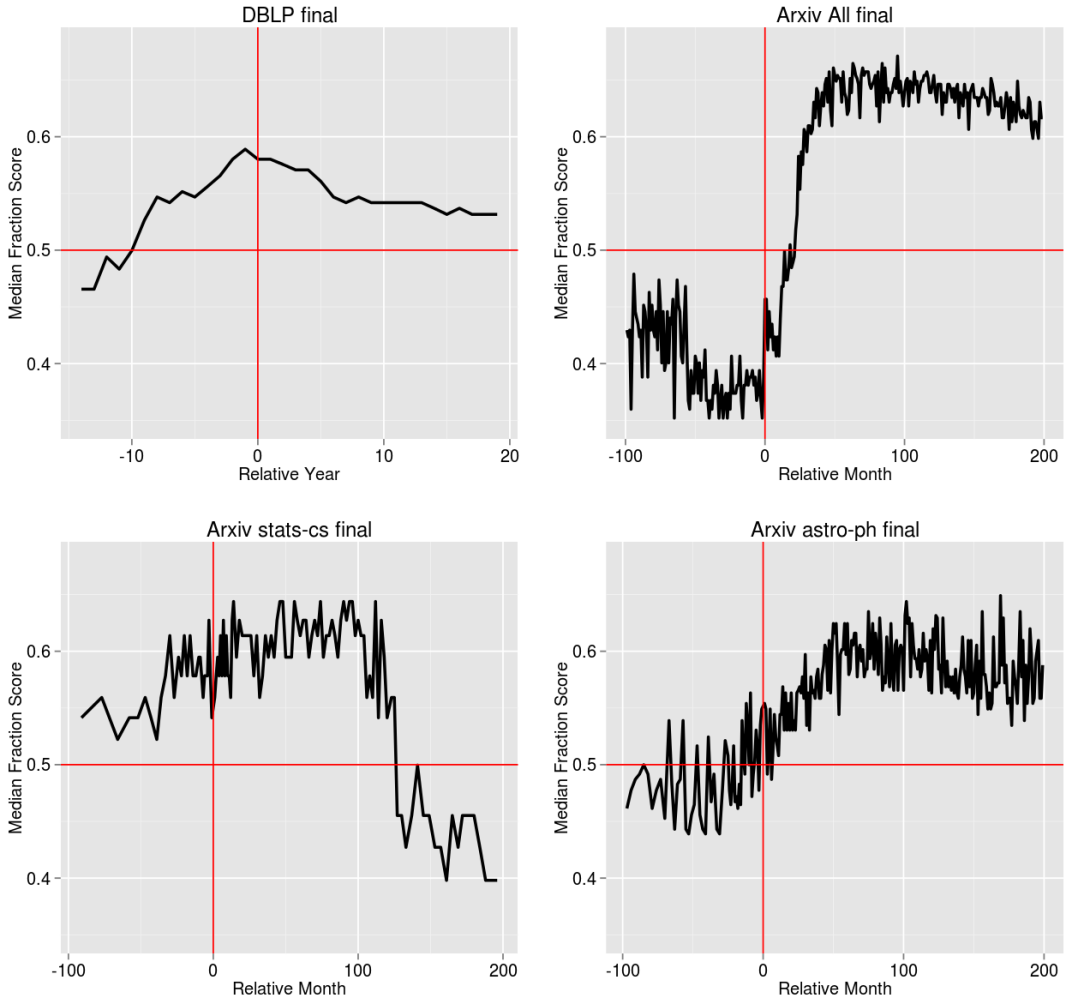


Figure 2.2: The status gradient for DBLP and Arxiv papers, as well as the stats-cs and astro-ph subsets of Arxiv, using final activity levels.

relative to their already high level of contribution to the site, the most active users are additionally adopting the trending words.

- However, there is an important transition in the curves right at relative time $t = 0$, the point at which the burst begins. For most of these communities there is a sharp drop, indicating that the aggregate final activity level of users engaging in the trend is abruptly reduced as the trend begins. Intuitively, this points to an influx of lower-activity users as the trend starts to become large. This forms interesting parallels with related phenomena in cases where users pursue content that has

become popular [5, 19].

This pair of properties — overrepresentation of high-activity users in trends (even relative to their general activity level); and an influx of lower-activity users at the onset of the trend — are the two dominant dynamics that the status gradient reveals. Relative to these two observations, we now identify a further crucial property, the distinction between producers and consumers.

2.5.2 Producers vs. Consumers

We noted that the academic domains we study exhibit a considerably different status gradient. On DBLP (Figure 2.2), the activity level of authors rises to a maximum very close to relative time $t = 0$, indicating an influx of high-activity users right at the start of a trend. Arxiv stats-cs shows the same effect, and the other Arxiv datasets show a time-shifted version of this pattern, increasing through time 0 and reaching a maximum shortly afterward. (This time-shifting of Arxiv relative to DBLP may be connected to the fact that Arxiv contains preprints while DBLP is a record of published work, which may therefore have been in circulation for a longer time before the formal date of its appearance.) This dramatic contrast to the status gradients in Figure 2.1 highlights the fact that there is no single "obvious" behavior at time $t = 0$, the start of the trend. It is intuitive that low-activity users should rush in at the start of a trend, as they do on Amazon, Reddit, and the beer communities; but it is also intuitive that high-activity users should arrive to capitalize on the start of a trend, as they do on DBLP and Arxiv. A natural question is therefore whether there is an underlying structural contrast between the domains that might point to further analysis.

Here we explore the following contrast. We can think of the users on Amazon, Red-

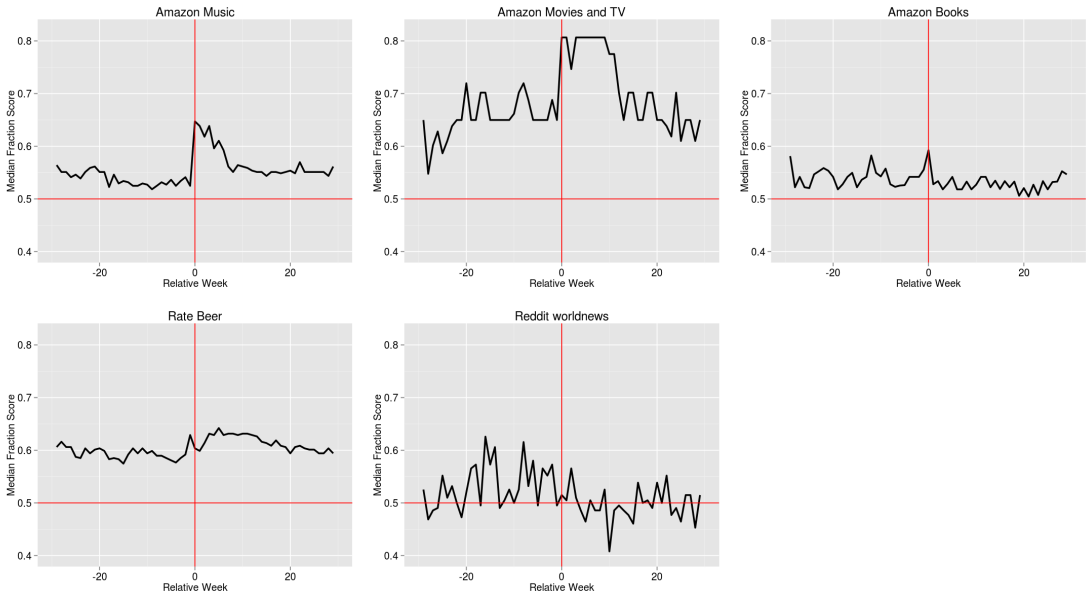


Figure 2.3: Status gradients for producers — brands on Amazon and the beer community, and domains for Reddit World News. As functions of time, these status gradients show strong contrasts with the corresponding plots for the activity levels of users (consumers).

dit, and the beer communities as *consumers* of information: they are reviewing or commenting on items (products on Amazon, generally links and news items on Reddit, and beers on the beer communities) that are being produced by entities outside the site. DBLP and Arxiv are very different: its bibliographic data is tracking the activities of *producers* — authors who produce papers for consumption by an audience. Could this distinction between producers and consumers be relevant to the different behaviors of the status gradients?

To explore this question, we look for analogues of producers in the domains corresponding to Figure 2.1: if the status gradient plots in that figure reflected populations of consumers, who are the corresponding producers in these domains? We start with Amazon; for each review, there is not just an *author* for the review (representing the consumer side) but also the *brand* of the product being reviewed (serving as a marker for the producer side). We can define activity levels for brands just as we did for users,

based on the total number of reviews this brand is associated with, and then use this in the Amazon data to compute status gradients for brands rather than for users.

The contrasts with the user plots are striking, as shown in Figure 2.3, and consistent with what we saw on DBLP and Arxiv: the status gradients for producers on Amazon go up at time $t = 0$, and for two of the three categories (Music and Movies/TV), the increase at $t = 0$ is dramatic. This suggests an interesting producer-consumer dynamic in bursts on Amazon, characterized by a simultaneous influx of high-activity brands and low-activity users at the onset of the burst: the two populations move inversely at the trend begins. Intuitively, the onset of a burst is characterized by producers of rising activity level moving in to provide content to consumers of falling activity level.

We can look for analogues of producers in the other two domains from Figure 2.1 as well. For Rate Beer, each review is accompanied by the brand of the beer, and computing status gradients for brands we find a mild increase at $t = 0$ here too — as on Amazon, contrasting sharply with the drop at $t = 0$ for the user population. For Reddit, it is unclear whether there is a notion of a "producer" as clean as brands in the other domains, but for Reddit World News, where most posts consist of a shared link, we can consider the domain of the link as a kind of producer of the information. The status gradient for domains on Reddit World News is noisy over time, but we see a generally flat curve at $t = 0$; while it does not increase at the onset of the trend, it again contrasts sharply with the drop at $t = 0$ in the user population.

Posters vs. Commenters

As a more focused distinction, we can also look at contrasts between different sub-populations of users on certain of the sites. In particular, since the text we study on

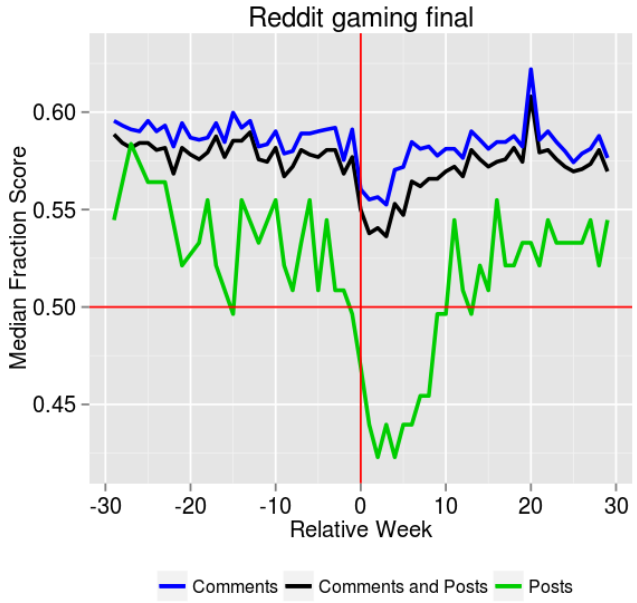


Figure 2.4: A comparison between the status gradients computed from posts, comments, and the union of posts and comments on a large sub-reddit (gaming) .

Reddit comes from threads that begin with a post and are followed by a sequence of comments, we can look at the distinction between the status gradients of posters and commenters.

We find (Figure 2.4) that high-activity users are overrepresented more strongly in the bursts in comments than in posts; this distinction is relatively minimal long before the burst, but it widens as the onset of the burst approaches, and the drop in the status gradient at $t = 0$ is much more strongly manifested among the posters than the commenters. This is consistent with a picture in which lower-activity users initiate threads via posts, and higher-activity users participate through comments, with this disparity becoming strongest as the trend begins.

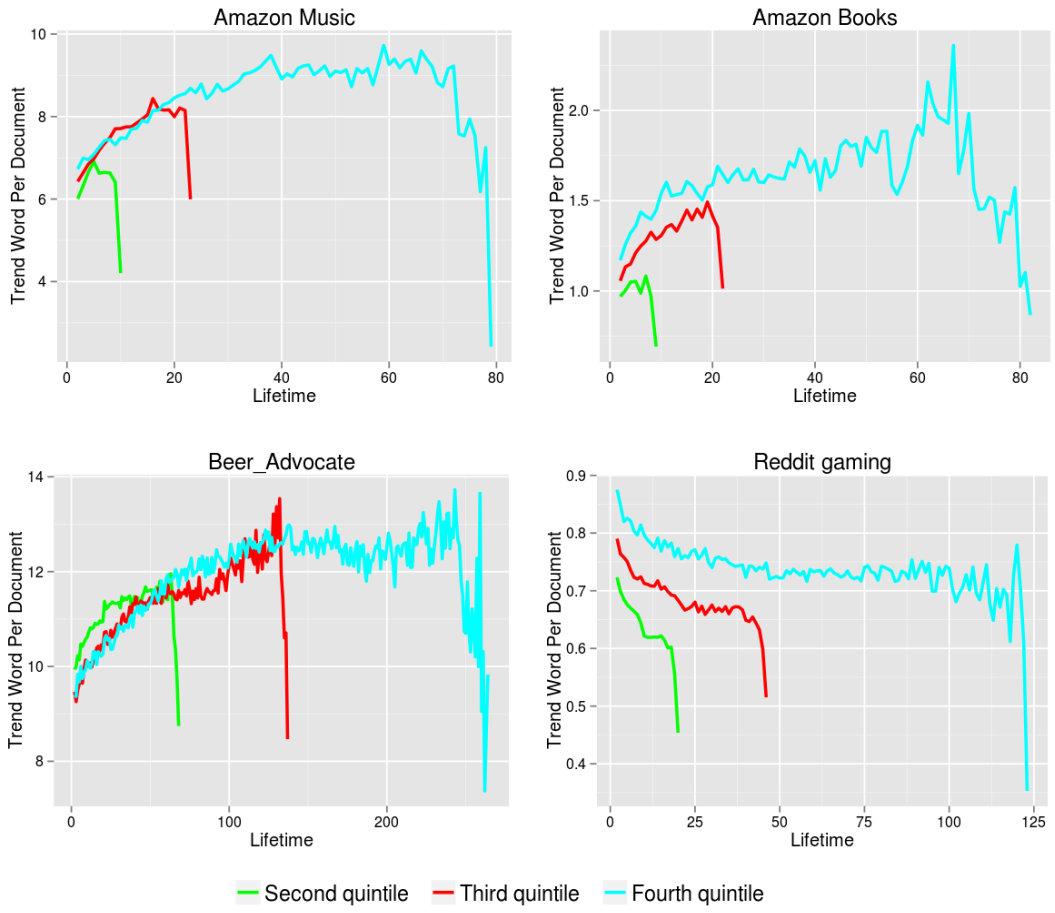


Figure 2.5: The average number of bursty words used per document, as a function of the author's life stage in the community.

Life stages

As a final point, we briefly consider a version of the dual question studied by Danescu-Niculescu-Mizil et al [25] — rather than tracking the life cycles of the words, as we have done so far, we can look at the life cycles of the users and investigate how they use bursty words over their life course on the site. One reason why it is interesting to compare to this earlier work using a similar methodology is that we are studying a related but fundamentally different type of behavior from what they considered. The word usage that they focused on can be viewed as *lexical innovations*, or novelties, in that they are

words that had never been used before at all in the community. Here, on the other hand, we are studying trending word usage through the identification of bursts — the words in our analysis might have been used a non-zero number of times prior to the start of the burst, but they grew dramatically in size when the burst began, thus constituting trending growth. It is not at all clear a priori that users' behavior with respect to bursty words over their lifetime should be analogous to their behavior with respect to novelties, but we can investigate this by adapting the methodology from Danescu-Niculescu-Mizil et al [25].

Here is how we set up the computation. First, we remove any authors (together with the documents they have written) if their final activity level is less than 10, since their life span is too short to analyze. Then, we find four cut-off values that divide authors into *quintiles* — five groups based on their final activity level such that each group has produced a fifth of the remaining documents. We focus on the middle three of these quintiles: three groups of different final activity levels who have each collectively contributed the same amount of content.

We then follow each author over a sequence of brief *life stages*, each corresponding to the production of five documents. For each life stage and each quintile we find the average number of bursty words per document they produce.

We find that the aggregate use of bursty words over user life cycles can look different across different communities. A representative sampling of the different kinds of patterns can be seen in Figure 2.5. For many of the communities, we see the pattern noted by Danescu-Niculescu-Mizil et al, but adapted to bursty words instead of lexical innovations — the usage increases over the early part of a user's life cycle but then decreases at the end. For others, such as Reddit gaming shown in the figure, users have the highest rate of adoption of bursty words at the beginning of their life cycles, and it decreases steadily from there. As with our earlier measures, these contrasts suggest the

broader question of characterizing structural differences across sites through the different life cycles of users and the trending words they adopt.

2.6 Conclusion

In this chapter, we have a proposed a definition, the *status gradient*, and shown how it can be used to characterize the adoption of a trend across a social media community's user population. In particular, this has allowed us to study the following contrast, which has proven elusive in earlier work: are trends in social media primarily picked up by a small number of the most active members of a community, or by a large mass of less central members who collectively account for a comparable amount of activity? Our goal has been to develop a clean, intuitive computational formulation of this question, in a manner that makes it possible to compare results across multiple datasets. We find recurring patterns, including a tendency for the most active users to be even further overrepresented in trends, and a contrast between the underlying dynamics for consumers versus producers of information.

CHAPTER 3

TRACING THE USE OF PRACTICES THROUGH NETWORKS OF COLLABORATION

An active line of research has used on-line data to study the ways in which discrete units of information—including messages, photos, product recommendations, group invitations—spread through social networks. There is relatively little understanding, however, of how on-line data might help in studying the diffusion of more complex *practices* — roughly, routines or styles of chapter that are generally handed down from one person to another through collaboration or mentorship. In this work, we propose a framework together with a novel type of data analysis that seeks to study the spread of such practices by tracking their syntactic signatures in large document collections. Central to this framework is the notion of an *inheritance graph* that represents how people pass the practice on to others through collaboration. Our analysis of these inheritance graphs demonstrates that we can trace a significant number of practices over long timespans, and we show that the structure of these graphs can help in predicting the longevity of collaborations within a field, as well as the fitness of the practices themselves.

3.1 Introduction

On-line domains have provided a rich collection of settings in which to observe how new ideas and innovations spread through social networks. A growing line of research has discovered principles for both the local mechanisms and global properties involved in the spread of pieces of information such as messages, quotes, links, news stories, and photos [4, 39, 54, 55, 53, 3, 20, 37, 12], the diffusion of new products through viral marketing [52], and the cascading recruitment to on-line groups [10, 6].

A common feature in these approaches has been to trace some discrete "unit of transmission" that can be feasibly tracked through the underlying system: a piece of text, a link, a product, or membership in a group. This is natural: the power of on-line data for analyzing diffusion comes in part through the large scale and fine-grained resolution with which we can observe things flowing through a network; therefore, to harness this power it is crucial for those things to be algorithmically recognizable and trackable. As a result, certain types of social diffusion have been particularly difficult to approach using on-line data—notably, a broad set of cascading behaviors that we could refer to as *practices*, which are a collection of styles or routines within a community that are passed down between people over many years, often through direct collaboration, mentorship or instruction. Particular stylistic elements involved in writing software, or performing music, or playing football, might all be examples of such practices in their respective fields. While complex practices are one of the primary modes studied by qualitative research in diffusion [80], the challenge for large-scale quantitative analysis has been both to recognize when someone has begun to adopt a practice, and also to identify how it was transmitted to them.

Tracking the Spread of Practices. A natural approach to tracking the spread of a practice is to find a concretely recognizable "tag" that tends to travel with the practice as it is handed down from one person to another, rendering its use and transmission easily visible. A beautiful instance of this strategy was carried out by David Kaiser in his analysis of the use of *Feynman diagrams* in physics [41]. Feynman diagrams were proposed by Richard Feynman as a way to organize complex physics calculations, and due to the technical sophistication involved in their use, the initial spread of Feynman diagrams within the physics community proceeded in much the style described above, with young researchers adopting the practice through collaboration with colleagues who

had already used it. In contrast to many comparable practices, Feynman diagrams had a distinctive syntactic format that made it easy to tell when they were being used. As a result, their spread could be very accurately tracked through the physics literature of the mid-20th-century. The result, in Kaiser's analysis, was a detailed map of how an idea spread through the field via networks of mentorship. As he writes:

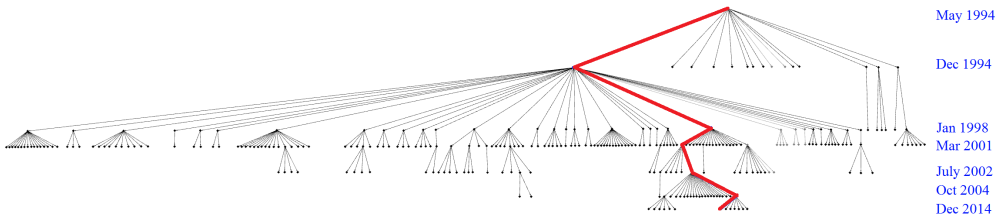
> *The story of the spread of Feynman diagrams reveals the work required to craft both research tools and the tool users who will put them to work. The great majority of physicists who used the diagrams during the decade after their introduction did so only after working closely with a member of the diagrammatic network. Postdocs circulated through the Institute for Advanced Study, participating in intense study sessions and collaborative calculations while there. Then they took jobs throughout the United States (and elsewhere) and began to drill their own students in how to use the diagrams. To an overwhelming degree, physicists who remained outside this rapidly expanding network did not pick up the diagrams for their research. Personal contact and individual mentoring remained the diagrams' predominant means of circulation even years after explicit instructions for the diagrams' use had been in print. [41]*

The Feynman diagram thus functions in two roles in this analysis: as an important technical innovation, and as a "tracking device" for mapping pathways of mentorship and collaboration. If we want to bring this idea to a setting with large-scale data, we must deal with the following question: where can we find a rich collection of such tracking devices with which to perform this type of analysis? We do not expect most objects in this collection to be technical advances comparable to the Feynman diagram, but we need a large supply of them, and we need to be able to mechanically recognize both their use and their spread.

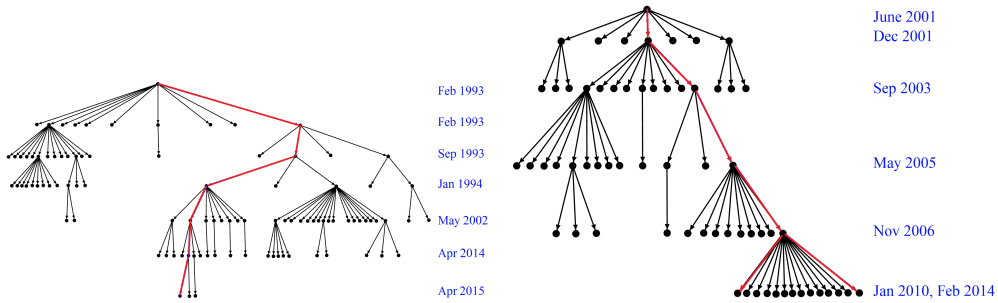
The present work: Diffusion of practices in academic writing. In this chapter, we describe a framework for tracking the spread of practices as they are passed down through networks of collaboration, and we demonstrate a number of ways in which our analysis has predictive value for the underlying system. We make use of a setting where practices have the recognizability that we need.—a novel dataset of latex macros in the e-print arXiv that we have recently developed.

In writing the LaTeX source for a paper, authors will often define one or more macros as a way to make the writing of the paper easier and more modular; as in a standard programming language, each instance of a macro instance is specified by a name (henceforth `name`) and a body definition (henceforth `body`) which defines the functionality of the macro. Each time the formatting software for the paper sees the name of the macro, it replaces it with the body of the macro in the text; thus, for example, the command, `\def\Reals{\mathbb{R}}` defines a macro, and whenever the author writes `\Reals` in the source file for the paper, the symbol $\mathbb{R}$ will appear in the outcome.

We will use this dataset in the next chapter for other purposes, specifically, treating macros names as instances of naming conventions. Macros in our context have a number of the key properties we need. First, a latex macro is something whose presence can be tracked as it spreads through the papers in the arXiv collection; we can thus see when an author first uses it, and when their co-authors use it. Second, while an arbitrary macro clearly does not correspond in general to an important technical innovation, a sufficiently complex macro often does encode some non-trivial technical shorthand within a concrete sub-field, and hence its use signifies the corresponding use of some technical practice within the field. And finally, there are several hundred thousand latex macros in papers on the arXiv, and so we have the ability to track a huge number of such diffusion events, and to make comparative statements about their properties.



(a) \hbox{$\rm\thinspace L_{\odot}$}



(b) \mbox{\boldmath Y}

(c) \mathrel{\mathpalette\@versim>}

Figure 3.1: Sample subsets of BFS trees for three different macros. At each depth we show the date when the highlighted author (node) uses the macro for the first time; the highlighted edge is a paper in which an author passes on the macro to an author (node) in the next level of the tree.

If we want to use macros to trace the diffusion of practices between collaborators, we first need to establish whether macros indeed spread via "inheritance" from co-authors: as with the Feynman diagram, can most of the initial set of uses of a macro trace a path back to a single early use through a chain of co-authorship? We find that this is true for a significant fraction of macros, by using an *inheritance graph* for each macro that records how each author's first use can be imputed to a co-authorship with an earlier user of the macro. Specifically, for each macro we can build a graph on the set of authors who have used it, and we include a directed *inheritance edge* from author u to author v if (i) u used the macro before v did, and (ii) v's first use of the macro is in a paper with u. We find that many of these inheritance graphs contain giant directed subtrees rooted at a single early use of the macro, indicating that a significant fraction of the users of the macro can indeed trace a direct path back to a single shared early ancestor under this inheritance relation.

These structures represent interesting instances of diffusion for several reasons. First, they are "organic" in a way that the spread of many on-line memes are not: when we study on-line diffusion in settings where a user's exposure to content is governed by a recommendation system or ranking algorithm, there is the added complexity that part of the diffusion process is being guided by the internals of the algorithms underlying the system. With macros in arXiv papers, on the other hand, while authors may use automated tools to format the source of their papers, there is relatively little influence from automated recommendations or rankings in the actual decisions to include specific macros. Second, we are studying processes here that play out over years and even decades; among other findings about the structure of our inheritance graphs, we observe that their diameters can take multiple years to increase even by one hop. We are thus observing effects that are taking place over multiple academic generations.

The present work: Estimating fitness. If these inheritance graphs—obtaining by tracing simple syntactic signatures in the source files of papers—are telling us something about the spread of practices through the underlying community, then their structural properties may contain latent signals about the outcomes of authors, topics, and relationships. In the latter part of this chapter, we show that this is the case, by identifying such signals built from the inheritance structures, and showing that they have predictive value.

As one instance, suppose we wish to estimate the future longevity of a collaboration between two authors u and v—that is, controlling for the number of papers they have written thus far, we ask how many papers they will write in the future. If (u, v) is an edge of the inheritance graph for some macro, does this help in performing such an estimate? One might posit that since this edge represents something concrete that u passed on to v in their collaboration, we should increase our estimate of the strength of the relationship

and hence its future longevity. This intuition turns out not to be correct on its own: the existence of a (u, v) edge by itself doesn't significantly modify the estimate. However, we find that something close to this intuition does apply. First, we note that since a (u, v) edge only means that a macro used by u showed up subsequently in a paper that u co-authored with v, it is providing only very weak information about v's role in the interaction. We would have a stronger signal if (u, v) were an *internal edge* of some inheritance graph, meaning that v has at least one outgoing edge; in this case, v was part of a paper that subsequently passed the macro on to a third party w. We find that if (u, v) is an internal edge of an inheritance graph, this does in fact provide a non-trivial predictive signal for increased longevity of the u-v collaboration; informally, it is not enough that u passed something on to v, but that v subsequently was part of the process of passing it on to a third party w. In fact, we find something more: when (u, v) is an edge that is not internal (so that u's passing on of the macro "ends" at v), it in fact provides a weak predictive signal that the collaboration will actually have slightly *lower* longevity than an arbitrary collaboration between two co-authors (again controlling for the number of joint papers up to the point of observation).

In what follows, we formalize this analysis and its conclusion. We also develop analyses through which macro inheritance can be used to help estimate the future longevity of an author—how many papers will they write in the future? —and the fitness of an individual macro itself—how many authors will use it in the future?

The remainder of the chapter is organized into three main sections. We first briefly describe the structure of the data and how it is used in our analyses. We then formally define the inheritance graphs and survey some of their basic properties. Finally, we analyze the relation between these inheritance structures and the longevity of co-authorships, authors, and macros.

3.2 Data Description

The dataset we study contains the macros used in over one million papers submitted to the e-print arXiv from its inception in 1991 through November 2015. The arXiv is a repository of scientific pre-prints in different formats, primarily in LaTeX. For a prefix of this time period, the ordering of the papers in our data is only resolved up to one-month granularity (the remainder is totally ordered), but our methods work with this level of granularity.

From the LaTeX source files we extract all macros defined by the most common methods, specifically `\def`, `\newcommand` and `\renewcommand`. This results in macros from over 400,000 papers. Note that we do not recursively substitute `names` that occur inside of another macro `body`. Macros have two major components, the `name` and the `body`. Whenever the author uses `\name` the LaTeX compiler replaces it with the `body` and compiles the text. In our study the `body` serves as the "tracking device" discussed in the introduction, for studying how a macro is passed between collaborators over time. In general, when we refer to a "macro", we mean a macro `body` unless specified otherwise. For our study we use macro `bodies` that have length greater than 20 characters, and which have been used by at least 30 different authors. We apply the length filter so that we can focus on macros that are distinctive enough that we expect them to move primarily through copying and transmission, rather than independent invention.
Table 3.1 summarizes basic statistics about our data.[1]

3.3 Inheritance Graphs

[1] Our macro dataset is available at `http://github.com/CornellNLP/Macros`; a repository of arXiv papers is available at `http://arxiv.org/help/bulk_data_s3`.

Number of papers with a macro	402,478
Number of macros defined	15,771,021
Number of unique macro `bodies`	2,586,548
Average number of `names` per `body`	1.1
Number of unique authors	168,451
Average author per paper	2.2

Table 3.1: Dataset details

Defining inheritance graphs. We begin by formally defining the *inheritance graphs* described in the introduction. For each macro m we create a graph (V_m, E_m) where V_m is the set of authors who have used macro m in at least one of their papers. We add a directed edge (u, v) to the edge set E_m if there is a paper that uses m with u and v as co-authors, such that (i) this is v's first use of m, but (ii) u has used m in at least one previous paper. This is the formal sense in which m is being passed from u to v: v's first use of m occurs in collaboration with u, a prior user of m. Note that there can be multiple edges leading into a single node. For instance take a paper with authors u, v and z that uses macro m, and assume that u and v have used m before but z is using it for the first time; then both the edges (u, z) and (v, z) are in the graph.

Now, if all authors of a paper p are using m for the first time, then the nodes corresponding to these authors will not have any incoming edges. (Nodes of this form are the only ones with no incoming edges.) For each such paper p, we replace the nodes corresponding to the authors of p with a single *supernode* corresponding to p. We will refer to this as a *source node*, and to the authors of p as *source authors*. The resulting graph, with supernodes for papers where no author has used the macro before, and with author nodes for all others, is the inheritance graph G_m for the macro m. Because the process of inheritance, as defined, goes forward in time, G_m is necessarily a directed acyclic graph (DAG).

Using these graphs we should be able to trace back a macro's life to its inception

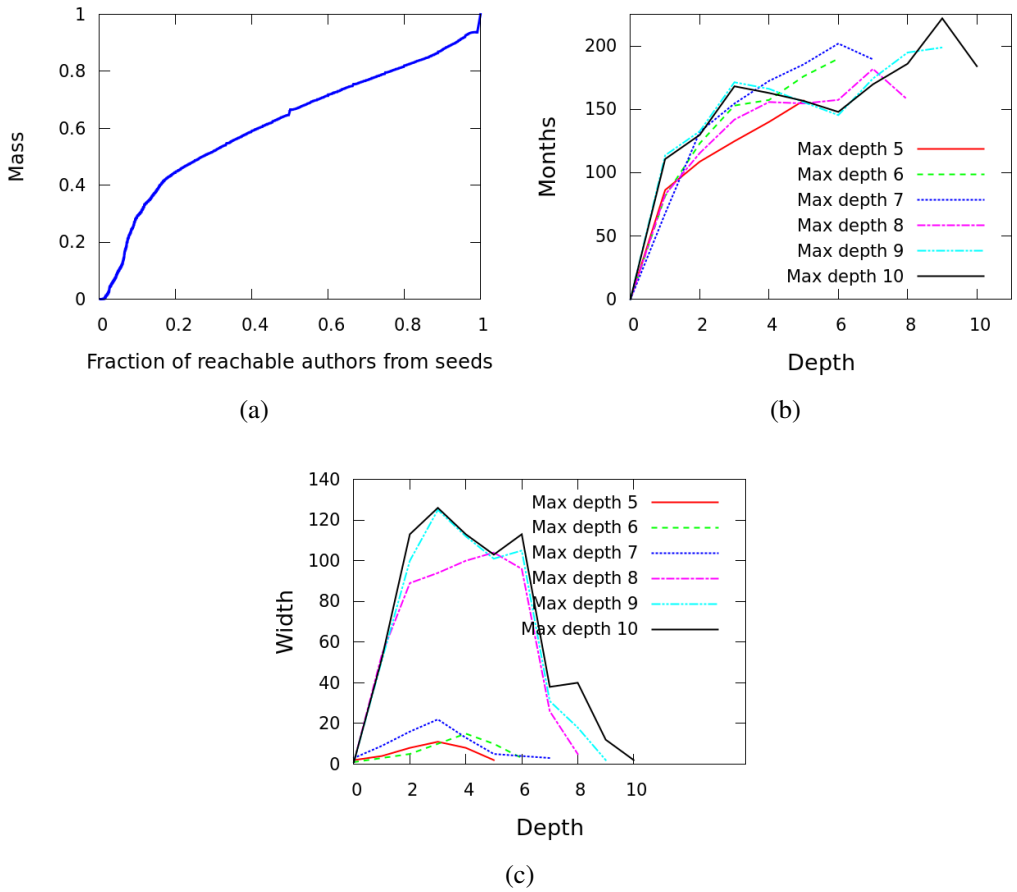


Figure 3.2: (a) The CDF for the ratio of the largest reachable set to the number of nodes in the graph. (b) The average number of months that pass from the date of appearance of the root paper to the date of appearance of nodes at a given depth, grouped by the maximum depth of the tree. (c) The average number of nodes in each depth, for the largest reachable set for each macro.

and to the authors who first used it. Note that there might be multiple source papers, and hence several groups of co-authors who independently serve as "origins" for the macro. For portions of the analysis where we are interested in looking at the number of authors who all follow from a single source paper, we will identify the source paper that has directed paths to the largest number of nodes in the graph G_m. We will refer to this as the *seed paper*, and to the set of authors of this paper as the *seed authors*. (Note that the seed paper might not be the chronologically earliest paper to use the macro m; it is simply the one that can reach the most other nodes.)

Analyzing the inheritance graphs.. Our dataset contains several hundred thousand different macros, and as a first step we analyze the properties of the graphs G_m that they produce. In Figure 3.1 we take three sample macros and show subsets of the breadth-first search (BFS) trees that are obtained starting from the seed paper. For example in Figure 3.1(a) the graph is created on the macro, `\hbox{$\rm\thinspace L_{\odot}$}` and the seed node is the paper **astroph/9405052** with authors Xavier Barcons and Maria Teresa Ceballos. The seed paper used this macro in 1994, and some of the nodes at depth 6 in the BFS tree are from 2014—a 20-year time span to reach a depth of 6 in the cascading adoption of the macro. This reinforces the sense in which we are studying cascades that play out on a multi-generational time scale of decades, rather than the time scale of minutes or hours that characterizes many on-line cascades. The seed node of Figure 3.1(b) is the paper **hep-th/0106008** with authors Selena Ng and Malcolm Perry, and the seed node of Figure 3.1(c) is the paper **hep-ph/9302234** with authors Jose R Lopez et al. Since all other nodes in these BFS trees have incoming edges, they all correspond to individual authors who enter the graph at their first adoption of the macro, whereas the root node corresponds to a single paper and to the contracted set of authors of this paper.

We now consider some of the basic properties of these inheritance graphs. First, each source paper has a *reachable set* in G_m—the set of nodes it can reach by directed paths—and recall that we defined the seed paper to be the source paper with the largest reachable set. In Figure 3.2(a) we observe that a non-trivial fraction of the macros have a seed paper whose reachable set is a large fraction of all the authors who eventually adopt the macro. This provides a first concrete sense in which the inheritance patterns contained in G_m represent a global structure that spans much of the use of the macro m.

In Figure 3.2(b) and 3.2(c) we show the properties of the graphs and nodes grouped based on the maximum depth of the BFS tree and the depth of the individual nodes.

Figure 3.2(b) shows the average time it takes for the macro to get from the root to the nodes in each depth grouped by the maximum depth of the tree. This figure shows how these cascades can take multiple years to add a single level of depth to the tree, and a decade or more to reach their eventual maximum depth. In Figure 3.2(c) we show the median width (number of nodes) of trees at each depth, again grouped by the maximum depth of the tree. Based on this plot we see that most of these trees have are narrow in their top and bottom layers, with fewer nodes, and are wider in the middle.

The plots thus far have been concerned with the global structure of the inheritance graph and its shortest paths as represented by breadth-first search trees. Now we take a deeper look at the properties of individual edges in the graph. For this we will first define the notions of local and global experience, and we will use these two terms throughout. At time t the *global experience* of an author is the number of papers the respective author has written. At time t the *local experience* of an author is defined with respect to a macro m and is the number of papers up to time t in which the author has used m. This is a version of the notion of local experience relative to an arbitrary term, as used in [76].

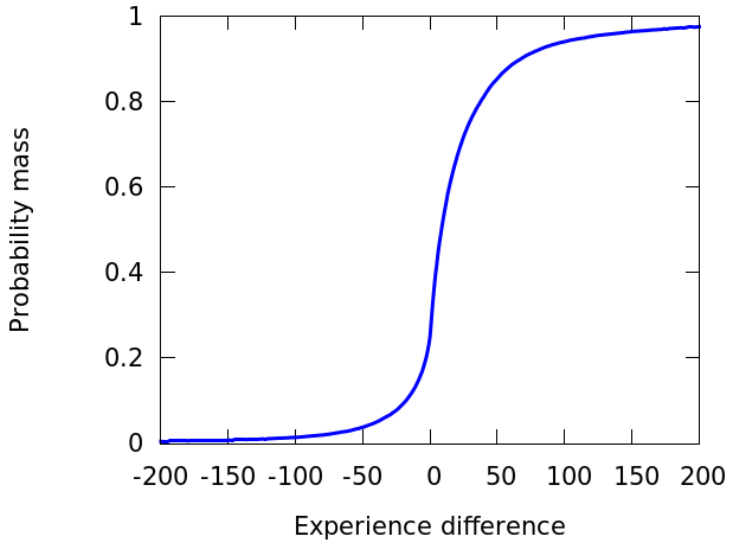


Figure 3.3: The Cumulative Distribution Function of the global experience difference between the source and destination of an edge.

Consider an edge (u, v) in the inheritance graph for a macro m. At the moment when the macro is passed from u to v, the local experience of v with respect to m is 0 by definition, and the local experience of u with respect to m is greater than 0. What do we expect about the global experience of these two nodes? To the extent that passing on a macro is a form of "teaching" from one person to another, we may expect the global experience of u (the "teacher") to be higher than the global experience of v (the "learner"). On the other hand, there is a history of sociological work in the diffusion of innovations suggesting that innovations often originate with outsiders who come from the periphery of the system [25, 60, 79, 86], which would be consistent with v having higher global experience than u. Figure 3.3 addresses this question by showing the cumulative distribution of the global experience difference between u and v. The median experience difference is clearly shifted in the positive direction, consistent with the "teacher" node u having the higher global experience in general.

3.4 Fitness

Now that we have some insight into how the information diffusion process unfolds in our data, we investigate whether these inheritance structures can provide predictive signal for the outcomes of co-authorships, authors, and the macros themselves. In all cases we will think in terms of the *fitness* of the object in question—the extent to which it survives for a long period of time and/or produces many descendants.

3.4.1 Fitness of collaborations

We start by considering the fitness of collaborations—given two authors u and v who have written a certain number of papers up to a given point in time, or perhaps who have not yet collaborated, can we use anything in the structure of macro inheritance to help predict how many more papers they will write in the future?

A natural hypothesis is that if v inherits macros from u, then this indicates a certain strength to the relationship (following the teacher-learner intuition above), and this may be predictive of a longer future history of collaboration. To examine this hypothesis, we perform the following computational test as a controlled paired comparison. We find pairs of co-authorships u-v and u'-v' with properties that (i) neither pair has collaborated before, (ii) their first co-authorship happens in the same month, (iii) (u, v) is an edge in an inheritance graph, and (iv) (u', v') is not. (Note that since we are looking at pairs of co-authorships, we are looking at four authors in total for each instance: u, v, u', and v'.) Now we can ask, aggregating over many such pairs of co-authorships, whether there is a significant difference in the future number of papers that these pairs of authors write together. (Since their initial co-authorships took place in the same month, they have a comparable future time span in which to write further papers.)

In fact, we find that there isn't a significant difference, at odds with our initial hypothesis about macro inheritance. However, there is more going on in the inheritance structure that we can take advantage of. We divide the edges of the inheritance graphs into two sets: *internal edges* (u, v), where the node v has at least one outgoing edge, and *terminal edges* (u, v), where the node v has no outgoing edge. Internal edges add extra structural information, since they indicate that not only u passed the macro m to v, but that v was then part of the process of passing m in a collaboration subsequent to the one in which they originally inherited it.

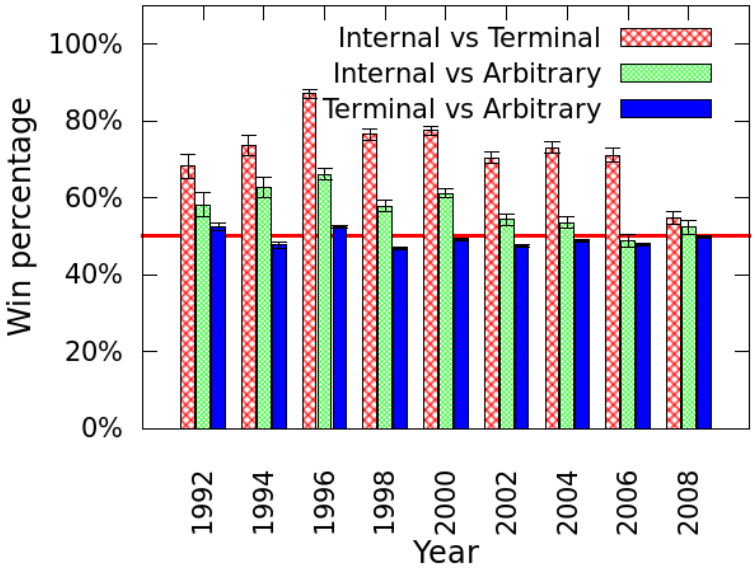


Figure 3.4: Comparison of three different co-authorship settings through different years in the data. The bars show the win percentage of the first of the two listed categories; e.g., the red bar indicates the percentage of times co-authors with internal edges end up writing more papers than the matched co-authors with terminal edges. The horizontal red line indicates the 50% baseline.

We find that the fitness of u-v co-authorships is significantly higher when (u, v) forms an internal edge, in contrast to the lack of effect when (u, v) is an arbitrary edge. We evaluate this using an extension of our previous paired comparison: in conditions (i)-(iv) above for forming pairs of co-authorships, we replace conditions (iii) and (iv) with the following:

- Internal edge vs. arbitrary co-authorship: (iii) (u, v) is an internal edge and (iv) (u', v') is not an edge.
- Internal edge vs. terminal edge: (iii) (u, v) is an internal edge and (iv) (u', v') is a terminal edge.
- Terminal edge vs. arbitrary co-authorship: (iii) (u, v) is a terminal edge and (iv) (u', v') is not an edge.

In each of these three settings, we look at the fraction of times that one of the categories produced the co-authorship with more future papers. In our paired setting, if we were to draw two co-authorships uniformly at random over all possible co-authorships (without regard to the type of the edge), there is a 50% chance that the first would produce the higher number of future papers. Thus, we can calibrate each of the three comparisons listed above using this 50% baseline. Figure 3.4 shows these results, grouped into two-year bins: we find that internal edges win a large fraction of the comparisons against each of the other two categories, whereas there is little difference between terminal edges and arbitrary co-authorships.

3.4.2 Fitness of authors

We now consider the fitness of the authors themselves; we will show that the way authors use macros can provide a weak but non-trivial signal about how many papers they will eventually write, a quantity that we refer to as the *fitness* of the author.

The particular property we consider is a type of "stability" in the usage of the macro. For a given macro `body`, there are many possible `names` that can be used for it, and authors differ in the extent to which their papers preserve a relatively stable choice of names for the same macro body: some almost always use the same name, while for other authors the name changes frequently. (For example, an author who almost always uses the name `\vbar` for the macro body `$\overline{v}$`, versus an author whose papers alternate between using `\vbar`, `\barv`, `\vb`, `\vbarsymb`, and others, all for this same macro body.) We could think of the first type of author as exerting more control over the source of her papers than the second type of author, and this distinction between the two types of authors—based on their behavior with respect to macros—naturally raises the

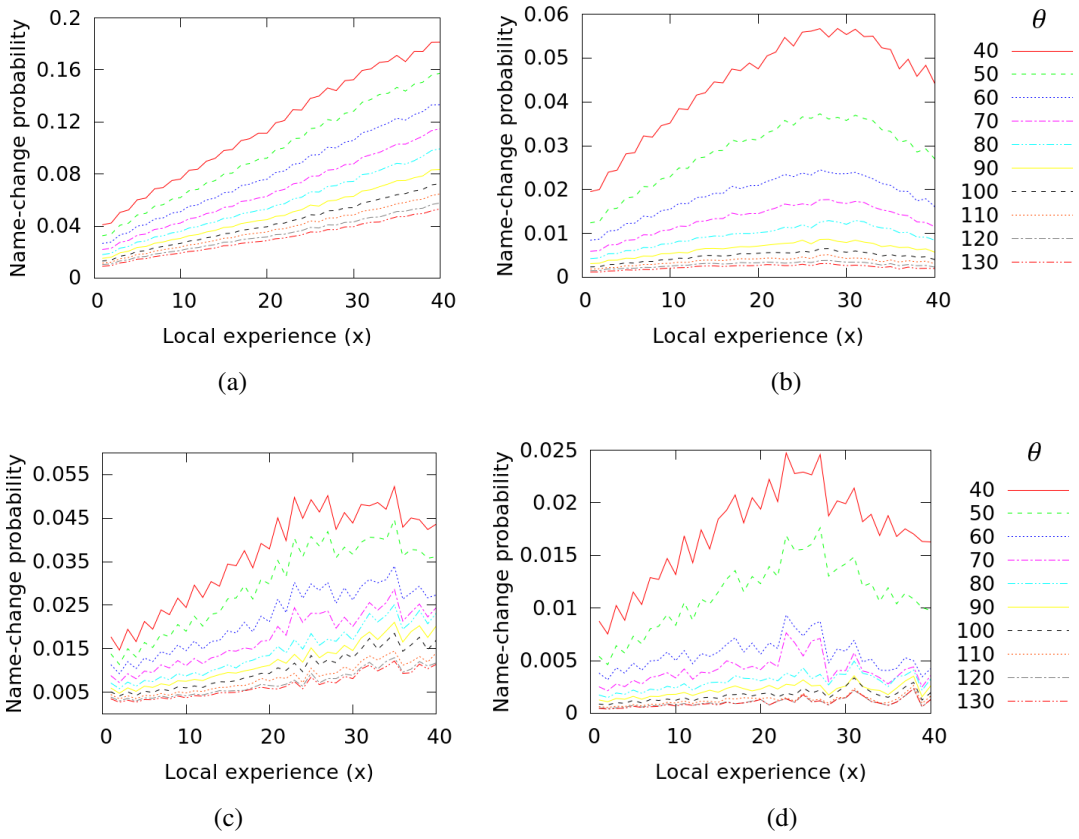
question whether the stability of macro names could provide predictive value for author fitness.

Here is how we formally define this measure. For a particular author a, we say they change the name of macro m on paper p if the previous time they used m's macro body, the name was different. Then, for a set of authors A and a set of macros M, we define $f(A, M, x)$ to be the probability of an author in A changing the name of a macro in M the x^{th} time they use it. We consider this *name-change probability*, $f(A, M, x)$ for $x \in [0, 40]$ and different groups of authors and macros. In particular we look at groups of authors that have more than θ papers in the entire corpus. We set θ to be $40, 50, \ldots, 130$ and we let M range over three possible sets: the set of all macros; the set of *wide-spread macros* (more than 250 authors use the macro body); and the set of *narrow-spread macros* (at least 20 authors used it and at most 250).

One source of variability in this analysis is that even once we fix the minimum number of papers θ written by an author a, as well as the usage number x of the macro m that we are considering, it is still possible that author a's x^{th} use of the macro might come toward the end of their professional lifetime or early in their professional lifetime. (It must come at the x^{th} paper they write or later, since they need time to have used the macro m a total of x times, but this is all we know.) It is easy to believe that authors who use a macro in their early life stages might exhibit different phenomena from those who use it in a later life stage. Therefore, in addition to the measures defined so far, we also consider analyses involving only the set of macro uses that come early in the authors' professional lifetime—specifically only macro uses that happen in the author's first 40 papers.

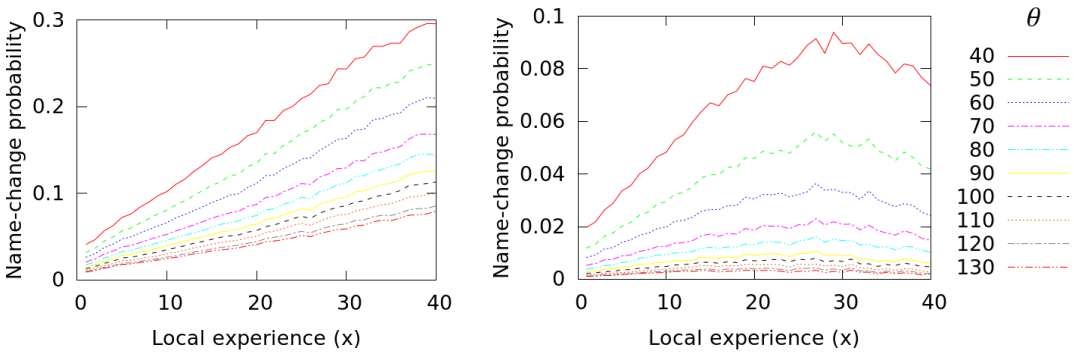
The results for all these settings are shown in Figure 3.4: the three sets of macros (all macros, wide-spread macros, and narrow-spread macros); for each of these sets, we

consider both the authors' full lifetimes and just their early life stages. In each case, the x-axis shows the number of macro uses (i.e. the authors' local experience with respect to the macro), and the different curves represent authors grouped by different values of the minimum number of papers θ.



This suggests that overtime authors build a certain "loyalty" to the names they have used consistently; this is consistent with our findings in the next chapter regarding the competition between macro-naming conventions.

But we also find something else: that (eventually) more prolific authors (larger θ) have a lower name-change probability (compare ordering of curves in each subplot of Figure 3.4. This suggests that the macro name change probability might be a signal with predictive value for author fitness (which, again, we define as the number of papers the author will eventually write).



(e) (f)

Figure 3.4: Each panel shows the probability an author changes the name of a macro on their x^{th} use of it. A single curve in each plot shows the set of all authors with at least θ papers, for θ equal to $40, 50, \ldots, 130$. Each row of panels corresponds to a different set of macros: the first row shows results for the set of all macros; the second for the set of narrow-spread macros; and the third for the set of wide-spread macros (as defined in the text). The left column of panels shows the analysis for each of these three sets over the authors' full professional lifetimes. The right column of panels shows the analysis for each of these three sets restricted to the authors' early life stages (first 40 papers only). Thus, the panels are (a) full lifetimes, all macros; (b) early life stages, all macros; (c) full lifetimes, narrow-spread macros; (d) early life stages, narrow-spread macros; (e) full lifetimes, wide-spread macros; (f) early life stages, wide-spread macros.

To test this idea, we set up an author fitness prediction task as follows. For a given minimum number of papers θ we consider the low-fitness authors to be the ones with fitness below the 20^{th} percentile and high-fitness authors to be those above the 80^{th} percentile. We then see whether simply using the frequency with which an author changes macro names in the first θ papers can serve as a predictor for this two-class problem: whether an author's fitness is below the 20th percentile or above the 80th percentile.

By using the probability of macro name changes, we are able to predict which of these two classes an author belongs to with a performance that exceeds the random baseline of 50% by a small but statistically significant amount. Figure 3.5 shows the performance for different values of θ. We emphasize that predicting an author's fitness is a challenging task for which one doesn't expect strong performance even from rich

Papers revealed (θ)	20'th Percentile	80'th Percentile
10	13	38
20	25	58
30	36	73
40	47	87
50	58	99

Table 3.2: Global experience thresholds used in defining the author fitness classes for each number θ of papers revealed.

feature sets; this makes it all the more striking that one can obtain non-trivial performance from the frequency of macro name changes, a very low-level property about the production of the papers themselves.

Moreover, for settings involving large values of θ the name-change probability is more predictive than an arguably more natural structural feature: the author's total number of co-authors (Figure 3.5). We also note that in such settings the name-change feature also outperforms other more direct macro-based features, such as the total number of macros used, or total number of distinct macro bodies used.

3.4.3 Fitness of macros

Finally, we consider the fitness of the macros themselves. We define the fitness of a macro to be the total number of authors who eventually use the macro body, and investigate which features are predictive of this variable.

We set up a prediction task as follows. We first find all macros that get adopted by at least k authors. Each of these macros has a fitness (of at least k), and we define $\sigma(k)$ to be the median of this multiset of fitness values: of all macros that reach at least k authors, half of them have a fitness of at most $\sigma(k)$, and half of them have a fitness of at

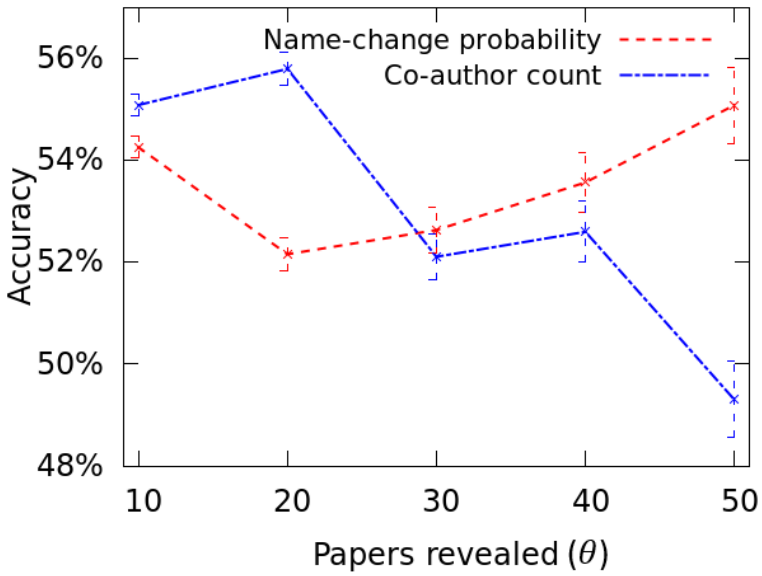


Figure 3.5: The accuracy of predicting the number of publications of an author given her first few papers, θ. We compare the performance of the name-change probability features with the features based on number of co-authors.

least $\sigma(k)$. In table 3.3 we report $\sigma(k)$ and the number of macro instances for different values of k.

k	$\sigma(k)$	Instances
40	98	49,415
80	156	30,107
120	242	20,119
160	340	14,662
200	437	11,794

Table 3.3: Summary of the macro fitness prediction dataset: For a macro that reaches at least k authors, the task is to predict whether it will eventually reach $\sigma(k)$ authors (the median fitness of such macros).

We can thus use $\sigma(k)$ to construct a balanced prediction task, in the style of the cascade prediction analyses from [20]. For a given macro that reaches at least k authors, we observe all the information on the papers and authors up to the point at which the k^{th} author adopts the macro, and the task is then to predict if this macro will eventually reach $\sigma(k)$ authors. We learn a logistic regression model for different values of k and

report the accuracy in Figure 3.6 on an 80-20 train-test split.[2]

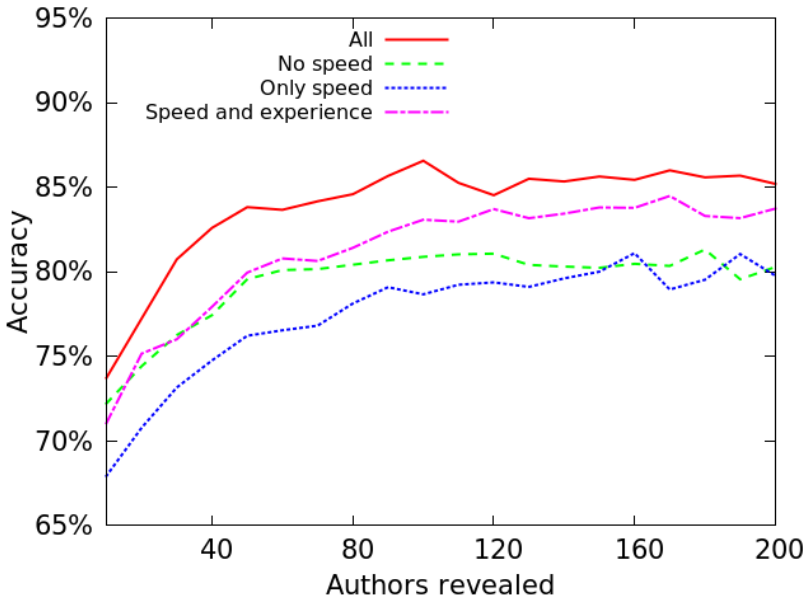


Figure 3.6: The accuracy of predicting how widely a macro spreads, using different subsets of features.

We use the following features.

- Features related to the *diffusion speed* of the macro: the number of papers that the macro needs in order to reach $\frac{k}{2}$ and k distinct authors; and the number of months that the macro needs in order to reach $\frac{k}{2}$ and k distinct authors.
- *Experience of the macro users*: the average usage experience of the first k authors who adopted it.
- *Structural features of the macro users*: the local and global clustering coefficients of the co-authorship graph on the first k authors to use the macro.
- *Structural features of the macro body*: the length of the macro body, the number of dollar signs in the macro (generally used for mathematical notation), the num-

[2]We can achieve a 1% to 4% better accuracy by using a non-linear classifier such as decision trees, but we opt to use the more interpretable model.

ber of non-alphanumerical characters, and the maximum depth of nested curly brackets.

In Figure 3.6 we show the prediction performance for different subsets of these features, as a function of k; note that performance increases with increasing k. As observed above, predicting macro fitness is a problem whose syntactic form is closely analogous to the prediction of cascade size for memes in social media [20]; given this, and the fact that the spread of macros plays out over so much longer time scales, and without the role of ranking or recommendation algorithms, it is interesting to note the similarities and contrasts in the prediction results. One of the most intriguing contrasts is in the role of diffusion speed features: for cascade prediction in social media, the speed features alone yielded performance almost matching that of the full feature set, and significantly outperforming the set of all non-speed features [20]. For our domain, on the other hand, the speed features perform 5-10% worse than the full feature set; they also perform worse for most values of k than the set of all non-speed features. This suggests that for macro fitness, the speed features are considerably less powerful than they are in the social media context, indicating that there may be more to be gained from the synthesis of a much broader set of features.

3.5 Conclusion

The spread of practices between collaborators is a challenging form of diffusion to track, since one needs to be able to recognize when someone has begun using a practice, and how it was conveyed to them. Motivated by work that used the Feynman diagram as an easily recognizable "tracer" of a complex practice [41], we track the spread of several hundred thousand macros through the papers of the e-print arXiv over a 25-year period.

Long macros often serve as technical shorthand within a defined sub-field, and their syntactic precision makes it easy to follow their flow through the collaboration network. We construct *inheritance graphs* showing how the macro spread between collaborators, and we find that many macros have a clear "seed set" of authors with the property that a large fraction of the subsequent users of the macro can trace a direct inheritance path back to this seed set. The resulting diffusion patterns are intriguing, in that they span multiple academic generations and several decades, and unlike cascades in social media, the spread of these macros takes place with very little influence from ranking or recommendation algorithms.

We also find that properties of macro inheritance provide signals that are predictive for larger-scale properties that have nothing to do with macros. These include predictions about the longevity of collaborations and the number of papers that an author will write over their professional lifetime on the arXiv.

CHAPTER 4

COMPETITION AND SELECTION AMONG CONVENTIONS

In many domains, a latent competition among different conventions determines which one will come to dominate. One sees such effects in the success of community jargon, of competing frames in political rhetoric, or of terminology in technical contexts. These effects have become widespread in the on-line domain, where the ease of information transmission makes them particularly forceful, and where the available data offers the potential to study competition among conventions at a fine-grained level.

In analyzing the dynamics of conventions over time, however, even with detailed on-line data, one encounters two significant challenges. First, as conventions evolve, the underlying substance of their meaning tends to change as well; and such substantive changes confound investigations of social effects. Second, the selection of a convention takes place through the complex interactions of individuals within a community, and contention between the users of competing conventions plays a key role in the convention's evolution. Any analysis of the overall dynamics must take place in the presence of these two issues.

In this chapter we study a setting in which we can cleanly track the competition among conventions while explicitly taking these sources of complexity into account. Our analysis is based on the spread of low-level authoring conventions in the e-print arXiv over 24 years and roughly a million posted papers: by tracking the spread of macros and other author-defined conventions, we are able to study conventions that vary even as the underlying meaning remains constant. We find that the interaction among co-authors over time plays a crucial role in the selection of conventions; the distinction between more and less experienced members of the community, and the distinction between conventions with visible versus invisible effects, are both central to the underlying

processes. Through our analysis we make predictions at the population level about the ultimate success of different synonymous conventions over time — and at the individual level about the outcome of "fights" between people over convention choices.

4.1 Introduction

Our work here begins by noting two central methodological challenges that arise in studying the evolution of conventions: one is an issue of content versus structure, and the other is an issue of local versus global effects.

• *Substantive shift.* First, it is possible for one convention to eclipse another because of a *substantive shift* — in which the substantive meaning of one convention has a relative advantage over the other. If we seek to understand the role that social structure plays, we must look for settings in which the competing conventions are essentially *synonymous* at the level of their substance. The work on this question to date has faced the challenge that in most natural settings it is hard to verify whether competing conventions are semantically equivalent, and thus to disentangle social effects from the relative advantage of one convention over another.

This is a distinction that is also familiar in studies of biological evolution, where the use of *neutral variation* — mutations that have minimal effect on an organisms's fitness — has come to be an enormously influential methodology for studying effects of population structure on evolution in the absence of overt selective pressure [44]. What is the analogue of neutral variation in the diffusion of on-line information?

• *Diffusion through interaction.* Second, much of the work on diffusion — both theoretically and via on-line data — has studied the global competition among conventions

via local mechanisms of contagion: these mechanisms posit that at a local level, the use of the convention spreads from one person to another, either probabilistically or through best-response behavior, and the competition among these contagion processes leads to the global outcome. But in most domains where non-trivial conventions are competing, the competition takes place not just globally but also locally through person-to-person interaction. Returning to our examples above: a single interaction between speakers in a dialogue or discussion, collaborators on a technical project, politicians framing a shared position, or artists performing a shared work may all implicitly involve competition between the conventions used by the participants in this interaction. The global outcome of the competition between two conventions may emerge from the results of thousands or millions of these micro-level competitions. This type of *diffusion through interaction* requires a fine-grained analysis of the local competition, rather than just a view of the local dynamics as concurrent contagion processes.

In the rest of the chapter, we propose an analysis framework for the competition and selection among conventions that explicitly addresses these issues of neutral variation and diffusion through interaction. We do this through a set of novel definitions and measures, together with a rich source of data that clearly display both notions at work.

The data we use in this chapter is very similar to the data used in Chapter 3. We study how low-level authoring conventions emerge through the collaborations among different overlapping sets of co-authors over a multi-year time span. The source files on the arXiv provide a detailed view into a wide range of such authoring conventions; we focus primarily on the role of author-defined macros in LaTeX as one abundant supply of conventions.

The appeal of focusing on macros is that they provide an extremely rich source of synonymous conventions in the social ecosystem of the arXiv. Whenever the `name` for a

macro changes while the `body` remains the same — for example, when someone chooses to use `\R` instead of `\Reals` for the symbol $\mathbb{R}$ — the author is settling on an arbitrary choice of convention while the underlying meaning remains constant. As argued above, this type of control for meaning is crucial if we want to study the social structure around convention change separately from substantive shifts in content; however, controlling for meaning is very hard to achieve unless one has an almost mechanistic specification of this meaning. Macros provide us with precisely such a specification in their `body`. Also, they are pervasive in arXiv: roughly 40% of all arXiv papers contain at least one user-defined macro.

Moreover, because papers on the arXiv are largely co-authored, the competition among synonymous macros is also a powerful setting in which to define and then study some of the basic properties of diffusion through interaction. Two macro `names` with the same `body` are competing not just globally based on their relative prevalence in the full population of papers, but also locally each time two people who follow different conventions come together to co-author a paper. Analyzing the history of the arXiv provides us with a way to study how such instance-by-instance competition plays out in the context of these larger diffusion processes.

The arXiv thus provides us with the ingredients for analyzing information diffusion in a way that addresses these methodological challenges. At the same time, we note that the arXiv is of course a controlled domain representing a single type of broad activity — scientific authorship — and as such our work is approaching these issues via a case study of this particular domain. It will be interesting to study how the observations here generalize to different contexts; our approach is set up to facilitate this by providing a road map for these types of analyses across domains.

Overview of results. We begin by using the controlled setting provided by our data to study the competition between synonymous conventions at a global level. A concrete way to formulate this question is to look at two competing `names` for the same macro `body` up to a certain point in time, and ask whether we can predict which `name` will become dominant at some point in the future. First, we find that properties of the `name` itself — e.g., features related to its orthography, since the meaning is fixed — do not seem to have any predictive power; the differences in the competing `names` for the same macro `body` appear to truly represent neutral variation, a fact that offers a striking opportunity to explore other features in the absence of selective pressure.[1]

We find, instead, that features related to the *experience* of a `name`'s early users — the number of previous papers that each has written on the arXiv — have significant predictive value for the question of whether a macro `name` will grow to become dominant. In general, `names` that eventually become dominant tend to start with an initial author population that is relatively "younger" (with lower experience), and then they successfully spread to "older" users. `Names` that don't achieve dominance are more associated with initial user populations that are older in aggregate, and also fail to spread to new adopters with higher experience. These hand-offs between different "generations" of people, and how they contribute to the success of a convention, is an interesting issue connected to the role of status in diffusion [25, 69], and the results from our data suggest interesting directions in which to explore these issues further.

We next ask how this competition plays out at a local level, dropping down from the global scale in order to explore diffusion through interaction. We develop a framework for analyzing the instance-by-instance competition that arises when authors fol-

[1]The fact that the `name` itself provides no predictive power may be in part a reflection of the fact that authors tend to choose reasonable and informative `names` for their macros; it is easy to imagine that a particularly inapt `name` could have more difficulties in its adoption, but this is not the situation that generally seems to apply.

lowing different synonymous conventions meet to collaborate. In particular, if authors *A* and *B* meet for the first time to write a two-authored paper, and they have previously used different macro `names` for the same macro `body`, how does the resulting "fight" over the choice of convention turn out? We find that the relative experience level of the two authors is again a highly informative property of the interaction; the author who is "younger" (with lower experience) tends to win these fights, with the probability of winning increasing as the gap in experience grows. Building a set of features based on experience — both numerically and through certain more complex structural analogues, such as the authors' graph-theoretic properties in the larger co-author network — we are able to develop methods for predicting the outcomes of these fights with non-trivial accuracy.

It is an interesting question to consider possible mechanisms for the dominance of younger participants in these instance-by-instance fights; a natural hypothesis is that they play a larger role in the detailed implementation of the paper, and hence have more control over definitional questions such as macros. Such a model would suggest the conjecture that younger co-authors should not necessarily win fights over questions that are less about low-level implementation and more about high-level, visible decisions where the status of the older co-author is arguably more implicated. We show that this indeed appears to be the case, by studying the latent competition between co-authors over conventions in the title of the paper, rather than the macros. We can think of the title as occupying the opposite end of the visibility spectrum from macro names, in that decisions about titling conventions are highly visible; and here, under a set of definitions that we formulate here, we find that the older co-author tends to win fights about titling conventions, with the effect increasing as the experience gap increases (here in the opposite direction from what we saw in fights over macro `names`). In summary, our results suggest the beginnings of a set of principles that could be summarized in carica-

ture as, "In a collaboration, the younger people win the invisible fights while the older people win the visible fights." We argue that developing this notion more deeply is an interesting direction for further research, and we point to some additional steps along these lines.

More broadly, our focus is on developing new definitions around some fundamental issues that have been difficult to address in diffusion and the selection of conventions — the role of neutral variation, represented through the properties of synonymous conventions, and the dynamics of competition not just at a global level but through the continuous low-level competition between users of different conventions as they interact in the system. We hope that our exploration of these definitions and concepts in a case study on the arXiv will indicate how such analyses can be carried more broadly across other domains as well.

4.2 Further Related Work

Words as conventions. One of the most widespread sources of conventions is in the choice of words used to refer to particular concepts. As noted above, our use of macros is designed to contrast with an inherent source of complexity in the analysis of conventions in natural language, namely the severe difficulty in controlling for the precise meaning of a concept as the words referring to it change.

Analysis of changes in language over long time periods has considered a dual problem to ours: how fixed linguistic constructs aquire new meanings. This has been undertaken recently in studies of historical shifts in word meanings [81, 61, 40] and grammatical constructions [67], relying on books and news data that span long periods of time. Related studies have been performed in the on-line domain, analyzing global changes in

the linguistic system of Twitter [32, 31, 36] and other on-line communities [25, 51].

Sociolinguistic studies of linguistic change have addressed changes in phonology and spelling that vary systematically across time [49], status [50] and region [66]. As such, these studies are similar to ours in that they also explore variations in for,m of conventions used to refer to fixed concepts (e.g., whether the final 'r' in 'car' is pronounced or not), albeit the discussions are generally limited to a handful of examples.

Diffusion of information and cultural items. As discussed in the introduction, there has been a long line of work studying the processes by which discrete units of information diffuse on-line; these include memes [57, 64], hashtags on Twitter [70, 71, 56, 58] and on-line news content [2, 16, 11]. A growing strand of research within this topic has considered the problem of predicting future popularity, with specific prediction studies involving downloadable content [78], quotes embedded in broader cultural contexts [24, 15], hashtags [83], and memes [20, 84].

Most of these previous studies could not control for the meaning of the convention or content that is spreading, with two notable exceptions. The first is a study on the emergence of the retweet convention on Twitter [46]; this represents an in-depth study of a single convention providing extended insights, in contrast to our study of thousands of distinct conventions and the common properties across them. The second is a study of competition between hashtags on Twitter that only differ in capitalization, suffixes, or relative levels of abbreviation (e.g., *#saveTheNationalHealthService* vs. *#savethenationalhealthservice* vs. *#savetheNHS*) [84]. Our setting allows for a more general way of identifying synonymous conventions, and for verifying that they are indeed synonymous. And perhaps more crucially, the study of hashtags in [84] showed that the orthography of different hashtags was in fact predictive of their success, establishing that

in fact these different versions of hashtags do not represent neutral variations as in our case, but instead variation that affects relative fitness.

Role of experience. Our work also explores the interplay between individuals' levels of experience and their roles in the diffusion of conventions, including the question of whether new conventions originate with younger members of the community, or whether the older members have a relative advantage in imposing their forms of conventions. Such questions about trade-offs based on experience and status in the diffusion of innovations has a long history of study in off-line domains [27, 68, 79, 60, 86, 18, 47], and more recently has been explored in on-line domains as well [25, 76]. However, these lines of work do not look at instances where synonymous conventions compete with each other, or where it is possible to see such competition playing out at a local level through person-to-person contention.

4.3 Data

The data used in this chapter is the data introduced in the previous chapter. Further details about the dataset can be found in Table 3.1.

4.4 Global Competition Between Conventions

We begin by considering the competition between conventions, in the form of macro names, at a global level. To give a concrete sense for the behavior we are interested in studying, here is a simple example of competition among macro names — one of many with a similar flavor on the arXiv. In March 1996, Luty, Schmaltz, and Terning posted a

paper to the arXiv, on an application of gauge theories in theoretical physics, in which they defined the macro name \Yfund to expand to a macro body representing a very simple instance of a combinatorial structure known as a Young tableau:

```
\raisebox{-.5pt}{\drawsquare{6.5}{0.4}}
```

This macro body was used again (with the same name \Yfund) in two more papers in May 1996, seven more in the remainder of 1996, and a steady stream of others after that. Of the first 30 uses of this macro body, all but three referred to it by the name \Yfund. (The other three used \fun, a name that never really caught on.) But then, in a paper in May 1998, Hanany, Strassler, and Uranga used the name \fund to refer to this macro body. A competition soon broke out between \Yfund and \fund, with \fund gradually becoming more prevalent. The macro body has by now appeared in over 600 papers on the arXiv; of the most recent 100 uses, 39 used \Yfund and 61 used \fund. Figure 4.1 shows how this changeover between the two macro names took place; the x-axis is a time axis, indexed in order of uses of the macro body, and the y-axis shows a sliding-window average of the fraction of authors using the names \Yfund and \fund as a function of time.

This type of dynamic has played out with many macros on the arXiv, and the point is not that the choice of macro name is consequential for the substance of the authors' research. In fact, the point is the opposite: changes in the macro name are a source of neutral variation, essentially incidental to the real progress of the community, and hence they let us probe the changeover dynamics in conventions at the level of individuals, their characteristics, and their interactions.

We also note that the competition underlying the changeover dynamics can take several different possible forms. It may be that authors following the two conventions

interact directly through co-authorship or other mechanisms. But it may also be that one convention overtakes another even without direct interaction between the followers of the two conventions, simply because one of the two conventions grows in adopters and usage significantly faster than the other. This too is a form of competition between the conventions, played out in their relative rates of growth.

Defining changeovers. We now describe how we identify a broad set of instances in which one macro `name` overtakes another. For parameters s, q, and θ, we find macro `bodies` that have at least s total occurrences, where there is a `name` N_e that is the most used name in the first q fraction of occurrences, and a different `name` N_ℓ is the most used in the last q fraction of occurrences. Moreover, each of N_e and N_ℓ is widely used in the sense that N_e is used by more than a θ fraction of authors who use the macro `body` in its first q fraction of occurrences, and that N_ℓ is used by more than a θ fraction of authors who use the macro `body` in its last q fraction of occurrences. In our analysis, we use $s = 100$, $q = 0.3$ and $\theta = 0.3$, although other choices of these parameters produce similar results.

If these properties are met for a given macro `body` β, we say that β undergoes a *changeover* from N_e to N_ℓ, and we refer to N_e as the ***e**arly name* for β, and N_ℓ as the ***l**ate name* for β. In Figure 4.2 and 4.3 we show some aggregate properties of the set of changeovers on the arXiv. First, consider a given macro `body` β with m_β occurrences; for any $0 \leq t_0 \leq t_1 \leq 1$, we define the interval $[t_0, t_1]$ in the `body`'s lifespan to be the set of papers indexed between $t_0 m_\beta$ and $t_1 m_\beta$ in the time-sorted ordering of papers using β. We will refer to quantities like t_0 and t_1 as "times," (or "macro life-stages") corresponding to a fraction of the way through a macro `body`'s lifespan on the arXiv. Now, if β undergoes a changeover, we define the function $f_\beta(t, t')$ to be the fraction of authors in the interval $[t, t']$ that use the early name N_e, and we define the function

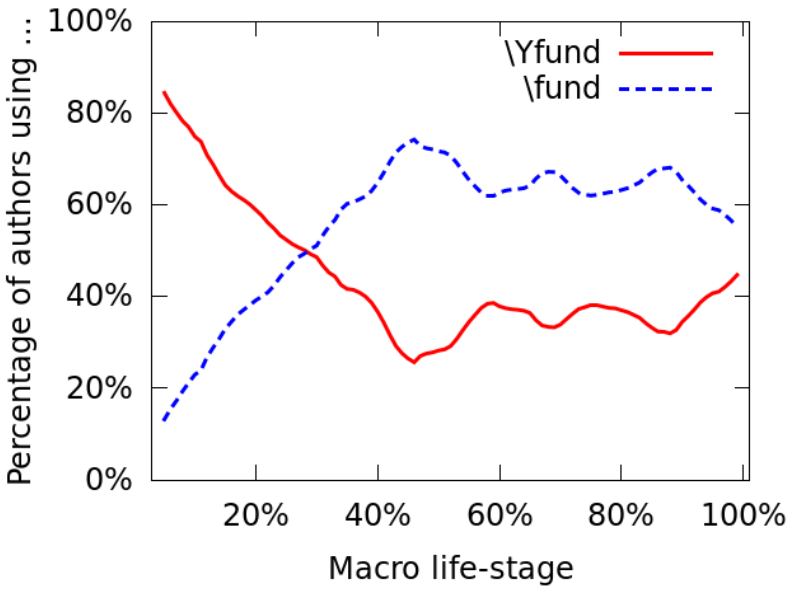


Figure 4.1: An example changeover: `\fund` surpasses the once-dominant `\Yfund` as the preferred `name` used to invoke Young tableau; y-axis indicates the percentage of users of each `name` out of all authors using the respective `body`.

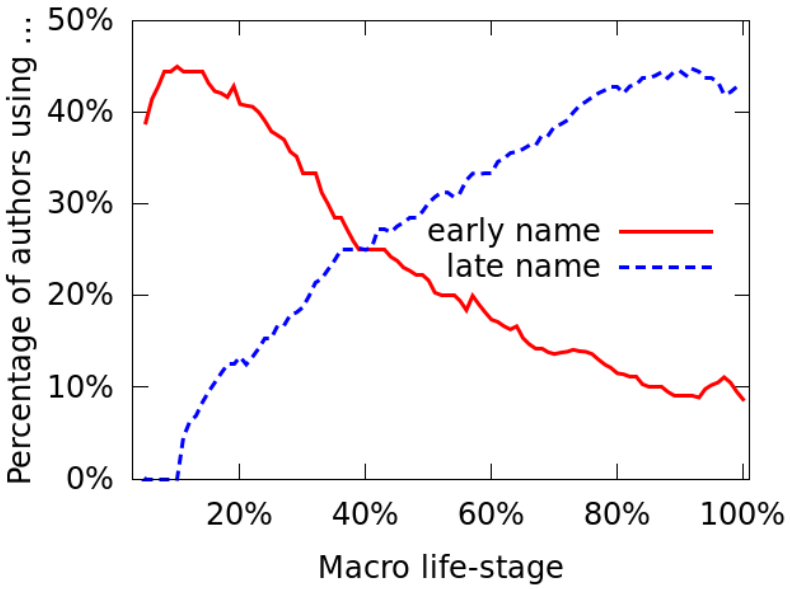


Figure 4.2: Aggregated temporal usage trends of early name (N_e) and late name (N_ℓ) for all macros undergoing changeovers; their crossing point is well before the middle of their lifespan.

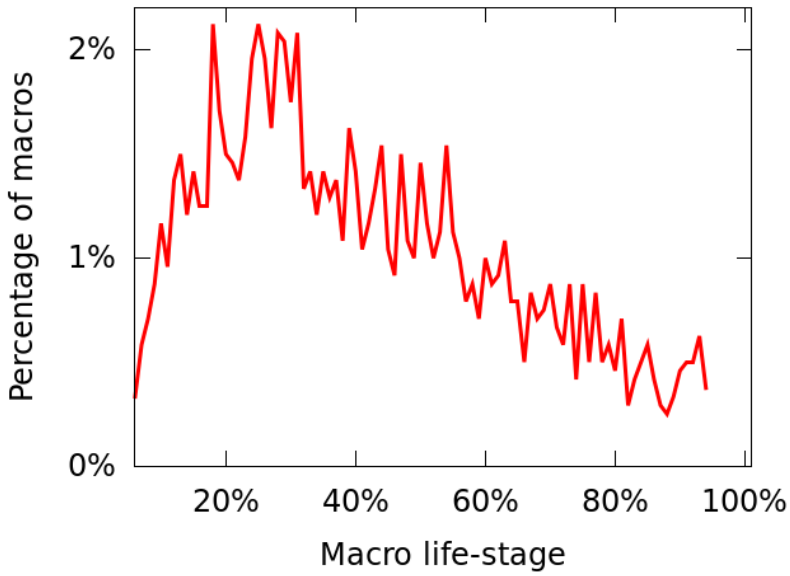


Figure 4.3: The distribution of crossing points (percentage out of all macros undergoing changeovers

$g_\beta(t, t')$ to be the fraction of authors in the interval $[t, t']$ that use the late name N_ℓ. We can turn these into single-variable functions by fixing an increment δ and defining $f_{\beta,\delta}(t) = f_\beta(t, t+\delta)$ and $g_{\beta,\delta}(t) = g_\beta(t, t+\delta)$; these are just the fractions of usage in length-δ intervals beginning at t.

Figure 4.2 fixes $\delta = 0.05$ and shows the median values of $f_{\beta,\delta}(t)$ and $g_{\beta,\delta}(t)$, as functions of t, aggregated over all β that undergo changeovers. It is intuitively sensible that $f_{\beta,\delta}(t)$ should be falling in t and $g_{\beta,\delta}(t)$ should be rising in t, since N_ℓ is in effect partially taking over from N_e. It is intriguing, however, that the shapes of the two curves are not symmetric in time—in that they cross well before the midway point at $t = 0.5$—considering that the definition of changeover is temporally symmetrical.

Figure 4.3 shows the distribution of these crossing points, over all β that undergo changeovers. For this plot, we formalize the crossing point as the minimum t such that $g_{\beta,\delta}(t') \geq f_{\beta,\delta}(t')$ for all $t' \in [t, t+.1]$, so as to require that the crossing persist for a non-

trivial interval of time. This plot too highlights the fact that the crossing tends to occur early in the usage of the macro `body` β, well before the midway point, although there is considerable diversity — for some macro `bodies`, the crossing point comes very late.

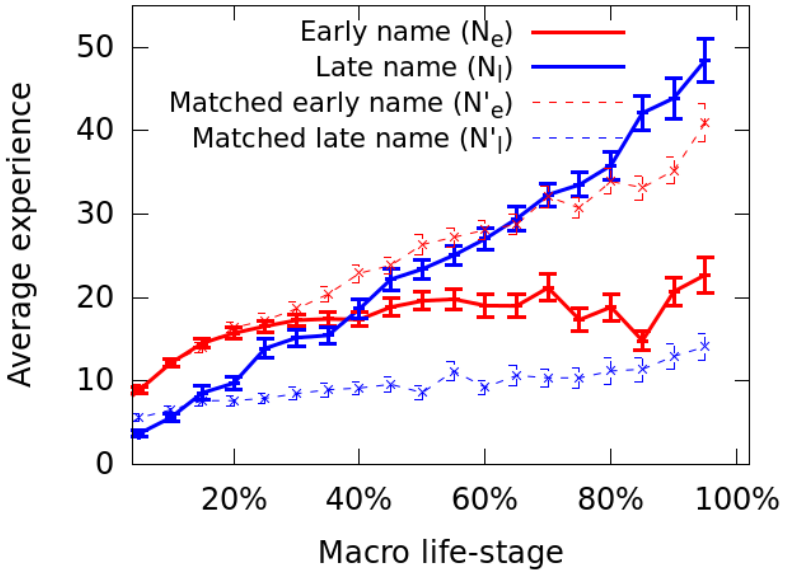
4.4.1 Properties of the authors in a changeover

We now examine the properties of the authors who use competing names in a changeover. In order to have a baseline for comparison, we pair macros undergoing changeovers with macros that do not undergo changeovers, but which have similar behavior up to their first q fraction of uses.

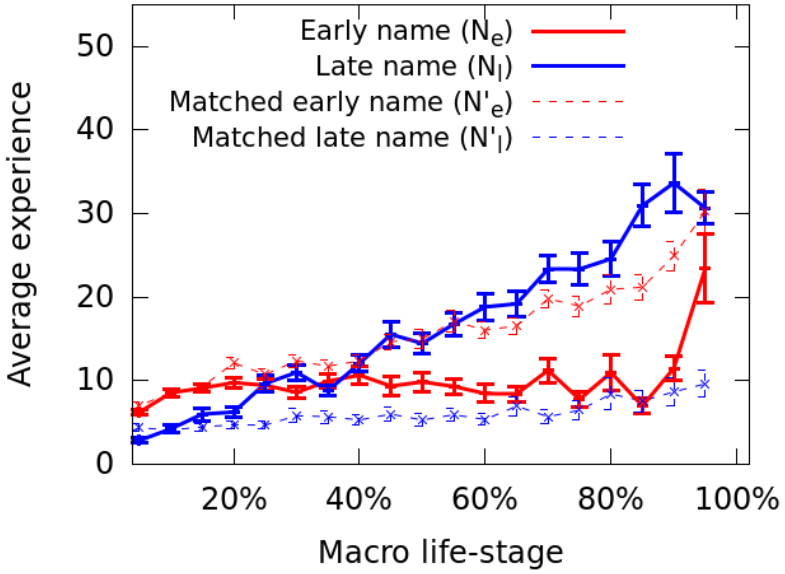
Thus, for each macro `body` β that undergoes a changeover, we find a macro `body` γ that does *not* undergo a changeover, for which (i) the total volumes of usage are approximately the same, $m_\beta \approx m_\gamma$, and (ii) there are two names N'_e and N'_ℓ for γ such that the prevalence of these two `names` in their early phases are approximately the same as N_e and N_ℓ respectively: $f_\beta(0, q) \approx f_\gamma(0, q)$ and $g_\beta(0, q) \approx g_\gamma(0, q)$.[2] We will refer to β and γ as a *matched pair* of macro `bodies`. Intuitively, from the perspective of their volume up to time q, the two `names` N_ℓ and N'_ℓ — for β and γ respectively — had very similar initial conditions, and hence we are forced to look at other properties to find a difference between them.

A key property that we consider is the *experience* of the authors using these names; recall that an author's experience at a given point in time is the number of papers they've written up to that time, which measures a kind of "age". (Accordingly, when we refer to authors as "younger" or "older," it is with respect to this measure of experience, and not to biological age.) Now, for a matched pair of macro `bodies` β and γ, we can look at

[2]The precise filter we use is to require that $m_\beta/m_\gamma \in [.91, 1.1]$ and $|f_\beta(0, q) - f_\gamma(0, q)|$ and $|g_\beta(0, q) - g_\gamma(0, q)|$ are both below .01.



(a)



(b)

Figure 4.4: Changeovers and user experience. (a) Average usage experience (b) Average adoption experience. When comparing `names` that eventually overtake their competitors (N_l, solid blue lines) with those that don't (N'_l, dashed blue lines), we observe that they tend to start with a younger user-base and then successfully transition to more experienced users.

the following quantities at a time $t \in [0, 1]$: the average *usage experience* of authors of each given name at time t, as well as the average *adoption experience* of users of each given name at time t; the former quantity aggregates over the experience of all users of the given name at time t, while the latter quantity aggregates only over the experience of authors using the given name *for the first time* at t.

In Figure 4.4 we show the average usage experience (left panel) and average adoption experience (right panel) for the four names N_e, N_ℓ, N'_e, and N'_ℓ, averaged over all matched pairs (β, γ). We notice a few respects in which these curves exhibit similar properties between β and γ: first, they all increase, which is natural since experience values are increasing as time runs forward on the arXiv. Moreover, the curves for N_ℓ and N'_ℓ start out below the curves for N_e and N'_e, which is consistent with the intuition that new terminology tends to start with more peripheral authors [76].

However, the curves for β and γ also differ in important ways, and this provides us with some insight into the differences between macro `bodies` β that undergo changeovers and macro `bodies` γ that don't. First, and most visibly, the name N_ℓ performs a major transition over its lifetime, going from authors with very low experience to authors with very high experience, while the experience of authors using N_e plateaus. Conversely, N'_ℓ fails to perform a corresponding transition, and remains concentrated on authors of low experience throughout its lifespan. In this sense, N_ℓ and N_e almost "change roles" as the plot progresses, with the curve for N_ℓ initially tracking N'_ℓ but eventually tracking N'_e, and the opposite holding for N_e. There is also a small but significant difference at the smallest values of t: the average experience for N_ℓ starts out lower than for N'_ℓ, a difference that may point to the value of low author experience in predicting the eventual success of a macro `name`. (We explore this further when we look at interaction dynamics in the next section.)

4.4.2 Predicting changeovers

We can also evaluate whether the properties we have assembled about the authors using competing names hold predictive power in the task of forecasting whether a changeover will occur. We formulate this as a prediction task, where for each matched pair of macro `bodies` β and γ — each involving two competing `names`— we try to predict early on which of them will undergo a changeover. Notice that because of the matching process, we have a balanced dataset where the two classes have the same aggregate characteristics in the early stages of their lifespans.

We first find that using only properties of the two competing macro `names` themselves — length, number of non-alphabetic characters, proportions of lowercase and uppercase characters — provides no predictive power. In other words, we can't predict simply from names like `\Yfund` and `\fund` which one will prevail. This reinforces the sense in which these truly represent neutral variations; the changes in name do not seem[3] to be fitness-enhancing on their own.

However, we get non-trivial predictive power when we add attributes of the authors using the `names` in their early stages. In particular, we define features based on the number of distinct authors using each `name` in the first q (= 0.3) fraction of the macro `body`'s lifespan, as well as the average usage experience and average adoption experience in windows of $[t, t + .05]$ for $t \in \{0, .05, .10, .15, .20, .25\}$.

Here and in the rest of the chapter, we perform logistic regression using features that are normalized using the z-score, with data that has balanced labels and 80/20 split on data for training and testing. The accuracies using different subsets of the features are presented in Table 4.1. The most important features are the average usage experience

[3]There is, however, the possibility that more complex `name` features could turn out to hold predictive power.

and the average adoption experience, with low values favoring changeovers: a `name` used by a younger generation is more likely to take over its competitor.

Feature set	**Accuracy**
Random baseline	50%
Name features	50%
Average usage experience features	58%
Average adoption experience features	60%
All author-based features	59%
All features	57%

Table 4.1: Accuracy of changeover prediction (± 4% confidence intervals for all rows).

4.5 Diffusion through Interaction: Dynamics of Local Competition

We now begin with a set of analyses designed to study how diffusion through interaction is taking place in our domain — the way in which competition over a set of conventions, in the form of different macro `names` for the same macro `body`, is taking place at a paper-by-paper level.

The "age" of the authors will again play an important role in these analyses, and we continue to use the *experience* of an author — the number of papers they have written — as a measure of age. Unless otherwise specified, when we are considering an author in the context of a particular paper they have written, we will be thinking about their experience at the moment this paper was written (as opposed to their eventual experience at the end of our dataset).

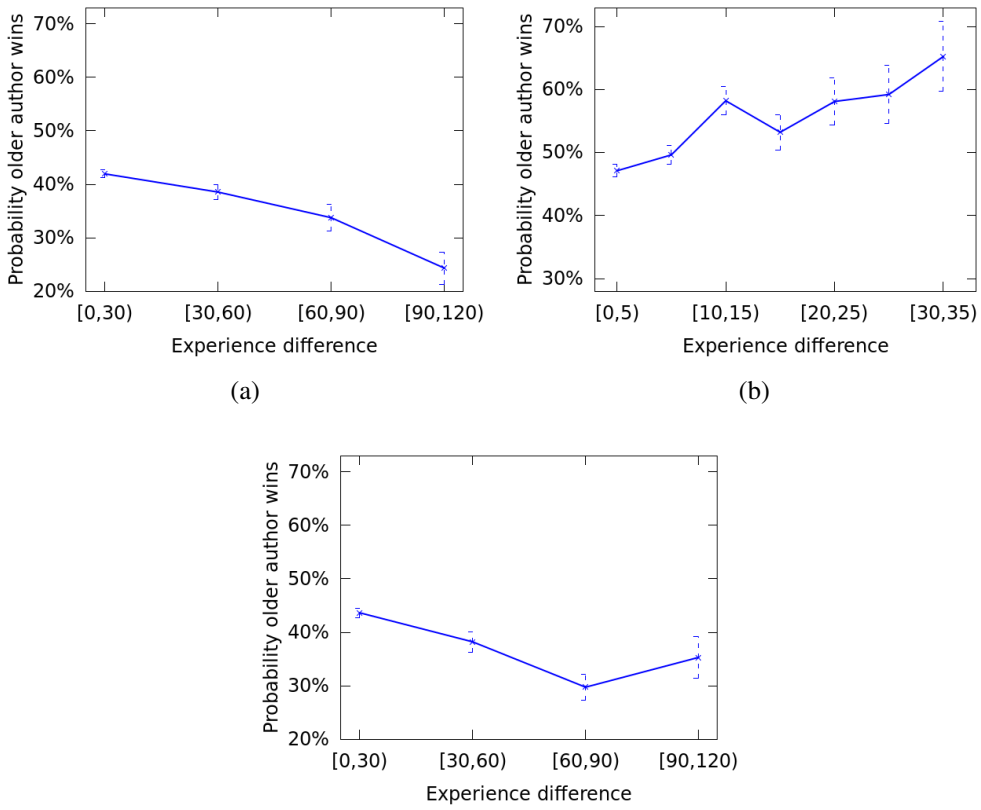


Figure 4.5: Percentage of fights won by the older author as a function of difference in experience. (a) Invisible fights: `name` (b) Visible fights: paper title (c) Low-visibility fights: `body`. The larger the experience gap, the more likely the younger author is to win the (invisible) macro `name` fights (a) and the (low-visibility) macro `body` fights (c); the opposite trend holds for the (much more visible) title fights (b).

4.5.1 Fights over macro names

We now consider the outcome of competition between two authors who have recently used different conventions for the same macro `body`. In order to have a consistent structured setting for such competition, we consider situations in which two authors A and B meet to write a paper, each having used a different `name` for the macro `body` β, and one of these two names is used in their co-authored paper. In this case, we say that a *fight* has occurred between A and B over the choice of name for `body` β, and the fight is won

by the author whose name is used in their joint paper. A basic question is to characterize and potentially predict the winner of a fight of this form — how does it depend on the properties of A and B, and potentially of the macro `body` and `name`?

More formally, we require that (i) two authors A and B write a paper P in which they are the only co-authors;[4] (ii) the paper P involves a macro `body` β that each author has used before; (iii) in their most recent uses of β (in earlier papers), A and B used names η_A and η_B with $\eta_A \neq \eta_B$; and (iv) one of η_A or η_B is used as the `name` for β in the new co-authored paper P. Further, in order to study macros that have non-trivial usage, we require that the macro `body` be used in the dataset by at least 30 distinct authors, and that the length of the body be at least 10 characters. Also, since certain pairs of authors might satisfy conditions (i)-(iv) many times, and there might be close alignment in the outcome of all their fights, we only consider a single fight for all pairs of papers that A and B co-author, selecting a chronologically earliest one.[5]

The left panel of Figure 4.5 shows that the younger author (the one with lower experience) wins these fights significantly more often than the older author, with the probability that the older wins decreasing as the experience difference between the two authors increases. These results are significant ($p < 0.05$)[6] using the full set of 1574 fight instances that passed the conjunction of all the filters described above, and was balanced to control for author-position effects, with the first and second authors winning an equal number of fights.

[4]For robustness, we also redo the analysis on three-author papers, examining fights between the second and the third authors and obtain qualitatively similar results.

[5]For the portion of our data in which we only have time granularity at the one-month level, we further restrict to instances in which A and B have no other papers in the month of their current co-authorship and their previous usage of β, making the temporal ordering unambiguous.

[6]Here and throughout the paper we use the binomial test for statistical significance.

A prediction task. We can use this distinction in experience to perform a prediction task: given an instance of a fight, and the past history leading up to it, how accurately can we predict who will win the fight?

The most basic approach to this would be to simply formulate a prediction task problem using two features: the experience of the first author and the experience of the second author, aiming to guess which of them will win the fights. The left panel of Figure 4.5 suggests that we would already be able to achieve non-trivial performance from just these two features, and we find that we achieve 58.2% accuracy using a logistic regression model. If we declare an instance to have label 0 when the first-listed author wins the fight, and label 1 when the second-listed author wins the fight, then we can interpret the coefficients for the experience of these two authors in the logistic regression model, shown in Table 4.2. Note that the positive coefficient for the first author (Experience 1) means that higher first-author experience produces an output of the logistic function that favors label 1, corresponding to the second author winning the fight. Conversely, the negative coefficient for the second author (Experience 2) means that higher second-author experience favors the label 0, corresponding to the first author winning. In consequence, for both authors higher experience works against them winning the fight, and lower experience tends to favor them in the prediction.

	Experience 1	Experience 2
Feature coefficient	0.40	-0.59

Table 4.2: Feature coefficients for predicting macro fights outcome when only using experience.

We can achieve non-trivially higher prediction performance by including a set of other natural features about the instance as follows. In particular, we use the following set of features.

Feature	Experience 1	Experience 2	Flexibility 1
Feature coefficient	0.48	-0.54	0.47
Feature	Flexibility 2	Degree 2	Betweenness 2
Feature coefficient	-0.54	-0.38	0.55

Table 4.3: Top 6 feature coefficients for predicting the outcome of macro fights.

- **Experience:** The number of papers the author has written prior to the fight.
- **Prior uses:** The number of prior papers in which the author has used the macro `body`.
- **Flexibility:** The fraction of consecutive uses of the `body` in which the author used two different `names`. (This is a measure of "flexibility" in that it shows the fraction of prior uses of the `body` in which the author changed `names` relative to their immediately preceding use.)
- **Degree:** We build a co-author graph on all authors who have used `body`, based only on papers that were written prior to the fight. This feature is the degree of the author in this co-author graph.
- **Betweenness:** The betweenness of the author in the co-author graph from the preceding point. (This is one measure of how central the author is in the graph.)
- **Properties of `name`:** The length of `name`.
- **Properties of `body`:** The length, number of non-alphabetic characters and the maximum depth of curly braces in `body`.

When we put all these features together, we are able to obtain 67.3% accuracy via logistic regression. The coefficients with the 6 largest absolute values in the model are shown in Table 4.3. Experience behaves as before; degree is naturally aligned with experience (since older authors have more time to acquire co-authors); and flexibility behaves as one would intuitively expect (since more flexible authors are more likely to

also change in the current fight). It is interesting that high betweenness helps predict that an author will win a fight, since the results on experience might have suggested the opposite intuition.

To summarize, as noted in the introduction of this chapter, it is interesting that the older author does not play the dominant role — relative to the younger author — in determining the outcome of the convention, given what we know about the tendency of high-status individuals to drive the outcomes of interactions [87]. A natural hypothesis is that the older author is ceding control over low-level decisions on conventions like macro `names` to the younger author. This suggests that we may arguably see a different outcome if we were to look at fights over decisions that were less low-level, more visible to readers, and hence more prominent. In such a case, where the status of the older author is more implicated by the outcome in the eyes of readers, is the older author more likely to win the fight? We now describe an analysis that addresses this contrast.

4.5.2 Visible fights

What would a more visible type of fight look like, and how does the young-old dynamic work in this case? We now formalize a set of fights involving one of the most visible decisions about a paper — the choice of title. For fights involving titles, we need a different set-up than the one we used for macros, and a more indirect one. In the case of macros, the convention was the choice of a `name` for a macro `body`, but since paper titles are generally written in a free-form manner, we need to decide what the space of possible conventions is.

In order to have a concrete definition to work with, and one that intuitively has a visible effect on the style of the title, we define conventions in the choice of title

based on the presence or absence of certain punctuation, formatting, or parts of speech. Specifically, for a given title, we ask whether it exhibits one of seven possible styles (not all mutually exclusive): Does it contain a colon, question mark, or mathematical notation; and is its first word a noun, verb, adjective, or determiner? For each of these seven questions, we say that a title is a *positive instance* of the corresponding style if it contains the indicated feature.

Now, if authors A and B write a paper together, how do we define the notion of a fight over one of these stylistic titling conventions? Asking about the immediately preceding title for each author produces data that is too sparse to get meaningful results, and so instead, we look at each author's lifetime tendency to use each of the stylistic conventions.

Defining a fight over the title. Specifically, let us fix one of the stylistic conventions σ defined above, and consider a two-authored paper in which the authors have never written a paper before.[7] Let E_y and E_o denote the experience of the younger and older authors on this paper respectively, and let P_y and P_o be the lifetime fraction of papers on which the younger and older authors used convention σ, only considering papers they did not write together.[8] Finally, we let I_σ be an indicator variable equal to 1 or 0 depending on the presence or absence of the convention in the paper.

Intuitively, we'd like to consider the value of I_σ in relation to which of P_y or P_o is larger. To have a meaningful baseline for comparison, we match each of our fights in pairs: for each fight given by $(E_y, E_o, P_y, P_o, I_\sigma)$, we find a fight using the same σ but a

[7] We also add some additional filters, including a sufficiently high experience for the older author, and at least 10 lifetime papers by the younger author that are not written with the older author.

[8] Since the younger author often has relatively few papers at the time of the fight, we use the set of all papers written by each author (not counting their joint papers) to determine these fractions. This uses information from the future beyond the paper itself, but note that we are not using this for a prediction task, only to determine the relative tendencies of the authors to use the convention over their lifetimes.

different paper, where the values of P_y and P_o are swapped, and where I_σ is inverted. That is, we find a fight $(E'_y, E'_o, P'_y, P'_o, I'_\sigma)$ with $P'_y \approx P_o$, and $P'_o \approx P_y$, and $I'_\sigma = 1 - I_\sigma$. Thus, we have a set of matched pairs, where in each pair, one of them has higher P_y, the other has higher P_o, and they differ on the presence or absence of σ.

What is the effect of this construction? If in each pair, the instance with higher P_y is always the one where $I_\sigma = 1$, it would mean that σ always occurs in the instances where the younger author has a higher tendency toward σ; in other words, we'd be able to perfectly infer the presence or absence of σ in each given pair from the relative values of P_y and P_o, with the younger author playing a greater role driving the presence of σ. If in each pair, the instance with higher P_o is always the one where $I_\sigma = 1$, we would again have perfect prediction with the older author driving the presence of σ. In general, we say that in a single pair, *low experience is dominant* if $I_\sigma = 1$ in the instance with higher P_y, and *high experience is dominant* if $I_\sigma = 1$ in the instance with higher P_o. If the I_σ values were assigned at random, we'd expect low experience and high experience to each be dominant in half the pairs. What do we see in the actual pairs?

We find that in approximately 57% of the pairs, high experience is dominant, which at the number of pairs we have is significant relative to a random assignment baseline with $p < .001$. Moreover, we can group the pairs into buckets based on $E_o - E_y$, and perform this analysis on each bucket separately. As we see in the middle panel of Figure 4.5, the extent to which high experience is dominant is increasing in the experience difference — the opposite effect from what we saw in the left panel for fights over macro `names`.

Thus, we have a concrete sense in which older authors are winning visible fights over features of the paper title, even though they are losing invisible fights over macro `names`.

4.5.3 Low visibility fights

We have now seen the outcomes of two types of fights representing opposite extremes of visibility — fights over macro names, which are essentially invisible to readers; and fights over stylistic conventions in a paper's title, which is extremely visible. Since younger authors tend to win the invisible fights and lose the visible fights in these formulations, it becomes natural to probe the spectrum of possible fights in between these extremes and thus gain more insight into how the outcome of a fight relates to its level of visibility.

This is largely an open question, but here we describe one initial investigation in this direction. Consider the case in which two authors meet to write a paper, and they each use the same macro `name` but for different `bodies`. For example, both authors might use `\eps`, but one uses it to mean `\epsilon` (ϵ) while the other uses it to mean `\varepsilon` (ε). Whose macro `body` will end up getting used in the paper they write together? We will call this a *macro-body fight*, and what's interesting is that it has exactly the structure of our earlier macro-name fights, except with the roles of the `name` and the `body` reversed: now the authors arrive with a shared `name` corresponding to different `bodies`, and this contention must be resolved. An important contrast, however, is that for many macro `bodies`, the outcome of this fight will be visible, albeit often at a very low level in the formatting and choice of symbols in the paper. We can therefore think of these as *low-visibility fights*, and can ask whether younger or older authors will tend to win them.

To explore this question, we need to deal with the fact that not all macro-body fights will have visible effects. Thus we select a small number of very common macro names where the effects of different `bodies` are generally visible in the paper. Specifically, we use the following names: `\proof`, `\eps` and `\Re`. For these fights we run the same

procedure as for macro-name fights, but we swap the roles of the `name` and `body` of the macro, and we remove the filter of length 20, since `names` and `bodies` are generally short in this case. With these three macro names we end up with 1092 fight instances. We observe that the young author wins 60% of the instances, suggesting that the pattern of outcomes is closer to what we saw in the invisible macro-name fights. Grouping the results by the difference in experience, we see in the right panel of Figure 4.5 that these low-visibility fights follow the same trend as the invisible fights.

4.6 Conclusion

Analyzing the competition among conventions has been a methodological challenge, because the substance underlying the convention generally changes, at least to some extent, together with the convention itself. We study a setting — macros on the e-print arXiv — where it is possible to fully control for the meaning of the convention (the `body` of the macro) even as the convention itself (the choice of `name`) is changing. In the resulting analysis, we focus on two main issues. First, we find that instances in which one macro `name` convention overtakes another are characterized by young initial users of the convention that ultimately succeeds, together with a transitional phase in which the successful convention spreads to older users. Second, we consider the local, instance-by-instance competition among pairs of authors who must resolve contention over the choice of convention in the process of writing a joint paper. In this type of *diffusion through interaction*, we find that younger authors tend to win fights (such as for macro `names`) that do not produce visible consequences, or produce low-visibility consequences, while older authors tend to win fights (such as for titling conventions) that produce highly visible consequences.

CHAPTER 5

CASCADES: A VIEW FROM AUDIENCE

Cascades on social and information networks have been a tremendously popular subject of study in the past decade, and there is a considerable literature on phenomena such as diffusion mechanisms, virality, cascade prediction, and peer network effects. Against the backdrop of this research, a basic question has received comparatively little attention: how desirable are cascades on a social media platform from the point of view of users? While versions of this question have been considered from the perspective of the *producers* of cascades, any answer to this question must also take into account the effect of cascades on their audience — the viewers of the cascade who do not directly participate in generating the content that launched it. In this work, we seek to fill this gap by providing a consumer perspective of information cascades.

Users on social and information networks play the dual role of producers and consumers, and in this chapter we focus on how users perceive cascades as consumers. Starting from this perspective, we perform an empirical study of the interaction of Twitter users with retweet cascades. We measure how often users observe retweets in their home timeline, and observe a phenomenon that we term the *Impressions Paradox*: the share of impressions for cascades of size k decays much more slowly than frequency of cascades of size k. Thus, the audience for cascades can be quite large even for rare large cascades. We also measure audience engagement with retweet cascades in comparison to non-retweeted or organic content. Our results show that cascades often rival or exceed organic content in engagement received per impression. This result is perhaps surprising in that consumers didn't opt in to see tweets from these authors. Furthermore, although cascading content is widely popular, one would expect it to eventually reach parts of the audience that may not be interested in the content. Motivated by the ten-

sion in these empirical findings, we posit a simple theoretical model that focuses on the effect of cascades on the audience (rather than the cascade producers). Our results on this model highlight the balance between retweeting as a high-quality content selection mechanism and the role of network users in filtering irrelevant content. In particular, the results suggest that together these two effects enable the audience to consume a high quality stream of content in the presence of cascades.

5.1 Introduction

When retweets were first introduced on Twitter, users expressed many such concerns [13]. A key question then is: what effect do cascades have on consumption behavior? A pithy answer is provided by the existence of networks with hundreds of millions of active users; this at least suggests that the effect of cascades is not as negative as users feared it to be.

One aspect of user consumption behavior is deeply intertwined with production in that production of content (via re-sharing) is also simultaneously consumption behavior. Production has been widely studied in the literature under the topics of information propagation and diffusion of content. We note however that this is only a part of consumption and some basic characteristics of consumption behavior have not been addressed to the best of our knowledge. For instance, although virality of content on Twitter has been extensively discussed [37, 88], we do not understand the view of virality from a consumer perspective: what fraction of tweets consumed by consumers on Twitter are viral? Do users engage with these more than with non-viral content in their home timeline? We emphasize that the consumer view could be quite different from the producer perspective for virality: even though a small fraction of tweets "go viral", a large fraction of the consumer experience on Twitter could still be shaped by viral content. This is because

when we think about the population of all *views* of tweets, we're sampling tweets in proportion to their popularity, and this sampling based on size leads to effects where a small number of items (extremely popular tweets in this case) can make up a large fraction of the sample [8].

We examine the above questions through empirical analysis of user behavior on Twitter. Each time a user views a tweet — referred to as an *impression* of the tweet — they can choose to engage with it through several means, including clicking on it, liking it, or retweeting it. Given observations of what tweets a user sees in their home timeline via tweet impression logs, as well as tweet engagement and sharing activity, we can piece together a consumer view of cascades on Twitter. Through this analysis, we observe that retweet cascades indeed occupy a substantial fraction (roughly a quarter) of a typical user's timeline, and 1 out of 3 impressions in the dataset we analyze are due to cascades. Thus, cascades have a substantial impact on the user experience at Twitter given their prevalence in users' home timelines. This impact is arising despite the fact that extremely few tweets generate large cascades; the point is that for a producer of content, it is very rare to see your tweet become viral, but for a consumer of content, much of your time is spent looking at viral content. We term this dichotomy the *Impressions Paradox*; it is a counter-intuitive contrast in two ways of looking at the same population, in the spirit of similar phenomena that arise because of sampling biased by size.

In light of our previous discussion, we note that this wide prevalence of retweets does indeed impose upon users content that they did not opt in to see. It is natural to wonder whether users respond negatively to this imposition on their home timeline, which they have carefully constructed through their choice of users to follow.

Analyzing user engagement with cascades provides a way to answer this question. In

particular, we compare user engagement probabilities (retweeting, liking and clicking) on retweeted content versus organic content (directly produced by a person the user follows). Our main finding here is that retweeted content rivals or exceeds the organic content in engagement. It is useful to consider this fact in the context of user fears of irrelevant content showing up in their timeline (even if it might be high quality; the best tweet on politics may be uninteresting to a user not interested in politics). Viewed in this light, our finding is perhaps quite unexpected. On the other hand, this result is exactly what one might expect if we think of retweets as a high quality tweet selection mechanism — users might only engage with the best tweets, so it is unsurprising that the best tweets get high engagement. Note however, that popular tweets also get viewed by many users, resulting in a very high number of impressions. Thus, even with an assumption that popular tweets have high quality, it seems unclear why they should get high engagement *per impression* as their growth in audience size might completely outpace the set of interested users.

In order to understand this effect quantitatively, we propose and analyze a simple theoretical model of retweeting behavior that teases apart these two effects. Our model is novel in that it inverts the traditional view of cascades as a tree being rooted at the author, to a tree that is centered at an arbitrary user — a member of the audience for cascades — who receives a mix of organic tweets and retweets. This model helps us quantify two metrics for a user's home timeline: precision (seeing content that is relevant or topical for users) and quality (highly engaging content for a topic). Intuitively, users would like to have a high precision and high quality home timeline where most of the content is relevant and highly engaging. In the presence of retweets, it seems unclear a priori whether the content will still be relevant, and further how would one quantify changes in tweet quality. Our analytical and simulation results show that it is indeed possible for users to have the best of both worlds by seeing high quality and relevant

retweets in their timeline. Furthermore, this model also helps us understand the *value* of retweets by quantifying a counterfactual world where retweets would not exist.

5.2 Related work

There has been extensive work on on-line information diffusion. This has included studies of news [2, 16, 11], recommendations [52], quotes [24], hashtags on Twitter [70, 83, 71, 56, 58], information flow on Twitter [88] and memes on Facebook [29, 20]. Past work has also investigated methodological issues including definitions of virality [37], the problem of prediction [20], the trade-off between precision and recall in cascading content [17], and the role of mathematical epidemic models [35].

In addition, it has been shown that only a very small fraction of cascades become viral [37] but the ones that do become viral cover a large/diverse set of users. In other words, if you are the source of a cascade you have a low chance of creating a viral cascade but, once we switch to the consumer's point of view we observe that a large fraction of a user's timeline is made up of these diffusing pieces of content. Another related theme on the work presented in this chapter has been the observation that a small number of "elite" users produce a substantial fraction of original content on Twitter [88]. As with other studies, this one also focused on active cascade participants, and our work is differentiated by the focus on cascade audience.

The primary focus of the body of prior work on cascades has been either on the source of the content or on the structural properties of the cascades themselves. In this chapter, we study the effect of different properties of the cascade tree, and the underlying follower graph, on the experience of the consumers of cascades. In particular we first show that although most tweets do not get re-shared but they are a significant fraction of

the content an average user reads. We also find that consumers prefer either very popular content or personalized content coming from users they opted to follow. Then we look at each consumer as an individual and show that different consumers might show different behavior but a single user is consistent on the type of content they like over several days. Finally, we complete our argument with a simple model that captures how the re-share mechanism features enhance the experience of consumers.

5.3 Empirical Analysis on Twitter

In order to analyze audience behavior with cascades, we undertook an empirical investigation on Twitter. A user u on Twitter typically spends the majority of their time on their personalized home page, called the home timeline, which primarily consists of a collection of *Tweets* from a set of users $F(u)$ that u chooses to follow. Since most content is consumed on the home timeline, we focus our analysis on user behavior in the home timeline. Further, to keep our analysis most interpretable, we ignore some products that rank the Twitter home timeline, such as "While you were away" [85], and focus exclusively on impressions of unranked tweets. These tweets are presented in a reverse-chronological fashion, and hence they allow us to study a version of the question that is independent of the ranking process (since ranking can have a large implication for the visibility and hence effect of cascades; see e.g. Facebook's work on this issue [33]).

We measure both views (or impressions) of tweets as well as user engagements with the tweets in our analysis. We define these terms in the sections below, but the goal of the analysis is to provide insight into the impact of cascades on consumer experience via impressions, and gauge their reception of this content via engagement. The dataset for this analysis was collected from Twitter logs during a 16 day period during summer

2016. Because of user privacy, we conduct all the analysis in a user-anonymized fashion, and present results from aggregate analysis. Note that for some of the plots we use a relative scale for the *y*-axis to anonymize actual values, as a relative comparison of values is the main goal for these plots.

5.3.1 Cascade Views

The first step in our empirical investigation is to understand whether cascades constitute a significant fraction of overall audience attention. Perhaps the simplest metric for measuring this is to understand the raw volume share in a user's home timelines. But before we proceed with that, we need to define what we mean by a tweet cascade and what constitutes a "view" of a tweet.

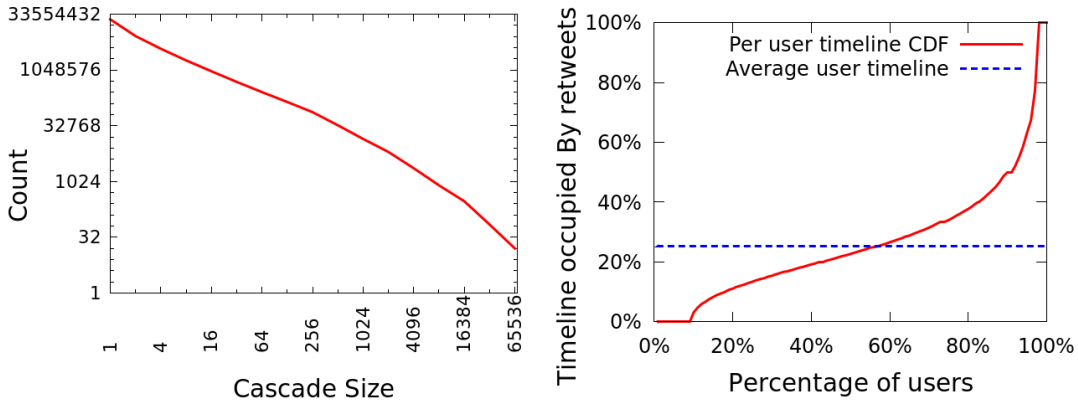


Figure 5.1: (a) The number of cascades in a log-log plot bucketed using the $\lfloor log_2 \rfloor$ function (b) The distribution of the fraction of home timeline impressions constituted by retweets, over all Twitter users. The horizontal line represents the average value.

In this work any tweet that is retweeted at least once is called a cascade as it has one retweeter (or adopter) other than the original tweeter. As has been noted in prior work, most cascades are shallow (star-like) and only a rare few go on to be "viral". We also refer to a *cascade size*, which is the number of eventual retweeters (or adopters) that the

tweet gathers, as measured after a few days of the original tweet. In order for our analysis to be valid the dataset must be large enough to catch cascades from a spectrum of sizes. We confirm this by visually plotting the cascade size distribution in Figure 5.1(a), which shows the average number of cascades with size between $[2^k, 2^{k+1})$ over the 16 days. Clearly, the data has enough cascades in each bucket even for a single day.

Next, we define a tweet view or an *impression* on the home timeline. The ideal measurement is to check that the user really "saw" the tweet, but in absence of that, we just measure whether the tweet stayed on the user's mobile screen for a large enough time. This filters out a variety of behaviors, and among them the common pattern of scrolling quickly through the home timeline, where the user just glances at a large number of tweets.

With these definitions we turn our attention to studying how much audience attention is commanded by cascades. Perhaps the most basic measurement to make is to measure what fraction of a users' home timeline impressions came from cascades that did not originate in the user's direct neighborhood. We find that 68% of all home timeline tweet impressions are from users' direct followings, and the remainder 32% come from cascades that originate from outside of a users' direct neighborhood. A different view of this overall statistic comes from looking at this from each individual user's perspective, through which we can measure what fraction of a user's timeline impressions come from retweets. The distribution of this quantity is shown in Figure 5.1(b), from which it is evident that for half of all users, approximately a quarter of their timeline consists of retweets. An additional dimension of tweet impressions coming from cascades is that these tweets bring in a fair bit of author diversity: 55% of unique authors who appear in a user's timeline are from outside the user's direct followings.

Given that nearly a quarter of a user's home timeline consists of cascades, it is natural

Distance	Impressions	Hop count	Impressions
1	68.86%	1	66.70%
2	30.53%	2	27.48%
3	0.59%	3	3.66%
4	10^{-3}%	4	1.09%
5	10^{-4}%	5	0.45%
6	4×10^{-5}%	6	0.23%
7	6×10^{-6}%	7	0.13%

Table 5.1: The percentage of tweets on a timeline based on their distance and hop-count to the receiver

to ask how these cascades reach the user. To provide insight into this question, we look at how far away the cascade originated, and how long it took to get to the user. For the former, we measure the network *distance* (shortest directed path) from the user to the cascade originator (the author of the root tweet in the cascade). Table 5.1 shows the percentage of tweet impressions that occur for each distance (out of all impressions), and from the data it is clearly visible that almost all impressions come from within distance 2 in the graph.

To understand the path the cascade took to reach a user, we reconstruct the cascade tree[1] and compute the number of hops on the tree that are between the user and the root of the tree; we refer to this quantity as the *hop-count*. The distribution of impressions w.r.t the hop-count is shown in Table 5.1. As is the case with distance, almost all impressions occur on hop counts 1 and 2. This data is all in agreement with prior work that has also commented on the vast majority of cascades being very shallow in terms of hop-count [37]. However, we do note that in contrast with distance, impressions for larger hop-counts don't quickly die down to zero. An obvious hypothesis for this is the possibility that some large cascades survive for a long time and hence reach users via all kinds of paths. This leads to the question of the impact of these large cascades from the

[1]The cascade is a directed acyclic graph from a user's perspective, but we can think of it as a tree by picking the first incoming edge for each node by time — that is also a close approximation to how Twitter treats retweets in practice.

perspective of impressions.

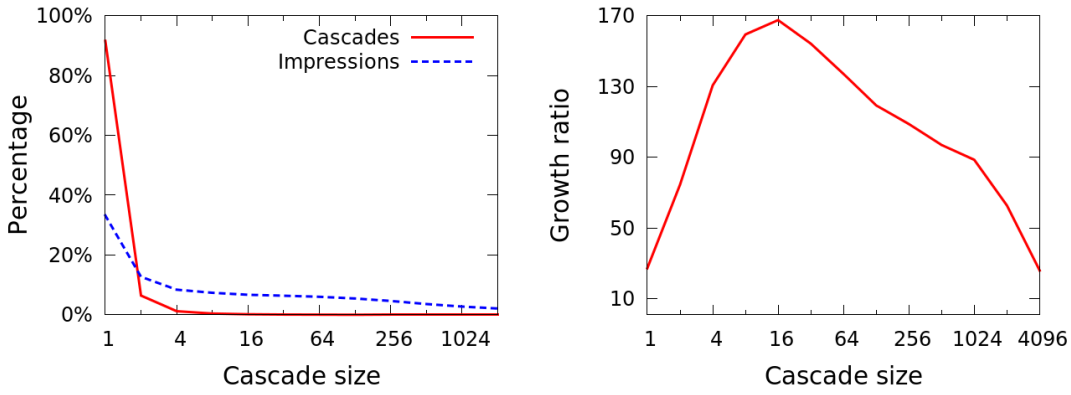


Figure 5.2: (a) Illustrating the *Impressions Paradox*: share of impressions for cascades of size k decays much more slowly than frequency of cascades of size k. Note that x-axis is log scale. (b) Cascade growth ratio is the ratio between number of impressions generated by these cascades to number of tweets generated by these cascades.

Recall that 1 out of 3 impressions arise from cascades, and in light of the previous discussion, we would like to understand *which* kinds of cascades are contributing to these impressions. It is useful to remember that large cascades are quite rare on Twitter, as illustrated in Figure 5.2(a): on a log-percentage plot, the fraction of tweets that get more than 8 retweets is less than 1%. However, since some cascades survive for a long time, it is natural to ask what is the share of impressions generated by these cascades. The share of impressions is also shown on the same Figure 5.2(a), and this presents a stark contrast from the probability of tweets generating a large cascade — even though 91% of tweets have a cascade size of 1, these generate only 33% of impressions coming from retweets, with the large cascades contributing a substantial fraction of impressions coming from retweets. We term this the *Impressions Paradox*: the share of impressions for cascades of size k decays much more slowly than frequency of cascades of size k.

We also note that for all cascades of a given size, one can compute a *cascade growth* metric: the ratio between number of impressions generated by these cascades to number

of tweets generated by these cascades. Intuitively, one would expect this ratio to be high for small cascades since almost every retweet brings in a large set of new audience members who haven't seen it by other means, while for larger cascades the gain in new audience per retweet might be lower since the presence of triangles means that new retweets may eventually reach people who have already seen it by other means. In fact, from Figure 5.2(b), we notice that the growth ratio for cascades hits a peak around size 32, and is the same for the smallest and largest cascades!

Thus, it seems clear that even though large cascades (especially "viral" ones) are infrequent on Twitter, they constitute a substantial fraction of audience attention. This observation leads to the question of how does the audience react to the presence of cascades in their home timeline. We address this question in the next section by analyzing user engagement.

5.3.2 Engagement with Cascades

Users on Twitter engage with tweets in a variety of ways, and we focus on the following engagements in our analysis: retweets (resharing the tweet with your followers), likes (previously known as favoriting), and clicks (either a click on a link/mention/hashtag in a tweet, or a visit to a "tweet details" page are considered as clicks). Together, these engagements provide a broad perspective on how users perceive the content as engagements are often a reflection of how much users enjoyed the content. This is not always the case though, and engagement is skewed by a range of factors: social acceptability, context, and inherent clickability (for instance, "clickbait" may have a high clickthrough rate) of content to name a few. Despite these shortcomings, the large scale of data analysis that we conduct does provide a directional guide on user enjoyment by measuring

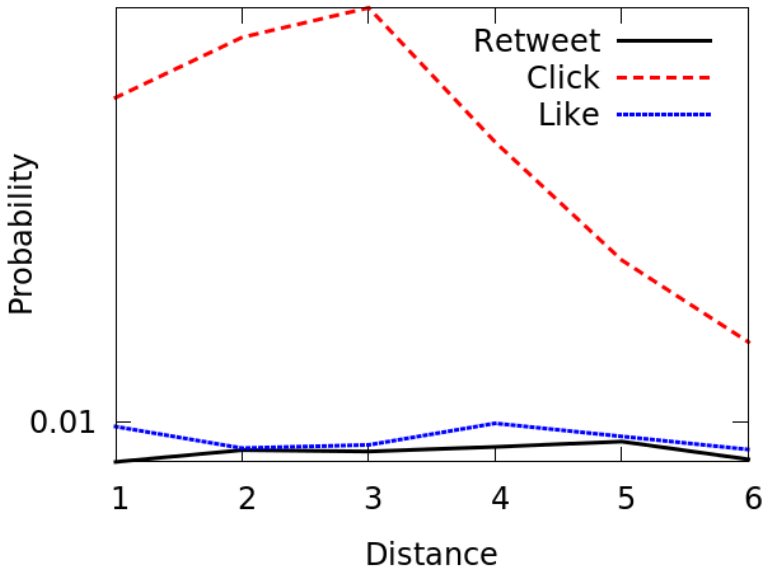


Figure 5.3: Probability of interaction based on distance. Note that the y-axis values are randomly pinned to 0.01.

engagement.

Engagement with cascades doesn't occur in a vacuum, and here we contrast engagement on cascades against the natural baseline of engagement on "organic" tweets, i.e. tweets authored directly by a user's followings. This comparison is readily obtained by measuring how the probability of retweets, likes and clicks varies with graph distance. These curves are presented in Figure 5.3, which shows that the various engagement measures behave quite differently. In particular, a tweet from a neighbor has a higher probability of receiving a like than tweets coming from users that are farther away in the network. On the other hand, a cascade tweet that originated outside of the user's direct network has a higher chance of getting retweeted and clicked. This perhaps is an indication that liking has a social element to it, and users tend to primarily like personalized content or tweets that come from their direct neighbors. On the other hand, a tweet that arrives in a cascade from outside the neighborhood is "retweetable" by definition, and hence just by this selection mechanism it increases its chances of getting retweeted

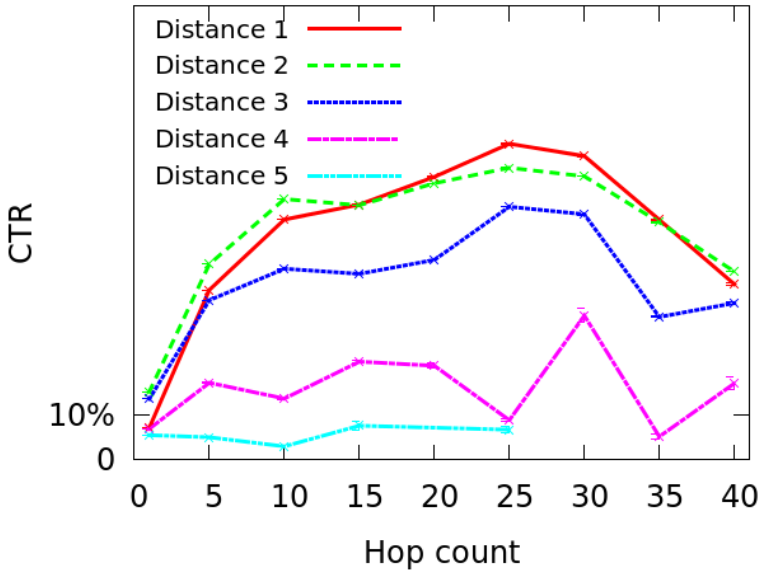


Figure 5.4: The click Through Rate (CTR) of tweets coming from different distances against different hop counts with errorbars. Note that the y-axis values are randomly pinned to 10%.

as compared to an average tweet that may not received any retweets at all. The click probability curve shows different behavior than both likes and retweets, and the probability of clicking on the tweet *increases* till distance 3. It is important to point out that to a consumer on Twitter, the only visible distinction in distance is whether it is 1 (direct connection) or greater (retweet). Thus, the difference between click probability in distances 2 and beyond likely comes from something other than user selectiveness between in and out of network content. An appealing hypothesis is that perhaps inherently clickable content travels farther on the network. We examine this next via a hop count analysis.

Recall that hop-count refers to the distance from the user to the author in the cascade tree. In order to understand the click probability variation, we turn to examining click-through rate (CTR) of tweets by hop-count — since this route is predominantly available to larger cascades, it provides us a way to measure how users react to large cascades

versus other smaller cascades. This data is presented in Figure 5.4, where there is a curve for a fixed distane showing the average CTR of tweets based on the hop-count value. There are several things to notice in this plot. First, observe that for a fixed hop-count, CTR generally decreases with distance — this clarifies that the increase with distance observed in Figure 5.3 is at least partially due to an effect that it to some extent like Simpson's paradox. We also note from Figure 5.4 that CTR generally increases with hop-count. Recall that higher hop counts are generally only available to large cascades, and hence the higher CTR indicates that popular content generates more clicks. We emphasize that this is not a causal statement, and in particular popular content might precisely be more popular *because* it generates more clicks.

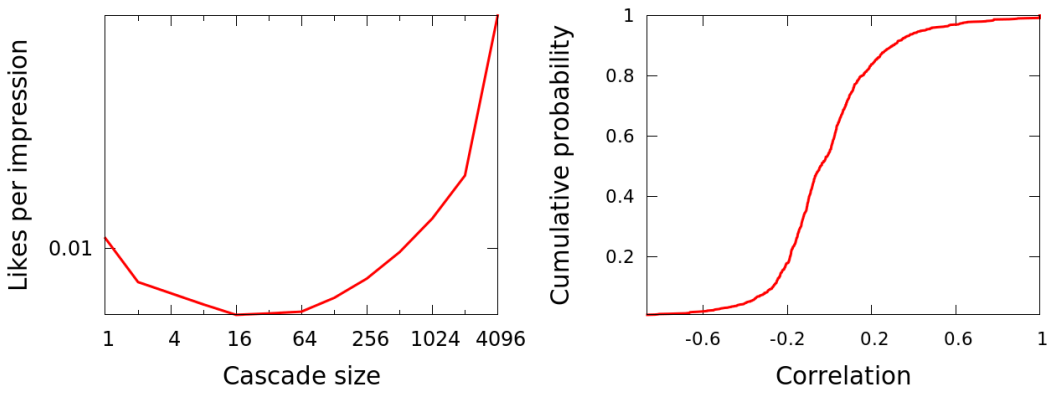
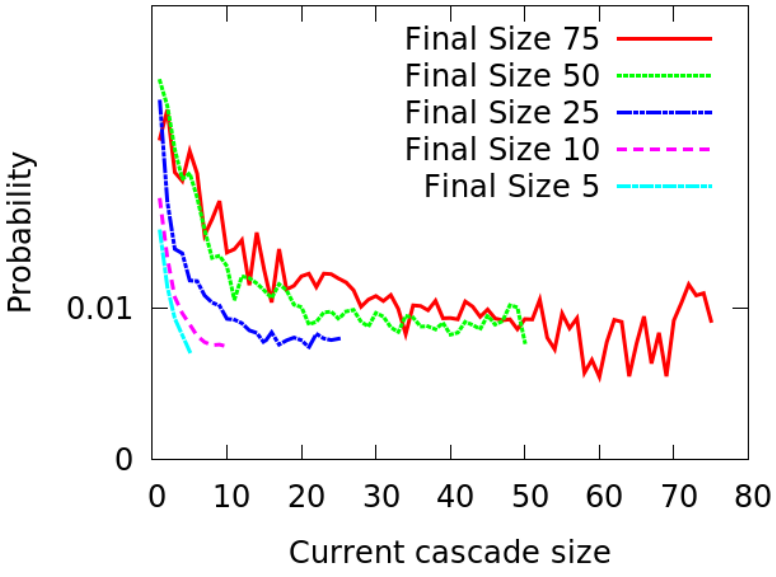


Figure 5.5: (a) Likes per impression for different cascade sizes bucketed by log base 2 (note that the y-axis values are randomly pinned to 0.01.) (b) Distribution over the correlation of different users, liking content based on its size.

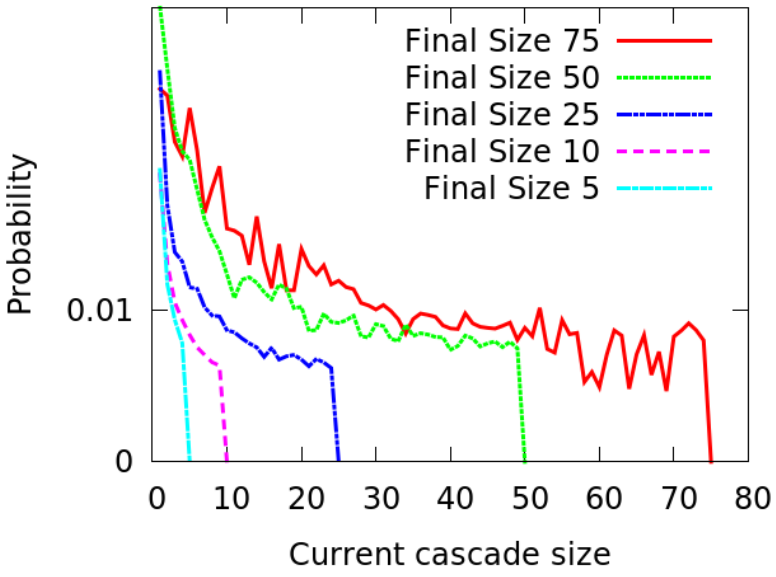
Given this data, we are left with the intriguing observation that users enjoy consuming (via liking/clicking) content both from their direct neighbors as well as from large cascades. But the combined effect of these two mechanisms is apriori unclear. In particular, since large cascades can travel farther from their source, does their overall appeal overcome the fact that their audience is not their direct neighbors? The most direct way to study this question is to look at like probability per tweet impression based on the

cascade size of the tweet, and Figure 5.5(a) shows the result of this investigation. Note that the cascades are bucketed as before, since the number of cascades with large sizes are rare. We observe from Figure 5.5(a) that cascades with very small and very large sizes have high like per impression rates. These stand in contrast with a smaller like per impression value for medium size cascades. We can thus see the two mechanisms mentioned above at play here: at small cascade sizes, content is being liked by neighbors (perhaps driven by a social aspect), while at large sizes content is likeable in general so most users enjoy consuming it. But there is an "uncanny valley" in the middle where not naturally likeable content reaches users who are not quite interested in it. The connection between Figure 5.2(b) and 5.5(a) is very interesting, and we leave further investigation of this connection as a direction for future work.

From 5.5(a), it seems clear that the most popular content, as indicated by size, has the highest like rate. Since users on Twitter can see the overall popularity of a tweet (the number of retweets and likes), size does provide a signaling mechanism that could potentially bias the highest like rate in the favor of popular tweets, fueling a rich-gets-richer effect. Another possibility is that tweets have an intrinsic "quality" that drives their popularity and higher size is partially a result of this quality (it could still involve other factors too, such as being lucky and receiving attention from popular users early in the tweet's lifetime). To disambiguate between these possibilities, we observe that over the lifetime of a cascade, different users view the tweet at different points in its popularity. This allows us to study whether users react differently to tweets that had different eventual sizes but were viewed at the same level of popularity by users. We can see from results in Figure 5.6 that the cascades that ended up with a larger eventual size had a higher like and retweet rate even earlier in their life. This provides some evidence that intrinsic quality of a tweet does contribute to its eventual popularity.



(a)



(b)

Figure 5.6: The behavior of consumers relative to the current popularity of the content, (a) Likes (b) Retweets.

Individual User Preferences

The analysis above suggests that globally popular content is also locally popular at an aggregate level for users. We now examine individual user variation for these preferences: do most users like globally popular content? Are users consistent in their preferences on local vs global content? In order to study these questions, we randomly selected 10000 *active* users on Twitter, where if a user had at least one interaction each day for more than 15 days out of a 16 day period in June, we count him/her as an active user.

Let us first turn to the question of local vs global preference for a user. To study this, for each user we compute the Pearson correlation coefficient between the final cascade size and like probability of all the tweets that the user viewed in this period. The distribution of these correlations over the 10000 users is shown in Figure 5.5(b). As we see from the plot, most users have a negative correlation, implying that they prefer personalized content. However, the correlation coefficient is low for these negatively correlated users, and if we only focus instead on users who have strong correlation then most of these are positively correlated. In either case, users do seem to exhibit a local/global content preference.

The local/global preference as expressed by a correlation is however rather weak, and leaves open the question of whether users are consistent in their preferences. Here, we study user consistency via the following analysis. We define the function $f(u, d, x)$, for user u, day d and any $x \in [0, 1]$, to be the fraction of likes produced by the x fraction of smallest cascades that user u sees on day d. If $f(u, d, 0.5) < 0.50$, it means that user u likes content that is part of smaller cascades at a greater rate than they like content that is part of larger cascades; we could think of this as the user liking personal content more than broadly-shared general content. The implication is the opposite if the inequality is reversed: if $f(u, d, 0.5) > 0.50$, then the user u like content that it part of large cascades

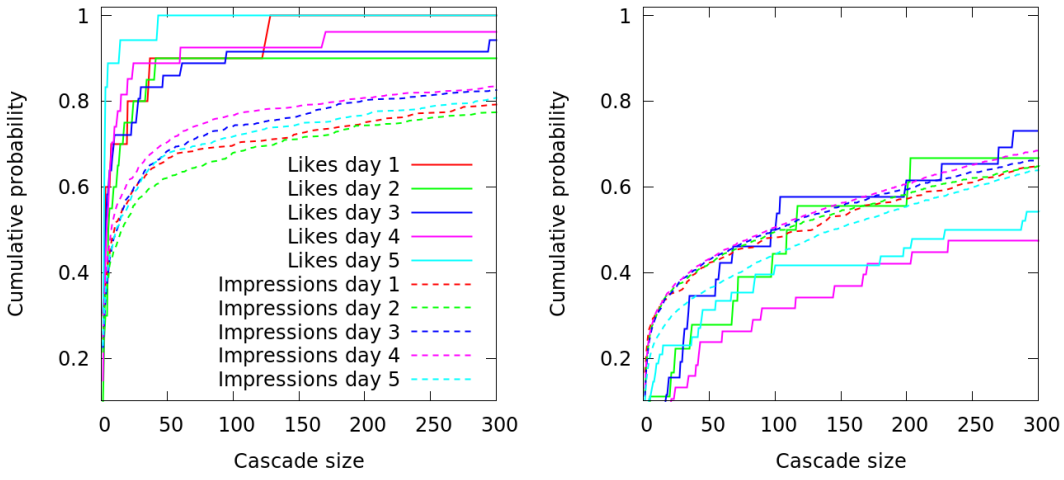
Consistent small cascade liker	19.4%
Consistent big cascade liker	47.1%
Indifferent	33.5%

Table 5.2: Fraction of each type of the three user.

(and hence more broadly-shared general content) at a greater rate. We operationalize this definition by identifying users who like small (respectively, large) cascades according to the condition $f(u,d,0.5) \geq 0.55$ (respectively $f(u,d,0.5) \leq 0.45$).

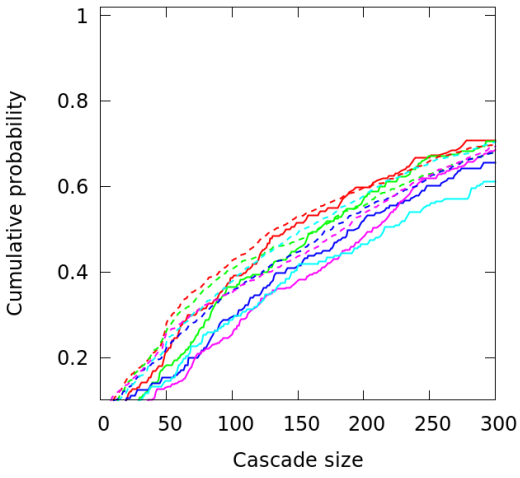
Further, if a user likes the same type of cascades at least 11 days over the 16 day period, we call her a consistent user (either a liker or large cascades or a liker of small cascades); otherwise we call her *indifferent*. With this definition of consistency, we find that more than 66% of our users are consistent. The fraction of users of each type can be seen in Table 5.2. As a simple baseline, note that if each user independently decided each day with uniform probability whether to like small cascades or large cascades, we'd expect only 21% of all users to be consistent, rather than over 66% as we find in the data.

We also provide a useful illustration of how these three user behaviors are distinct. We define $i(u,d,x)$ ($l(u,d,x)$) to be the fraction of impressions (likes) that user u had on day d and the cascade sizes were less than x. We then show how these two functions behave for all three types of users (prefers personal content, broader content or indifferent) in Figure 5.7. We draw the function $i(u,d,x)$ with dashed line and $l(u,d,x)$ with solid lines. This data clearly indicates that users do indeed have consistently different preferences, which might stem from using Twitter for different purposes. We leave a more in-depth investigation of this to future work, but note that Twitter does exhibit both social and information network structural properties indicating the presence of multiple usage scenarios [63].



(a)

(b)



(c)

Figure 5.7: Three samples of different groups of users: (a) Small cascade liker (b) Large cascade liker (c) Indifferent. Note that figure (b) and (c) have the same legends as figure (a).

We now summarize our overall empirical findings: cascades have a large audience on Twitter, larger than one might expect just based on cascade occurrence (cf. the *Impressions Paradox*). Furthermore, users seem to like these cascades, even relative to organic content. However, these observations seem to be somewhat at odds with each other: since a lot of users see large cascades, either cascading content is always appreciated by a large majority or somehow it reaches predominantly users who would enjoy it. The latter possibility is what we explore in the remainder of this work, with the lens of a simple theoretical model.

5.4 Modeling Cascades For Audience

The empirical analysis presented earlier leads to the conclusion that a quarter of a typical user's timeline consists of retweets, and users find this content engaging. As we've alluded to before, this presents a conflict between the view of users choosing exactly what content they want to see (by following a set of users) and engaging content being bubbled up through the network via retweets. In this section, we posit a simple theoretical model that shows this conflict can be resolved in a natural manner by users being selective about content they retweet. We emphasize that the model is purposely barebones so that it can provide stylized insight into why cascades turn out to be relevant for users.

The main idea of our model is to think of an *inverted* tree where a member of the audience (a consumer of cascades) is the root, in contrast to the standard view of the cascade originator as the root node. This allows us to examine the path that content takes to arrive at the audience member at the root node. The model includes a notion of *topic*, which governs whether a given user is interested in a given tweet. We can now

define the model formally using these notions.

Each user u in the network has a set of topics they are interested in: $I_u \subset I$, where $|I_u| = d$ and the set I contains the universe of topics ($|I| = D$). For a fixed arbitrary user a, we consider their two-hop neighborhood (recall from Table 5.1 that vast majority of the retweets arrive from two hops away): a follows the set of users $B = \{b_1, \ldots, b_k\}$, and each user b_i follows $C_i = \{c_{i1}, \ldots, c_{ik'}\}$, with the entire second hop neighborhood being denoted by $C = \bigcup_{i=1}^{k} C_i$. In this simple network, we assume that if a user u follows user v, then u is interested in a significant fraction of v's topics. More formally, we assume that for a fixed constant $0 < \alpha < 1$, $|I_u \cap I_v| \geq \alpha d$. We'll shortly address the question of how this extended neighborhood is generated, but first we will focus on the cascade process given such a network.

Given a network that has content being produced and consumed on it, for each user u we will have a "home timeline" which is populated by content produced by the set of users that u follows; we denote the set of tweets in u's home timeline by TL_u. We now specify the content production process. First, we link users' topical interests to tweets by assigning a single topic i_t to each tweet t, where i_t is selected from the author's list of topics. We also assume that tweet t has an intrinsic *quality* q_t, which is in line with our observations from Figure 5.6. For tweet production, we assume simplistically that at all users produce tweets at each epoch as follows. First, each user u creates an original candidate tweet t_o on one of the topics $i_{t_o} \in I_u$, with a quality q_{t_o} drawn from a specified distribution $\mathcal{D}$, but doesn't publish this candidate to its followers yet. The user picks her tweet that she will publish as follows: she considers her own tweet and her home timeline TL_u — consisting of tweets produced in the previous epoch by the users she follows — and then among the tweets t in this set for which $i_t \in I_u$, she selects the tweet t_h of highest quality q_{t_h}

This setting reflects the fact that users can both participate in cascades as well as produce original content. Further, the model allows for user curation for cascades that biases towards participation in higher quality cascades. The goal of the model is to capture the consumer viewpoint on cascades, and the consumer view is governed by whether the content in their home timeline is high quality, and also on whether it is on a topic that is interesting to them. We formally define these metrics for the home timeline of a given user a as follows:

Definition 5.4.1 (Precision) *We define* precision *for a user u as the fraction of tweets in u's timeline that u would be interested in:* $\text{Precision}_u = \sum_{t \in \text{TL}_u} \frac{|\{i_t \in I_u\}|}{|\text{TL}_u|}$.

Definition 5.4.2 (Quality) *We define the timeline* quality *for a user u as the average quality of all tweets in u's home timeline: quality$(u) = \sum_{t \in \text{TL}_u} \frac{q(t)}{|\text{TL}_u|}$.*

Definition 5.4.3 (Timeline Utility (TLU)) *If δ is the quality coefficient for tweets that are not on topic for the consumer then the overall* timeline utility *of the home timeline for a given user u, can be defined as*

$$\text{TLU}_u = \sum_{t \in \text{TL}_u} g(t, u) \cdot q_t.$$

Here $g(t, u) = 1$ if $i_t \in I_u$, and $g(t, u) = \delta \leq 1$ otherwise. Thus, TLU_u increases if there are high-quality tweets and decreases if there are off-topic tweets.

Thus, the model aims to capture the effect of cascades on the twin goals of having users enjoy both high quality and precise content. Intuitively, retweets seem to filter for high quality tweets but the effect on precision is less clear. In fact, the exact effect on precision depends on network topology. We now define how the network is created, starting with a simple model that is analytically tractable. We'll later define a more complex model on which we simulated the model.

Possibly the simplest way to construct a network is to build a tree. We proceed by having given the node a for which we want to study the above metrics, and her corresponding topical interests. Then we pick k random interest sets that satisfies the homophily condition (of α fraction interest overlap), and designate those to be the interest sets of the nodes in B. We then repeat the same procedure for each b_i and create k neighbors for each b_i while also satisfying the homophily restrictions. Given this network generation model, now we can analyze the effect of cascades on previously defined quality and precision metrics.

For the theoretical analysis we look into a basic setting where $\delta = 1$. In the next subsection we remove these restrictions and report the results on the simulation. We begin by noting that based on the graph generation process, the distribution of topics in C_i on the topics in I_{b_i} is uniform. Now, in every epoch but the first (once retweets start moving through the network), the view of a node b_i is as follows. If b_i receives a tweet that she is not interested in, that has no chance of being propagated. Otherwise, she will keep the tweet as a candidate in the maximization step. Now, we know that the distribution of the tweet topics coming from $\{c_{i1}, \ldots, c_{ik}\}$ is uniform on I_{b_i}, so we observe that the model is equivalent to b_i itself generating many tweets and only publishing the best. If we look at the process in this manner, then it is clear that the expected precision does not change when we introduce retweets. However, with this procedure the quality of tweets will go up. It is easy to see that the expected quality of a single tweet is less than the expectation of the maximum of independent draws from the same distribution. We formalize the gain in quality for two specific distributions, and note that the analysis can be extended to other distributions.

In particular we investigate the two cases where $\mathcal{D}$ is either a uniform distribution over $[0, 1]$ or an exponential distribution with rate λ. We use $\mathcal{G}$ to refer to the distribution

of the maximum of k draws from $\mathcal{D}$. A question we want to answer is; how much do retweets help the quality of the timeline to increase? Let X be a random variable drawn from $\mathcal{D}$ and Y be a random variable drawn from $\mathcal{G}$. With our notation we are interested in $\frac{\mathbb{E}[Y]}{\mathbb{E}[X]}$. It is well known that the mean of a uniform distribution over $[0, 1]$ is $\frac{1}{2}$ and the mean of a exponential distribution of rate λ is $\frac{1}{\lambda}$. Now we state two well-known lemmas about the maximum of a set of independent draws from a fixed distribution [72].

Lemma 5.4.4 *The expectation of the maximum of k i.i.d draws from a uniform distribution over* $[0, 1]$ *is* $\frac{k}{k+1}$.

Lemma 5.4.5 *The expectation of the maximum of k i.i.d draws from an exponential distribution of rate* λ *is* $\frac{H_k}{\lambda}$*, where* H_k *is the k-th harmonic number.*

We can now state the main theoretical result on the effect of retweets on quality. We skip by proof but note that it is easily obtained from the above two lemmas.

Theorem 5.4.6 *The multiplicative increase in quality* $\frac{\mathbb{E}[Y]}{\mathbb{E}[X]}$ *in the scenario where* $\mathcal{D}$ *is uniform is* $\frac{2k}{k+1}$ *and when* $\mathcal{D}$ *is an exponential distribution with rate* λ*, it is* H_k.

The theorem formalizes the gain in quality that comes from having cascades in the network, which is a counterfactual that is not easily observable in the Twitter network[2]. To gain some sense of the scale of this increase, let us consider a network where the average degree of nodes is 50 (This number is a lower-bound for the Twitter network). By plugging in 50 for k we see that the uniform and exponential distribution model yield a 96% and 350% increase in quality, respectively! This shows how valuable retweets can be to a network, as illustrated by our stylized model. We re-emphasize that we do not

[2]This remains hard to measure even via experimentation as there is a large network effect to contend with.

claim our theorem to be representative of the gain in the Twitter setting — the goal of this modeling exercise is to shed light on the effect of cascades on the audience, and we can quantify that effect with the help of a formal model. We now explore generalizing this model by removing some assumptions.

5.4.1 Model Extensions

The above model is quite simple and makes a few strong assumptions in order to be analytically tractable. In this section, we explore the effect of removing or generalizing these assumptions via simulating the model. In particular, there are two obvious ways we can generalize the model.

First, we previously assumed that the two-hop graph was a tree, but now we generalize that to the following two graphs:

- **Tree Contracted model:** Here, we create the graph using the Tree model but contract all the nodes in layer B and C that have the same interest set into one single node.
- **k-NN:** Here, we create a number of nodes[3] with their own interest set, and each person follows the k people that have the most common interests with them. (We break ties in this choice of k at random.)

Second, we can change various aspects of the model in the previous section, which includes both changing existing parameter values and generalization of behavior from the model in the previous section. The generalizations we examine are as follows:

- δ: Recall that if a user receives an off-topic tweet of quality q, then she sees that

[3]For our simulations, we use 4×10^5 nodes.

as a tweet with quality $q\delta$. This parameter is used for both in finding the TLU and the retweeting procedure, and was previously set to 1. Here, we explore the effect of using other values for δ.

- $k' = \frac{|C|}{|B|}$: This parameter allows us to control the average degree of the neighbors of the target node to her degree. We note that in our simulations, the results stabilize once this ratio is above 20, and this value is above 20 for Twitter in particular.
- **The self-interest factor (p)**: We introduce a new parameter to allow users the flexibility to tweet original content instead of simply retweeting others. We stipulate that in each epoch, a user tweets her own tweet with probability p, and otherwise, retweets one of the tweets that she has received. Also, when retweeting the user differentiates between the tweets from the people she follows and the people that she does not follow (We take following to be a proxy for knowing). Thus, once she decides to retweet someone, with probability p she picks the highest quality tweet created by one of her immediate followees, otherwise she picks the highest quality tweet from the pool of tweets coming from more than one hop away from her.

  Note that as we increase p, this method creates a bias towards (i) creating her own organic tweet, and (ii) while retweeting, she prioritizes her immediate neighbors' tweet over tweets coming from deeper in the network. This creates a mechanism that constructs timelines which most of its content is from at most two hops away.

The simulations results from varying the network model and the above parameters are all shown in Figure 5.8. From the results, we observe that the following holds unless δ is close to 1[4]: by introducing retweets in the network, the precision remains essentially the same and the quality goes up, leading to a higher TLU. This is quite consistent with the theorem in the previous section and shows that the observations in the previous

[4]Recall that δ being close to 1 would imply users not distinguishing between on and off topic tweets, and hence it makes sense to focus on results where δ is smaller than 1.

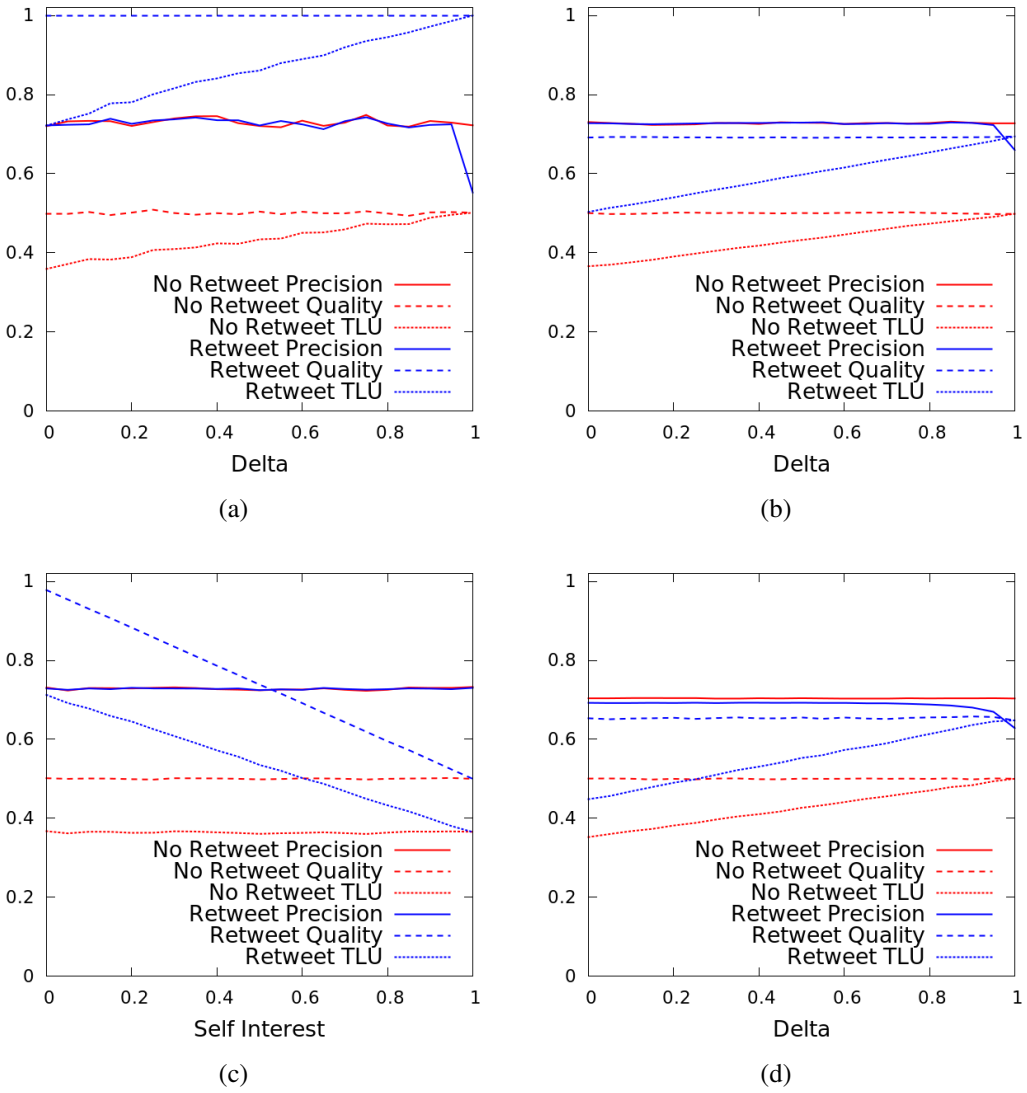


Figure 5.8: Precision, Quality and TLU over different models and parameters set. (a) Simple tree model (b) Tree Contracted model with p = 0.6 (c) Tree Contracted model with $\delta = 0$ (d) k-NN model with p=0.6.

section are somewhat robust to the specific network model and parameters.

5.5 Conclusion

Information cascades on networks affect not only the users who actively participate in them, but also the audience who are consequently exposed to the cascades. Our work provides a view of cascades from the point of view of the audience, which has not received much attention to the best of our knowledge. Our findings related to the *Impressions Paradox* provides a novel perspective for future work on the effect of cascades on their audience.

The theoretical model presented here is quite simplistic, but it does highlight the crucial role of cascade participants as gatekeepers of precision.

Healthy dynamics on social networks requires both active producers and engaged consumers. We believe our work provides a novel and useful consumer counterpoint to the extensive literature on the role that producers play in cascades.

CHAPTER 6

FUTURE WORK

Many of the valuable data sources that has been studied in this thesis, have just been created and we are just getting started on understanding interactions on these online social platforms. Moving forward there are two natural ideas that is worth pursuing. First being, now that we know more about interactions among people on online social networks we can go back to the original question and see if our findings hold in the offline world. Second, is taking the work shown in each of the chapters in this thesis and, generalizing, completing and verifying them.

The work shown in chapter 2 proposed an approach that is suitable in many contexts, therefore it suggests a wide range of directions for further work. In particular, we showed how the activity level of users participating in a trend changes over time, but there are many parameters of the trend that vary as time unfolds, and it would be interesting to track several of these at once and try to identify relationships across them. It would also be interesting to try incorporating the notion of the status gradient into formulations for the problem of starting or influencing a cascade, building on theoretical work on this topic [28, 43].

In chapter 3 we shed light on how innovations spread through networks of collaboration. First, it would be interesting to develop a comparative analysis between the structure of inheritance graphs that we defined and the corresponding structures for the diffusion of on-line memes. Are there systematic ways in which the two types of diffusion patterns differ, and can these be connected to differences in the underlying mechanisms? Second, we believe that there may well be additional links between inheritance structures and prediction problems for the trajectory of the overall system; for example, can we evaluate the future course of larger sub-areas based on the inheritance patterns

that exhibit? And finally, identifying "tracers" for complex practices is a style of analysis that can be applied in other domains as well; as we broaden the set of contexts in which we can perform this type of analysis, we may better understand the ways in which the flow of practices helps reinforce and illuminate our understanding of large collaborative communities.

Chapter 4 proposed clean methodology to study conventions that are synonymous which enables a number of further directions for research. First, there are other domains where it should be possible to control for the meaning of a convention, for example in repositories of source code where naming conventions can change while the behavior of the code remains constant. It is an interesting question to see whether similar phenomena hold in the dynamics of conventions there. More generally, it is an interesting question to look for additional structure beyond the changeover and fight dynamics presented in this thesis on the competition between these conventions. And finally, it is intriguing to consider using the mathematics developed around the theory of neutral variation [44] to begin developing models for the evolution of synonymous conventions over time.

Chapter 5 tries to reveal why users on social networks do not get overwhelmed or annoyed by the content they receive after introducing resharing mechanisms. We used empirical experiments and theoretical models to achieve this goal. In addition to further generalizations of the model, this work raises several additional open questions for future work. Clearly, not all users retweet on topic, but does the network rewire itself (via audience following/unfollowing) so that precision remains high? Furthermore, an aspect of cascades we did not discuss here is their usefulness as a discovery mechanism; can effective discovery coexist with high precision in the network?

BIBLIOGRAPHY

[1] Eric Abrahamson and Lori Rosenkopf. Social network effects on the extent of innovation diffusion: A computer simulation. *Organizational Science*, 8(3), 1997.

[2] Lada A. Adamic and Natalie Glance. The political blogosphere and the 2004 US election: Divided they blog. In *International workshop on link discovery*, 2005.

[3] Lada A. Adamic, Thomas M. Lento, and Andrew T. Fiore. How you met me. In *Proc. 6th International Conference on Weblogs and Social Media*, 2012.

[4] 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.

[5] Jon Aizen, Dan Huttenlocher, Jon Kleinberg, and Antal Novak. Traffic-based feedback on the Web. *PNAS*, 101, 2004.

[6] Ashton Anderson, Daniel P. Huttenlocher, Jon M. Kleinberg, Jure Leskovec, and Mitul Tiwari. Global diffusion via cascading invitations: Structure, growth, and homophily. In *Proc. 24th International World Wide Web Conference*, pages 66–76, 2015.

[7] Sinan Aral, Lev Muchnik, and Arun Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *PNAS*, 2009.

[8] Richard Arratia and Larry Goldstein. Size bias, sampling, the waiting time paradox, and infinite divisibility: when is the increment independent? Technical Report 1007.3910, arxiv.org, October 2007.

[9] W. Brian Arthur. Competing technologies, increasing returns, and lock-in by historical events. *Economic Journal*, 99(394):116–131, March 1989.

[10] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks: Membership, growth, and evolution. In *Proc. 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.

[11] Eytan Bakshy, Solomon Messing, and Lada Adamic. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 5 2015.

[12] Nicola Barbieri, Francesco Bonchi, and Giuseppe Manco. Topic-aware social influence propagation models. *Knowledge and information systems*, 37(3):555–584, 2013.

[13] Lisa Barone. Why twitter's new retweet feature sucks. `http://outspokenmedia.com/social-media/twitters-new-retweet-feature-sucks/`. 2009, Accessed: 2016-08-20.

[14] Marshall H. Becker. Sociometric location and innovativeness: Reformulation and extension of the diffusion model. *American Sociological Review*, 35(2), 1970.

[15] Michael Bendersky and David A. Smith. Cll wkshp. at naacl, 2012.

[16] Jonah Berger and Katherine L. Milkman. What makes online content viral? *Journal of Marketing Research*, 2012.

[17] Reza Bosagh Zadeh, Ashish Goel, Kamesh Munagala, and Aneesh Sharma. On the precision of social and information networks. In *Proceedings of the first ACM conference on Online social networks*, pages 63–74. ACM, 2013.

[18] Ronald S. Burt. Structural holes and good ideas. *American Journal of Sociology*, 110(2), 2004.

[19] John W. Byers, Michael Mitzenmacher, and Georgios Zervas. The groupon effect on yelp ratings: a root cause analysis. In *Proceedings of EC*, 2012.

[20] Justin Cheng, Lada A. Adamic, P. Alex Dow, Jon M. Kleinberg, and Jure Leskovec. Can cascades be predicted? In *Proc. 23rd International World Wide Web Conference*, pages 925–936, 2014.

[21] Riley Crane and Didier Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *PNAS*, 2008.

[22] Mary Crossan and Marina Apaydin. A multi-dimensional framework of organizational innovation: A systematic review of the literature. *Journal of Management Studies*, 47(6), 2010.

[23] Richard L. Daft. A dual-core model of organizational innovation. *Academy of Management Journal*, 21(2), 1978.

[24] Cristian Danescu-Niculescu-Mizil, Justin Cheng, Jon Kleinberg, and Lillian Lee. You had me at hello: How phrasing affects memorability. In *Proceedings of the ACL*, 2012.

[25] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. No country for old members: User lifecycle and linguistic change in online communities. In *Proc. 22nd International World Wide Web Conference*, 2013.

[26] G. de Tarde and E.W.C. Parsons. *The Laws of Imitation*. H. Holt, 1903.

[27] Paul Deutschmann. Communication and adoption patterns in an Andean village. Technical report, Programa Interamericano de Información Popular, 1962.

[28] Pedro Domingos and Matthew Richardson. Mining the network value of customers. In *Proceedings of ACM SIGKDD*, 2001.

[29] P Alex Dow, Lada A Adamic, and Adrien Friggeri. The anatomy of large facebook cascades. In *Proceedings of ICWSM*, 2013.

[30] David Easley and Jon Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press, 2010.

[31] Jacob Eisenstein. Phonological factors in social media writing. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 11–19, Atlanta, 2013.

[32] Jacob Eisenstein. What to do about bad language on the internet. In *Proceedings of NAACL-HLT*, pages 359–369, 2013.

[33] Facebook. News Feed FYI: helping make sure you Don't miss stories from friends. http://newsroom.fb.com/news/2016/06/news-feed-fyi-helping-make-sure-you-dont-miss-stories-from-friends/. Accessed: 2016-10-24.

[34] Matthew Gentzkow and Jesse Shapiro. What drives media slant? Evidence from u.s. daily newspapers. *Econometrica*, 78(1):35–71, 2010.

[35] Ashish Goel, Kamesh Munagala, Aneesh Sharma, and Hongyang Zhang. A note on modeling retweet cascades on twitter. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 119–131. Springer International Publishing, 2015.

[36] Rahul Goel, Sandeep Soni, Naman Goyal, John Paparrizos, Hanna Wallach, Fernando Diaz, and Jacob Eisenstein. The social dynamics of language change in online networks. *SocInfo*, 2016.

[37] Sharad Goel, Ashton Anderson, Jake M. Hofman, and Duncan J. Watts. The structural virality of online diffusion. *Management Science*, 62(1):180–196, 2016.

[38] Sharad Goel, Duncan Watts, and Daniel Goldstein. The structure of online diffusion networks. In *Proceedings of EC*, 2012.

[39] Daniel Gruhl, R. V. Guha, David Liben-Nowell, and Andrew Tomkins. Information diffusion through blogspace. In *Proc. 13th International World Wide Web Conference*, 2004.

[40] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *ACL*, 2016.

[41] David Kaiser. Physics and Feynmans diagrams. *American Scientist*, 93:156–165, Mar-Apr 2005.

[42] Elihu Katz and Paul Lazarsfeld. *Personal Influence*. Free Press, 1955.

[43] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of ACM SIGKDD*. ACM, 2003.

[44] Motoo Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, 1983.

[45] Jon Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of ACM SIGKDD*, 2002.

[46] Farshad Kooti, Haeryun Yang, Meeyoung Cha, Krishna P. Gummadi, and Winter A. Mason. Proceedings of ICWSM, 2012.

[47] David Krackhardt. Organizational viscosity and the diffusion of innovations. *Journal of Mathematical Sociology*, 1997.

[48] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. On the bursty evolution of blogspace. In *Proceedings of WWW*, 2003.

[49] William Labov, Ingrid Rosenfelder, and Josef Fruehwald. One hundred years of sound change in philadelphia: Linear incrementation, reversal, and reanalysis. *Language*, 89(1):30–65, 2013.

[50] William L. Labov. *The Social Stratification of English in New York City*. Washington, D.C.: Center for Applied Linguistics, 1966.

[51] Wan Shun Eva Lam. *Language Socialization in Online Communities*, pages 2859–2869. Springer US, 2008.

[52] Jure Leskovec, Lada Adamic, and Bernardo Huberman. The dynamics of viral marketing. *ACM Transactions on the Web*, 2007.

[53] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proc. 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009.

[54] Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst. Cascading behavior in large blog graphs. In *Proc. SIAM International Conference on Data Mining*, 2007.

[55] David Liben-Nowell and Jon Kleinberg. Tracing information flow on a global scale using Internet chain-letter data. *Proc. Natl. Acad. Sci. USA*, March 2008.

[56] Yu-Ru . R. Lin, Drew Margolin, Brian Keegan, Andrea Baronchelli, and David Lazer. #bigbirds never die: Understanding social dynamics of emergent hashtag. In *ICWSM*, 3 2013.

[57] Avishay Livne, Matthew P Simmons, Eytan Adar, and Lada A Adamic. The party is over here: Structure and content in the 2010 election. *ICWSM*, 11:17–21, 2011.

[58] Suman Kalyan Maity, Ritvik Saraf, and Animesh Mukherjee. # bieber+# blast=# bieberblast: Early prediction of popular hashtag compounds. In *Proceedings of CSCW*, 2016.

[59] Julian John McAuley and Jure Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *Proceedings of WWW*, 2013.

[60] Milbrey W. McLaughlin. The Rand change agent study revisited: Macro perspectives and micro realities. *Educational Researcher*, 19(9), 1990.

[61] Sunny Mitra, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee, and Pawan Goyal. That's sick dude!: Automatic identification of word sense change across different timescales. In *ACL*, 2014.

[62] G.A. Moore. *Crossing the Chasm, 3rd Edition: Marketing and Selling Disruptive Products to Mainstream Customers*. Collins Business Essentials. HarperCollins, 2014.

[63] Seth A Myers, Aneesh Sharma, Pankaj Gupta, and Jimmy Lin. Information network or social network?: the structure of the twitter follow graph. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 493–498. ACM, 2014.

[64] Elisa Omodei, Thierry Poibeau, and Jean-Philippe Cointet. Multi-level modeling of quotation families morphogenesis. In *SocialCom*. IEEE, 2012.

[65] Fred C. Pampel. Inequality, diffusion, and the status gradient in smoking. *Social Problems*, 2002.

[66] Claudia Peersman, Walter Daelemans, Reinhild Vandekerckhove, Bram Vandekerckhove, and Leona Van Vaerenbergh. The effects of age, gender and region on non-standard linguistic variation in online social networks. *arXiv preprint arXiv:1601.02431*, 2016.

[67] Florent Perek. Vector spaces for historical linguistics: Using distributional semantics to study syntactic productivity in diachrony. In *ACL*, 2014.

[68] Everett Rogers. Characteristics of agricultural innovators and other adopter categories. Technical Report 882, Agricultural Experimental Station, Wooster OH, 1961.

[69] Everett Rogers. *Diffusion of Innovations*. Free Press, fourth edition, 1995.

[70] 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 WWW*, pages 695–704, 2011.

[71] Daniel M Romero, Chenhao Tan, and Johan Ugander. On the interplay between social and topical structure. In *Proceedings of ICWSM*, 2013.

[72] Sheldon M. Ross. *Introduction to Probability Models, Ninth Edition*. Academic Press, Inc., Orlando, FL, USA, 2006.

[73] Rahmtin Rotabi, Cristian Danescu-Niculescu-Mizil, and Jon Kleinberg. Competition and selection among conventions. In *Proc. 26th International World Wide Web Conference*, 2017.

[74] Rahmtin Rotabi, Cristian Danescu-Niculescu-Mizil, and Jon Kleinberg. Tracing the use of practices through networks of collaboration. *arXiv preprint arXiv:1703.09315*, 2017.

[75] Rahmtin Rotabi, Krishna Kamath, Jon Kleinberg, and Aneesh Sharma. Cascades: A view from audience. In *Proceedings of the 26th International Conference on World Wide Web*, pages 587–596. International World Wide Web Conferences Steering Committee, 2017.

[76] Rahmtin Rotabi and Jon Kleinberg. The status gradient of trends in social media. In *Proc. 10th International Conference on Weblogs and Social Media*, pages 319–328, 2016.

[77] Bryce Ryan and Neal C Gross. The diffusion of hybrid seed corn in two iowa communities. *Rural sociology*, 8(1):15, 1943.

[78] Matthew Salganik, Peter Dodds, and Duncan Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311:854–856, 2006.

[79] George Simmel. *The Sociology of Georg Simmel*. Free Press (translated by Kurt H. Wolf), 1908.

[80] David Strang and Sarah Soule. Diffusion in organizations and social movements: From hybrid corn to poison pills. *Annual Review of Sociology*, 24:265–290, 1998.

[81] Nina Tahmasebi, Thomas Risse, and Stefan Dietze. Towards automatic language evolution tracking, a study on word sense tracking. In *Workshop on Knowledge Evolution and Ontology Dynamics (EvoDyn 2011)*, 2011.

[82] Chenhao Tan and Lillian Lee. All who wander: On the prevalence and characteristics of multi-community engagement. In *Proceedings of WWW*, 2015.

[83] Oren Tsur and Ari Rappoport. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *Proc. 5th ACM International Conference on Web Search and Data Mining*. ACM, 2012.

[84] Oren Tsur and Ari Rappoport. Don?'t let me be #misunderstood: Linguistically motivated algorithm for predicting the popularity of textual memes. In *International Conference on Weblogs and Social Media*, 2015.

[85] Twitter. While you were away... `https://blog.twitter.com/2015/while-you-were-away-0`. Accessed: 2016-10-24.

[86] Thomas Valente. Network interventions. *Science*, 337(49), 2012.

[87] David Willer (editor). *Network Exchange Theory*. Praeger, 1999.

[88] Shaomei Wu, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Who says what to whom on twitter. In *Proceedings of the WWW*, 2011.

[89] Jaewon Yang and Jure Leskovec. Patterns of temporal variation in online media. In *Proceedings of WSDM)*, 2011.

[90] H. Peyton Young. The economics of convention. *Journal of Economic Perspectives*, 10(2):105–122, Spring 1996.