COMPUTATIONAL APPROACHES TO UNDERSTANDING HUMAN BEHAVIOR FROM ONLINE SOCIAL INTERACTIONS: LANGUAGE AND COMMUNITIES

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

Chenhao Tan

August 2016

COMPUTATIONAL APPROACHES TO UNDERSTANDING HUMAN BEHAVIOR FROM ONLINE SOCIAL INTERACTIONS: LANGUAGE AND COMMUNITIES

Chenhao Tan, Ph.D.

Cornell University 2016

Online socio-technical systems, such as Wikipedia and Facebook, offer massive amounts of data for researchers to study human behavior, while posing great challenges and opportunities in building these systems. For instance, individuals may find it difficult to make their messages heard online, and service providers struggle to figure out a robust way to organize various communities. Accordingly, this thesis investigates computational approaches in two broad and crucial directions: the effect of wording on social interaction and multi-community engagement.

Although language is an important channel for online social interactions, it is unclear how wording choices impact the communicative goals of a speaker, due to many confounding factors such as who the speaker is and what the topic is about. As a result, current systems that assist people in writing only correct simple grammatical errors. In this thesis, we take advantage of massive online social interactions to investigate the role of wording while controlling for important confounding factors. First, we conduct natural experiments on Twitter to study the effect of wording on message propagation, controlling for the author and the topic. We demonstrate that wording indeed matters and build classifiers that outperform humans in predicting which tweet within a pair will be retweeted more. Second, we develop a large-scale study for another com-

mon communicative goal, i.e., making arguments to change another person's opinion. By comparing similar counterarguments to the same original opinion, we show that language factors play an essential role. In particular, the interplay between the language of the opinion holder and that of the counterargument provides highly predictive cues of persuasiveness.

The second subject of this thesis is motivated by the existence of multiple communities. Nowadays many websites allow users to self-organize into communities and these websites consequently provide massive data to study multi-community engagement quantitatively. We first study this issue from the perspective of individual users by examining users' life trajectories across multiple communities. Using datasets from Reddit and DBLP, we provide the first characterization of users' multi-community engagement. For instance, in contrast with the "getting old and settling down" hypothesis, users in our data *continually explore* new communities. Users' wandering behavior can also be used to predict their future activity levels. From the perspective of community organizers, it remains an open question how to design the space of communities: what constitutes a community, how different communities relate to each other, and how to keep communities flourishing. As an initial effort in this direction, we investigate the relationship between highly related communities and create a taxonomy to distinguish different types of highly related communities. One interesting finding regarding users' behavior is: for several types of highly-related community pairs, after a newer community is created, users that engage in the newer community tend to be more active in their original community than users that do not explore, even when controlling for previous level of engagement.

BIOGRAPHICAL SKETCH

Chenhao Tan grew up in Jingdezhen, China, known as the "Porcelain Capital". He did not pick up crafting or art as a kid, but was intrigued by the beauty of math and the "competition" between humans instead.[1] After finishing high school, he went to Beijing for undergraduate studies at Tsinghua University. Driven by the two interests above, he studied Computer science, a similar subject to mathematics in his opinion, and Economics, a discipline that studies human behavior in a scientific way. He built predictive models on social networks in his undergraduate honor thesis. He decided to receive more schooling by pursuing a doctoral degree at the Department of Computer Science at Cornell University. In Ithaca, he obtained a comprehensive education both in research and in life over six wonderful years, four of which were spent at Telluride house. After graduation, he will strive to continue this journey to develop knowledge on human behavior in a computational way.

[1] It is up to interpretation whether this is a fortunate or unfortunate fact.

Dedicated to all sentient beings

ACKNOWLEDGEMENTS

First of all, I would like to express my deepest gratitude to my advisor and friend, Lillian Lee. What she has taught me is beyond the many hours that we spent elbow to elbow working together. She has been a true role model, which I believe is hard to describe verbally, and thus I hope to demonstrate through my work and life in the years to come.

Another great fortune during my PhD is to be able to work with Jon Kleinberg and Sendhil Mullainathan, two MacArthur award winners. Witnessing their geniuses is a great experience and makes a best way to learn. They have always been a great source of inspiration.

In addition, this PhD experience cannot be great without my wonderful collaborators, colleagues and friends: Robert Kleinberg who agreed to advise me in my minor, applied mathematics; excellent researchers that I have co-authored with, Lada Adamic, Claire Cardie, Ed Chi, Eunsol Choi, Cristian Danescu-Niculescu-Mizil, Adrien Friggeri, Evgeniy Gabrilovich, Bo Gao, Jack Hessel, David Huffaker, Long Jiang, Isabel Kloumann, Arnd Christian Konig, Gueorgi Kossinets, Tao Lei, Ping Li, Tian Li, Quan Lin, Bin Lu, Michael Macy, Vlad Niculae, Bo Pang, Daniel M. Romero, Alex Smola, Jimeng Sun, Jie Tang, Wenbin Tang, Benjamin K. Tsou, Johan Ugander, Fengjiao Wang, Shaomei Wu, Ming Zhou and Thomas Zimmermann; friends that I hang out in the same office, Shrutarshi Basu, Ivaylo Boyadzhiev, Kyle Croman, Ian Lenz, Yixuan Li, Antonio Marcedone, Stavros Nikolaou, Jon Park, Maithra Raghu, Amit Sharma, Adith Swaminathan, Zhiyuan Teo, Lu Wang, Andreas Veit and Yexiang Xue; the great NLP seminar that I am going to miss, Xilun Chen, Yao Cheng, Liye Fu, Arzoo Katiyar, Moontae Lee, Karthik Raman, Mats Rooth, Myle Ott, Alexandra Schofield, Tianze Shi, Kai Sun, Bishan Yang, Ainur Yessenalina and Justine

Zhang; undergraduates that I worked with, Eunsol Choi and Kelvin Luu; classmates that spent essentially the entire PhD together with me, Shuo Chen and Wenlei Xie; Becky Stewart, who helped with almost everything and made sure that my stay in the US is legitimate.

Most of my time outside research has been spent at Telluride house, where I participated in experimental democratic education. In additional to the unique educational experience, I made some of my best friends. Here are some examples without listing everyone: Paulina Aroch, Holly Case, Duane Corpis, Albert Chu, Nancy Elshami, Hu Fu, Chong Guo, Jacob Krell, Daniel Marshall, Sadev Parikh, Rick Peng, Karl Pops, Srinath Reddy and Zbigniew Truchlewski. I learned a great deal from/with them, ranging from federalist papers to "useless" pop culture.

Last but not least, my parents, Yueming Tan and Shaolan Zhang, have provided me with great support. I especially appreciate the freedom that they have given me to pursue my ideal throughout my life. Special thanks to Ningzi Li, who has been brave and supporting me through rough times.

The work presented in this thesis was supported in part by the National Science Foundation grants ISS-0910664 and IIS-1016099, grants from Google and Yahoo!, a Yahoo! Key Scientific Challenges award and a Facebook Fellowship.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The widespread use of online socio-technical systems has offered researchers massive amounts of data on human behavior, which has contributed to an emerging field, "computational social science" (Lazer et al., 2009). For the first time, researchers can measure pairwise distances between people in social networks at a global scale (Ugander et al., 2011): we are better connected (4.7 steps) than "six degrees of separation" (Travers and Milgram, 1969). Researchers can also see how a user adapts her language to a community over her tenure within a community (Danescu-Niculescu-Mizil et al., 2013). Interestingly, users are able to adapt their language when they are new to the community but as they get "older", they deviate from the community in terms of language uses.

In addition to enabling a myriad of exciting studies on human behavior, it is important to keep in mind that these socio-technical systems are constantly evolving "organisms". Users are producing new content, forming new relationships, and reporting new complaints; while service providers are working night and day to tweak these systems for better user experiences. On the one hand, users ask for improvements on current systems, or even completely new systems. For instance, cyberbullying has become a serious problem for the youth (Smith et al., 2008). In general, it is difficult to have a civilized discussion online as everyone seems angry on the Internet (Natalie Wolchover). As a result, it is an important challenge to figure out what factors may contribute to a healthy discussion environment online. Another example is that crime related advertisements were more likely to show up when one searched African-American names (Sweeney, 2013). It is increasingly important to understand and avoid

such bias or discrimination in online socio-technical systems, as more and more people rely on search engines and social media platforms to access information (Bakshy et al., 2015; Pew Research Center). On the other hand, design choices made by service providers can lead to significant consequences: a simple feature on Facebook that allows users to claim that he or she voted can increase voter turnout in elections and potentially impact election outcomes (Bond et al., 2012); with the launch of Digg version 4, disgruntled users declared a "quit Digg day" on August 30, 2010 (Wikipedia), and started "the great Digg migration" to Reddit (ForWhatReason).

Furthermore, it is important to note that as data collection from offline activities is increasingly common, similar opportunities to study human behavior and improve policymaking arise in offline settings. If one draws an analogy between service providers and governments, users and citizens, many opportunities and challenges above have counterparts in offline settings. Although the main subject of this thesis is online social interactions, it is useful to think about these questions in both online and offline settings. More discussion will be presented in Chapter 6.

In light of these opportunities and challenges, this thesis focuses on two important components in socio-technical systems: 1) the effect of wording on social interaction and 2) multi-community engagement.

1.1 Wording matters

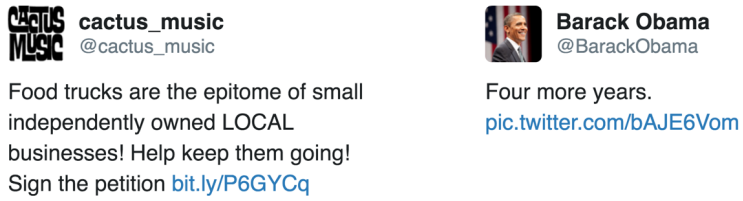
Language is an important channel through which we communicate with each other. As it is a powerful tool that is almost entirely under a writer or a speaker's

control, a lot of researchers have been working on understanding the art of rhetoric since Aristotle (350 BC). However, we still lack a quantitative understanding of the effect of wording on various communicative goals. There are many open questions towards building a system that can reason about the effect of language or even provide suggestions for people to improve their wording.

To demonstrate the complexity of this problem, let us consider a person who runs a campaign for refugees on social media and hopes to obtain as many shares as possible. She may spend hours crafting a tweet. Does it matter at all? Is it possible to build systems that can help individual users craft better messages for communication or persuasion?

Researchers have studied retweeting behavior, or message popularity in general and identified important factors that affect message popularity (Artzi et al., 2012; Bakshy et al., 2011; Borghol et al., 2012; Guerini et al., 2011, 2012; Hansen et al., 2011; Hong et al., 2011; Lakkaraju et al., 2013; Milkman and Berger, 2012; Ma et al., 2012; Petrović et al., 2011; Romero et al., 2013; Suh et al., 2010; Sun et al., 2013; Tsur and Rappoport, 2012). The factors include properties of the author and the topic of the message. The insights of these studies make it easy to guess which tweet would be retweeted more in Figure 1.1a: Obama is a prominent figure and presidential election was a popular topic, so one would infer that the tweet on the right would be shared more. However, these insights are not helpful for this user's decision making. She cannot get a million followers overnight, and talking about presidential election does not improve conditions for food trucks. What is in her control is the way that she delivers the message.

This suggests a different way to ask the question: which tweet will be



(a) A pair of tweets without control

(b) A topic- and author-controlled pair

Figure 1.1: The question of which one will be retweeted more is much easier to answer without control in the top figure than with control in the bottom figure. However, insights that are obtained from the easier question do not help the user compose a tweet for her goal.

retweeted more in the pair after controlling for **the author and the topic** (Figure 1.1b). This question can offer useful insights for an individual to phrase her message, but it is much harder to answer. In Chapter 2, we demonstrate how to develop an algorithm to answer such questions for topic- and author-controlled pairs more accurately than humans.

In general, this pairwise formulation can be viewed as a preliminary way to build systems that help people improve their writing. Message propagation on Twitter is one example of many social interactions in reality. In addition to short tweets, we may write an extensive argument in the hope of changing someone's opinions, or give a speech in a recruitment meeting to argue for a candidate, or explain a complicated concept to a student. In this thesis, we will also present a study on persuasive arguments (Chapter 3).

1.2 A world of communities

There are various communities in the offline world, ranging from a local community on organic food to a global community that supports female education. As the Internet develops, multiple communities arise in online socio-technical systems as well. These communities can be created for a variety of purposes. Some may be a projection of offline communities, e.g., an online community can be created for Cornell University. Some may be founded to focus on a particular subtopic, e.g., ukpolitics can be created to focus on politics in the UK. One interesting reason is that conflicts within a community lead to the emergence of new communities. Zachary studied the fission of a Karate club into two subgroups and found that the flow of sentiments and information across the ties in a social network can accurately predict future memberships in the new subgroups (Zachary, 1977). It is intriguing to consider whether communities are founded due to conflicts in online settings. If online communities indeed split into multiple ones, it remains an open question whether it is possible to foresee a splintering that is about to happen and even prevent the split if it hurts the community or recommend the split if it benefits both subgroups. Also, instead of focusing on the engagement of most users, small subgroups after splintering may reflect the state of marginalized groups.

In addition to investigating each community as a unit, another important perspective is to study each individual user on their navigation among multiple communities. Although it is common that multiple communities coexist, previous studies have mostly focused on user behavior in a single community. As a result, basic questions remain unanswered: do users keep exploring new communities or do they stay in a relatively stable set of communities?

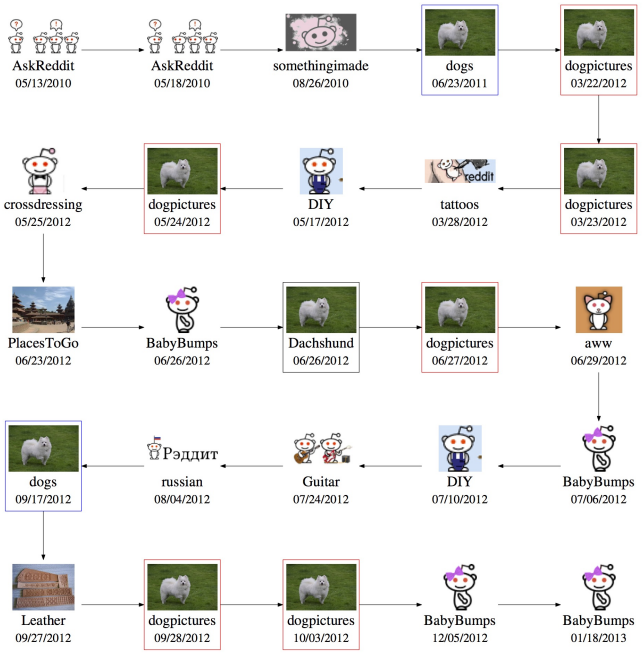


Figure 1.2: The first 25 communities to which an example user posted.

Now many multi-community sites, such as Reddit, 4chan, Wikia and Facebook groups, provide researchers with the opportunity to answer such questions. Figure 1.2 presents the first 25 communities to which an example user posted. It shows that this particular user wandered among many different communities. Such life trajectories of users moving between communities open up interesting questions: what is the relationship between different communities that a user explores? Will a user stop exploration or fail to adapt to new culture as they become senior members in communities? Finally, is wandering behavior related to a user's future activity levels in communities? An understanding of these issues can offer insights for service providers to design and improve socio-technical systems.

1.3 Organization and contributions

In order to study language and multiple communities, we have conducted a series of studies and they are organized into four chapters (we use tags for each chapter):

- (Chapter 2, language, message propagation, twitter). As illustrated in Section 1.1, we aim to identify the effect of wording, because wording is one of the few factors directly under an author's control, in contrast with identity, status, topic, etc. Although we cannot create a parallel universe, we are able to conduct natural experiments on Twitter thanks to a surprising observation that it is common for the same user to post different tweets containing the same URL.[1] We perform a battery of experiments to seek generally-applicable, non-Twitter-specific features of more successful phrasings. We investigate the utility of features like informativeness, resemblance to headlines, and conformity to the community norm in language use. We develop a classifier that outperforms average humans in distinguishing the tweet that was retweeted more.
- (Chapter 3, language, arguments, ChangeMyView). Changing someone's opinion is arguably one of the most important challenges in social interaction. The underlying process proves difficult to study: it is hard to know how someone's opinions are formed and whether and how someone's views are shifted. Fortunately, ChangeMyView, an active community on Reddit, provides a platform where users present their own opinions and reasoning, invite others to contest them, and acknowledge when the ensuing discussions change their original views. This allows us to examine the

[1] When this study was done, Twitter feeds were in reverse chronological order.

effect of phrasing in winning arguments. We show that language factors play an essential role. In particular, the interplay between the language of the opinion holder and that of the counterargument provides highly predictive cues of persuasiveness. We also investigate how interaction mechanisms in ChangeMyView other than wording choices affect the success of an argument, which demonstrates the complexity of real world scenarios.

- (Chapter 4, multi-community, life trajectory). Although analyzing user behavior within individual communities is an active and rich research domain, people usually interact with multiple communities both on- and offline. Now as large social-media platforms allow users to easily form and self-organize into interest groups or communities, massive datasets are available to study multi-community engagement. We examine three aspects of multi-community engagement: the sequence of communities that users post to, the language that users employ in those communities, and the feedback that users receive, using longitudinal posting behavior on Reddit as our main data source, and DBLP for auxiliary experiments. We find that in contrast with the "getting old and setting down" hypothesis, users continually explore new communities, and users that end up leaving the communities explore less. We also demonstrate the effectiveness of the features drawn from the wandering trajectories in predicting users' future level of activity.

- (Chapter 5, highly related communities, science, BadScience). In a world of communities, highly related communities can arise for many reasons. Two communities both focusing on wallpapers may have been created without knowing each other; a subgroup may become independent for the purpose of concentrated discussions, e.g., cooking vs. baking; as in the

famous study about the fission of a Karate club (Zachary, 1977), a community may split into two because of a conflict. We investigate the relationships between highly related communities using data from reddit.com consisting of 975M posts and comments spanning an 8-year period. We identify a set of typical affixes that users adopt to create highly related communities and build a taxonomy of affixes. One interesting finding regarding users' behavior is: for several types of highly-related community pairs, after a newer community is created, users that engage in the newer community tend to be more active in their original community than users that do not explore, even when controlling for previous level of engagement.

In Chapter 6, I conclude my thesis with thoughts on future work.

CHAPTER 2

WORDING MATTERS: MESSAGE PROPAGATION

2.1 Brief overview

Consider a person trying to spread an important message on a social network. He/she can spend hours trying to craft the message. Does it actually matter? While there has been extensive prior work looking into predicting popularity of social-media content, the effect of wording per se has rarely been studied since it is often confounded with the popularity of the author and the topic. To control for these confounding factors, we take advantage of the surprising fact that there are many pairs of tweets containing the *same* url and written by the *same* user but employing different wording. Given such pairs, we ask: which version attracts more retweets? This turns out to be a more difficult task than predicting popular topics. Still, humans can answer this question better than chance (but far from perfectly), and the computational methods we develop can do better than both an average human and a strong competing method trained on non-controlled data.

Most of the contents in this chapter are published in Tan et al. (2014). It is joint work with Lillian Lee and Bo Pang.

2.2 Introduction

How does one make a message "successful"? This question is of interest to many entities, including political parties trying to *frame* an issue (Chong and

Druckman, 2007), and individuals attempting to make a point in a group meeting. In the first case, an important type of success is achieved if the national conversation adopts the rhetoric of the party; in the latter case, if other group members repeat the originating individual's point.

The massive availability of online messages, such as posts to social media, now affords researchers new means to investigate at a very large scale the factors affecting message propagation, also known as adoption, sharing, spread, or virality. According to prior research, important features include characteristics of the originating author (e.g., verified Twitter user or not, author's messages' past success rate), the author's social network (e.g., number of followers), message timing, and message content or topic (Artzi et al., 2012; Bakshy et al., 2011; Borghol et al., 2012; Guerini et al., 2011, 2012; Hansen et al., 2011; Hong et al., 2011; Lakkaraju et al., 2013; Milkman and Berger, 2012; Ma et al., 2012; Petrović et al., 2011; Romero et al., 2013; Suh et al., 2010; Sun et al., 2013; Tsur and Rappoport, 2012). Indeed, it's not surprising that one of the most retweeted tweets of all time was from user BarackObama, with 40M followers, on November 6, 2012: "Four more years. [link to photo]".

Our interest in this work is the effect of alternative message *wording*, meaning *how* the message is said, rather than what the message is about. In contrast to the identity/social/timing/topic features mentioned above, wording is one of the few factors directly under an author's control when he or she seeks to convey a **fixed** piece of content. For example, consider a speaker at the ACL business meeting who has been tasked with proposing that Paris be the next ACL location. This person cannot on the spot become ACL president, change the shape of his/her social network, wait until the next morning to speak, or

Table 2.1: Topic- and author-controlled (TAC) pairs. Topic control = inclusion of the same URL. n_i gives the number of the retweets for t_i.

author	tweets	#retweets
natlsecuritycnn	t_1: FIRST ON CNN: After Petraeus scandal, Paula Broadwell looks to recapture 'normal life.' **http://t.co/qy7GGuYW**	$n_1 = 5$
	t_2: First on CNN: Broadwell photos shared with Security Clearance as she and her family fight media portrayal of her **[same URL]**	$n_2 = 29$
ABC	t_1: Workers, families take stand against Thanksgiving hours: **http://t.co/J9mQHiIEqv**	$n_1 = 46$
	t_2: Staples, Medieval Times Workers Say Opening Thanksgiving Day Crosses the Line **[same URL]**	$n_2 = 27$
cactus_music	t_1: I know at some point you've have been saved from hunger by our rolling food trucks friends. Let's help support them! **http://t.co/zg9jwA5j**	$n_1 = 2$
	t_2: Food trucks are the epitome of small independently owned LOCAL businesses! Help keep them going! Sign the petition **[same URL]**	$n_2 = 13$

campaign for Rome instead; but he/she can craft the message to be more humorous, more informative, emphasize certain aspects instead of others, and so on. In other words, we investigate whether a different choice of words affects message propagation, *controlling for user and topic*: would user BarackObama have gotten significantly more (or fewer) retweets if he had used some alternate wording to announce his re-election?

Although we cannot create a parallel universe in which BarackObama tweeted something else[1], fortunately, a surprising characteristic of Twitter allows us to run a fairly analogous *natural experiment*: external forces serendipitously provide an environment that resembles the desired controlled setting (DiNardo, 2008). Specifically, *it turns out to be unexpectedly common for the same user to post different tweets regarding the same URL* — a good proxy for fine-grained topic[2] — within a relatively short period of time.[3] Some example pairs are

[1]Cf. the Music Lab "multiple universes" experiment to test the randomness of popularity (Salganik et al., 2006).

[2]Although hashtags have been used as coarse-grained topic labels in prior work, for our purposes, we have no assurance that two tweets both using, say, "#Tahrir" would be attempting to express the same message but in different words. In contrast, see the same-URL examples in Table 2.1.

[3]Moreover, Twitter presents tweets to a reader in strict chronological order, so that there are no algorithmic-ranking effects to compensate for in determining whether readers saw a tweet.

shown in Table 2.1; we see that the paired tweets may differ dramatically, going far beyond word-for-word substitutions, so that quite interesting changes can be studied.

Looking at these examples, can one in fact tell from the wording which tweet in a topic- and author-controlled pair will be more successful? The answer may not be a priori clear. For example, for the first pair in the table, one person we asked found t_1's invocation of a "scandal" to be more attention-grabbing; but another person preferred t_2 because it is more informative about the URL's content and includes "fight media portrayal". In an Amazon Mechanical Turk (AMT) experiment (§2.5), we found that humans achieved an average accuracy of 61.3%: not that high, but better than chance, indicating that it is somewhat possible for humans to predict greater message spread from different deliveries of the same information.

Buoyed by the evidence of our AMT study that wording effects exist, we then performed a battery of experiments to seek generally-applicable, non-Twitter-specific features of more successful phrasings. §2.6.1 applies hypothesis testing (with Bonferroni correction to ameliorate issues with multiple comparisons) to investigate the utility of features like informativeness, resemblance to headlines, and conformity to the community norm in language use. §2.6.2 further validates our findings via prediction experiments, including on completely fresh held-out data, used only once and after an array of standard cross-validation experiments.[4] We achieved 66.5% cross-validation accuracy and 65.6% held-out accuracy with a combination of our custom features and bag-

And, Twitter accumulates retweet counts for the entire retweet cascade and displays them for the original tweet at the root of the propagation tree, so we can directly use Twitter's retweet counts to compare the entire reach of the different versions.

[4] And after crossing our fingers.

of-words. Our classifier fared significantly better than a number of baselines, including a strong classifier trained on the most- and least-retweeted tweets that was even granted access to author and timing metadata.

2.3 Related work

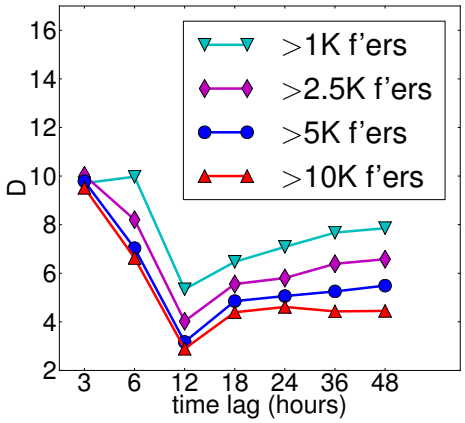
The idea of using carefully controlled experiments to study effective communication strategies dates back at least to Hovland et al. (1953). Recent studies range from examining what characteristics of *New York Times* articles correlate with high re-sharing rates (Milkman and Berger, 2012) to looking at how differences in description affect the spread of content-controlled videos or images (Borghol et al., 2012; Lakkaraju et al., 2013). (Simmons et al., 2011) examined the variation of quotes from different sources to examine how textual memes mutate as people pass them along, but did not control for author. Predicting the "success" of various texts such as novels and movie quotes has been the aim of additional prior work not already mentioned in §2.2 (Ashok et al., 2013; Louis and Nenkova, 2013; Danescu-Niculescu-Mizil et al., 2012a; Pitler and Nenkova, 2008; McIntyre and Lapata, 2009). There have been few large-scale studies exploring wording effects in a both topic- and author-controlled setting. Employing such controls, we find that predicting the more effective alternative wording is much harder than the previously well-studied problem of predicting popular content when author or topic can freely vary.

Related work regarding the features we considered is deferred to §2.6.1 (features description). Follow-up studies have explored factors such as phonetics (Guerini et al., 2015).

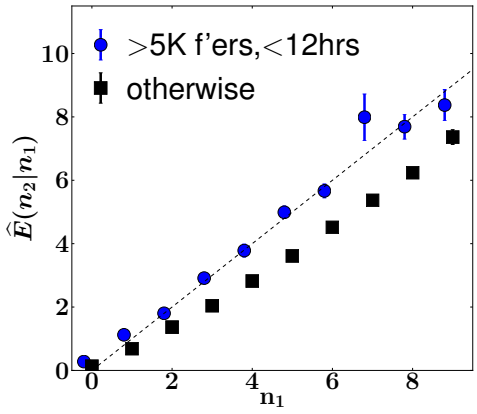
2.4 Data

Our main dataset was constructed by first gathering 1.77M topic- and author-controlled (henceforth *TAC*) tweet pairs differing in more than just spacing. The total excludes: tweets containing multiple URLs; tweets from users posting about the same URL more than five times (since such users might be spammers); the third, fourth, or fifth version for users posting between three and five tweets for the same URL; retweets (as identified by Twitter's API or by beginning with "RT @"); non-English tweets. We accomplished this by crawling timelines of 236K user ids that appear in prior work (Kwak et al., 2010; Yang and Leskovec, 2011) via the Twitter API. This crawling process also yielded 632K TAC pairs whose only difference was spacing, and an additional 558M "unpaired" tweets; as shown later in this work, we used these extra corpora for computing language models and other auxiliary information. We applied non-obvious but important filtering — described later in this section — to control for other external factors and to reduce ambiguous cases. This brought us to a set of 11,404 pairs, with the *gold-standard* labels determined by which tweet in each pair was the one that received more retweets according to the Twitter API. We then did a second crawl to get an additional 1,770 pairs to serve as a held-out dataset. The corresponding tweet IDs are available online at `http://chenhaot.com/pages/wording-for-propagation.html`. (Twitter's terms of service prohibit sharing the actual tweets.)

Throughout, we refer to the textual content of the earlier tweet within a TAC pair as t_1, and of the later one as t_2. We denote the number of retweets received by each tweet by n_1 and n_2, respectively. We refer to the tweet with higher (lower) n_i as the "better (worse)" tweet.



(a) For *identical* TAC pairs, retweet-count deviation vs. time lag between t_1 and t_2, for the author follower-counts given in the legend.



(b) Avg. n_2 vs. n_1 for identical TAC pairs, highlighting our chosen time-lag and follower thresholds. Bars: standard error. Diagonal line: $\widehat{E}(n_2|n_1) = n_1$.

Figure 2.1: (a): The ideal case where $n_2 = n_1$ when $t_1 = t_2$ is best approximated when t_2 occurs within 12 hours of t_1 and the author has at least 10,000 or 5,000 followers. (b): in our chosen setting (blue circles), n_2 indeed tends to track n_1, whereas otherwise (black squares), there's a bias towards retweeting t_1.

Using "identical" pairs to determine how to compensate for follower-count and timing effects. In an ideal setting, differences between n_1 and n_2 would be determined solely by differences in wording. But even with a TAC pair, retweets might exhibit a temporal bias because of the chronological order of tweet presentation (t_1 might enjoy a first-mover advantage (Borghol et al., 2012) because it is the "original"; alternatively, t_2 might be preferred because retweeters consider t_1 to be "stale"). Also, the number of followers an author has can have complicated indirect effects on which tweets are read.

We use the 632K TAC pairs wherein t_1 and t_2 are *identical*[5] to check for such confounding effects: we see how much n_2 deviates from n_1 in such settings, since if wording were the only explanatory factor, the retweet rates for identical

[5]Identical up to spacing: Twitter prevents exact copies by the same author appearing within a short amount of time, but some authors work around this by inserting spaces.

tweets ought to be equal. Figure 2.1a plots how the time lag between t_1 and t_2 and the author's follower-count affect the following deviation estimate:

$$D = \sum_{0 \le n_1 < 10} |\widehat{E}(n_2|n_1) - n_1|,$$

where $\widehat{E}(n_2|n_1)$ is the average value of n_2 over pairs whose t_1 is retweeted n_1 times. (Note that the number of pairs whose t_1 is retweeted n_1 times decays exponentially with n_1; hence, we condition on n_1 to keep the estimate from being dominated by pairs with $n_1 = 0$, and do not consider $n_1 \geq 10$ because there are too few such pairs to estimate $\widehat{E}(n_2|n_1)$ reliably.) Figure 2.1a shows that the setting where we (i) minimize the confounding effects of time lag and author's follower-count and (ii) maximize the amount of data to work with is: when t_2 occurs within 12 hours after t_1 and the author has more than 5,000 followers. Figure 2.1b confirms that for identical TAC pairs, our chosen setting indeed results in n_2 being on average close to n_1, which corresponds to the desired setting where wording is the dominant differentiating factor.[6]

Focus on meaningful and general changes. Even after follower-count and time-lapse filtering, we still want to focus on TAC pairs that (i) exhibit significant/interesting textual changes (as exemplified in Table 2.1, and as opposed to typo corrections and the like), and (ii) have n_2 and n_1 sufficiently different so that we are confident in which t_i is better at attracting retweets. To take care of (i), we discarded the 50% of pairs whose similarity was above the median, where similarity was tf-based cosine.[7] For (ii), we sorted the remaining pairs by

[6] We also computed the Pearson correlation between n_1 and n_2, even though it can be dominated by pairs with smaller n_1. The correlation is 0.853 for "> 5K f'ers, <12hrs", clearly higher than the 0.305 correlation for "otherwise".

[7] Idf weighting was not employed because changes to frequent words are of potential interest. Urls, hashtags, @-mentions and numbers were normalized to [url], [hashtag], [at], and [num] before computing similarity.

$n_2 - n_1$ and retained only the top and bottom 5%.[8] Moreover, to ensure that we do not overfit to the idiosyncrasies of particular authors, we cap the number of pairs contributed by each author to 50 before we deal with (ii).

2.5 Human accuracy on TAC pairs

We first ran a pilot study on Amazon Mechanical Turk (AMT) to determine whether humans can identify, based on wording differences alone, which of two topic- and author- controlled tweets is spread more widely. Each of our 5 AMT tasks involved a disjoint set of 20 randomly-sampled TAC pairs (with t_1 and t_2 randomly reordered); subjects indicated "which tweet would other people be more likely to retweet?", provided a short justification for their binary response, and clicked a checkbox if they found that their choice was a "close call". We received 39 judgments per pair in aggregate from 106 subjects total (9 people completed all 5 tasks). The subjects' justifications were of very high quality, convincing us that they all did the task in good faith[9]. Two examples for the third TAC pair in Table 2.1 were: "[t_1 makes] the cause relate-able to some people, therefore showing more of an appeal as to why should they click the link and support" and, expressing the opposite view, "I like [t_2] more because [t_1] starts out with a generalization that doesn't affect me and try to make me look like I had that experience before".

If we view the set of 3900 binary judgments for our 100-TAC-pair sample as

[8] For our data, this meant $n_2 - n_1 \geq 10$ or ≤ -15. Cf. our median number of retweets: 30.

[9] We also note that the feedback we got was quite positive, including: "...It's fun to make choices between close tweets and use our subjective opinion. Thanks and best of luck with your research" and "This was very interesting and really made me think about how I word my own tweets. Great job on this survey!". We only had to exclude one person (not counted among the 106 subjects), doing so because he or she gave the same uninformative justification for all pairs.

constituting independent responses, then the accuracy for this set is 62.4% (rising to 63.8% if we exclude the 587 judgments deemed "close calls"). However, if we evaluate the accuracy of the *majority* response among the 39 judgments per pair, the number rises to 73%. The accuracy of the majority response generally increases with the dominance of the majority, going above 90% when at least 80% of the judgments agree (although less than a third of the pairs satisfied this criterion).

Alternatively, we can consider the average accuracy of the 106 subjects: 61.3%, which is better than chance but far from 100%. (Variance was high: one subject achieved 85% accuracy out of 20 pairs, but eight scored below 50%.) This result is noticeably lower than the 73.8%-81.2% reported by (Petrović et al., 2011), who ran a similar experiment involving two subjects and 202 tweet pairs, but where the pairs were *not* topic- or author-controlled.[10]

We conclude that even though propagation prediction becomes more challenging when topic and author controls are applied, humans can still to some degree tell which wording attracts more retweets. Interested readers can try this out themselves at `http://chenhaot.com/retweetedmore/quiz`.

2.6 Experiments

We now investigate computationally what wording features correspond to messages achieving a broader reach. We start (§2.6.1) by introducing a set of generally-applicable and (mostly) non-Twitter-specific features to capture our

[10]The accuracy range stems from whether author's social features were supplied and which subject was considered.

Table 2.2: Notational conventions for tables in §2.6.1.

One-sided paired t-test for feature efficacy		*One-sided binomial test for feature increase (Do authors prefer to 'raise' the feature in t_2?)*
↑↑↑↑: p<1e-20	↓↓↓↓: p>1-1e-20	YES : t_2 has a higher feature score than t_1, α = .05
↑↑↑ : p<0.001	↓↓↓ : p>0.999	NO : t_2 has a lower feature score than t_1, α = .05
↑↑ : p<0.01	↓↓ : p>0.99	(x%): $\%(f_2 > f_1)$, if sig. larger or smaller than 50%
↑ : p<0.05	↓ : p>0.95	
∗: passes our Bonferroni correction		

intuitions about what might be better ways to phrase a message. We then use hypothesis testing (§2.6.1) to evaluate the importance of each feature for message propagation and the extent to which authors employ it, followed by experiments on a prediction task (§2.6.2) to further examine the utility of these features.

2.6.1 Features: efficacy and author preference

What kind of phrasing helps message propagation? Does it work to explicitly ask people to share the message? Is it better to be short and concise or long and informative? We define an array of features to capture these and other messaging aspects. We then examine (i) how effective each feature is for attracting more retweets; and (ii) whether authors prefer applying a given feature when issuing a second version of a tweet.

First, for each feature, we use a one-sided paired t-test to test whether, on our 11K TAC pairs, our score function for that feature is larger in the better tweet versions than in the worse tweet versions, for significance levels α = .05, .01, .001, 1e-20. Given that we did 39 tests in total, there is a risk of obtaining

Table 2.3: Explicit requests for sharing (where only occurrences POS-tagged as verbs count, according to the (Gimpel et al., 2011) tagger).

	effective?	author-preferred?
rt	↑↑↑↑ *	——
retweet	↑↑↑↑ *	YES (59%)
spread	↑↑↑ *	YES (56%)
please	↑↑↑ *	——
pls	↑	——
plz	↑↑	——

false positives due to multiple testing (Dunn, 1961; Benjamini and Hochberg, 1995). To account for this, we also report significance results for the conservatively Bonferroni-corrected ("BC") significance level α = 0.05/39=1.28e-3.

Second, we examine author preference for applying a feature. We do so because one (but by no means the only) reason authors post t_2 after having already advertised the same URL in t_1 is that these authors were dissatisfied with the amount of attention t_1 got; in such cases, the changes may have been specifically intended to attract more retweets. We measure author preference for a feature by the percentage of our TAC pairs[11] where t_2 has more "occurrences" of the feature than t_1, which we denote by "$\%(f_2 > f_1)$". We use the one-sided binomial test to see whether $\%(f_2 > f_1)$ is significantly larger (or smaller) than 50%.

Not surprisingly, it helps to ask people to share. (See Table 2.3; the notation for all tables is explained in Table 2.2.) The basic sanity check we performed here was to take as features the number of occurrences of the verbs 'rt', 'retweet', 'please', 'spread', 'pls', and 'plz' to capture explicit requests (e.g. "please retweet").

[11] For our preference experiments, we added in pairs where $n_2 - n_1$ was not in the top or bottom 5% (cf. §2.4, meaningful changes), since to measure author preference it's not necessary that the retweet counts differ significantly.

Table 2.4: Informativeness.

	effective?	author-preferred?
verb	↑↑↑↑ *	YES (56%)
noun	↑↑↑↑ *	——
adjective	↑↑↑ *	YES (51%)
adverb	↑↑↑ *	YES (55%)
proper noun	↑↑↑ *	NO (45%)
number	↑↑↑↑ *	NO (48%)
hashtag	↑	——

Informativeness helps. (Table 2.4) Messages that are more informative have increased *social exchange value* (Homans, 1958), and so may be more worth propagating. One crude approximation of informativeness is length, and we see that length helps.[12] In contrast, (Simmons et al., 2011) found that shorter versions of memes[13] are more likely to be popular. The difference may result from TAC-pair changes being more drastic than the variations that memes undergo.

A more refined informativeness measure is counts of the parts of speech that correspond to content. Our POS results, gathered using a Twitter-specific tagger (Gimpel et al., 2011), echo those of Ashok et al. (2013) who looked at predicting the success of books. The diminished effect of hashtag inclusion with respect to what has been reported previously (Suh et al., 2010; Petrović et al., 2011) presumably stems from our topic and author controls.

Be like the community, and be true to yourself (in the words you pick, but not necessarily in how you combine them). (Table 2.5) Although distinctive messages may attract attention, messages that conform to expectations might be more easily accepted and therefore shared. Prior work has explored this tension: Lakkaraju et al. (2013), in a content-controlled study, found that the

[12]Of course, simply inserting garbage isn't going to lead to more retweets, but adding more information generally involves longer text.

[13]Memes represent short quoted texts that act as signature of topics or events.

Table 2.5: Conformity to the community and one's own past, measured via scores assigned by various language models.

	effective?	author-preferred?
twitter unigram	↑↑↑ *	YES (54%)
twitter bigram	↑↑↑ *	YES (52%)
personal unigram	↑↑↑ *	YES (52%)
personal bigram	———	NO (48%)

more up-voted Reddit image titles balance novelty and familiarity; Danescu-Niculescu-Mizil et al. (2012a) (henceforth DCKL'12) showed that the memorability of movie quotes corresponds to higher lexical distinctiveness but lower POS distinctiveness; and Sun et al. (2013) observed that deviating from one's own past language patterns correlates with more retweets.

Keeping in mind that the authors in our data have at least 5000 followers[14], we consider two types of language-conformity constraints an author might try to satisfy: to be similar to what is normal in the Twitter community, and to be similar to what his or her followers expect. We measure a tweet's similarity to expectations by its score according to the relevant language model, $\frac{1}{|T|}\sum_{x\in T}\log(p(x))$, where T refers to either all the unigrams (unigram model) or all and only bigrams (bigram model).[15] We trained a Twitter-community language model from our 558M unpaired tweets, and personal language models from each author's tweet history.

Imitate headlines. (Table 2.6) News headlines are often intentionally written to be both informative and attention-getting, so we introduce the idea of scoring by a language model built from New York Times headlines.[16]

[14]This is not an artificial restriction on our set of authors; a large follower count means (in principle) that our results draw on a large sample of decisions whether to retweet or not.

[15]The tokens [at], [hashtag], [url] were ignored in the unigram-model case to prevent their undue influence, but retained in the bigram model to capture longer-range usage ("combination") patterns.

[16] To test whether the results stem from similarity to *news* rather than headlines per se, we

Table 2.6: LM-based resemblance to headlines.

	effective?	author-preferred?
headline unigram	↑↑	YES (53%)
headline bigram	↑↑↑↑ *	YES (52%)

Table 2.7: Retweet score.

	effective?	author-preferred?
rt score	↑↑ *	NO (49%)
verb rt score	↑↑↑↑ *	——
noun rt score	↑↑↑ *	——
adjective rt score	↑	YES (50%)
adverb rt score	↑	YES (51%)
proper noun rt score	———	NO (48%)

Use words associated with (non-paired) retweeted tweets. (Table 2.7) We expect that provocative or sensationalistic tweets are likely to make people react. We found it difficult to model provocativeness directly. As a rough approximation, we check whether the changes in t_2 with respect to t_1 (which share the same topic and author) involve words or parts-of-speech that are associated with high retweet rate in a very large separate sample of unpaired tweets (retweets and replies discarded). Specifically, for each word w that appears more than 10 times, we compute the probability that tweets containing w are retweeted more than once, denoted by $rs(w)$. We define the *rt score* of a tweet as $max_{w \in T} rs(w)$, where T is all the words in the tweet, and the *rt score* of a particular POS tag z in a tweet as $max_{w \in T \& \text{tag}(w)=z} rs(w)$.

Include positive and/or negative words. (Table 2.8) Prior work has found that including positive or negative sentiment increases message propagation (Milkman and Berger, 2012; Godes et al., 2005; Heath et al., 2001; Hansen et al., 2011). We measured the occurrence of positive and negative words as determined by

constructed a NYT-text LM, which proved less effective. We also tried using Gawker headlines (often said to be attention-getting) but pilot studies revealed insufficient vocabulary overlap with our TAC pairs.

Table 2.8: Sentiment (contrast is measured by presence of both positive and negative sentiments).

	effective?	author-preferred?
positive	↑↑↑ *	——
negative	↑↑↑ *	——
contrast	↑↑↑ *	——

Table 2.9: Pronouns.

	effective?	author-preferred?
1st person singular	———	YES (51%)
1st person plural	———	YES (52%)
2nd person	———	YES (57%)
3rd person singular	↑↑	YES (55%)
3rd person plural	↑	YES (58%)

the connotation lexicon of (Feng et al., 2013) (better coverage than LIWC (Pennebaker et al., 2007), a commonly used lexicon dictionary). Measuring the occurrence of both *simultaneously* was inspired by Riloff et al. (2013).

Refer to other people (but not your audience). (Table 2.9) First-person has been found useful for success before, but in the different domains of scientific abstracts (Guerini et al., 2012) and books (Ashok et al., 2013). In contrast with prior studies, we find that tweets that were retweeted more correlate with more third-person pronouns.

Generality helps. (Table 2.10) DCKL'12 posited that movie quotes are more shared in the culture when they are general enough to be used in multiple contexts. We hence measured the presence of indefinite articles vs. definite articles.

The easier to read, the better. (Table 2.11) We measure readability by using Flesch reading ease (Flesch, 1948) and Flesch-Kincaid grade level (Kincaid et al., 1975), though they are not designed for short texts. We use negative grade level so that a larger value indicates easier texts to read.

Table 2.10: Generality.

	effective?	author-preferred?
indefinite articles (a,an)	↑↑↑ *	——
definite articles (the)	———	YES (52%)

Table 2.11: Readability.

	effective?	author-preferred?
reading ease	↑↑	YES (52%)
negative grade level	↑	YES (52%)

Final question: Do authors prefer to do what is effective? Recall that we use binomial tests to determine author preference for applying a feature more in t_2. Our preference statistics show that author preferences in many cases are aligned with feature efficacy. But there are several notable exceptions: for example, authors tend to increase the use of @-mentions and 2nd person pronouns even though they are ineffective. On the other hand, they did not increase the use of effective ones like proper nouns and numbers; nor did they tend to increase their rate of sentiment-bearing words. Bearing in mind that changes in t_2 may not always be intended as an effort to improve t_1, it is still interesting to observe that there are some contrasts between feature efficacy and author preferences.

2.6.2 Predicting the "better" wording

Here, we further examine the collective efficacy of the features introduced in §2.6.1 via their performance on a binary prediction task: given a TAC pair (t_1, t_2), did t_2 receive more retweets?

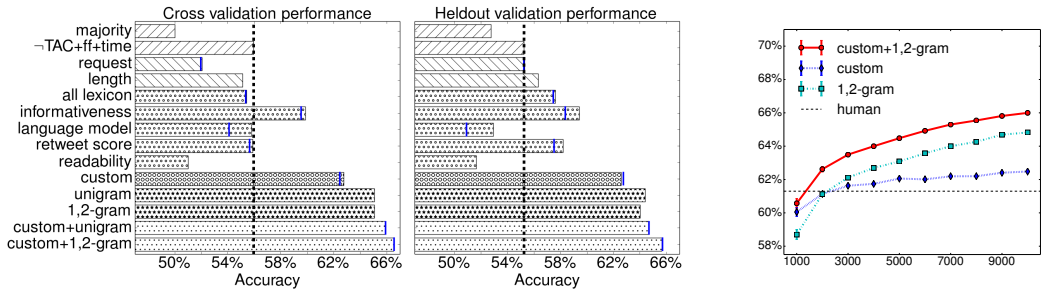
Our approach. We group the features introduced in §2.6.1 into 16 lexicon-based features (Table 2.3, 2.8, 2.9, 2.10), 9 informativeness features (Table 2.4), 6 lan-

guage model features (Table 2.5, 2.6), 6 rt score features (Table 2.7), and 2 readability features (Table 2.11). We refer to all 39 of them together as *custom* features. We also consider tagged bag-of-words ("BOW") features, which includes all the unigram (word:POS pair) and bigram features that appear more than 10 times in the cross-validation data. This yields 3,568 unigram features and 4,095 bigram features, for a total of 7,663 so-called *1,2-gram features*. Values for each feature are normalized by linear transformation across all tweets in the training data to lie in the range [0, 1].[17]

For a given TAC pair, we construct its feature vector as follows. For each feature being considered, we normalize its value to [0, 1] for each tweet in the pair and take the difference as the feature value for this pair. We use L2-regularized logistic regression as our classifier, with parameters chosen by cross validation on the training data. (We also experimented with SVMs. The performance was very close, but mostly slightly lower.)

A strong non-TAC alternative, with social information and timing thrown in. One baseline result we would like to establish is whether the topic and author controls we have argued for, while intuitively compelling for the purposes of trying to determine the best way for a given author to present some fixed content, are really necessary in practice. To test this, we consider an alternative binary L2-regularized logistic-regression classifier that is trained on unpaired data, specifically, on the collection of 10,000 most retweeted tweets (gold-standard label: positive) plus the 10,000 least retweeted tweets (gold-standard label: negative) that are neither retweets nor replies. Note that this alternative thus is granted, by design, roughly *twice* the training instances that our clas-

[17] We also tried normalization by *whitening* (Krizhevsky, 2009), but it did not lead to further improvements.



(a) Cross-validation and heldout accuracy for various feature sets. Blue lines inside bars: performance when custom features are restricted to those that pass our Bonferroni correction (no line for readability because no readability features passed). Dashed vertical line: ¬TAC+ff+time performance.

(b) Cross-validation accuracy vs data size. Human performance was estimated from a disjoint set of 100 pairs (see §2.5).

Figure 2.2: Accuracy results. Pertinent significance results are as follows. In cross-validation, custom+1,2-gram is significantly better than ¬TAC+ff+time (p=0) and 1,2-gram (p=3.8e-7). In heldout validation, custom+1,2-gram is significantly better than ¬TAC+ff+time (p=3.4e-12) and 1,2-gram (p=0.01) but not unigram (p=0.08), perhaps due to the small size of the heldout set.

sifiers have, as a result of having roughly the same number of tweets, since our instances are pairs. Moreover, we additionally include the tweet author's follower count, and the day and hour of posting, as features. We refer to this alternative classifier as *¬TAC+ff+time*. (Mnemonic: "ff" is used in bibliographic contexts as an abbreviation for "and the following".) We apply it to a tweet pair by computing whether it gives a higher score to t_2 or not.

Baselines. To sanity-check whether our classifier provides any improvement over the simplest methods one could try, we also report the performance of the majority baseline, our request-for-sharing features, and our character-length feature.

Performance comparison. We compare the accuracy (percentage of pairs whose labels were correctly predicted) of our approach against the competing methods. We report 5-fold cross validation results on our balanced set of 11,404 TAC

pairs and on our completely disjoint heldout data[18] of 1,770 TAC pairs; this set was never examined during development, and there are no authors in common between the two testing sets.

Figure 2.2a summarizes the main results. While ¬TAC+ff+time outperforms the majority baseline, using all the features we proposed beats ¬TAC+ff+time by more than 10% in both cross-validation (66.5% vs 55.9%) and heldout validation (65.6% vs 55.3%). We outperform the average human accuracy of 61% reported in our Amazon Mechanical Turk experiments (for a different data sample); ¬TAC+ff+time fails to do so.

The importance of topic and author control can be seen by further investigation of ¬TAC+ff+time's performance. First, note that it yields an accuracy of around 55% on our alternate-version-selection task,[19] even though its cross-validation accuracy on the larger most- and least-retweeted unpaired tweets averages out to a high 98.8%. Furthermore, note the superior performance of unigrams trained on TAC data vs ¬TAC+ff+time — which is similar to our unigrams but trained on a larger but non-TAC dataset that included metadata. Thus, TAC pairs are a useful data source even for non-custom features. (We also include individual feature comparisons later.)

Informativeness is the best-performing custom feature group when run in isolation, and outperforms all baselines, as well as ¬TAC+ff+time; and we can

[18] To construct this data, we used the same criteria as in §2.4: written by authors with more than 5000 followers, posted within 12 hours, $n_2 - n_1 \geq 10$ or ≤ -15, and cosine similarity threshold value the same as in §2.4, cap of 50 on number of pairs from any individual author.

[19] One might suspect that the problem is that ¬TAC+ff+time learns from its training data to over-rely on follower-count, since that is presumably a good feature for non-TAC tweets, and for this reason suffers when run on TAC data where follower-counts are by construction non-informative. But in fact, we found that removing the follower-count feature from ¬TAC+ff+time and re-training did not lead to improved performance. Hence, it seems that it is the non-controlled nature of the alternate training data that explains the drop in performance.

see from Figure 2.2a that this is not due just to length. The combination of all our 39 custom features yields approximately 63% accuracy in both testing settings, significantly outperforming informativeness alone ($p<0.001$ in both cases). Again, this is higher than our estimate of average human performance.

Not surprisingly, the TAC-trained BOW features (unigram and 1,2-gram) show impressive predictive power in this task: many of our custom features can be captured by bag-of-word features, in a way. Still, the best performance is achieved by combining our custom and 1,2-gram features together, to a degree statistically significantly better than using 1,2-gram features alone.

Finally, we remark on our Bonferroni correction. Recall that the intent of applying it is to avoid false positives. However, in our case, Figure 2.2a shows that our potentially "false" positives — features whose effectiveness did not pass the Bonferroni correction test — actually do raise performance in our prediction tests.

Size of training data. Another interesting observation is how performance varies with data size. For $n = 1000, 2000, \ldots, 10000$, we randomly sampled n pairs from our 11,404 pairs, and computed the average cross-validation accuracy on the sampled data. Figure 2.2b shows the averages over 50 runs of the aforementioned procedure. Our custom features can achieve good performance with little data, in the sense that for sample size 1000, they outperform BOW features; on the other hand, BOW features quickly surpass them. Across the board, the custom+1,2-gram features are consistently better than the 1,2-gram features alone.

Top features. Finally, we examine some of the top-weighted individual fea-

tures from our approach and from the competing ¬TAC+ff+time classifier. The top three rows of Table 2.12 show the best custom and best and worst unigram features for our method; the bottom two rows show the best and worst unigrams for ¬TAC+ff+time. Among custom features, we see that community and personal language models, informativeness, retweet scores, sentiment, and generality are represented. As for unigram features, not surprisingly, "rt" and "retweet" are top features for both our approach and ¬TAC+ff+time. However, the other unigrams for the two methods seem to be a bit different in spirit. Some of the unigrams determined to be most poor only by our method appear to be both surprising and yet plausible in retrospect: "icymi" (abbreviation for "in case you missed it") tends to indicate a direct repetition of older information, so people might prefer to retweet the earlier version; "thanks" and "sorry" could correspond to personal thank-yous and apologies not meant to be shared with a broader audience, and similarly @-mentioning another user may indicate a tweet intended only for that person. The appearance of [hashtag] in the best ¬TAC+ff+time unigrams is consistent with prior research in non-TAC settings (Suh et al., 2010; Petrović et al., 2011).

2.7 Conclusion

In this work, we conducted the first large-scale topic- and author-controlled experiment to study the effects of wording on information propagation.

The features we developed to choose the better of two alternative wordings posted better performance than that of all our comparison algorithms, including one given access to author and timing features but trained on non-TAC data,

Table 2.12: Features with largest coefficients, delimited by commas. POS tags omitted for clarity.

Our approach	
best 15 custom	twitter bigram, length (chars), rt (the word), retweet (the word), verb, verb retweet score, personal unigram, proper noun, number, noun, positive words, please (the word), proper noun retweet score, indefinite articles (a,an), adjective
best 20 unigrams	rt, retweet, [num], breaking, is, win, never, ., people, need, official, officially, are, please, november, world, girl, !!!, god, new
worst 20 unigrams	:, [at], icymi, also, comments, half, ?, earlier, thanks, sorry, highlights, bit, point, update, last, helping, peek, what, haven't, debate
¬TAC+ff+time	
best 20 unigrams	[hashtag], teen, fans, retweet, sale, usa, women, butt, caught, visit, background, upcoming, rt, this, bieber, these, each, chat, houston, book
worst 20 unigrams	:, ..., boss, foundation, ?, ~, others, john, roll, ride, appreciate, page, drive, correct, full, ', looks, @ (not as [at]), sales, hurts

and also bested our estimate of average human performance. According to our hypothesis tests, helpful wording heuristics include adding more information, making one's language align with both community norms and with one's prior messages, and mimicking news headlines. Readers may try out their own alternate phrasings at http://chenhaot.com/retweetedmore/ to see what a simplified version of our classifier predicts.

In future work, it will be interesting to examine how these features generalize to longer and more extensive arguments. Moreover, understanding the underlying psychological and cultural mechanisms that establish the effectiveness of these features is a fundamental problem of interest.

CHAPTER 3

WORDING MATTERS: WINNING ARGUMENTS

3.1 Brief overview

Changing someone's opinion is arguably one of the most important challenges of social interaction. The underlying process proves difficult to study: it is hard to know how someone's opinions are formed and whether and how someone's views shift. Fortunately, ChangeMyView, an active community on Reddit, provides a platform where users present their own opinions and reasoning, invite others to contest them, and acknowledge when the ensuing discussions change their original views. In this chapter, we study these interactions to understand the mechanisms behind persuasion.

We find that persuasive arguments are characterized by interesting patterns of interaction dynamics, such as participant entry-order and degree of back-and-forth exchange. Furthermore, by comparing similar counterarguments to the same opinion, we show that language factors play an essential role. In particular, the interplay between the language of the opinion holder and that of the counterargument provides highly predictive cues of persuasiveness. Finally, since even in this favorable setting people may not be persuaded, we investigate the problem of determining whether someone's opinion is susceptible to being changed at all. For this more difficult task, we show that stylistic choices in how the opinion is expressed carry predictive power.

Most contents of this chapter are published in Tan et al. (2016). This is joint work with Vlad Niculae, Cristian Danescu-Niculescu-Mizil and Lillian Lee.

3.2 Introduction

Changing a person's opinion is a common goal in many settings, ranging from political or marketing campaigns to friendly or professional conversations. The importance of this topic has long been acknowledged, leading to a tremendous amount of research effort (Cialdini, 1993; Dillard and Shen, 2014; Eagly and Chaiken, 1993; Petty and Cacioppo, 2012; Popkin, 1994; Reardon, 1991). Thanks to the increasing number of social interactions online, *interpersonal persuasion* has become observable at a massive scale (Fogg, 2008). This allows the study of interactive persuasion *in practice, without elicitation*, thus bypassing some limitations of laboratory experiments and leading to new research questions regarding dynamics in real discussions. At the same time, the lack of the degree of experimental control offered by lab trials raises new methodological challenges that we address in this work.

It is well-recognized that multiple factors are at play in persuasion. Beyond (*i*) the characteristics of the arguments themselves, such as intensity, valence and framing (Althoff et al., 2014; Bailey et al., 2014; Bryan et al., 2011; Burgoon et al., 1975; Hullett, 2005), and (*ii*) social aspects, such as social proof and authority (Chaiken, 1987; Cialdini et al., 1999; Mitra and Gilbert, 2014), there is also (*iii*) the relationship between the opinion holder and her belief, such as her certainty in it and its importance to her (Petty et al., 1997; Pomerantz et al., 1995; Tormala and Petty, 2002; Zuwerink and Devine, 1996). Thus, an ideal setting for the study of persuasion would allow access to the reasoning behind people's views in addition to the full interactions. Furthermore, the outcome of persuasion efforts (e.g., which efforts succeed) should be easy to extract.[1]

[1] One might think that the outcome is trivially "no one ever changes their mind", since people can be amazingly resistant to evidence contravening their beliefs (Chambliss and Garner,

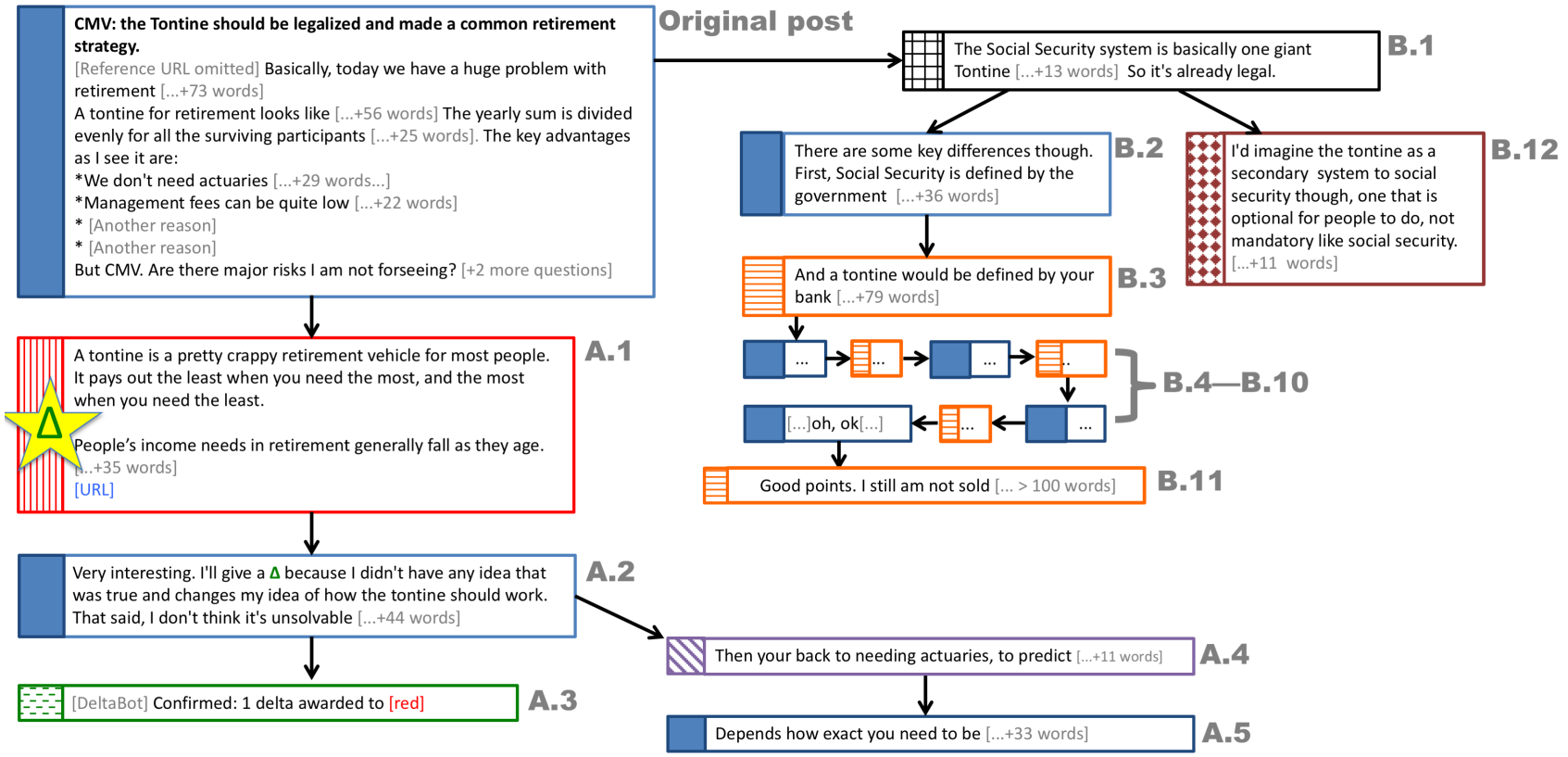


Figure 3.1: A fragment of a "typical" `/r/ChangeMyView` *discussion tree*—typical in the sense that the full discussion tree has an average number of replies (54), although we abbreviate or omit many of them for compactness and readability. Colors indicate distinct users. Of the 17 replies shown (in our terminology, every node except the original post is a reply), the OP explicitly acknowledged only one as having changed their view: the starred reply A.1. The explicit signal is the "Δ" character in reply A.2. (The full discussion tree is available at https://www.reddit.com/r/changemyview/comments/3mzc6u/cmv_the_tontine_should_be_legalized_and_made_a/.)

One forum satisfying these desiderata is the active Reddit subcommunity `/r/ChangeMyView` (henceforth CMV).[2] In contrast to general platforms such as Twitter and Facebook, CMV requires posters to state the reasoning behind their beliefs and to reward successful arguments with explicit confirmation. Moreover, discussion quality is monitored by moderators, and posters commit to an openness to changing their minds. The resulting conversations are of reasonably high quality, as demonstrated by Figure 3.1, showing the top portion of a discussion tree (an original post and all the replies to it) about legalizing the "tontine".[3] In the figure, Reply B.1 branches off to an extended back-and-forth between the blue original poster (OP) and the orange user; as it turns out, neither ends up yielding, although both remain polite. Reply A.1, on the other hand, is successful, as the OP acknowledges at A.2. The example suggests that content and phrasing play an important role (A.1 does well on both counts), but also that interaction factors may also correlate with persuasion success. Examples include time of entry relative to others and amount of engagement: the discussion at B.1 started earlier than that at A.1 and went on for longer.

Outline and highlight reel. This work provides three different perspectives on the mechanics of persuasion. First, we explore how interaction dynamics are associated with a successful change of someone's opinion (Section 3.4). We find (example above to the contrary) that a challenger that enters the fray before another tends to have a higher likelihood of changing the OP's opinion; this holds even for first-time CMV challengers, and so is not a trivial consequence of more

1996; McRaney, 2011; Nyhan and Reifler, 2010). But take heart, change does occur, as we shall show.

[2] `https://reddit.com/r/changemyview`

[3] It is not necessary for the reader to be familiar with tontines, but a brief summary is: a pool of money is maintained where the annual payouts are divided evenly among all participants still living.

experienced disputants contriving to strike first. Although engaging the OP in some back-and-forth is correlated with higher chances of success, we do not see much OP conversion in extended conversations. As for opinion conversion rates, we find that the more participants there are in the effort to persuade the OP, the larger the likelihood of the OP changing her view; but, interestingly, the relationship is sublinear.

Besides interaction dynamics, language is a powerful tool that is in the full control of the challengers. In Section 3.5 we explore this perspective by tackling the task of predicting which of two *similar* counterarguments will succeed in changing the same view. By comparing similar arguments we focus on the role of stylistic choices in the presentation of an argument (identifying reasoning strategies is a separate problem we do not address). We experiment with style features based solely on the counterargument, as well as with features reflecting the interplay between the counterargument and the way in which the view is expressed. Style features and interplay features both prove useful and outperform a strong baseline that uses bag-of-words. In particular, interplay features alone have strong predictive power, achieving an improvement of almost 5% in accuracy over the baseline method (65.1% vs 59.6%) in a *completely fresh* heldout dataset. Our results also show that it is useful to include links as evidence—an interesting contrast to studies of the *backfire effect*: "When your deepest convictions are challenged by contradictory evidence, your beliefs get stronger" (Chambliss and Garner, 1996; McRaney, 2011; Nyhan and Reifler, 2010). However, it hurts to be too intense in the counterargument. The feature with the most predictive power of successful persuasion is the dissimilarity with the original post in word usage, while existing theories mostly study matching in terms of attitude functions or subject self-discrepancy (Petty and Wegener,

1998; Tykocinski et al., 1994).

In the majority of cases, however, opinions are not changed, even though it takes courage and self-motivation for the original poster to post on CMV and invite other people to change her opinion. Can we tell whether the OP is unlikely to be persuaded from the way she presents her reasoning? In Section 3.6, we turn to this challenging task. In our pilot study, humans found this task quite difficult in a paired setting and performed no better than random guessing. While we can outperform the random baseline in a realistic imbalanced setting, the AUC score is only 0.54. Our feature analysis is consistent with existing theories on self-affirmation (Cohen et al., 2000; Correll et al., 2004) and shows that malleable beliefs are expressed using more self-confidence and more organization, in a less intense way.

While we believe that the observations we make are useful for understanding persuasion, we do not claim that any of them are causal explanations.

In Section 3.7, we discuss other observations that may open up future directions, including attempts to capture higher-level linguistic properties (e.g., semantics and argument structure); Section 3.8 summarizes additional related work and Section 3.9 concludes.

(OP) Title: I believe that you should be allowed to drive at whatever speed you wish as long as you aren't driving recklessly or under extenuating circumstances CMV.
I think that if you feel comfortable driving 80 mph or 40 mph you should be allowed to do so, as long as you aren't in a school or work zone, etc. because there are a lot more risks in those areas. I think when you're comfortable driving you will be a better driver, and if you aren't worrying about the speed limit or cops you are going to be more comfortable. However, I think that you should only be allowed to drive at whatever speed you wish as long as you aren't driving recklessly. If you're weaving in and out of traffic at 90, you probably shouldn't be allowed to go 90, but if you just stay in the fast lane and pass the occasional person I don't think there is a problem. CMV.

↓ ↓

(C1) Some issues with this:

1. Who's to say what is reckless driving? Where do you draw the line? Speed is the standard that ensures we know what is considered to be reckless. The idea of driving any speed you want creates a totally subjective law.
2. How do you judge whether to pass other drives and such? There are a lot of spatial awareness issues with the roads being so unpredictable.
3. How do you expect insurance and courts to work out who's at fault for an accident?

A: "Yeah this guy was going 100 mph!"

B: "But I wasn't driving recklessly - you were!"

It's simply not realistic and creates some serious legal issues.

(C2) They're many issues I have with this idea but I'll start with the most pressing one. Think of the amount of drivers you pass by every day. Imagine all of them going at whatever speed they choose. How would this work? You cannot have a driver going 35 and a driver who wants to go 65 in the same lane.

Now lets take this onto the highway and you can see how horrific this could get quickly. They're too many drivers out on the road for everyone to choose there own speed.

Speed limits protect us all because it gives us a reasonable expectation in whatever area we're driving in. Have you ever been on the highway being a driver going 40mph? If you're doing the speed limit (65) you catch up to them so fast you barely have time to react before an accident occurs. You aren't expecting this low speed when everyone is going at similar speeds to yours.

Drivers need to know the speed expectations so they can drive and react accordingly. If everyone goes at whatever speed they want it will only cause many many accidents.

Figure 3.2: An *original post* and a pair of *root replies* **C1** and **C2** contesting it, where **C1** and **C2** have relatively high vocabulary overlap with each other, but only one changed the OP's opinion. (Section 3.5 reveals which one.)

3.3 Dataset

We draw our data from the `/r/ChangeMyView` subreddit (CMV), which has over 211,000 subscribers to date. It is self-described[4] as "dedicated to the civil discourse [sic] of opinions". CMV is well-suited to our purposes because of its setup and mechanics, the high quality of its arguments, and the size and activity of its user base. We elaborate below.

The mechanics of the site are as follows. Users that "accept that they may be wrong or want help changing their view" submit *original posts*, and readers are invited to argue for the other side. The original posters (*OPs*) explicitly recognize arguments that succeed in changing their view by replying with the *delta* (Δ) character (an example is node A.2 in Figure 3.1) and including "an explanation as to why and how" their view changed. A Reddit bot called the DeltaBot confirms deltas (an example is A.3 in Figure 3.1) and maintains a leaderboard of per-user Δ counts.[5] The experimental advantages of this setup include:
(1) Multiple users make different attempts at changing the same person's mind on the same issue based on the same rationale, thus controlling for a number of variables but providing variation along other important aspects. Figure 3.2, for example, presents in full two counter-arguments, C1 and C2. They both respond to the same claims, but differ in style, structure, tone, and other respects.
(2) The deltas serve as explicit persuasion labels that are (2a) provided by the actual participants and (2b) at the fine-grained level of individual arguments, as opposed to mere indications that the OP's view was changed.

[4] Quotations here are from the CMV wiki.

[5] Although non-OPs can also issue deltas, in this work, we only count deltas given by a user in their OP role. A consequence is that we only consider discussion trees where the OP's Reddit account had not been deleted—i.e., the original post is not attributed to the ambiguous name "[deleted]"— at the time of crawl.

(3) The OP has, in principle, expressed an openness to other points of view, so that we might hope to extract a sufficient number of view-changing examples. These advantages are not jointly exhibited by other debate sites, such as CreateDebate.com, ForandAgainst.com, or Debate.org.

The high quality of argumentation makes CMV a model site for seeing whether opinion shifts can at least occur under favorable conditions. Moderators enforce CMV rules, making sure that OPs explain why they hold their beliefs and do so at reasonable length (500 characters or more), and that OPs engage in conversation with challengers in a timely fashion. Other rules apply to those who contest the original post. There are rules intended to prevent "low effort" posts, such as "Posts that are only a single link with no substantial argumentation", but "Length/conciseness isn't the determining [criterion]. Adequate on-topic information is."[6] Figure 3.2 shows an example where indeed, the OP described their point in reasonable detail, and the responders raised sensible objections.

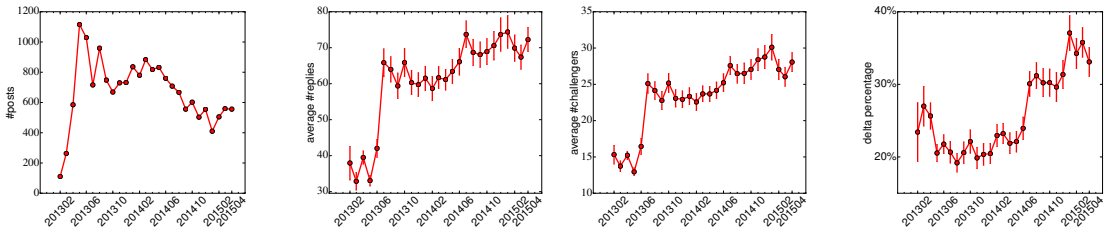
The high amount of activity on CMV means that we can extract a large amount of data. We process all discussion trees created at any time from January 2013, when the subreddit was created, to August 2015, saving roughly the final 4 months (May–August 2015) for held-out evaluation. Some size statistics are given in Table 3.1. Monthly trends are depicted in Figure 3.3:[7] after the initial startup, activity levels stabilize to a healthy, stable growth in average number of replies and challengers, as, gratifyingly, do OP conversion rates, computed as the fraction of discussion trees wherein the OP awarded a Δ (Fig-

[6]It is worth noting that, as in many online communities, not all these rules were in place at the site's creation. It is a separate and interesting research question to understand what effects these rules have and why they were put in place. The currently enforced set of rules is available at https://www.reddit.com/r/changemyview/wiki/rules.

[7]We omit the first month as the DeltaBot may not have been set up.

Table 3.1: Dataset statistics. The disjoint training and test date ranges are 2013/01/01–2015/05/07 and 2015/05/08–2015/09/01.

	# discussion trees	# nodes	# OPs	# uniq. participants
Training	18,363	1,114,533	12,351	69,965
Heldout	2,263	145,733	1,823	16,923



(a) Number of posts per month.

(b) Average no. replies per post.

(c) Average no. challengers per post.

(d) Average delta percentage.

Figure 3.3: Monthly activity over all full months represented in the training set. The *delta percentage* is the fraction of discussion trees in which the OP awarded a delta.

ure 3.3d). For posts where the OP gave at least one delta, the OP gave 1.5 deltas on average. This dataset is available at https://chenhaot.com/pages/changemyview.html.

3.4 Interaction dynamics

Changing someone's opinion is a complex process, often involving repeated interactions between the participants. In this section we investigate the relation between the underlying dynamics and the chances of "success", where "success" can be seen from the perspective of the challenger (did she succeed in changing the OP's opinion?), as well as from that of the set of challengers (did anyone change the OP's view?).

In order to discuss the relation between interaction dynamics and success, we now introduce corresponding terminology using the example illustrated in Figure 3.1:

- An original statement of views (*original post*) together with all the replies form a *discussion tree*.
- A direct reply to an original post is called a *root reply* (A.1 and B.1 in Figure 3.1). The author of a root reply is a *root challenger*.
- A *subtree* includes a root reply and all its children (B.1–B.12 form one of the two subtrees in Figure 3.1).
- A *path* constitutes all nodes from root reply to a leaf node. Figure 3.1 contains four paths: P_1: A.1, P_2: A.1, A.2, A.4, A.5, P_3: B.1–B.11 and P_4: B.1, B.12. Note that when a Δ is awarded, the DeltaBot automatic reply (A.3) and the OP's post that triggers it (A.2) are not considered part of the path.

In order to focus on discussions with non-trivial activity, in this section we only consider discussion trees with at least 10 replies from challengers and at least one reply from the OP.

3.4.1 Challenger's success

A challenger is successful if she manages to change the view of the OP and receive a Δ. We now examine how the interaction patterns in a discussion tree relate to a challenger's success.

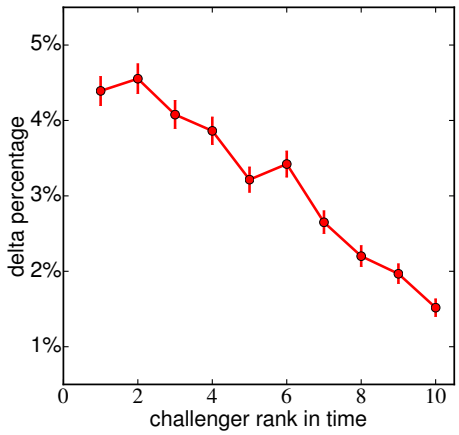
Entry time. How does the time when a challenger enters a discussion relate to her chances of success? A late entry might give the challenger time to read attempts by other challengers and better formulate their arguments, while an early entry might give her the first-mover advantage.[8] Even for original posts that eventually attract attempts by 10 unique challengers, the first two challengers are 3 times more likely to succeed as the 10^{th} (Figure 3.4a).

One potential explanation for this finding is that dedicated expert users are more likely to be more active on the site and thus see posts first. To account for this, we redo the analysis only for users that are participating for the first time on CMV. We observe that even after controlling for user experience, an earlier entry time is still more favorable.

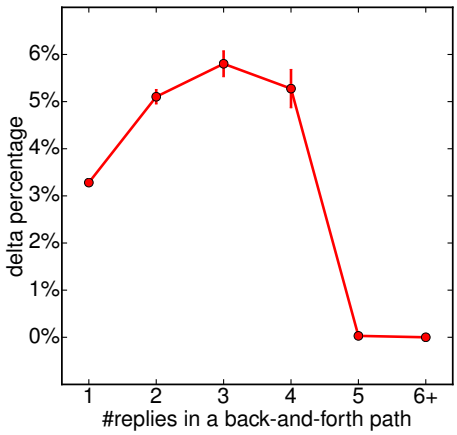
Back-and-forth. After entering a discussion, the challenger can either spend more effort and engage with the OP in a back-and-forth type of interaction or call it quits. Figure 3.4b shows the relation between the likelihood of receiving a Δ and degree of back-and-forth, defined as the number of replies the root challenger made in a path involving only her and the OP.[9] We observe a non-monotonic relation between back-and-forth engagement and likelihood of success: perhaps while some engagement signals the interest of the OP, too much engagement can indicate futile insistence; in fact, after 5 rounds of back-and-forth the challenger has virtually no chance of receiving a Δ.

[8] Note that although reply display order is affected by upvotes, entry time is an important factor when the OP follows the post closely.

[9] If a subtree won a Δ, we only consider the winning path; otherwise, other conversations would be mistakenly labeled unsuccessful. For instance, the path A.1, A.2, A.4, A.5 in Figure 3.1 is not considered.



(a) Delta ratio vs. entry order.



(b) Delta ratio vs. degree of back-and-forth exchanges.

Figure 3.4: Figure 3.4a shows the ratio of a person eventually winning a delta in a post with at least 10 challengers depending on the order of her/his entry. *Early entry is more likely to win a delta.* Figure 3.4b presents the probability of winning a delta given the number of comments by a challenger in a back-and-forth path with OP. With 6 or more replies in a back-and-forth path, *no* challengers managed to win a delta among our 129 data points (with 5 replies, the success ratio is 1 out of 3K). In both figures, error bars represent standard errors (sometimes 0).

3.4.2 OP's conversion

From the perspective of an original post, conversion can happen when any of the challengers participating in the discussion succeeds in changing the OP's view. We now turn to exploring how an OP's conversion relates to the volume and type of activity her original post attracts.

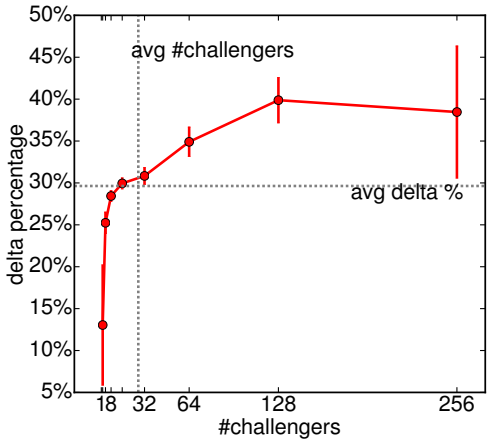
Number of participants. It is reasonable to expect that an OP's conversion is tied to the number of challengers (Chaiken, 1987; Cialdini et al., 1999). For instance, the OP might be persuaded by observing the sheer number of people arguing against her original opinion. Moreover, a large number of challengers will translate into a more diverse set of arguments, and thus higher likelihood that the OP will encounter the ones that best fit her situation. Indeed, Fig-

ure 3.5a shows that the likelihood of conversion does increase with the number of unique challengers. Notably, we observe a saturation in how much value each new challenger adds beyond a certain point.

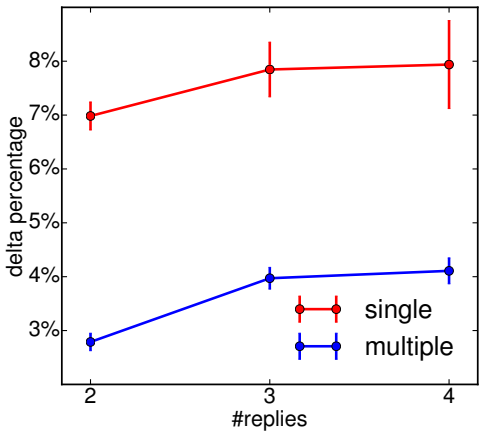
Sheer number of challengers or diversity of counterarguments? To distinguish between the two possible explanations proposed in the previous paragraph, we control for the diversity of counterarguments by focusing only on subtrees, in which challengers generally focus on the same argument. To make a fair comparison, we further control the number of total replies in the subtree. In light of Figure 3.4b, we only consider subtrees with between 2 and 4 replies. Figure 3.5b shows that single-challenger subtrees consistently outperform multiple-challenger subtrees in terms of conversion rate. This observation suggests that the sheer number of challengers is not necessarily associated with higher chances of conversion. The fact that multiple-challenger subtrees are less effective might suggest that when talking about the same counterargument, challengers might not be adding value to it, or they might even disagree (e.g., B.12 vs. B.2 in Figure 3.1); alternatively, root replies that attract multiple challengers might be less effective to begin with.

3.5 Language indicators of persuasive arguments

The interaction dynamics studied in the previous section are to a large extent outside the challenger's influence. The language used in arguing, however, is under one's complete control; linguistic correlates of successful persuasion can therefore prove of practical value to aspiring persuaders. In order to understand what factors of language are effective, we set up paired prediction tasks



(a) Delta percentage vs. number of unique challengers.



(b) Single-challenger subtree vs. multiple-challenger subtree controlled by the number of replies.

Figure 3.5: Probability that a submitted view will be changed, given (a) the total number of unique challengers binned using $\log_2$, and (b) the number of replies in a subtree.

to explore the effectiveness of textual discussion features, in the context of CMV.

3.5.1 Problem setup

In order to study an individual's success in persuasion, we consider the collection of arguments from the same person in the same line of argument. We focus on arguments from root challengers since the root reply is what initiates a line of argument and determines whether the OP will choose to engage. We define all replies in a path by the root challenger as a *rooted path-unit*, e.g., reply A.1 and B.1 in Figure 3.1.

As shown in Section 3.4, situations where there is more than one reply in a rooted path-unit correspond to a higher chance that the OP will be persuaded. So, while the challenger's opening argument should be important, statements made later in the rooted path-unit could be more important. To distinguish

these two cases, we consider two related prediction tasks: *root reply*, which only uses the challenger's opening argument in a rooted path-unit, and *full path*, which considers the text in all replies within a rooted path-unit.

In response to the same original post, there are many possible ways to change someone's view. We aim to find linguistic factors that can help one formulate her/his argument, rather than to analyze reasoning strategies.[10] Hence, for each rooted path-unit that wins a Δ, we find the rooted path-unit in the same discussion tree that did not win a Δ but was the most "similar" in topic. We measure similarity between rooted path-units based on Jaccard similarity in the root replies after removing stopwords (as defined by Mallet's dictionary (McCallum, 2002)):

$$\text{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|},$$

where A, B are the sets of words in the first reply of each of the two rooted path-units. This leads to a balanced binary prediction task: which of the two lexically similar rooted path-units is the successful one? With this setup, we attempt to roughly de-emphasize *what* is being said, in favor of *how* it is expressed.

We further avoid trivial cases, such as replies that are not arguments but clarifying questions, by removing cases where the root reply has fewer than 50 words. In order to make sure that there are enough counterarguments that the OP saw, motivated by the results in Section 3.4.2, we also require that there are at least 10 challengers in the discussion tree and at least 3 unsuccessful rooted path-units before the last reply that the OP made in the discussion tree.

[10] That is an intriguing problem for future work that requires a knowledge base and sophisticated semantic understanding of language.

In an ideal world, we would control for both length (Danescu-Niculescu-Mizil et al., 2012a) and topic (Jaech et al., 2015; Tan et al., 2014), but we don't have the luxury of having enough data to do so. In our pilot experiments, annotators find that Jaccard-controlled pairs are easier to compare than length-matched pairs, as the lexical control is likely to produce arguments that make similar claims. Since length can be predictive (for instance, **C2** won a Δ in Figure 3.2), this raises the concern of false positive findings. Hence we develop a post-mortem "dissection" task (labelled *root truncated*) in which we only consider the root reply and truncate the longer one within a pair so that both root replies have the same number of words. This forcibly removes all length effects.

Disclaimer: Features that lose predictive power in the *root truncated* setting (or "reverse direction"[11]) are not necessarily false positives (or non-significant), as truncation can remove significant fractions of the text and lead to different distributions in the resultant dataset. Our point, though, is: if features retain predictive power *even in* the root truncated settings, they must be indicative beyond length.

We extract pairs from the training and heldout periods respectively as training data (3,456 pairs) and heldout testing data (807 pairs). Given that our focus is on language, we only use text-based features in this section.[12] In preprocessing, we remove explicit edits that users made after posting or commenting, and convert quotations and URLs into special tokens.

[12] An entry order baseline only achieves 54.3% training accuracy.

Table 3.2: Significance tests on interplay features. Features are sorted by average p-value in the two tasks. In all feature testing tables, the number of arrows indicates the level of p-value, while the direction shows the relative relationship between positive instances and negative instances, ↑↑↑↑: $p < 0.0001$, ↑↑↑: $p < 0.001$, ↑↑: $p < 0.01$, ↑: $p < 0.05$. T in the *root reply* column indicates that the feature is also significant in the *root truncated* condition, while T^R means that it is significant in *root truncated* but the direction is reversed.

Feature name	*root reply*	*full path*
reply frac. in all	↓↓↓↓(T)	↓↓↓↓
reply frac. in content	↓↓↓↓(T)	↓↓↓↓
OP frac. in stopwords	↑↑↑↑(T^R)	↑↑↑↑
#common in stopwords	↑↑↑↑(T^R)	↑↑↑↑
reply frac. in stopwords	↓↓↓↓	↓↓↓↓
OP frac. in all	↑↑↑↑(T^R)	↑↑↑↑
#common in all	↑↑↑↑(T^R)	↑↑↑↑
Jaccard in content	↓↓↓↓(T)	↓↓↓↓
Jaccard in stopwords	↑↑↑↑(T^R)	↑↑↑↑
#common in content	↑↑↑↑(T^R)	↑↑↑↑
OP frac. in content	↑ (T^R)	↑↑↑↑
Jaccard in all	↓ (T)	

3.5.2 Features

In order to capture characteristics of successful arguments, we explore two classes of textual features: (Section 3.5.2.1) features that describe the interplay between a particular challenger's replies and the original post, and (Section 3.5.2.2) features that are solely based on his/her replies. We present those features that are statistically significant in the training data under the paired t-test with Bonferroni correction for multiple comparisons.

3.5.2.1 Interplay with the original post: Table 3.2

The context established by the OP's statement of her view can provide valuable information in judging the relative quality of a challenger's arguments. We

capture the interplay between arguments and original posts through similarity metrics based on word overlap.[13] We consider four variants based on the number of unique words in common between the argument (A) and the original post (O):

- number of common words: $|A \cap O|$,
- reply fraction: $\frac{|A \cap O|}{|A|}$,
- OP fraction: $\frac{|A \cap O|}{|O|}$,
- Jaccard: $\frac{|A \cap O|}{|A \cup O|}$.

While stopwords may be related to how challengers coordinate their style with the OP (Danescu-Niculescu-Mizil et al., 2012b; Niederhoffer and Pennebaker, 2002), content words can be a good signal of new information or new perspectives. Thus, inspired by previous results distinguishing these vocabulary types in studying the effect of phrasing (Tan et al., 2014), for each of the four variants above we try three different word sets: stopwords, content words and all words.

The features based on interplay are all significant to a certain degree. Similar patterns occur in *root reply* and *full path*: in number of common words and OP fraction, persuasive arguments have larger values because they tend to be longer, as will be shown in Section 3.5.2.2; in reply fraction and Jaccard, which are normalized by reply length, persuasive arguments are more dissimilar from the original post in content words but more similar in stopwords. Keeping in mind that the pairs we compare are chosen to be similar to each other, our analy-

[13] We also tried *tf-idf*, topical, and word embedding–based similarity in cross validation on training data. We defer discussion of potentially useful features to Section 3.7.

sis indicates that, under this constraint, persuasive arguments use a more different wording from the original post in content, while at the same time matching them more on stopwords.

If we instead use truncation to (artificially) control for reply length, persuasive arguments present lower similarity in all metrics, suggesting that effects might differ over local parts of the texts. However, it is consistent that successful arguments are less similar to the original post in content words.

3.5.2.2 Argument-only features: Table 3.3

We now describe cues that can be extracted solely from the replies. These features attempt to capture linguistic style and its connections to persuasion success.

Number of words. A straightforward but powerful feature is the number of words. In both *root reply* and *full path*, a larger number of words is strongly correlated with success. This is not surprising: longer replies can be more explicit (O'Keefe, 1997, 1998) and convey more information. But naïvely making a communication longer does not automatically make it more convincing (indeed, sometimes, more succinct phrasing carries more punch); our more advanced features attempt to capture the subtler aspects of length.

Word category–based features. As suggested by existing psychology theories and our intuitions, the frequency of certain types of words may be associated with persuasion success. We consider a wide range of categories (see Section 3.10 for details), where for each, we measure the raw number of word occurrences and the length-normalized version.

Word score–based features. Beyond word categories, we employ four scalar word-level attributes (Brysbaert et al., 2014; Warriner et al., 2013):

- Arousal captures the intensity of an emotion, and ranges from "calm" words (*librarian, dull*) to words that excite, like *terrorism* and *erection*.
- Concreteness reflects the degree to which a word denotes something perceptible, as opposed to abstract words which can denote ideas and concepts, e.g., *hamburger* vs. *justice*.
- Dominance measures the degree of control expressed by a word. Low-dominance words can suggest vulnerability and weakness (*dementia, earthquake*) while high-dominance words evoke power and success (*completion, smile*).
- Valence is a measure of how pleasant the word's denotation is. Low-valence words include *leukemia* and *murder*, while *sunshine* and *lovable* are high-valence.

We scale the four measures above to lie in $[0, 1]$.[14] We extend these measures to texts by averaging over the ratings of all content words. Table 3.3 shows that it is consistently good to use calmer language. Aligned with our findings in terms of sentiment words (Section 3.10), persuasive arguments are slightly less happy. However, no significant differences were found for concreteness and dominance.

[14]While the resources cover most common words, out-of-vocabulary misses can occur often in user-generated content. We found that all four values can be extrapolated with high accuracy to out-of-vocabulary words by regressing on dependency-based word embeddings (Levy and Goldberg, 2014) (median absolute error of about 0.1). Generalizing lexical attributes using word embeddings was previously used for applications such as figurative language detection (Tsvetkov et al., 2014).

Characteristics of the entire argument. We measure the number of paragraphs and the number of sentences: persuasive arguments have significantly more of both. To capture the lexical diversity in an argument, we consider the *type-token ratio* and *word entropy*. Persuasive arguments are more diverse in *root reply* and *full path*, but the *type-token ratio* is surprisingly higher in *root truncated*: because of correlations with length and argument structure, lexical diversity is hard to interpret for texts of different lengths. Finally, we compute Flesch-Kincaid grade level (Kincaid et al., 1975) to represent readability. Although there is no significant difference in *root reply*, persuasive arguments are more complex in *full path*.

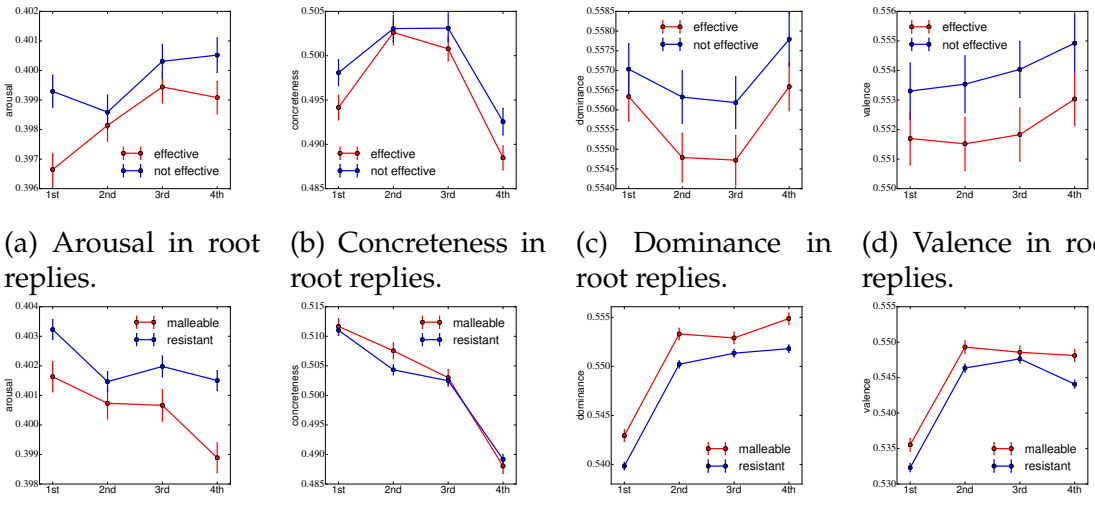
Formatting. Last but not least, discussions on the Internet employ certain writing conventions enabled by the user interface. Since Reddit comments use Markdown[15] for formatting, we can recover the usage of bold, italic, bullet lists, numbered lists and links formatting.[16] While these features are not applicable in face-to-face arguments, more and more communication takes place online, making them highly relevant. Using absolute number, most of them are significant except numbered lists. When it comes to normalized counts, though, only italicizing exhibits significance.

3.5.2.3 They hold no quarter, they ask no quarter

Understanding how a line of argument might evolve is another interesting research problem. We investigate by quartering each argument and measuring certain feature values in each quarter, allowing for finer-grained insight into argument structure.

[15] https://daringfireball.net/projects/markdown/

[16] We also consider numbered words (*first*, *second*, *third*, etc.) as the textual version of numbered lists.



(a) Arousal in root replies.

(b) Concreteness in root replies.

(c) Dominance in root replies.

(d) Valence in root replies.

(e) Arousal in original posts.

(f) Concreteness in original posts.

(g) Dominance in original posts.

(h) Valence in original posts.

Figure 3.6: Style features in different quarters. The first row shows how arousal, concreteness, dominance and valence change in different quarters of the root reply, while the second row shows the same features in the original posts. The descending concreteness trend suggests that opinions tend to be expressed in a particular-to-general way; replies notably differ by having both the opening and the closing be abstract, with a concrete middle. These differences are indicative of the functions that the two forms of utterances serve: a CMV rule is that original posts should not be "like a persuasive essay". Error bars represent standard errors.

Word score–based features in quarters. (Figure 3.6) With the exception of arousal, effective arguments and ineffective arguments present similar patterns: the middle is more concrete and less dominant than the beginning and end, while valence rises slightly over the course of an argument. We also see interesting differences in psycholinguistic patterns between original posts and replies. (We defer detailed discussion to Section 3.6.) In terms of arousal, however, successful arguments begin by using calmer words.

Interplay with the original post. (Figure 3.7) To capture partial overlap and possible divergence from the OP's view, we divide both the original post and the rooted path-unit into quarters, and measure similarity metrics between all

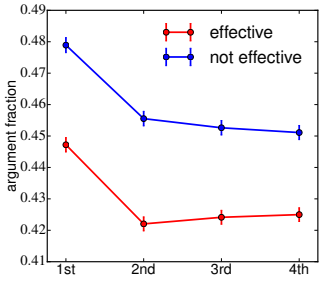


Figure 3.7: Similarity between each quarter of an argument and the entire original post.

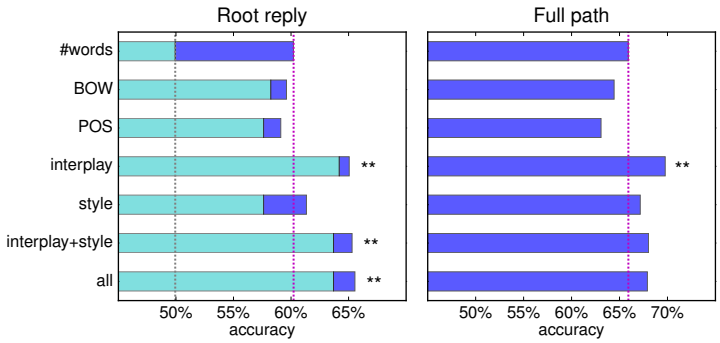


Figure 3.8: **Prediction results.** The cyan fraction in the left figure shows the performance in *root truncated*, and the purple bar shows the performance in *root reply*. The magenta line shows the performance of *#words* in *root reply*, while the gray line shows the performance of *#words* in *root truncated*, which is the same as random guessing. The figure on the right gives the performance in *full path* (the magenta line gives the performance of *#words*). The number of stars indicate the significance level compared to the *#words* baseline according to McNemar's test. (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.)

subdivisions (including the full unit).[17] Since the reply fraction in content words is the most significant interplay feature, in Figure 3.7 we only show the fraction of common content words in different quarters of replies vs. the original post. Both effective and ineffective arguments start off more similar with the original post; effective arguments remain less similar overall.

3.5.3 Prediction results

We train logistic regression models with ℓ_1 regularization on the training set and choose parameters using five cross-validation folds, ensuring that all pairs of

[17] In prediction, we also take the maximum and minimum of these quarter-wise measures as an order-independent way to summarize fragment similarity.

arguments that share the same OP are in the same fold.[18] All features are standardized to unit variance, and missing values are imputed using the training fold sample mean. We evaluate using pairwise accuracy in the *heldout dataset*, where we restricted ourselves to *a single experimental run* (after holding our collective breath) to further reduce the risk of overfitting. The results are, in fact, in line with what we describe in the training-data analysis here.

Feature sets. As shown in Section 3.5.2, the number of words is very predictive, providing a strong baseline to compare against. Bag-of-words features (*BOW*) usually provide a strong benchmark for text classification tasks. We restrict the size of the vocabulary by removing rare words that occurred no more than 5 times in training and ℓ_2-normalize term frequency vectors. Since part-of-speech tags may also capture properties of the argument, we also use normalized term frequency vectors by treating part-of-speech tags as words (*POS*). Features in Section 3.5.2.1 are referred to as *interplay*; features in Section 3.5.2.2 constitute the feature set *style*. Finally, we use a combination of style and interplay, as well as a combination that includes all the above features (*all*). Note that style and interplay are dense and very low-dimensional compared to *BOW*.

Interplay with the OP plays an essential role. (Figure 3.8) *#words* is indeed a very strong baseline that achieves an accuracy of 60% in *root reply* and 66% in *full path*. As a sanity check, in *root truncated*, it indeed gets only 50%. In comparison, *BOW* achieves similar performance as *#words*, while *POS* gives even worse performance. However, interplay features lead to a 5% absolute improvement over the *#words* baseline in *root reply* and *full path*, and a 14% absolute improvement in *root truncated*. In fact, the performance of interplay is already close to

[18] We also tried ℓ_2 regularization, random forests and gradient boosting classifiers and found no improvement beyond the cross-validation standard error.

using the combination of interplay and style and using all features. In *root truncated*, although the performance of style features drops significantly, interplay achieves very similar performance as in *root reply*, demonstrating the robustness of the interplay features.

3.6 "Resistance" to persuasion

Although it is a good-faith step for a person to post on CMV, some beliefs in the dataset are still "resistant" to changes, possibly depending on how strongly the OP holds them and how the OP acquired and maintained them (Pomerantz et al., 1995; Tormala and Petty, 2002; Zuwerink and Devine, 1996). Since CMV members must state their opinion and reasons for it in their own words, we can investigate differences between how resistant and malleable views are expressed. In this section, we seek linguistic and style patterns characterizing original posts in order to better understand the mechanisms behind attitude resistance and expression, and to give potential challengers a sense of which views may be resistant before they engage.

However, recognizing the "malleable" cases is not an easy task: in a pilot study, human annotators perform at chance level (50% on a paired task to distinguish which of two original posts is malleable). In light of our observation that persuasion is unsuccessful in 70% of the cases from Section 3.4, we set up an imbalanced prediction task. We focus on cases where at least 10 challengers attempt counterarguments, and where the OP replied at least once,[19] alleviating the concern that an opinion appears resistant simply because there was little

[19] Although in preprocessing we replaced all explicit edits, we also remove all posts containing the word *"changed"*, to avoid including post-hoc signals of view change.

effort towards changing it. This brings us 10,743 original posts in the training data and 1,529 original posts in the heldout data. We then analyze systematic expression patterns that characterize malleable beliefs and that signal open-mindedness.

3.6.1 Stylistic features for open-mindedness

We employ the same set of features from Section 3.5.2.2 to capture the characteristics of original posts. Among them, only a handful are significantly predictive of malleability, as shown in Table 3.4.

Personal pronouns and self-affirmation. First person pronouns are strong indicators of malleability, but first person plural pronouns correlate with resistance. In psychology, self-affirmation has been found to indicate open-mindedness and make beliefs more likely to yield (Cohen et al., 2000; Correll et al., 2004). Our result aligns with these findings: individualizing one's relationship with a belief using first person pronouns affirms the self, while first person plurals can indicate a diluted sense of group responsibility for the view. Note that it was also found in other work that openness is negatively correlated with first person singular pronouns (Pennebaker and King, 1999).

Table 3.4: Opinion malleability task: statistically significant features after Bonferroni correction.

Feature name	More malleable?
#1st person pronouns	↑↑↑↑
frac. 1st person pronoun	↑↑↑↑
dominance	↑↑↑↑
frac. 1st person plural pronoun	↓↓↓
#paragraphs	↑↑
#1st person plural pronoun	↓↓
#bolds	↑
arousal	↓
valence	↑
bullet list	↑

Formatting. The use of more paragraphs, bold formatting, and bulleted lists are all higher when a malleable view is expressed. Taking more time and presenting the reasons behind an opinion in a more elaborated form can indicate more engagement.

Word score–based features. Dominance is the most predictive of malleability: the average amount of control expressed through the words used is higher when describing a malleable view than a resistant one. The same holds for happiness (captured by valence). In terms of arousal, malleable opinions are expressed significantly more serenely, ending on a particularly calm note in the final quarter, while stubborn opinions are expressed with relatively more excitement.

3.6.2 Prediction performance

We use weighted logistic regression and choose the amount and type of regularization (ℓ_1 or ℓ_2) by grid search over 5 cross-validation folds. Since this is an imbalanced task, we evaluate the prediction results using the area under the

ROC curve (AUC) score. As in Section 3.5, we use the number of words as our baseline. In addition to the above features that characterize language style (*style*), we use bag-of-words (*BOW*), part-of-speech tags (*POS*) and a full feature set (*all*). The holdout performance is shown in Figure 3.9.

The classifiers trained on bag of words features significantly outperforms the #*words* baseline. Among words with largest coefficients, resistant views tend to be expressed using more decisive words such as *anyone*, *certain*, *ever*, *nothing*, and *wrong*, while *help* and *please* are malleable words. The *POS* classifier significantly outperforms random guessing, but not the baseline. Nevertheless, it yields an interesting insight: comparative adjectives and adverbs are signs of malleability, while superlative adjectives suggest stubbornness. The full feature set (*all*) also significantly outperform the #*words* baseline. The overall low scores suggest that this is indeed a challenging task for both humans and machines.

3.7 Further discussion

Here we discuss other observations that may open up avenues for further investigation of the complex process of persuasion.

Experience level. Beyond the interactions within a discussion tree, CMV is a community where users can accumulate experience and potentially improve their persuasion ability. Figure 3.10a shows that a member's success rate goes up with the number of attempts made. This observation can be explained by at least two reasons: the success rate of frequent challengers improves over time, and/or frequent challengers are better at persuasion from the beginning. To disentangle these two possible reasons, for challengers who attempted to change

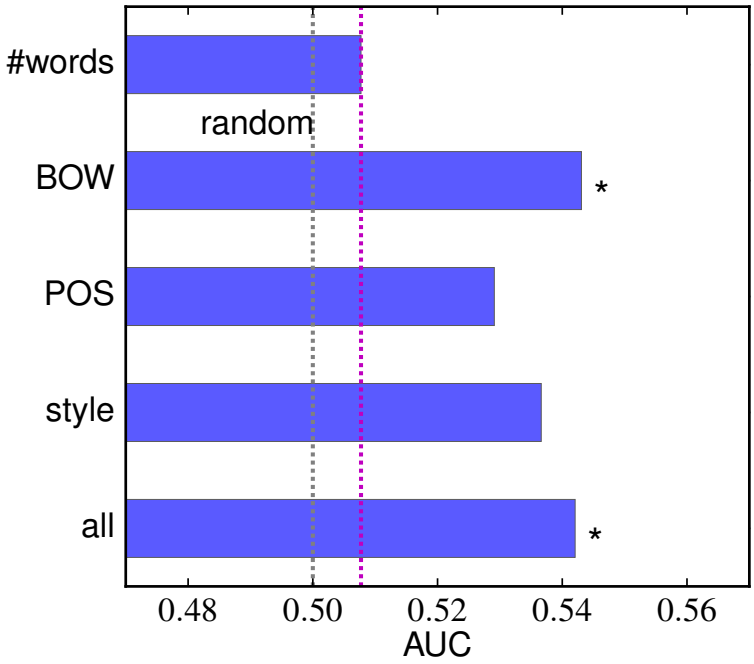
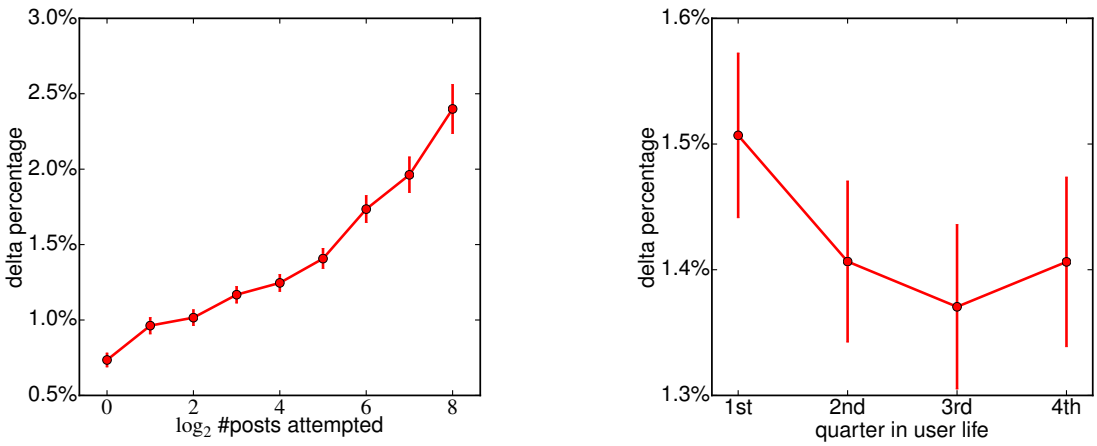


Figure 3.9: **Opinion malleability prediction performance**: AUC on the heldout dataset. The purple line shows the performance of *#words*, while the gray line gives the performance of random guessing. The *BOW* and *all* feature sets perform significantly better than the *#words* baseline, according to one-sided paired permutation tests. *BOW*, *POS*, *style* and *all* outperform random guessing using bootstrapped tests. (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.)

at least 16 views, we split all the attempts into 4 equal chunks sorted by time. Figure 3.10b presents how the success rate changes over a challenger's life, suggesting that the success rate of frequent challengers does not increase.[20] It is worth noting that this lack of apparent improvement might be explained by a gradual development of a "taste" for original posts that are harder to address (McAuley and Leskovec, 2013). Such community dynamics point to interesting research questions for future work.

Attempts to capture high-level linguistic properties. We experimented with a broader set of features in cross validation. One important class are attempts to

[20] In terms of the correlation between previous success (lifetime deltas) and success rate, the result is similar: beyond 4–5 deltas there is no noticeable increase.



(a) Delta ratio vs. $\log_2$ #posts.

(b) No growth over life.

Figure 3.10: Effect of experience.

capture the semantics of original statements and arguments. We experimented with using topic models (Blei et al., 2003) to find topics that are the most malleable (*topic: food, eat, eating, thing, meat* and *topic: read, book, lot, books, women*), and the most resistant (*topic: government, state, world, country, countries* and *topic: sex, women, fat, person, weight*). However, topic model based features do not seem to bring predictive power to either of the tasks. For predicting persuasive arguments, we attempted to capture interplay with word embeddings for text similarity using both the centroid distance and the word mover's distance (Kusner et al., 2015). Both distances proved predictive by themselves, but were not able to improve over the features presented in the work in cross validation. More generally, better semantic models applicable to online discussions could open up deeper investigations into effective persuasion strategies.

Sequential argument structure. Another promising direction is to examine the structure of arguments via the sequence of discourse connectors. For instance, we can recover interesting structures such as "$first_0$–but_1–$because_2$" and "now_1–$then_2$–$instead_3$", where the subscripts indicate which quarter the discourse con-

nector occurred in. These features did not perform well in our tasks due to low recall, or lack of argumentative structure in the data, but they deserve further exploration.

3.8 Additional related work

A few lines of research in natural language processing are related to our work. Argumentation mining focuses on fine-grained analysis of arguments and on discovering the relationships, such as support and premise, between different arguments (Lippi and Torroni, 2016; Mochales and Moens, 2011). Studies have also worked on understanding persuasive essays (Farra et al., 2015; Persing and Ng, 2015; Stab and Gurevych, 2014), opinion analysis in terms of agreement and ideology (Thomas et al., 2006; Somasundaran and Wiebe, 2010; Hasan and Ng, 2014; Sridhar et al., 2015; Rosenthal and McKeown, 2015) and semantic frames in political debates (Cano-Basave and He, 2016). Another innovative way of using Internet data to study mass persuasion is through AdWords (Guerini et al., 2010). In the context of argumentation, similar to the theme of the previous chapter, Zhang and Litman (2016) examine the argumentative purposes of revisions.

3.9 Conclusion

In this work, in order to understand the mechanisms behind persuasion, we use a unique dataset from `/r/ChangeMyView`. In addition to examining interaction dynamics, we develop a framework for analyzing persuasive arguments

and malleable opinions. We find that not only are interaction patterns connected to the success of persuasion, but language is also found to distinguish persuasive arguments. Dissimilarity with the wording in which the opinion is expressed turns out to be the most predictive signal among all features. Although members of CMV are open-minded and willing to change, we are still able to identify opinions that are resistant and to characterize them using linguistic patterns.

There are many possible extensions to our approach for representing arguments. In particular, it would be interesting to model the framing of different arguments and examine the interplay between framing of the original post and the replies. For instance, is benefit-cost analysis the only way to convince a utilitarian?

Furthermore, although this novel dataset opens up potential opportunities for future work, other environments, where people are not as open-minded, can exhibit different kinds of persuasive interactions; it remains an interesting problem how our findings generalize to different contexts. It is also important to understand the effects of attitude change on actual behavior (Petty et al., 1997).

Finally, beyond mechanisms behind persuasion, it is a vital research problem to understand how community norms encourage such a well-behaved platform so that useful rules, moderation practices, or even automated tools can be deployed in future community building.

3.10 Appendix

In this section we explain our features based on word categories.

- (In)definite articles (inspired by Danescu-Niculescu-Mizil et al. (2012a)). These are highly correlated with length, so they are both highly significant in terms of absolute numbers. However, in terms of word ratios, definite articles (e.g., "the" instead of "a") are preferred, which suggests that specificity is important in persuasive arguments.
- Positive and negative words. We use the positive and negative lexicons from LIWC (Pennebaker et al., 2007). In absolute numbers, successful arguments are more sentiment-laden in both *root reply* and *full path*. When truncating, as well as when taking the frequency ratio, persuasive opening arguments use *fewer* positive words, suggesting more complex patterns of positive emotion in longer arguments (Hullett, 2005; Wegener and Petty, 1996).
- Arguer-relevant personal pronouns. We consider 1st person pronouns (*me*) 2nd person pronouns (*you*) and 1st person plural pronouns (*us*). In both *root reply* and *full path*, persuasive arguments use a significantly larger absolute number of personal pronouns.
- Links. Citing external evidence online is often accomplished using hyperlinks. Persuasive arguments use consistently more links, both in absolute and in per-word count. We make special categories for interesting classes of links: those to `.com` and `.edu` domains, and those to PDF documents. Maybe due to high recall, `.com` links seem to be most powerful. Features

based on links also tend to be significant even in the *root truncated* condition.

- Hedging. Hedges indicate uncertainty; an example is "It could be the case". Their presence might signal a weaker argument (Durik et al., 2008), but alternately, they may make an argument easier to accept by softening its tone (Lakoff, 1975). We curate a set of hedging cues based on (Hanauer et al., 2012; Hyland, 1998). Hedging is more common in persuasive arguments under *root reply* and *full path*.
- Examples. We consider occurrences of "for example", "for instance", and "e.g.". The absolute number of such example markers is significantly higher in persuasive arguments.
- Question marks. Questions can be used for clarification or rhetorical purposes. In terms of absolute number, there are more in *root reply* and *full path*. But when it comes to ratio, if anything, it seems better to avoid using question marks.
- Quotations. One common practice in argumentation is to quote the other party's words. However, this does not seem to be a useful strategy for the root reply.

Table 3.3: Argument-only features that pass a Bonferroni-corrected significance test. Features are sorted within each group by average p-value over the two tasks. Due to our simple truncation based on words, some features, such as those based on complete sentences, cannot be extracted in *root truncated*; these are indicated by a dash. We remind the reader of the *root truncated* disclaimer from Section 3.5.

Feature name	*root reply*	*full path*
#words	↑↑↑↑	↑↑↑↑
Word category–based features		
#definite articles	↑↑↑↑	↑↑↑↑
#indefinite articles	↑↑↑↑	↑↑↑↑
#positive words	↑↑↑↑(T^R)	↑↑↑↑
#2nd person pronoun	↑↑↑↑	↑↑↑↑
#links	↑↑↑↑(T)	↑↑↑↑
#negative words	↑↑↑↑	↑↑↑↑
#hedges	↑↑↑↑	↑↑↑↑
#1st person pronouns	↑↑↑↑	↑↑↑↑
#1st person plural pronoun	↑↑↑↑	↑↑↑↑
#`.com` links	↑↑↑↑(T)	↑↑↑↑
frac. links	↑↑↑↑(T)	↑↑↑↑
frac. `.com` links	↑↑↑↑(T)	↑↑↑↑
#examples	↑	↑↑↑↑
frac. definite articles	↑ (T)	↑↑
#question marks	↑ —	↑↑↑↑
#PDF links	↑	↑↑↑
#`.edu` links		↑
frac. positive words	↓	
frac. question marks	—	↓
#quotations		↑↑↑↑
Word score–based features		
arousal	↓ (T)	↓↓↓
valence	↓	
Entire argument features		
word entropy	↑↑↑↑	↑↑↑↑
#sentences	↑↑↑↑—	↑↑↑↑
type-token ratio	↓↓↓↓(T^R)	↓↓↓↓
#paragraphs	↑↑↑↑—	↑↑↑↑
Flesch-Kincaid grade levels	—	↓↓↓
Markdown formatting		
#italics	↑↑↑↑—	↑↑↑↑
bullet list	↑↑↑↑—	↑↑↑↑
#bolds	↑↑ —	↑↑↑↑
numbered words	↑	↑↑↑↑
frac. italics	↑ —	↑

CHAPTER 4

MULTIPLE COMMUNITIES: ALL WHO WANDER

4.1 Brief overview

Although analyzing user behavior *within* individual communities is an active and rich research domain, people usually interact with *multiple* communities both on- and off-line. How do users act in such multi-community environments? Although there are a host of intriguing aspects to this question, it has received much less attention in the research community in comparison to the intra-community case. In this work, we examine three aspects of multi-community engagement: the *sequence of communities* that users post to, the *language* that users employ in those communities, and the *feedback* that users receive, using longitudinal posting behavior on Reddit as our main data source, and DBLP for auxiliary experiments. We also demonstrate the effectiveness of features drawn from these aspects in predicting users' future level of activity.

One might expect that a user's trajectory mimics the "settling-down" process in real life: an initial exploration of sub-communities before settling down into a few niches. However, we find that the users in our data continually post in new communities; moreover, as time goes on, they post increasingly evenly among a more diverse set of smaller communities. Interestingly, it seems that users that eventually leave the community are "destined" to do so from the very beginning, in the sense of showing significantly different "wandering" patterns very early on in their trajectories; this finding has potentially important design implications for community maintainers. Our multi-community perspective also allows us to investigate the "situation vs. personality" debate from language

usage across different communities.

Most of the contents in the chapter are published in Tan and Lee (2015). This is joint work with Lillian Lee.

4.2 Introduction

树挪死，人挪活 (*People, unlike trees, thrive on relocation*).

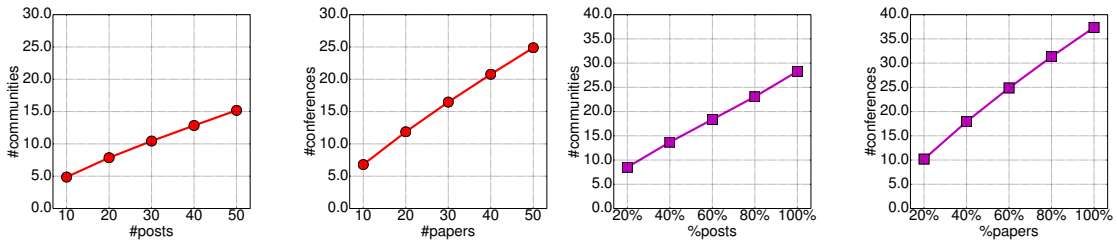
—A Chinese saying

How people behave *within* a given community is a profound and broad question that has inspired work ranging from basic social-science research (e.g., Shaw (1971)) to the design of online social systems (e.g., Kraut and Resnick (2012)). However, many settings offer an array of *multiple* possible interest subgroups for users to engage in. In the offline world, for example, within the bounds of a single college campus, students can get involved with a variety of clubs, organizations, and social circles. And in the online case, there are many multi-community sites, such as Reddit, 4chan, Wikia, and StackExchange, all of which host a slew of topic-based sub-discussion forums. As the results in this chapter show, multi-community settings exhibit many interesting and useful properties that are not manifested in within-community situations, and so *our main goal is to demonstrate that multi-community engagement is an exciting and underexploited research area*: we believe that such work will shed additional light on human behavior and on the design of social-media systems.

To demonstrate, we first tackle a seemingly foregone conclusion: that, analogously to the human life course (Bühler, 1935; Erikson and Erikson, 1998), a

person first passes through an "adolescent" phase of trying out many different interests before "settling down". Indeed, the best-paper award at WWW 2013 was given to an excellent within-community study (Danescu-Niculescu-Mizil et al., 2013) demonstrating (among other things) that users' language use becomes more inflexible and out-of-step with the community's over time. But, contrary to this expectation, we find that even people with long histories of participation in a global community *continually* try out new sub-communities. Figure 4.1 depicts this for two very different settings: for Reddit and for the universe of computer-science conferences given by DBLP, the latter choice inspired by Backstrom et al. (2006). Note that despite their very different timescales (one can post to Reddit at any time, but submission deadlines only roll around every so often) and barriers to entry (conferences have gate-keepers, whereas posting on Reddit can be done essentially at will), they exhibit the same qualitative behavior. On average, Redditors post to 5 communities in their first 10 posts and then post to 2.5 *new* communities every 10 posts, while researchers publish at 5 *new* venues every 10 papers (Fig. 4.1a and 4.1b). These exploration trends continue over the users' lifetimes (Fig. 4.1c, 4.1d). Thus, while within a single community "all users die old" (Danescu-Niculescu-Mizil et al., 2013), it seems that a multi-community setting keeps users young by offering them choices to explore as an alternative to opting out entirely.

Having established the prevalence of "wandering" behavior, we are led to investigate a host of related phenomena. *We believe that these phenomena are interesting in their own right, and at times quite surprising. Moreover, we also demonstrate that our findings inspire new kinds of features that are strongly predictive of users' future level of activity.*



(a) 1st 50, Reddit (b) 1st 50, DBLP (c) lifetime, Reddit (d) lifetime, DBLP

Figure 4.1: Mean number of **unique** communities (subreddits for Reddit, conferences for DBLP) where people make their temporally first x contributions (left-hand plots) or their first x percent of contributions (right-hand plots), for "long-lived" people (50+ contributions overall). For Reddit (respectively, DBLP), contributions = posts (papers). Standard-error intervals are depicted, but very small, and trends for the median are consistent with the mean. Note that the left-hand plots depict long timespans: the average time to accumulate 50 contributions is 456.0 days on Reddit, 15.6 years on DBLP.

Example of a Redditor's first 50 subreddits, in the order posted to, first-time communities underlined: skyrim, aww, skyrim, aww, pics, aww, aww, pics, WTF, aww, pics, WTF, pokemontrades, funny, pokemontrades, pics, aww, AskReddit, pics, pokemon, fashion, AskReddit, aww, Scotland, fashion, aww, Scotland, pics, keto, keto, Fitness, keto, skyrim, pokemon, cats, aww, aww, pokemon, Scotland, AskReddit, fashion, keto, pokemon, ketouk, Scotland, keto, pics, ketouk, funny, gamecollecting.
Two DBLP examples: the set of venues of *James Harland*'s first 50 papers: LPAR, ACE, NACLP, TABLEAUX, DALT, ECOWS, CADE, Australian Joint Conference on Artificial Intelligence, IAT, ICLP, ICSOC, ILPS, "Workshop on Programming with Logic Databases (Book), ILPS", Future Databases, AAMAS, ACSE, EDBT, JICSLP, ACSC, ACAL, SAC, AAMAS (1), PRICAI, Computational Logic, CLIMA, ECAI, AMAST, ISLP, "Workshop on Programming with Logic Databases (Informal Proceedings), ILPS", KR, CATS.
Jure Leskovec's: INFOCOM, HT, AAAI, PKDD, ICDE, ECCV (4), KDD, ICDM, UAI, NIPS, ICML, CHI, VLDB, WWW, EC, WAW, WSDM, ICWSM, PAKDD, CIKM-CNIKM, JCDL, SDM, WWW (Companion Volume).

Organization, further highlights and design implications. In Sections 4.3 and 4.4, we propose an analysis framework and investigate three aspects of users' community trajectories: the communities they post to (§4.4.1), the language they use within a community (§4.4.2), and the feedback they receive from other members of the community (§4.4.3). Consistently, we see that — again, in contrast to the "older people become less adventurous" hypothesis — our users appear to continually seek out new and different communities, and adopt the language

characteristics of the new communities. In particular, an important problem in understanding users' language uses is to measure the divergence between different language models. We propose using various vocabularies and observe interesting differences even focusing on stopwords.

Another interesting point, albeit arguably less surprising, is that they tend to move to smaller communities (a fact noted by Redditors[1]), which might be a signal to site designers to make sure to offer a menu of narrowly-targeted options for users to choose from (or to ensure that sub-groups can arise organically).

Finally, a complete surprise is that for users who made at least 50 posts, the patterns exhibited by those who end up departing the site altogether are *already* significantly different from those users who end up staying by *their first 10 posts*. The fact that future abandonment can be detected so early should be of interest to administrators of social-media systems. But, there is an unexpected factor potentially making this discrimination difficult: in our data, the eventually departing users are often most similar *not* to the least active users in our study, but to the *most* active users. We conjecture that our "dying" users are actively striving to remain engaged, but are not quite managing to explore enough to make their overall posting experience satisfactory. A design implication might be to include mechanisms in one's site that more proactively suggest new, diverse sub-communities for posting.

In Section 4.5, we show that the aforementioned differences in patterns are not "mere" correlations, but do indeed serve as features that are effective at predicting future activity level.

Again, our overall goal is to encourage further work on multi-community

[1]One comment: "the longer you are on reddit, the more you get pulled into smaller subs".

settings. As a spur to the imagination, and as a demonstration that this research domain is rich with possibilities, in this chapter, we discuss in sections 4.6 and 4.7 two additional questions that arise. First, what makes a user abandon a community and move on to new ones? We see that the positivity of initial feedback correlates with what groups users choose to return to, a finding that contradicts recent results on the power of negative feedback (Cheng et al., 2014), albeit for commenting instead of posting. Second, we make a foray into the "situation vs. personality" debate in psychology (Kenrick and Funder, 1988; Donellan et al., 2009): how much of our behavior is determined by fixed personality traits, versus how much is variable and influenced by the specific situation at hand? We consider this question from a linguistic perspective, and determine that *even after topic-specific vocabulary is discarded* (after all, it wouldn't be interesting to find that people use gym-related words at the gym that they don't use at work), users *do* employ different language patterns in different communities. This means that they are able to adapt even into "maturity". In the next chapter, we will explore this issue from the perspective of communities, in particular, why users create highly related communities.

4.3 Experimental setup

In the following, we first describe the data that we use and then propose an analysis framework for capturing the temporal dynamics of multi-community engagement.

4.3.1 Datasets.

The main dataset used in this chapter is drawn from Reddit, a very active community-driven platform for submitting, commenting on, and rating posts[2] (Singer et al., 2014). Reddit is organized into thousands of topic-based, user-created discussion forums called "subreddits", which users can post to essentially at will (modulo spam filtering, rate limits, and deletion of posts by moderators). Other users can "upvote" or "downvote" posts; the difference between the number of upvotes and the number of downvotes, a difference that we henceforth refer to as *feedback*, is readily available.[3]

Relying primarily on RedditAnalytics[4], in February 2014 we collected all 76.6M posts ever submitted to Reddit since its inception except posts by bots and banned users, together with their associated feedback values. We discarded the last month of posts, since their feedback values might not have had sufficient time to converge.

Since we need our users' community trajectories to be long enough to be *able* to exhibit significant wandering (whether or not they actually do), the set of users we consider are those who have made at least 50 posts, following the choice in Danescu-Niculescu-Mizil et al. (2013). We focus on the 157K *50+ posters* who first posted between January 2008 and January 2012 so that we have at least two years' worth (2012-2014) of observations for each of them. We chose to start from January 2008 because users were granted the ability to create their

[2] A Reddit post consists at a minimum of a title that serves as anchor-text for a link. The link may be to an offsite item ("link post") or to some text that the post's author places on Reddit ("text post"). The dataset with more detailed explanation is available at https://chenhaot.com/pages/multi-community.html.

[3] The actual number of upvotes or downvotes is purposely inaccessible: http://bit.ly/1xrciQY.

[4] http://redditanalytics.com/

own subreddits at will then. Not only are the 50+ posters good objects of study because we have a lot of data on their behavior, but they also play a major role in determining the character of Reddit because they made 63% of the posts written by users who first posted in the time period under consideration.[5] The caveats of focusing on users who made at least 50 posts will be discussed later in this chapter.

In order to ensure that our findings generalize beyond Reddit, we also consider a (more) physical-world multi-community situation: the set of conferences in computer science. Conferences generally correspond to topic areas within CS, and each can be thought of as representing a social group, at least to some degree. In this setting, we take "posting" to mean publishing a paper. We use the DBLP database[6] to find what papers appeared in which conferences, and refer to the resultant dataset as "DBLP". For DBLP, we do not consider an analog of Reddit's feedback, although citation or download counts could be used in future work.

It is important to note that program committees play a huge role in determining an author's conference trajectory. This property makes DBLP a less suitable domain for the questions of user choice that we focus on in this work. We thus place our DBLP trajectory results in the Appendix (§4.10).

Statistics on the 50+ posters in Reddit and DBLP are given in Table 4.1.

Note 1: how we define "posting". In this work, we use the term *posting* to refer to submitting an item to be voted or commented upon, the same as posts in

[5]Cross-posting (posting the same URL to multiple subreddits, with or without a title change) accounts for only 3% of the posts from the users that we consider in this chapter — only 1.77% if we only consider their first 50 posts.

[6]http://dblp.uni-trier.de

	Reddit	DBLP
Average number of posts	152.04	86.30
Median	89.	71.
Avg. no. of communities	28.85	38.08
Median	26.	34.
Mean avg. time gap btwn posts	10.47 days	3.36 mos

Table 4.1: Statistics for 50+ posters (157K in Reddit, 10K in DBLP).

CMV in the previous chapter. We distinguish posting from *commenting on posts* for several reasons. First, posting is important for site designers to encourage since the site will presumably die without fresh conversation-starters. Second, posting is not affected by a confounding factor that commenting is subject to: Reddit influences commenting by how it presents potential targets for comments (e.g., by ranking them, or featuring targets on the Reddit home page). Nonetheless, looking at commenting in multi-community environments is an interesting direction for future research. We conjecture that it would lead to new findings since, for example, we do know that top posters are generally not top commenters, and vice versa.[7]

4.3.2 Analysis framework.

We now set up terminology and concepts that facilitate discussion of users' trajectories among communities.

For each post by a given user, we store the timestamp, *time*, and the *community* (sometimes *C* for short). For Reddit data, we also store the post's *feedback* as of February 2014 and its *words* (the anchor-text plus any text written by the user, all tokenized and part-of-speech tagged using the Stanford NLP package[8]).

[7] http://bit.ly/1tendtD
[8] http://nlp.stanford.edu/software/corenlp.shtml

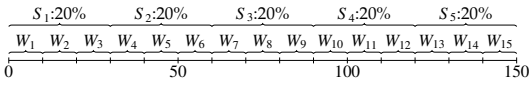
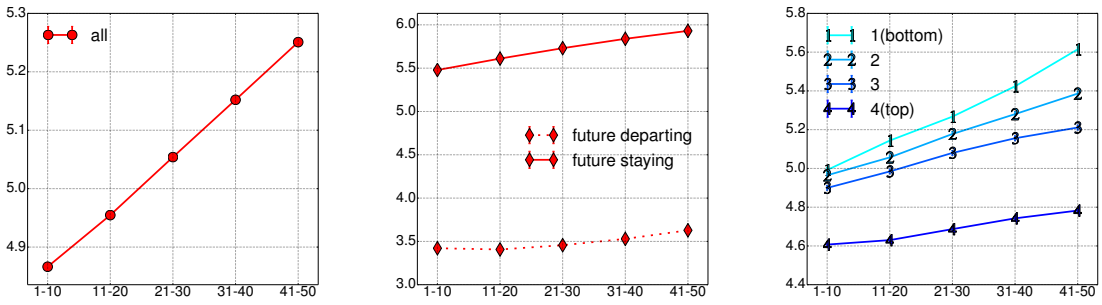


Figure 4.2: Illustration of windows and stages for window size $w = 10$, number of stages $S = 5$, number of posts $T = 150$, number of windows $T_w = 15$. W_i is a window; S_i is a stage.

Several of the questions we are interested in pertain to properties of subsequences of trajectories. For example, suppose we want to know whether users are visiting a broader set of communities over time; one way to check is to look at how many communities they engaged with in their first w posts versus in their last w posts. Therefore, a basic element in our analysis is a *window*. Let variable t index the posts made by a user u, and suppose u has made T posts altogether. We split the entire index sequence $1, \ldots, T$ into non-overlapping consecutive windows W_i of size w, where i ranges from 1 to $T_w \stackrel{def}{=} \lfloor T/w \rfloor$. For example, in Fig. 4.2, W_6 would be the integers in the range $[51, 60]$. We use $w = 10$ throughout this chapter. Our Reddit results were insensitive to choices of w, although the results on DBLP are contingent on the choice of w, which may be due to the substantial effort required to publish papers.

We define functions F on windows W_i to summarize properties of that window and track how these properties change over time. We use two ways to define F. One way is to directly define F based on the entire window, for example, $F(W_i) = |\{C_t : t \in W_i\}|$, the number of unique communities in W_i. The other way is to define a function f for each index t — for example, $f(t)$ could be the number of words in the t^{th} post — and let $F(W_i)$ be induced by f's average value over the indices in W_i: $F(W_i) = \frac{1}{w} \sum_{t \in W_i} f(t)$.

Given a window size w and a function of interest, F, we take two perspectives to track the trajectory of F: a *full-life view (all the user's posts)* and a *fixed-*



(a) Overall trend

(b) Future departing status

(c) Future activity quartile

Figure 4.3: Number of unique communities per window. x-axis: each of the first 5 windows. y-axis: number of unique communities appearing in the corresponding window. In Fig. 4.3b and Fig. 4.3c, users are categorized by their future state *after* the initial 50 posts. Standard-error intervals are depicted, but very small.

Note 2: y-axes scales, and other considerations regarding subsequent figures. Since many of the figures in Section 4.4 tend to support the same overall point as in Figure 4.3, we make the subsequent figures relatively small (labeling the y-axes in the captions), but use the same x-axis, legends, and line styles in all of them.
As in Figure 4.3, each of the other figures in Section 4.4 consists of three sub-figures. In each, we scale the y-axes according to the corresponding data's range in order to show significant changes (all figures show standard-error bars, which are tiny). But it should be noted that the lines when averaging over all users (leftmost sub-figure in the figures) would usually look flatter if plotted on the graphs that divide users by departure status (middle sub-figures) or activity quartile (rightmost sub-figures).

prefix view (50 posts). The rationales are as follows:

The first perspective, *full-life*, tracks users' entire lifetimes. Because the value of some functions is affected by choice of window size (e.g., the number of unique communities), we still fix the window size in the full-life view, but set an additional parameter S of the number of life stages that we want to examine, where each life stage contains the same number of windows, as depicted in Fig. 4.2. For each stage, we compute the average value over the windows in that stage.

A slight problem with the full-life view is that for different users, the value of the same life stage (say, the first 10% of one's life) may be based on a significantly different number of posts (say, 10 for one user but 100 for another). The full-life view also includes information about the entirety of the user's life, and thus is not appropriate for prediction settings (for example, one does not ordinarily know at the time what percent of one's life has already passed). Thus we also take a *fixed-prefix* view, where *only* the initial 50 posts are examined. (Recall from the caption of Fig. 4.1 that this encompasses a long time span on average.) Thus, the same amount of data is used for every user and the induced features are valid for predicting future behavior. We will focus on the fixed-prefix view for now since the fixed-prefix view can be used to forecast future activity levels, and place some full-life-view results in the Appendix (§4.10).

Future activity level. We further relate our analysis to users' future activity level, since future activity level is a useful quantity to predict. We employ two different ways to categorize users' future commitment: the two-way classification of whether a user eventually abandons the global community altogether or not, and a 4-way split based on the relative number of posts that a user eventually makes over his/her lifetime, as follows.

- Departing status. To determine which users should be considered to have abandoned the site, we define a date (specifically, 6 months before January 2014) as the start-of-future (SOF). We define *departing* users as those who stopped posting as of SOF; we define *lasting* users as non-departing users who additionally post at least once in the first 3 months and at least once in the second 3 months since SOF, so that they are consistently "active". There are 43,910 departing users and 75,708 lasting users. Note that they

all made at least 50 posts before SOF.

- Activity quartile. We split users into four quartiles based on the number of posts that they make in their entire life after the initial 50 posts. (As it happens, the lasting/departing ratio is higher in the the higher-activity quartile.)

4.4 Trajectory properties

We have established in Fig. 4.1 that users do constantly "wander around" in multi-community environments. In this section, we apply the framework proposed in §4.3 to explore three aspects of this wandering process: (§4.4.1) the communities users post to; (§4.4.2) the language users employ in each community; (§4.4.3) the feedback that users receive from other community members. In §4.5, we will further validate the effectiveness of features based on these properties in prediction tasks.

4.4.1 Multi-community aspects

We have shown in §4.2 that users on average consistently post to 2.5 new communities every 10 posts (Fig. 4.1). But what else characterizes their patterns of movement among communities? The answers to this question have the design implications outlined in §4.2.

Section summary. *We find that over time, users span more communities every 10 posts, "jump" more, and concentrate less.*[9] *They enter smaller and less similar commu-*

[9] The continual exploration is not simply an effect of the introduction of new communities

nities. Eventually-departing users seem consistently less "adventurous" than lasting users even, notably, from the very beginning. Curiously, eventually-departing users act similarly to users in the top activity quartile.

In the following, we explain the metrics for understanding these properties and discuss related theories.

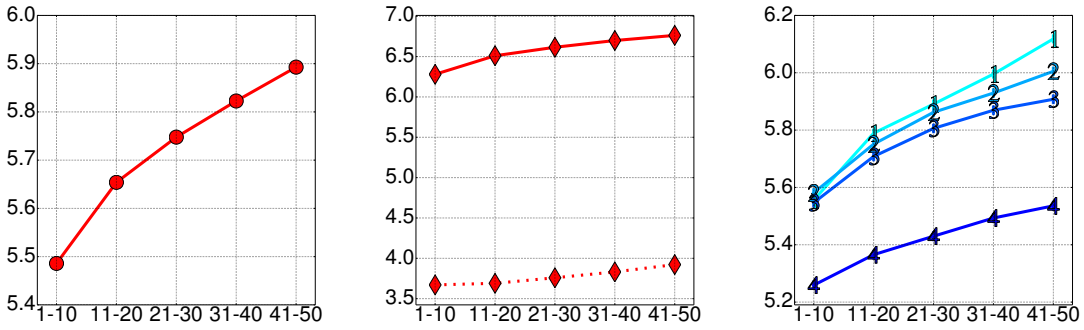
Users span more and more unique communities in a window, but relatively speaking, departing users span fewer unique communities. Figure 4.3 shows the per-window number of unique communities that users post to. The actual number is interesting: in Fig. 4.1, users post to 2.5 new communities every 10 posts; here on average, users post to around 5 communities every 10 posts, and thus only around 2.5 of them are ones that they have ever posted to. Given that users have more potential communities to go back to over time, this suggests that they do not tend to return to some previous communities. More discussion as to why users return to certain communities will be presented in §4.6.

Users "jump" between communities more and more "frequently", but departing users do so at around half the "rate". (Fig. 4.4) To understand how often users "jump", we count the number of "jumps" that users make per window. Formally, define $F(W_i) = \sum_{t,t+1 \in W_i} I(C_t \neq C_{t+1})$, where $I(x)$ is the indicator function: $I(x) = 1$ if x is true, 0 otherwise.

Note that the number of unique communities in a window of 10 does not determine how often users "jump". Given a window size of 10, users can jump as many as 9 times; given that users on average span 5 communities in a window, users can jump as few as 4 times. In fact, users make around 5.8 "jumps" per 10

over time. For instance, although new communities or options also emerge in real life, people seem to settle down and do not explore much.

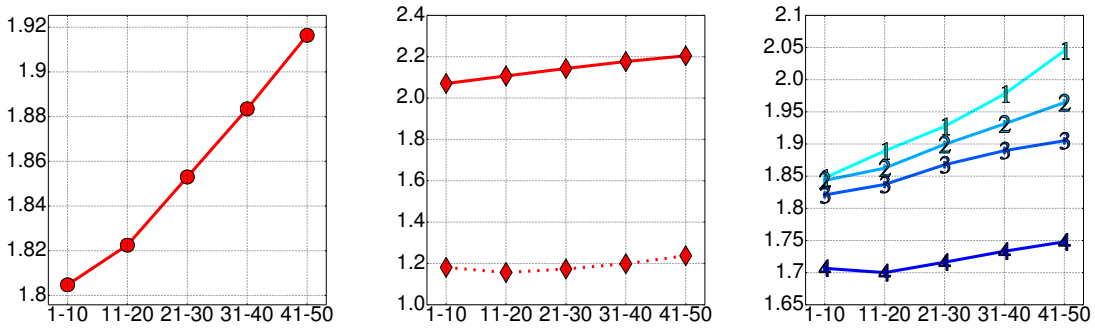
posts.



(a) Overall trend (b) Departing status (c) Activity quartile

Figure 4.4: Number of "jumps".

Users spread their posts out more and more evenly, but relatively speaking, departing users focus more. (Fig. 4.5) We employ entropy as a metric for concentration, following Adamic et al. (2008). Entropy is based on the probability of a community appearing in a window W_i, $p_c = \frac{1}{w} \sum_{t \in W_i} I(C_t = c)$, and is defined as $-\sum_c p_c \log_2 p_c$ for W_i. It is an information-theoretic measure that grows as the intra-window community-posting distribution approaches the uniform distribution (minimum concentration) (Shannon, 1948). The same qualitative results hold if we use the Gini-Simpson index ($1 - \sum_c p_c^2$), a commonly used metric in ecology for species concentration (Gini, 1912; Simpson, 1949). An alternative hypothesis regarding the difference in activity quartiles is that there isn't really a difference, but perhaps users in the higher-activity quartile make several posts in a single community where a lower-activity user makes just one, e.g., $C_1C_1C_1C_2C_2C_2$ vs. C_1C_2. If this were so, we would observe a lower entropy simply due to accidentally choosing a window size that is small relative to the average burst size. However, we verified that this "burstiness" hypothesis does not hold, since the higher-activity users only change communities about 0.5 fewer times than lower-activity ones.

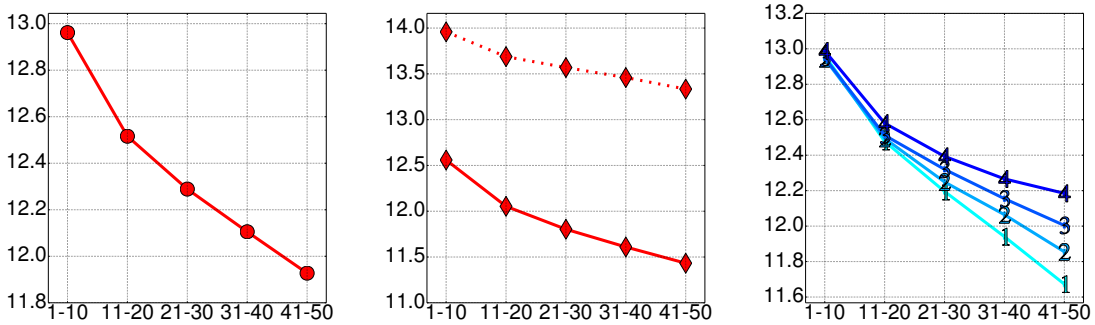


(a) Overall trend (b) Departing status (c) Activity quartile

Figure 4.5: Entropy of community-posting distribution.

Users enter smaller-looking communities (fewer posts per month), but relatively speaking, departing users prefer larger communities. (Fig. 4.6) Engaging with different communities entails a choice between communities of different sizes. A large community can encompass diverse community purposes and member preferences, leading to broader appeal, but at the same time, a large size may dilute personal connection and lead to more conflicts (Ren et al., 2007). Or, size might not have any effect at all. However, Reddit does not provide directly applicable metrics for community size: the number of subscribers or those "online now" can consist mostly of passive observers. To study this question, we set $f(t)$ to log of the number of posts made by the user in the community in month t as a simple metric of how "large" the active portion of a community looks to an incoming user. We observe similar trends when extracting the number of users who posted in a month as the metric.

We note that with respect to this metric of community size, the full-life view, shown in the Appendix (Fig. 4.16a), differs from the fixed-prefix perspective plotted above. In the full-life view, the higher-activity quartile users eventually enter smaller communities than lower-activity quartile users. It seems that they just move more slowly to such communities.

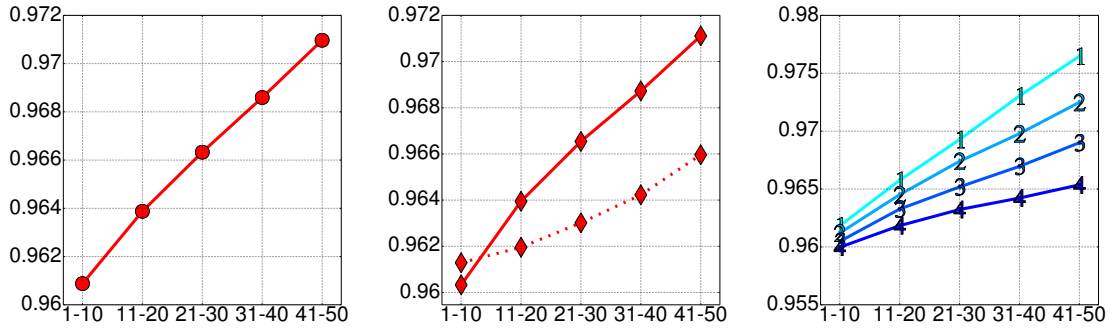


(a) Overall trend (b) Departing status (c) Activity quartile

Figure 4.6: Average $\log_2$(number of monthly posts in communities that a user posts to). Note that it is *not* the case that big subreddits are being abandoned as a whole: despite the availability over time of more and more small subreddits, the number of posts in the popular subreddits continues to increase.

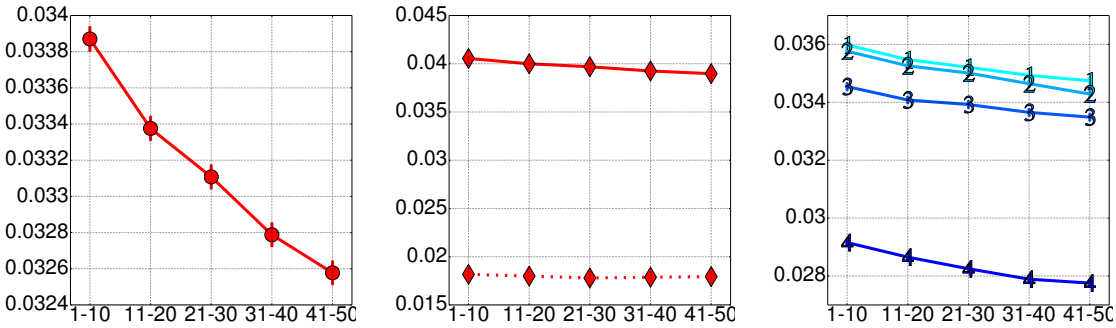
Users post to less similar communities over time, but relatively speaking, departing users prefer more similar ones. (Fig. 4.7) One hypothesis for how people select new communities is that they explore similar communities to those they have visited in the past, because they want more exposure to topics that they are already interested in. On the other hand, perhaps they choose new communities because their interests have changed, implying that they would choose more different communities.

We measure the dissimilarity between communities C_1 and C_2 based on poster overlap, restricting attention to just those communities with at least 1000 posts to ensure sufficient data. Denoting the set of users who ever posted in a community C as U_C, our measure is $1 - \frac{|U_{C_1} \cap U_{C_2}|}{|U_{C_1} \cup U_{C_2}|}$. Note that the dissimilarity between two communities is computed based on their eventual poster set, since we want to capture the "actual", eventual relationship between the two, and so does not change over time. For a window W_i, the overall community dissimilarity $F(W_i)$ is defined as the average of all the pairwise dissimilarities between the communities that the user posted at during that window W_i.



(a) Overall trend (b) Departing status (c) Activity quartile

Figure 4.7: Community dissimilarity based on poster overlap.



(a) Overall trend (b) Departing status (c) Activity quartile

Figure 4.8: Percentage of first singular person pronouns.

The same trends hold if we measure language dissimilarity between communities using the KL-divergence between community language models.

Different activity quartiles. For *all* of the above metrics, users of different *future* activity quartiles manifest significant differences even in their very earliest behavior, although the differences are not as dramatic as those between departing users and lasting users. The curves for the different quartiles always appear in either the order 1,2,3,4 or 4,3,2,1, and the highest-activity quartile curves are always the closest to those for departing users.

4.4.2 Language aspects

The second aspect that we examine is the language that users employ within communities. This examination, and the formulation we apply below, are inspired by Danescu-Niculescu-Mizil et al. (2013), which found that in single-community settings, users first pass through an "adolescent" phase where they learn linguistic norms, but after this phase stop adapting to new norms and become increasingly distant from the community. Our results indicate that this is *not* the case in the multi-community setting. Rather, with respect to part-of-speech tags or stopwords, users do not move farther and farther away from the community distribution; and when (frequent) content words are included, users seem to "stay young", continuously growing closer to the community's language. Surprisingly, departing users are better mimics of the community's language than lasting users are. The bulk of this section provides the experimental evidence, based on various forms of cross-entropy, from which we draw these conclusions.

Additionally, we, like Danescu-Niculescu-Mizil et al. (2013), find that the usage of 1st-person-singular pronouns (e.g., I, me) declines over time,[10] which has been argued to indicate a greater sense of community affiliation (Chung and Pennebaker, 2007; Sherblom, 2009). However, upon closer inspection, the fact that departing posters use these words *less* frequently than those users who end up staying seems problematic for such theories — although one could speculate that the cause is that our departing users start out with strong affiliation needs but become disappointed. These results are shown in Figure 4.8.

Cross-entropy with vocabulary-varying language models. We use cross-

[10] Acronyms such as "TIL" (for "today I learned") were not included.

entropy to measure the distance between (a language model constructed from) a user's t^{th} post and a language model built from all the posts in the corresponding community, C, in that same month $m(t)$. Importantly, we will compute these models based on various choices of vocabulary V; this will reveal that although users' topical-word usage grows closer and closer to that of the community's, their usage in part-of-speech tags and stopwords stabilizes in terms of distance from the community's.

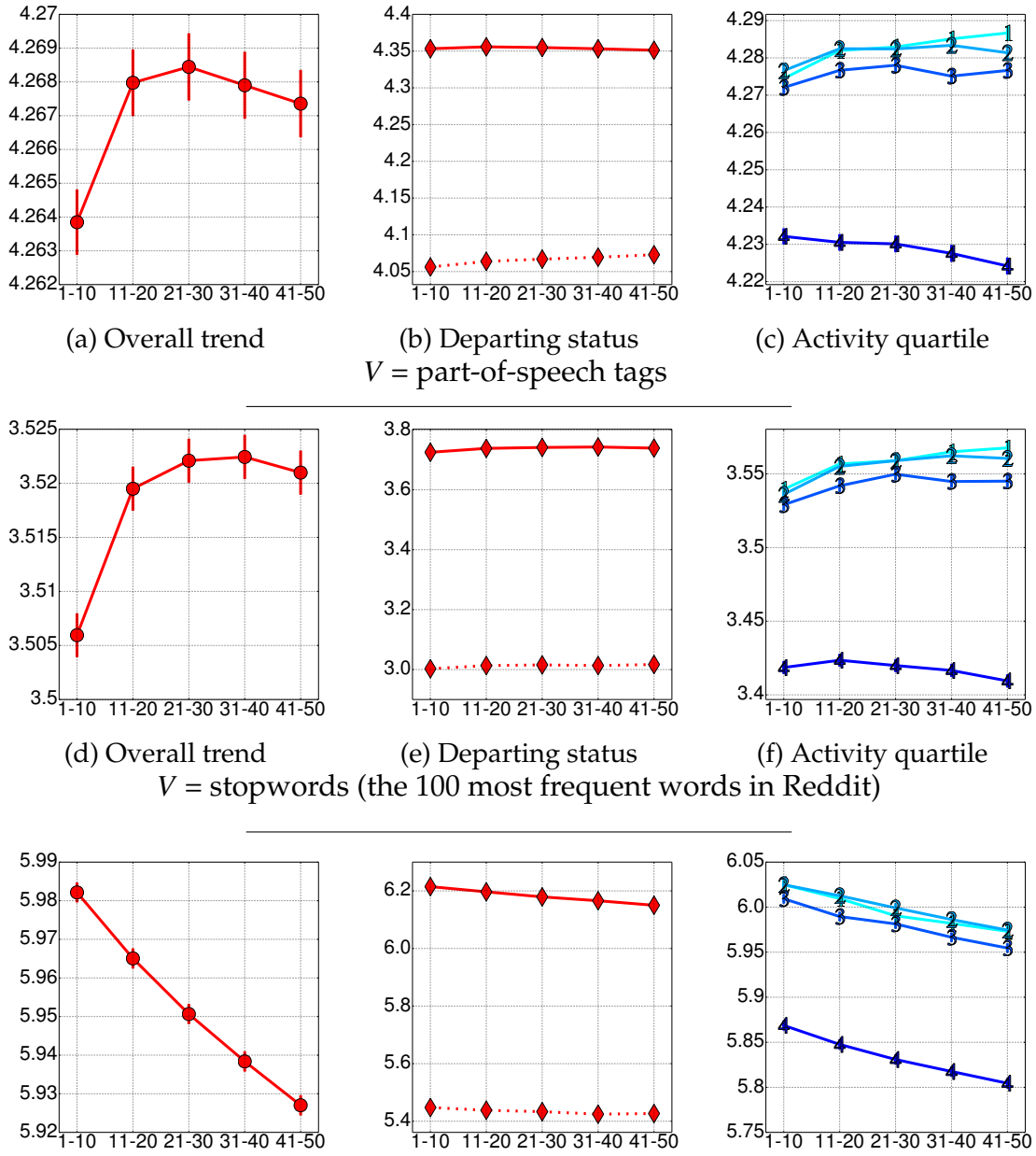
The first step of our V-dependent language-model construction is to replace every instance of any word not in V with the new token "<RARE>". Next, we define the community-based language model to be the distribution over $v \in V \cup \{\text{<RARE>}\}$ given by setting p^C to the relative frequency of v in the concatenation $words^{C,m(t)}$ of all the posts in C during the month $m(t)$. Then, we measure the cross-entropy by

$$f(t) = \frac{1}{|words_t|} \sum_{v \in words_t} \log_2 \frac{1}{p^C(v)}.$$

(This equation shows why we do not need to smooth the community language model: since $words_t$ is a component of $words^{C,m(t)}$, $p^C(v) > 0$ for $v \in words_t$.)

With all of this in hand, Figure 4.9 depicts representative evidence for the conclusions we drew at the beginning of this section. Specifically, the evidence consists of cross-entropy values for V chosen to be 46 parts-of-speech tags, the most frequent 100 words in Reddit, or the most frequent 1000 words in Reddit. Trends for V set to the 500 or 5000 most frequent words are similar to the most frequent 1000 words.

Technical aside: the potentially confounding factor of rare words interacting with community posting volume. We also used a "full" vocabulary that contains all words that appear more than 100 times in Reddit (180K types), but do



(a) Overall trend (b) Departing status (c) Activity quartile

V = part-of-speech tags

(d) Overall trend (e) Departing status (f) Activity quartile

V = stopwords (the 100 most frequent words in Reddit)

(g) Overall trend (h) Departing status (i) Activity quartile

V = stopwords + content words (1000 most frequent words)

Figure 4.9: Distance from the community language model. The rows indicate different choices of vocabulary V.

not show the results here. This is due to the fact that for large vocabulary sizes, what appears to be differences in language matching can actually be merely a side-effect of one class of users posting in more-voluble communities. The ar-

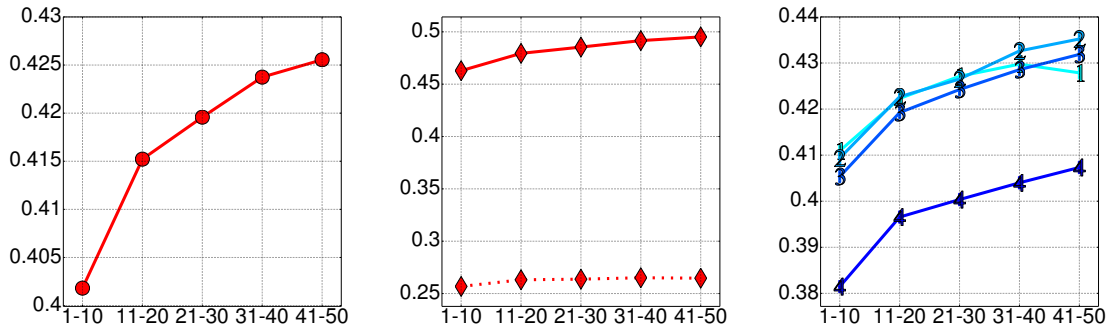
gument runs as follows. The full vocabulary allows for many words v' with low frequency in the community — say, 1 — to contribute to the cross-entropy computation. The probability estimate $p^C(v')$ for such words is $1/|words^{C,m(t)}|$ (where t is chosen appropriately). So, in groups where $|words^{C,m(t)}|$ is large, the contribution of such v' to the cross entropy is bigger than it would be for sub-communities where $|words^{C,m(t)}|$ is small. This concern cannot be alleviated simply by sub-sampling a community's posts, since the true root of the problem is rare words, not just the length and number of posts in the community per se.

4.4.3 Feedback aspects

A final question that Reddit data allow us to easily answer is, how are users received by other members of the community? For each post, Reddit provides the difference between the number of upvotes and number of downvotes. Because the average value of this difference can vary among different communities, we measure the feedback that users get by the relative position of this difference among all posts in the community that month, i.e., how often the posts made by a user outperform the "median post" in a community. For each index t, we define $f(t)$ as $I(feedback_t > median(C_t, m(t)))$, where $median(C, m)$ represents the median vote difference in community C in month m.

Surprisingly, the feedback that 50+ posters receive is *continually* growing more positive, although the rate slows over time (Fig. 4.10). However, the growth is small compared to the drastic differences between departing users and lasting users. Even departing users get more-positive feedback over time, but the increase is not as great as for lasting users. Users in the top activity quar-

tile also fare worse, although as shown in the relative perspective (Fig. 4.16b), they catch up in the later stages of their life. The results are consistent if we measure how often posts outperform 75% of the community's posts.



(a) Overall trend (b) Departure status (c) Activity quartile

Figure 4.10: Success rate at outperforming the median vote difference.

4.4.4 Recap

In all three aspects that we examined, users with different future activity levels manifest significant differences in their trajectories of multi-community engagement. Interestingly, users that eventually depart seem "destined" to do so even from the very beginning, since the curves for the departing vs. lasting users generally start out apart and maintain or increase that distance over time. Meanwhile, there are smaller but significant differences in these metrics between users at different activity quartiles. It is important to note that some metrics can be correlated (e.g., number of unique communities and entropy). However, none of the metrics determines another, so we believe discussing each one of them was valuable.

Another interesting phenomenon we consistently observe is that for all our metrics, users in the top activity quartile are the closest to the departing users

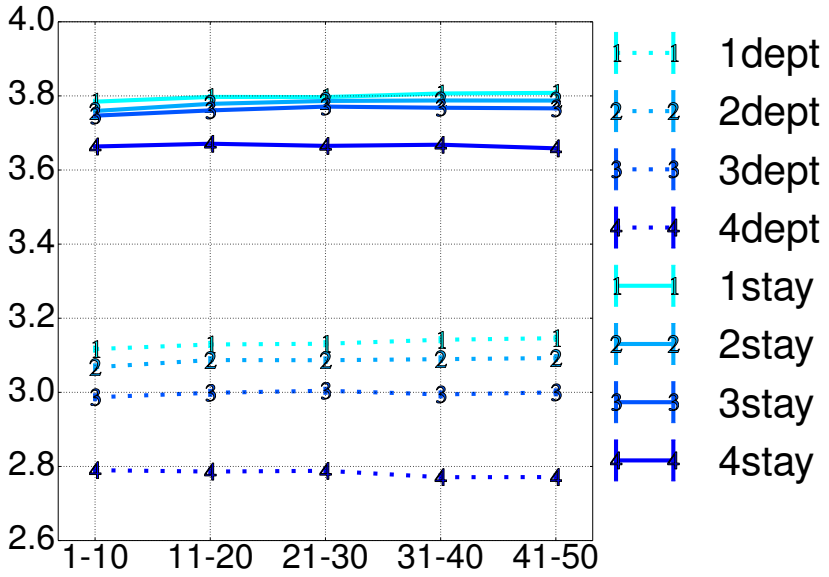


Figure 4.11: Interplay between departure status and activity quartiles. y-axis: distance from the corresponding monthly language model when setting the vocabulary to the 100 most frequent words. *i*dept refers to departing users in the *i*-th quartile; *i*stay refers to lasting users in the *i*-th quartile.

in the first 50 posts (a direct comparison for language is shown in Fig. 4.11).

4.5 Predicting departure and activity levels

We have now seen many properties of multi-community engagement that correlate with user activity. To examine the effectiveness of these properties in prediction, we set up two different prediction tasks that correspond to how we measure users' future activity level in §4.3:

- Future departure status. In this task, we predict whether users abandon Reddit in the future. We use F1 for evaluation, with the minority class (departing users, as defined in §4.3) as the target class. We use weighted L2-regularized logistic regression as classifier.

- Future total number of posts. This is a regression task where the goal is, for a given user, to estimate $\log_2$(future number of posts). We employ L2-regularized support vector regression, and measure performance by root mean squared error (RMSE).

Each instance consists of a user's first 50 posts.

Baseline and features. We consider the following feature sets, where for window-based features we set the window size $w = 10$, thus deriving 50/10 = 5 values.[11]

- Average time-gap between posts. Danescu-Niculescu-Mizil et al. (2013) states that this is an effective feature used in prior work on churn prediction (Dror et al., 2012; Yang et al., 2010). *Thus, this feature by itself serves as our (strong) baseline.*
- Multi-community aspects (henceforth "sub info"). This includes number of unique communities, number of "jumps", entropy, and Gini-Simpson index based on the user's community-posting distribution, as well as mean log "apparent" community size as defined in §4.4.1. Similarity between communities is not used because information about the future is incorporated in the way we compute it.
- Language aspects ("lang" for short). This includes cross-entropy with the monthly community language model for the following choices of vocabulary: part-of-speech tags; the top 100, 500, 1000, 5000, 10000 most frequent

[11] Alternatively, one could set $w = 50$, thus extracting features from all 50 posts in a single batch. This approach turns out to be poorer than using 5 windows because trend information is not captured.

words; and the full vocabulary as defined in §4.4.2. Additionally, we include the proportion of 1st-person-singular pronouns and post length in words.

- Feedback aspects. This includes the fraction of posts that outperform 50% and 75% of all of the corresponding month's worth of the community's posts in terms of positivity of feedback. Refer back to §4.4.3 for more information.

For entropy, Gini-Simpson index, and number of unique communities, we include the value for all 50 posts, since for these features, the values for all 50 posts are not simply the average of the values from 5 windows of 10 posts. We also use the index of the window with the largest value and the smallest value as features, following Danescu-Niculescu-Mizil et al. (2013). All features are linearly scaled to $[0, 1]$ based on training data.

Experiment protocol. In both tasks, we perform 30 randomized trials. In each trial, we randomly draw 20,000 users from our dataset as training data and a distinct set of 5,000 users as testing data. We use 5,000 users from the training data as validation set. We use LIBLINEAR (Fan et al., 2008) in all prediction tasks. For significance testing, we employ the paired Wilcoxon signed rank test (Wilcoxon, 1945).

The standard procedure for generating learning curves would be to only look at the *first* x posts as x varies, $x = 10, 20, 30, 40, 50$. A non-obvious but ultimately fruitful idea we introduce here is to contrast the effectiveness of the information in the early part of each 50-post instance with that of the late part of the 50-post instance. That is, we compare the performance if we use the *first* ("fst" in our plots) x posts with the performance of using the *last* x posts.

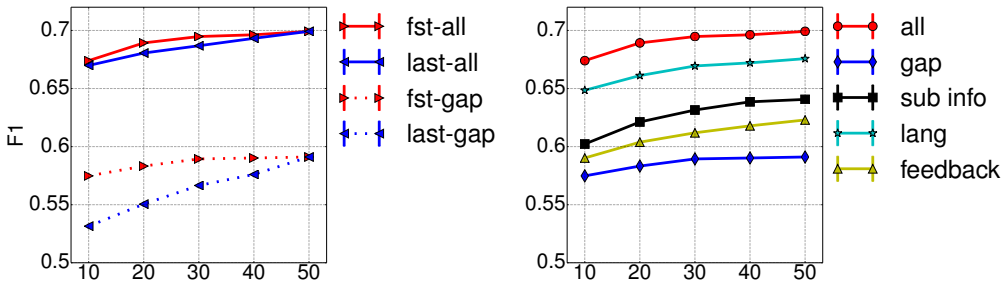
(One might expect later periods to be more predictive, given that they are more recent. But surprisingly, we will see that when we predict departure status, we find that earlier information is more useful, which again suggests that departing users are "destined" to leave from the very beginning.)

4.5.1 Predicting departing status

Basic comparisons. (Figure 4.12a) Using all features outperforms a strong baseline that uses time-gap features by 18.3% — the difference between an F1 of .699 and an F1 of .591 — which shows the effectiveness of features drawn from multi-community engagement.

The performance of the first x posts is always above that of the last x posts. This suggests that the initial information is more predictive of eventual departure. Note that for 50+ posters, departure is quite "far away" from the initial posts. In fact, using all features drawn from only the first 10 posts outperforms time-gap features extracted from all 50 posts. Thus it may be very important for designers of social systems to make sure that users start well, perhaps through positive feedback or by recommending communities to post in (which can differ from the communities one might recommend that a user reads).

Feature-set analysis. (Figure 4.12b) In predicting departure, it is most useful to know how well users match a community's language. The second most useful features are the patterns of community visitation. Language-matching, community-trajectory, and community-feedback features all outperform time-gap information, which suggests that how users interact with different communities is more predictive than activity rate in predicting whether 50+ users will



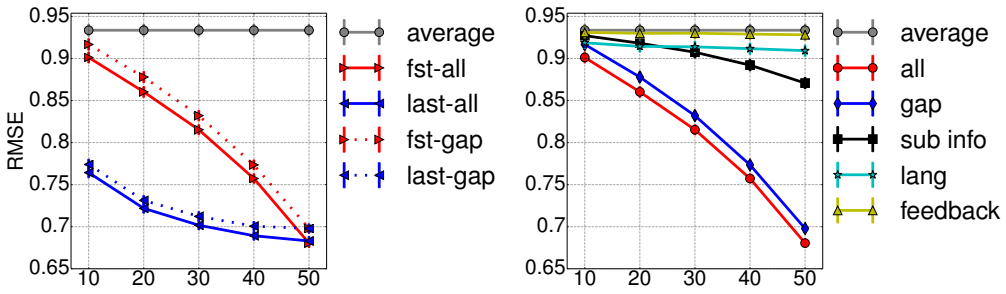
(a) Window comparison (b) Feature comparison

Figure 4.12: Results for predicting departing status. y-axis: F1 measure. In Fig. 4.12a, the dashed lines show the performance of the baseline, timing-based features; the solid lines show the performance of using all features. Red lines show the performance using the first x posts, while blue lines show the performance using the last x posts. Fig. 4.12b: performance of different feature sets. All differences for 50 posts are statistically significant according to the Wilcoxon signed rank test ($p < 0.001$).

leave.

4.5.2 Predicting activity quartile

Comparisons with the baseline. (Figure 4.13a, 4.13b) In contrast to the case just discussed of predicting departure status, time-gap between posts is a much stronger feature in predicting future total number of posts. This is plausible because for these 50+ posters, time-gaps in posting determine how many posts that people can physically make. However, adding all the features based on multi-community engagement still improves the performance over timing information to a statistically significant degree. Prior work has shown that adding language features can lead to big improvements over timing-based features (Danescu-Niculescu-Mizil et al., 2013); the relatively small improvement in our experiment may be due to the fact that the datasets in Danescu-Niculescu-Mizil et al. (2013) have a longer history than ours.



(a) Window comparison (b) Feature comparison

Figure 4.13: Results for predicting $\log_2$(future total number of posts). y-axis: RMSE, the smaller the better. The line styles are the same as in Fig. 4.12. "Average" shows a baseline that always predits the mean value in the training data. All differences for 50 posts are statistically significant according to the Wilcoxon signed rank test ($p < 0.001$).

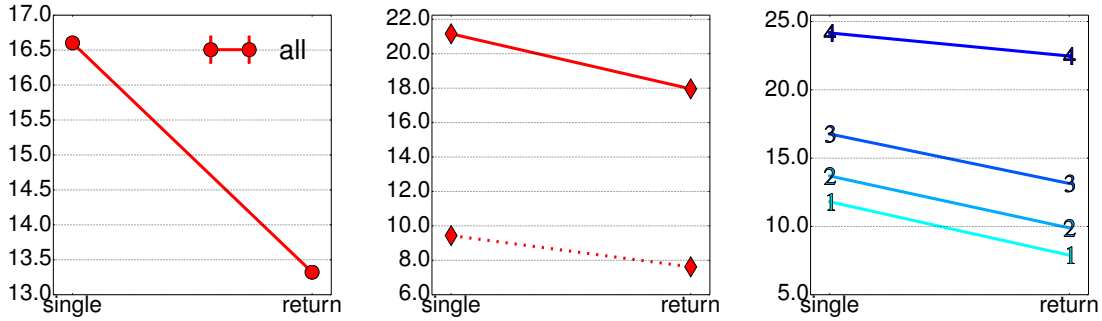
Also, using the last x posts is much more effective than using the first x posts. There thus seems to be different factors affecting 50+ posters with respect to deciding whether to remain in a community versus deciding to be highly active in it.

4.6 When do users abandon their posts?

We have already seen that (our) users constantly try out new communities, but we have not yet addressed a related question of practical importance to community maintainers, as well as of inherent social-scientific interest: how much and why do users *abandon* communities?

We can frame the "how much" issue succinctly by asking the following question. Suppose we partition the set of communities a user visits into (1) those that he or she abandons after just a single post, and (2) those that he or she posts at least twice to. Which set — the single-post communities or multiple-posts

communities, is larger, on average? We claim that the answer is not a priori obvious[12]. But the data shows that users rack up more abandoned communities than return engagements, as depicted in the figure below. This suggests that although users are constantly willing to post to new groups, they are often only giving these new groups one shot.



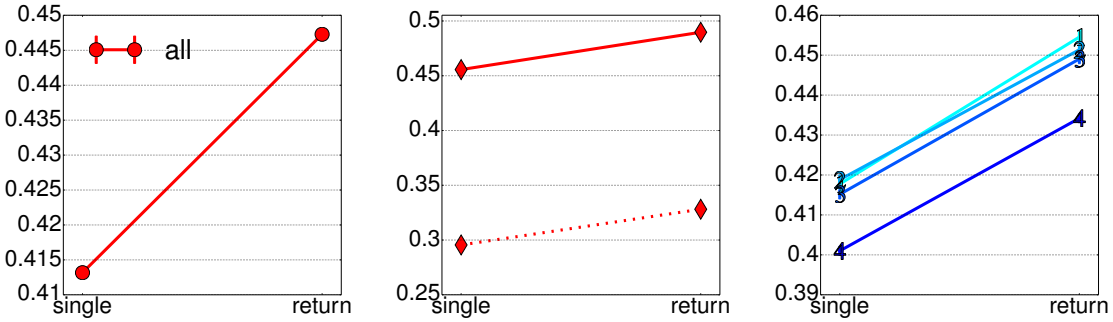
(a) Overall (b) Departing status (c) Activity levels

Figure 4.14: Comparison of the average number of communities where a user posts only once vs. more than once.

What is happening in the single-post communities that causes a user to stop posting in them immediately? We find that positivity of feedback (in Reddit, the difference in upvotes and downvotes) may play a substantial role, as shown by the figure below. Figure 4.15 is based on the *very first* post that a user makes in every community they posted in; it plots the percentage of such first posts that received a feedback score above that of the median feedback score in the respective community.

One reason that this is interesting to note is that our results contrast with previous findings of the power of *negative* feedback for predicting repeated commenting (Cheng et al., 2014); we conjecture that the difference is due to different impulses driving posting vs. commenting behavior.

[12]Recall the title of Duncan Watts' recent book "Everything Is Obvious: *Once You Know the Answer".



(a) Overall (b) Departing status (c) Activity levels

Figure 4.15: Users get better feedback for the first post in the communities that they eventually returned to than for the communities that they ended up making only a single post in. y-axis: average fraction of a user's post with feedback score better than the community's median. We exclude users that have only single-post communities or only multiple-posts communities, thus controlling for individual-user characteristics to some extent. All differences between connected points are statistically significant according to the paired t-test ($p < 0.001$).

4.7 Do users speak differently in different communities?

So far we have revealed interesting and sometimes arguably counterintuitive properties of multi-community engagement, and demonstrated that they are effective cues in predicting a user's future activity level. But an additional fascinating and orthogonal question is: when users participate in multiple communities, to what degree are their actions stable *across* settings? To look at this question is to contribute another piece of evidence to the "situation vs. personality" debate (Kenrick and Funder, 1988; Donellan et al., 2009): how much of our behavior is determined by fixed personality traits, versus how much is variable and influenced by the specific situation at hand? Or, to put it a bit more dramatically, are you fundamentally the same person at work as you are at the gym?

Here, we study the question with respect to language use. The overall mes-

sage is, *even after topic-specific vocabulary is discarded* (after all, it wouldn't be interesting to find that people use gym-related words at the gym that they don't use at work), individuals *do* employ different language patterns in different communities. The way we determine this is conceptually straightforward: we check whether it's possible to tell which community a user's posts come from based just on the distribution of stopwords or non-content-words within their posts.

Specifically, given a vocabulary V of non-content words, we create classification instances from the 227K triples that exist in our data consisting of (1) a user u, (2) words of u's first 25 posts in some community C_1, and (3) words of u's first 25 posts in a different community C_2. We compute the cross entropy of each post against the corresponding monthly language models, over the restricted vocabulary V, constructed from each of the two communities C_1 and C_2. We divide these 25 posts into windows of 5 posts and take the average cross entropy in each window, in order to be more robust and potentially capture trends, but it simplifies exposition to think of just a single post. Add-1/|V| smoothing is applied to all language models concerned. We then use these non-content-word cross-entropies as features to guess which of (2) and (3) came from community C_1 in a binary classification task: we concatenate features from (2) and (3) to form either $[(2), (3)]$ or $[(3), (2)]$, and label the former as positive and the latter as negative.

We run experiments for several choices of V: parts-of-speech, the 100 most frequent words in Reddit, and the 500 most frequent words in Reddit. The first two choices definitely do not include topic-specific words, and the latter will not include many (there are 180K words in the full Reddit vocabulary), and so

these choices may be taken to represent a user's language *style* (Argamon and Levitan, 2005; Danescu-Niculescu-Mizil et al., 2011). If the user's style does not change from community to community, then the cross-entropy features mentioned above will not be helpful for determining that item (1) comes from C_1 and not C_2; thus, accuracy at matching language model to community would be 50%. But, as shown below, the average accuracies, utilizing logistic classification, of 30 random-split experiments (10K tuples for training and development, 2500 for testing) for each choice of V are (statistically) significantly above 50%:

V	accuracy
parts of speech	62.5%
most frequent 100 words	56.0%
most frequent 500 words	61.4%

4.8 Related work

Anthropologists, psychologists and sociologists have looked at some questions regarding multi-community engagement, often in the context of interaction with new social circles or cultures (Bühler, 1935; Hurtado, 1997; Berry, 1997). Recently, computer scientists have turned to examining multi-community engagement data available online (Backstrom et al., 2006; Adamic et al., 2008; Vasilescu et al., 2013a, 2014; Lakkaraju et al., 2013; Guimarães et al., 2015). Our work differs by focusing on the following specific problems: (a) characterizing full community-trajectory sequences, as opposed to looking at pairwise community transitions (Backstrom et al., 2006; Vasilescu et al., 2013a, 2014); (b) revealing how properties of these trajectories correlate with a user's future cross-community activity — we incorporate but also go beyond language-based fea-

tures, as inspired by previous within-community work (Danescu-Niculescu-Mizil et al., 2013; Rowe, 2013), and timing-based features (Dror et al., 2012); (c) considering the effect of each community's positive and negative feedback, which may shed light on why users choose some communities over others.

Researchers have also been working on predicting users' survival (also known as churn prediction) (Dasgupta et al., 2008; Dror et al., 2012; Yang et al., 2010) and activity level (De Choudhury et al., 2010; Zhu et al., 2013). They focus on the single-community setting. A number of studies examined community-level evolution or the success of individual communities (often websites) (Iriberri and Leroy, 2009; Kairam et al., 2012; Ludford et al., 2004; Zhu et al., 2014a,b), whereas our work focuses on the life cycle of users.

4.9 Concluding discussion

Summary. We have investigated properties of multi-community engagement; this is a setting that has not received much computational research attention before, and yet is important because it encompasses many online and physical situations. In this first large-scale study of the phenomenon, we have found a number of sometimes counterintuitive but robust properties — some involving choice of community, some involving language use within communities, and some involving feedback from communities — revolving around the discovery that users "wander" and explore communities to a greater extent than might have been previously suspected.

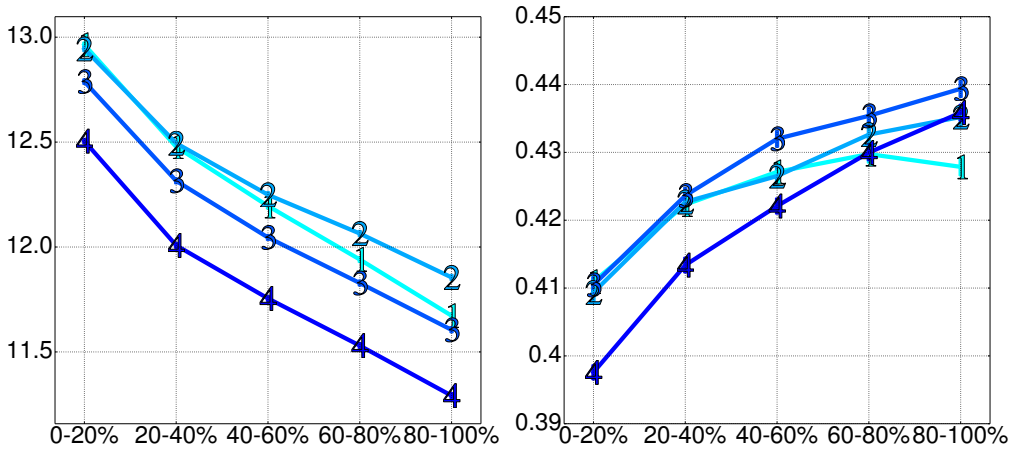
Limitations and further directions. We focused on posting, but commenting

and other related behaviors are very interesting subjects for future study. Our study is quantitative and observational. Qualitative studies, or controlled experiments regarding the design implications in §4.2, can further improve our understanding.

It is important to note that our study is limited to "50+ posters" so that we would have enough history per user to observe a relatively long trajectory. This is an unusually engaged group of users that comprises 5.9% of our users. We have not addressed the question of how multi-community engagement is exhibited by users who are not as active.

The notion of considering users to exist in a multi-community setting can in principle be extended to looking at user behavior across multiple websites or apps. With the advent and adoption of multiple-website services such as OpenID, observing users at that scale of multi-community engagement may well become quite important in the future.

There are many more challenging questions that arise from taking a multi-community perspective. For example, are the particularly nomadic treated differently? What is multi-community engagement like in real life, considering the cost of switching? How can we extend current theories and principles in community design to a multi-community setting? Further understanding of these questions is crucial for on- and off-line community design and an exciting direction for future work. In the next chapter, we will explire



(a) (b)

Figure 4.16: Comparison of different Reddit activity quartiles from the full-life perspective. (a): mean $\log_2$(monthly number of posts). (b): fraction of posts that outperform the median value of feedback positivity in the corresponding month and community.

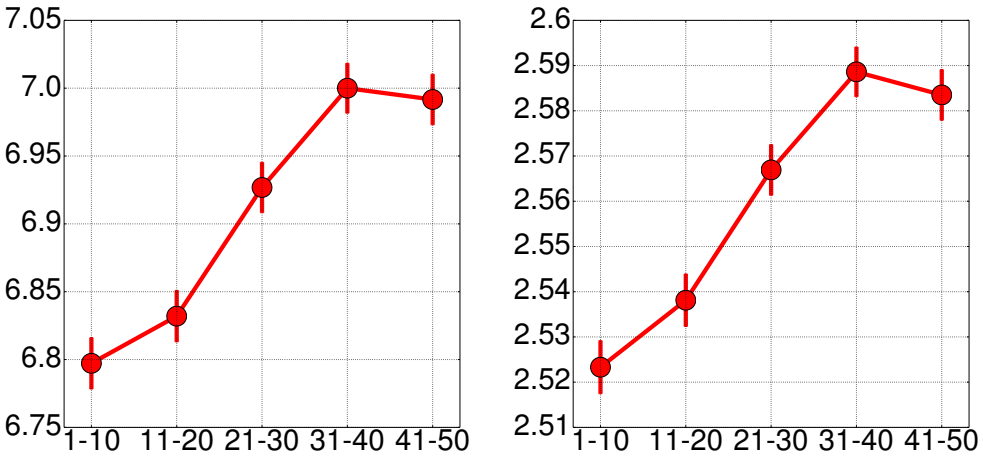
4.10 Appendix

Full-life view for users in Reddit. In general, the overall trends and differences between departing users and staying users are the same as in the fixed-prefix view. But in terms of activity quartiles, there are some interesting differences. For example, the ordering of the activity quartiles with respect to mean $\log_2$(number of posts that month) completely reverses itself (compare Fig. 4.16a to Fig. 4.6c). For feedback, as users receive better feedback over time, users in the top activity quartile receive worse feedback in the beginning and catch up later in their life (Fig. 4.16b). These results are natural consequences of the trend developing over time. This suggests that the trends that we observe are robust over user life.

Fixed-prefix view for researchers in DBLP. In DBLP, authors span more conferences per window over time (Fig. 4.17a) in an increasingly scattered fashion

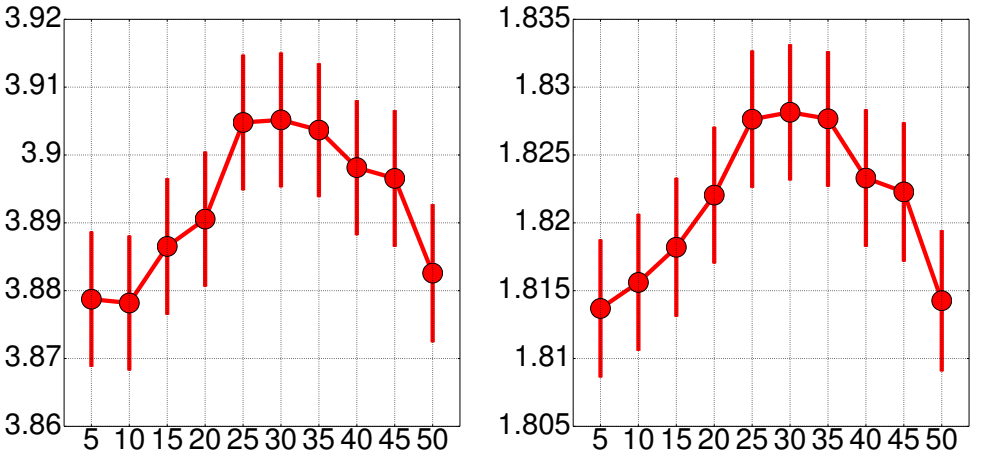
(Fig. 4.17b), but in contrast to Reddit, there is saturation in the last two windows. Perhaps this suggests that as researchers become very senior, they publish more papers in some favorite set of venues.

When a very small window size is considered (w=5), the number of unique conferences and within-window entropy first increase and then decrease (Fig. 4.17c and 4.17d). But, changing the window size does *not* affect our central observation in Fig. 4.1 that 50+ researchers are publishing in new conferences at a relatively consistent rate over the years.



(a) (b)

Window size of 10.



(c) (d)

Window size of 5.

Figure 4.17: Fixed-prefix view for researchers in DBLP. (a,c): number of unique conferences per window. (b,d): entropy of the conference publishing distribution per window.

CHAPTER 5

MULTIPLE COMMUNITIES: A STORY OF HIGHLY RELATED COMMUNITIES

5.1 Brief overview

When large social-media platforms allow users to easily form and self-organize into interest groups, highly related communities can arise. For example, the Reddit site hosts not just a group called food, but also HealthyFood, foodhacks, foodporn, and cooking, among others.[1] Are these highly related communities created for similar classes of reasons (e.g., *true* to distinguish one as a better community and *advice* to focus on helping fellow members)? How do users allocate attention between such close alternatives when they are available or emerge over time? Are there different types of relations between close alternatives such as sharing many users vs. a new community drawing away members of an older one vs. a splinter group failing to cohere into a viable separate community? We investigate the interactions between highly related communities using data from `reddit.com` consisting of 975M posts and comments spanning an 8-year period. We identify a set of typical affixes that users adopt to create highly related communities and build a taxonomy of affixes. One interesting finding regarding users' behavior is: after a newer community is created, for several types of highly-related community pairs, users that engage in a newer community tend to be *more active* in their original community than users that do not explore, even when controlling for previous level of engagement.

Most of the contents in this chapter are published in Hessel et al. (2016). This

[1] Throughout this chapter, we use sans-serif fonts for group names.

is joint work with Jack Hessel and Lillian Lee.

5.2 Introduction

Social networks are in constant flux, with new communities forming and old communities dying over time. On websites such as Facebook and Reddit, users have complete freedom to create communities at their own discretion. This has led to a very large number of communities arising organically from user initiative, for a variety of reasons. One reason is to create divisions that satisfy the need to better organize discussions; in fact, community design theory argues that "a growing Web community needs subdivisions which might be represented as towns, neighborhoods, topics, categories, conferences, or channels, depending on your metaphor" (Kim, 2000; Jones and Rafaeli, 2010). Or, new groups can develop because of religious, political, or other schisms; online examples include groups whose very names attempt to connote superiority to others, e.g., the subreddits trueatheism vs. atheism. Other reasons surely exist. The tremendous reach of modern social media provides researchers much greater data to examine these social processes at scale.

An interesting and frequently occurring version of the group creation process is that a new concept or culture may gain in popularity and, in a meme-like fashion, draw users to create a new community by using that concept as *an affix*[2] of their community name. For example, on Facebook, after the creation of the OMG Confessions group, anonymous confession pages with names combining a college with the word *confession* or *confessional* proliferated to the degree that

[2] An affix is either a prefix or a suffix.

Table 5.1: The 10 most common Reddit group-name affixes.

Affix	Example	# Pairs
s	auto, autos	63
porn	space, spaceporn	26
circlejerk	hiphop, hiphopcirclejerk	23
ask	science, askscience	21
shitty	ideas, shittyideas	17
music	running, runningmusic	17
help	tech, techhelp	11
2	dota, dota2	9
true	atheism, trueatheism	9
learn	math, learnmath	9

one can now find a confession page for almost every university campus. (Birnholtz et al. (2015) examine what kind of questions people ask on such pages.) Table 5.1 shows some examples from Reddit: the second column shows pairs of subcommunities where the name of one is a modified form of the other (ignore the third column for now).[3]

In this work, we investigate highly related communities that are based on affixes. An understanding of these highly related communities may help community organizers identify subtopics in a community and create an appropriate subdivision to cultivate focused discussions, or monitor subgroups that potentially feel marginalized or underserved, and decide whether to change community norms or create a dedicated community for that subgroup.

Despite the ubiquity of such affixes, and their appeal as easily-identifiable (albeit sometimes imperfect) instances of the important phenomenon of highly related communities, little is known about canonical affixes and the activity in

[3] An additional, whimsical example from Reddit is *random_acts_of_*, indicating people asking for or sending free things to others. Instantiations include random_acts_of_pizza, random_acts_of_amazon, and random_acts_of_books. Althoff et al. (2014) used random_acts_of_pizza to study effective ways to ask for a favor.

the resultant highly related communities. For instance, are neighborhoods, topics, and channels enough to capture all possible affixes? Are there classes of affixes that are generally applicable? Perhaps different affixes behave in different ways. Moreover, once a highly related community is created, how does it interact with the existing community? Will it overtake it? Will the two share the same user base? One of our goals is to analyze user behavior in the existing community *after they participate in the new community.*

Organization and contributions. In this chapter, we construct a dataset from Reddit and present the first large-scale study on the coexistence of highly related communities. Details about the dataset are introduced in "Dataset Description".

Our first contribution is to characterize the space of *affixes*. We build a taxonomy of common affixes that users adopt to create highly related communities. For instance, we identify a category of "parody" affixes (*circlejerk*, *shitty*, *funny*, *lol*, *bad*). This category generally shares the same user base with its corresponding unaffixed community. On the other hand, we identify a category of "derivative" affixes (*meta*, *anti*, *srs*, *post*, *ex*) that likely attract different user bases. Surprisingly, a non-trivial fraction of affixed communities exist before the unaffixed ones. Also, an interesting class of *spinoff* communities arises where early participants in the new community come from the existing community.

Our second contribution is to introduce a framework for analyzing users who try out spinoff communities (dubbed "explorers") and comparing them to "nonexplorers" who never leave the original subreddit. We make the surprising observation that in multiple classes of affixes, users who explore spinoff communities are *more active in the original communities after exploring* when compared to similarly active users who never tried the alternative. This resonates with the

findings in the previous chapter that users who "wander" to different (potentially completely unrelated) groups tend to stay active longer on the site as a whole. Our observations may suggest that spinoff communities generally serve a complementary rather than competitive role in multi-community settings.

Finally, we summarize related work and offer some concluding thoughts.

5.3 Dataset Description

Our starting point for understanding highly related communities, affixes, spinoffs, nonexplorers, and explorers is an examination of *topically related communities*. As such, we compile a dataset from `reddit.com`, a site where users are allowed to create communities called subreddits at their discretion and that we have studied in previous chapters. Users can name the subreddits that they create so that like-minded people can identify them effectively. As a result of unmoderated creation and limitless naming possibilities, there are a wide variety of subreddits on Reddit, e.g., funny, worldnews, politics, IAmA, todayilearned, etc. On these subreddits, users submit link-based posts or text-based posts, comment on others' posts, and up/down vote posts and comments. We construct a dataset that includes all activities on Reddit from its inception until 2014, an 8-year period, by combining two data sources: a post dataset that was organized in the previous chapter, and all comments data extracted by Jason Baumgartner.[4] We focus on communities that are active and that enjoy a reasonable number of users. Specifically, we require all communities to include at least 300 unique users that made posts. This left us with just under 5.7K communities.

[4]Information is available at https://pushshift.io. The dataset in the previous chapter was also originally extracted by Jason Baumgartner.

Table 5.2: Summary statistics for our Reddit corpus. Posts are from the previous chapter and include all posts on Reddit from its inception in 2006 to February, 2014. All comments on these posts up until November 2014 were drawn from Jason Baumgartner's comment dataset.

Data type	count
Subreddits	5,692
Posts	88M
Comments	887.5M

Table 5.2 presents basic statistics of this dataset.[5] The metadata for the Reddit conversation trees that we used here is available for download.[6]

As discussed in the introduction, user-defined subreddit names are an important indicator of relationships between highly related communities (e.g., food vs. HealthyFood). We first retrieve all possible pairs of communities where one community name is the other's suffix or prefix, ignoring case (food is the suffix of HealthyFood, ignoring case). We refer to the difference between the names in a pair as the *affix*. For instance, *healthy* is the affix in the pair food vs. HealthyFood. There are around 4K such pairs over our dataset.

Using common affixes as a starting point allows us to discuss the the space of possible highly related communities. For example, this framing allows us to make statistical observations about all pairs with *healthy* or *true* as affixes. Note that we omit some interesting highly related communities pairs by focusing on affixed pairs. One example is TwoXChromosomes, a very popular "subreddit ... intended for women's perspectives," and TrollXChromosomes, its satirical counterpart.

[5]The statistics reported here include posts and comments made by users who deleted their accounts and banned accounts.

[6]http://goo.gl/sHUfhC

Identifying topically related communities. Unsurprisingly, not all pairs of communities identified through affixes are actually highly related communities. An example is "ru" and "rum;" the first one is a Russian community while the second one is about the liquor. In order to quantify subreddit similarity, we compute the content similarity between pairs of communities. As suggested in Singer et al. (2014), subreddits can focus on text posts in addition to link-based posts. Therefore, we employ a method that can account for either link-dominant or text-dominant subreddits. Specifically, we use Jaccard similarity between the set of links to capture similarity based on links,[7] and use Jensen-Shannon divergence between topic distributions (derived from a topic model trained on 6.6M text posts) to capture similarity based on text, following Hessel et al. (2015). Since these two metrics are not comparable by raw value, we compute the full background distribution of topic similarity scores based on all 1.62M possible pairs of the 5.7K communities in our dataset and compute the percentile of each affix pair in each distribution.

We consider a pair of communities to be topically related if either link similarity is above the 90th percentile *or* topical similarity based on text is above the 90th percentile. Accounting for our definition of topical similarity yields just over 1.7K pairs from our original set of 4K.

The last step of our preprocessing is to identify generalizable affixes that are commonly used in these highly related communities. We count the frequency of affixes and keep affixes that occur at least three times, so that all affixes in the final dataset carry a general meaning (it is not possible to make general statements about affixes that only occur once). This step brings us to 99 affixes

[7] Jaccard similarity is defined as $\frac{A \cap B}{A \cup B}$, where A and B are the set of links from two subreddits respectively. We have used Jaccard similarity with a different definition in Chapter 3.

Table 5.3: A taxonomy of affixes.

Adjective-like	
"better"	true, plus
"parody"	circlejerk, shitty, funny, lol, bad
"derivative"	post, ex, meta, anti, srs
"genre"	classic, fantasy, indie, folk, casual, dirty, classic, metal, academic, 90s, free, social
"nsfw"	nsfw_, nsfw, asian, trees, gonewild, gw, r4r, tree
Verb-like	
"learning, improvement"	ask, help, learn, advice, hacks, stop
"action"	exchange, randomactsof, trade, trades, classifieds, market, swap, random_acts_of_, requests, invites, builds, making, mining, craft
Noun-like	
"place"	uk, reddit, chicago, us, dc, steam, canada, american, boston, android, online, web
"medium"	porn, pics, music, memes, videos, vids, comics, apps, games, gaming, game
"subject"	science, news, dev, servers, tech, tv, guns, recipes, city, u, college, man, girls
Minor	
"equivalent, competition"	s, al, ing, the, alternative
"generation"	2, 3, 4, 5
"modifier"	ism, n, an

and 572 pairs of highly related communities distributed between them.

5.4 Characterizing affixes

The goal of this section is to explore the types of canonical affixes that users on Reddit utilize. To accomplish this exploration, we first build a taxonomy of common affixes to better understand their basic properties and relationships.

Next, we explore the temporal characteristics of the pairs. In general, we observe an accelerating culture of creating highly related communities, meaning that highly related communities are being created at increasing rates. We also observe that, in most cases, the affixed community in a pair was created after the unaffixed one, even though there is a non-trivial fraction that went the other way, e.g., ukpolitics and uspolitics both existed before politics. We further explore whether the newer community "overtakes" the older one in popularity. We then offer potential rationales that may help explain the surprising finding that a *quarter of the newer communities are more active.* The final characteristic that we examine is whether the newer community actually shares a user base with the older one, at least when the new one is forming. Despite the high similarity both in community name and in content, almost half of newer subreddits in pairs are not, in fact, born out of their older partners.

5.4.1 The space of affixes

In order to achieve a basic understanding of what canonical affixes users adopt to create new communities, we first build a taxonomy of the 99 affixes from the dataset section in Table 5.3.

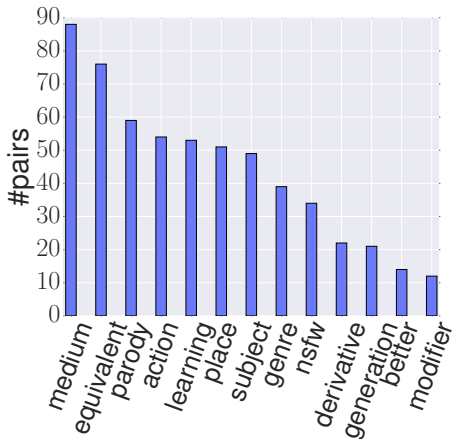
We start with a coarse structure based on part-of-speech. Among the adjective-like, the largest category is based on "genre", e.g., rock vs. classicrock. Some other very interesting classes also arise: "better", which indicates a certain level of superiority (e.g., atheism vs. trueatheism); communities dedicated to "parody" where users are likely aware of the culture in the unaffixed one (e.g., history vs. badhistory); and "derivative", which probably attracts a very different

audience (e.g., war vs. antiwar). In fact, *anti* and *meta* can be recursive, e.g., jokes, antijokes and antiantijokes.

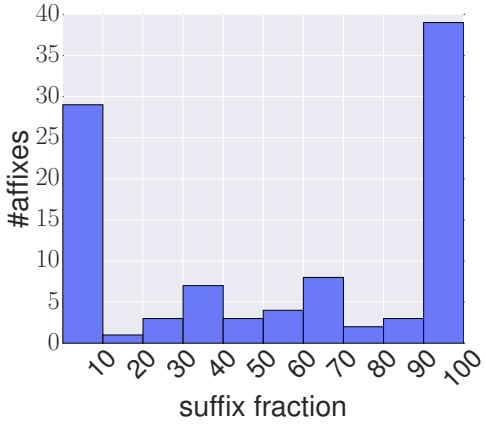
Among the verb-like affixes, a class of self-improvement or learning communities exists, e.g., programming vs. learnprogramming. In "actions", there are many exchange related affixes, including *trades* (e.g., pokemon vs. pokemontrades) and *swap* (e.g., scotch vs. scotchswap). Altruistic behavior signified by *random_acts_of_* (e.g., pizza vs. random_acts_of_pizza) has been studied specifically in Althoff et al. (2014).

The noun-like affixes closely match the conceived metaphor of splitting space in community design theory (Kim, 2000). Indeed, we see a group of affixes based on "place", such as *uk* (e.g., politics vs. ukpolitics). "Medium", named for the *medium* of the content, e.g., *pics*, *vids*, etc. is another common category, including *videos* (e.g., cat vs. catvideos). The last one is based on "subject", such as *recipes* (e.g., vegan vs. veganrecipes). Noun-like affixes are probably used to encourage better discussions. These communities do not necessarily share similar users; e.g., people who are interested in veganrecipes may not be vegans; people who are invested in ukpolitics may not care about politics in general.

Surprisingly, there is a class of relatively minor changes that can cause community pairs to differ significantly. An example of "modifier" is *ism*, as in vegetarian vs. vegetarianism, which align in topic but likely attract different people. Another interesting class is "equivalent". One example is wallpaper vs. wallpapers: these two subreddits have indistinguishable (to us) content and thousands of members each, yet the moderator sets are disjoint, and neither mentions the other in their respective extensive and overlapping lists of related subreddits. In cases like this, the newer community may be created without knowing about



(a) #Pairs by category.



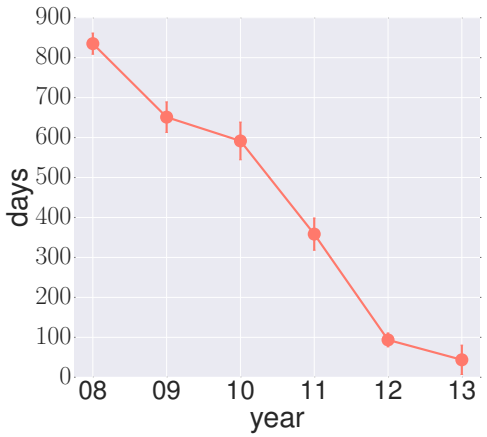
(b) Histogram of fraction as suffixes for all affixes.

Figure 5.1: (a) Medium is the most frequent affix, while modifier is the least. (b) Two distinct types of affixes exist: suffix-dominant and prefix-dominant.

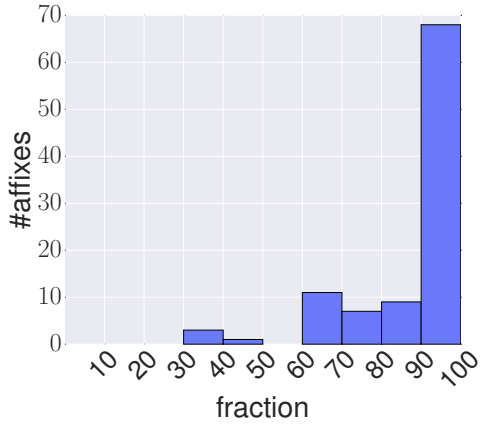
the older one, although in other cases there may be known prior interactions; for example, Politic was created because some users do not like the rules in politics.

Although some decisions in our taxonomy are arbitrary, we consider it useful and meaningful to get an overall sense of possible affixes. All affixes we consider seem to be generalizable changes that one can make with some community name to obtain another community name.

Frequency of affixes. Next, we examine the frequency of affixes. Table 5.1 presents the 10 most common affixes and Figure 5.1a shows the frequency by category. The most common affix is simply the character *s*, which suggests that it is perhaps common for "redundant" communities to be created. The most common category is "medium" with 88 pairs, while the least common one is "modifier" with 12 pairs. There is some variation in frequency within each category. For instance, one interesting observation is that although *porn* and *pics* both fall in "medium" and indicate a related picture-driven community, *porn* has more than 4 times more highly related communities (33 vs 7).



(a) Average gap for pairs grouped by the creation year of the older community.



(b) Histogram of fraction where the affixed community was created later.

Figure 5.2: (a) The newer related community is more and more quickly over the years. (b) For most affixes, the affixed community was created later, though there are many counterexamples.

Position of affixes in community name. As shown in Figure 5.1b, most but not all affixes are either suffix-dominant or prefix-dominant. Overall, "generation", "medium", and "modifier" tend to be used as suffixes, while "genre", "derivative", and "place" are usually used as prefixes. "parody" and "nsfw" can be used either way, for example, *funny* in videos vs. funnyvideos and Guildwars2 vs. guildwars2funny.

5.4.2 Temporal Relationships within Pairs

It is always possible to determine which community in a pair was created earlier. The first characteristic that we examine is the gap between the creation time of two communities in a pair. The overall average gap since 2008 is 749 days, when users on Reddit were first allowed to create their own communities. If we compute the average gap grouped by the creation year of the older community,

in Figure 5.2a we see a consistent trend that the newer community is created more and more quickly over the years. This suggests that there may be an accelerating culture of creating highly related communities over time, or that as there are more users on Reddit, affixed communities arise more quickly.

For most affixes, the community with the affix was newer. We further examine whether the newer community within a pair is the affixed one. This is indeed the case in 86% of our pairs. However, if we change our focus from pairs to affixes, we find that for 33% of the affixes, there was at least one instance where the affixed version was actually *created before* the "original" (see Figure 5.2b).

The four affixes for which the affixed version of the community more often exists first are *ing*, *al*, *ism* and *s*; these are mostly in the "equivalent/competition" class in Table 5.3. As a result, we observe phenomena like different communities focusing on exactly the same thing (e.g., wallpaper vs. wallpapers) or two communities eventually deciding to explicitly merge into one (e.g., wedding vs. weddings). Communities with different foci but similar names might also fall into this category, such as vegetarian vs. vegetarianism.

These four affixes do not cover all possible cases where the affixed was created earlier. For instance, twincitiessocial was created before twincities.

5.4.3 Does the New Overtake the Old?

Another important characteristic is how active the newer community is compared to the older one after its inception.

Newer communities tend to be less active, but in a quarter of pairs, the newer

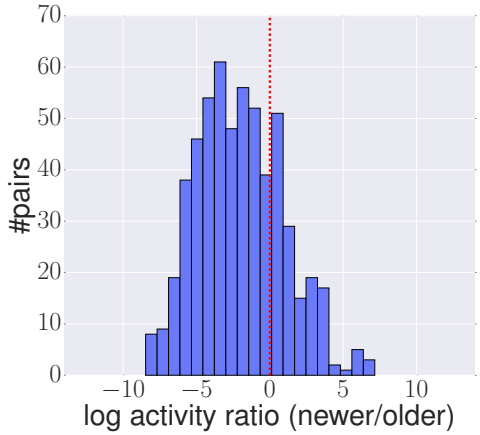
one is more active. We compute the log ratio in activity level (the total number of comments plus the total number of posts) between the newer community and the older community with add-one smoothing, only considering actions *after the newer community was created* so that we compare pairs during the same time period. According to this metric, a positive value means more activity in the newer community and a negative value means less activity in the newer community. Figure 5.3a demonstrates there is a trend that affixed versions of communities tend to be less active. The mean log ratio is -2.0, which suggests that new community is usually 13.5% as active as the older one. However, a nontrivial fraction of newer communities (25.7%) are more active.

A closer look at the more active newer communities. It's somewhat surprising that 25.7% of newer communities overtake their established counterparts. Why does this occur? Figure 5.3 presents examples of possible reasons that the younger community might surpass its older counterpart.

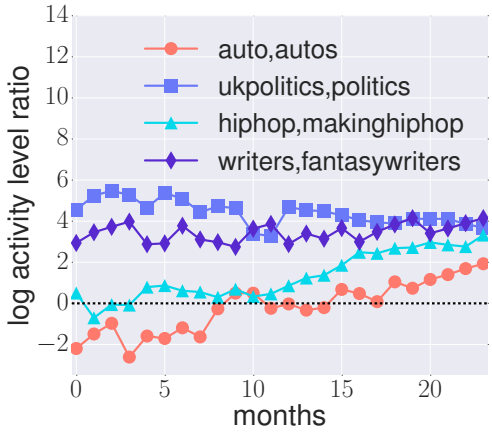
The first reason is that the affix represents something that naturally appeals to more people. One example is writers vs. fantasywriters. As soon as fantasywriters was created, its activity level was more than 7 times as great as that in writers. Here are top 3 affixes that consistently lead to more activity: *the* (e.g., stopgirl vs. thestopgirl), *ex* (e.g., mormon vs. exmormon), *steam* (e.g., deals vs. steamdeals).

Second, the newer community may be "equivalent" to the older one and the newer one may win the competition. For example, in the case of auto vs. Autos, it took Autos a while to exceed the activity level in auto, but Autos is now much more popular (see Figure 5.3b).

Third, the newer community may actually be the non-modified one (14%



(a) Histogram of log activity level ratio between the newer community and the old one.

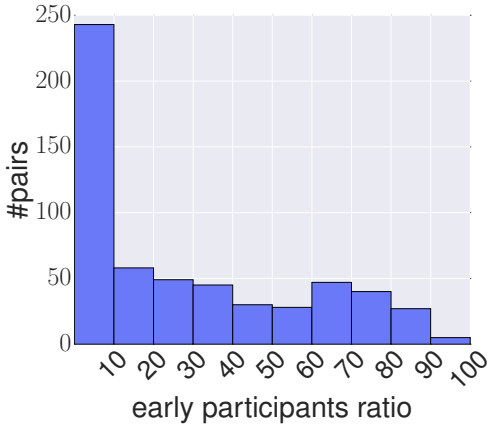


(b) Case study on pairs in which the newer one has more activity, where activity is binned on a month-to-month basis.

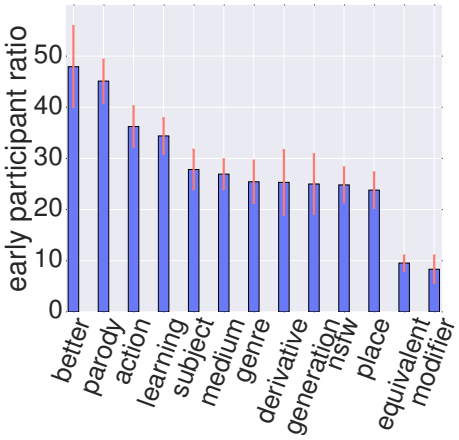
Figure 5.3: (a) The older community tend to have a higher level of activity. (b) Examples of different reasons that the newer one can have more activity. It shows how the log activity level ratio changes over time since the newer one was created in the first two years.

of pairs have this property, as we earlier observed) and the newer one might achieve popularity because it is more general. For instance, politics is more popular than ukpolitics despite the later's earlier founding. In this case, as soon as politics was created, its activity level exceeded ukpolitics.

The fourth and relatively rare reason is that the older one may have a large competitor, in other words, the newer one may originate from an even bigger community than the older one. An example is hiphop vs. makinghiphop. makinghiphop started at a similar size as hiphop but exceeded hiphop significantly later. Although hiphop and makinghiphop are both active, there is a much larger hiphop-related community on Reddit, hiphophead. makinghiphop might actually originate from hiphophead instead of hiphop.



(a) Histogram of the fraction of early participants in the new community that were from the old community over all pairs.



(b) Average fraction of early participants in the new community that were from the old community sorted by categories.

Figure 5.4: (a) Surprisingly, the majority of highly related communities do not share more than 10% of early participants. (b) "Better" has the highest average early participant ratio, while "modifier" has the lowest.

5.4.4 Where are early participants in the new communities from?

The last reason in the above discussion leads to a natural question: where are the participants in the newer community from? Are they from the older one in a pair? This question is difficult to answer, as a subreddit may establish its own identity and unique audience over time, even if it was born out of an existing community. If we simply look at the overlap between two communities over all users, we may mistakenly believe that they have never shared the user base as a result of a large number of later users. We thus focus on the first n participants in the newer community (the *early participants*) and compute the fraction of them that were also members of the older community. A user is considered a member of the old community if they took any action in the old community within the last 30 days prior to interacting with the new community. We refer to this metric

as "early participant fraction". While we present results for $n = 100$, similar results hold for different n.

Almost half of highly related communities do not really share early participants. As shown in Figure 5.4a, surprisingly, the majority of newer subreddits in highly related communities pairs are not "founded" by members of the older community. For example, only 7 of the first 100 participants in makinghiphop were members of hiphop.

Figure 5.4b presents the average early participant ratio for all categories in Table 5.3. It shows that "better", "parody", "action" and "learning" usually attract members from the older community. It also partly demonstrates why we obtain such a low average early participant ratio. "equivalent" and "modifier" appear more than likely to attract completely different participants, e.g., vegetarian vs. vegetarianism. We also notice significant differences even within a single category. One notable example is *meta* (65.8% from the original community) vs. *ex* (1% from the original community).

5.4.5 From Highly Related Communities to Spinoffs

Thus far, we have explored the complex space of possible affixes, and the highly related communities that are created through them. We find that a non-trivial fraction of the new communities were not the affixed ones, or did not share the same user base of the older one. For these pairs, it is unclear whether the new community is a subdivision of the old one, or whether users in the existing community are affected by the new one's presence. In order to better understand how users in the existing community may behave *after exploring the new commu-*

nity, we will focus on a subset of highly related communities called *spinoffs* in the remainder of this chapter.

5.5 Spinoffs: Substitutions or Complements?

We now formally define *spinoff* communities. First: the newer of the two pairs in a highly related community is a *spinoff* if it satisfies the following properties: 1) more than 10% of the first 100 early participants in the newer community are members of the older community; and 2) the newer community is the affixed one, so that it is likely to represent a specialization or some other topic of interest. We will sometimes refer to a pair of highly related communities that contain a spinoff as a *spinoff pair*.

In this section, we investigate how a user's behavior within the older subreddit is affected once they try out the newer spinoff: do such users get "distracted" by the new one, or does the new subcommunity complement the old one? Phrased differently, do users tend to decrease, increase, or not change their activity levels in the original community after trying the spinoff?

Surprisingly, we find that users who explore the spinoff generally become *more* active in the *original* community. Furthermore, with respect to the taxonomy we developed in Table 5.3, the magnitude of this trend *depends on the type of affix*: larger in "action", "better", and "parody", smaller in "medium", and *negative* in "nsfw". Finally, it seems that this complementary effect is more prominent for users with lower activity levels, although there is less data to compare users with different activity levels, and results vary depending on specific pairs.

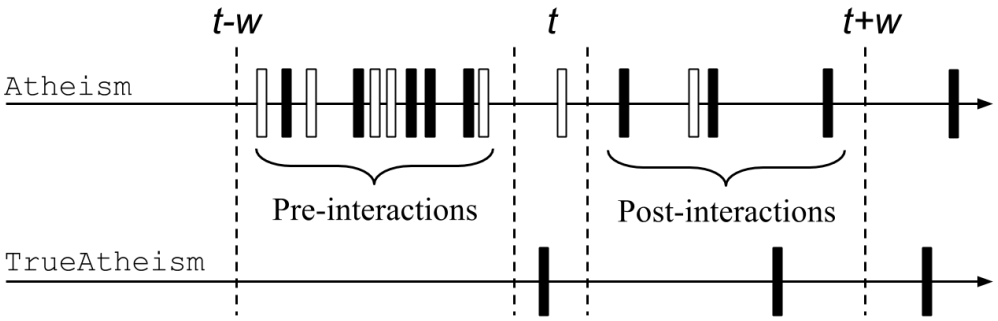


Figure 5.5: Schematic of the exploration experiment setup. TrueAtheism is a spinoff of Atheism, and the activity of two users is shown over time. Each box represents an interaction. With respect to the two subreddits shown, the dark user is an explorer, and the light user is a nonexplorer. Time t is the time of the dark user's first interaction with the spinoff subreddit. Here, the number of pre-interactions for both the dark and light users is 5. The dark user has 3 post-interactions, whereas the light user only has one.

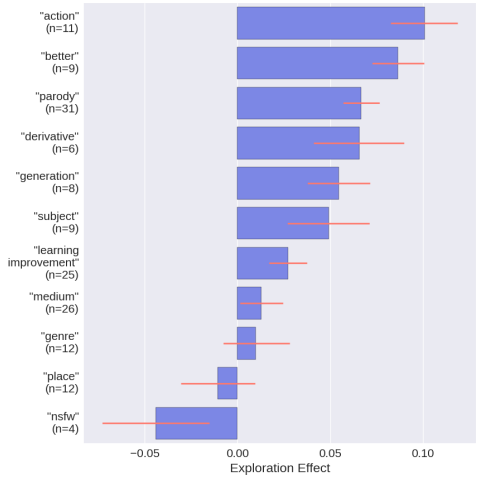
Disclaimer: *we do not make any claims of causality given the observational nature of our dataset.*

5.5.1 Experiment setup

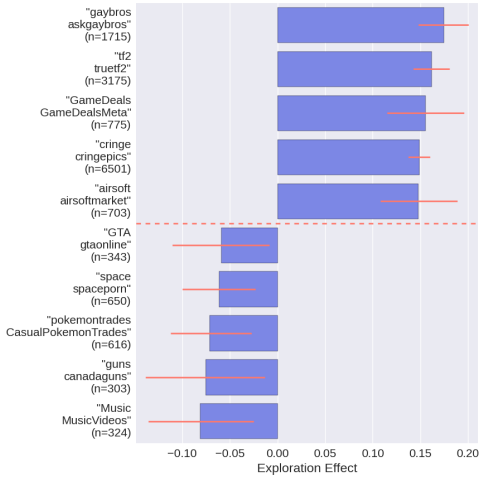
To understand user behavior in the *original* community *after* participating in the spinoff community, we propose an experiment framework in which we first pair "an explorer" and a "similar" "nonexplorer" in the original community. After identifying this pair of users, we compare their behavior pattern after the explorer first participated in the spinoff community, as illustrated in Figure 5.5.

Specifically, for each spinoff pair (e.g., Atheism vs. TrueAtheism in Figure 5.5), we define *explorers* as users who were active in the original community in a time window before their first participation in the spinoff community.[8] The darker user in Figure 5.5 is an example. We denote the time of her first interac-

[8]Participation and being active are both defined as either posting or commenting.



(a) Exploration effect by category (n is the number of sampled pairs, only $n \geq 4$ is shown)



(b) Top/bottom 5 exploration effect by pair (n is the number of sampled users)

Figure 5.6: Difference between explorers and nonexplorers in the fraction of users that become more active in post-interactions (in the older community) compared to pre-interactions. Larger values indicate more activity from explorers. (a) categories from our taxonomy and (b) specific pairs. Error bars represent 95% CIs.

tions in the spinoff community as t, and refer to her interaction in the original community from $t - w$ to t as *pre-interactions* and her interaction in the original community from t to $t + w$ as *post-interactions*. We consider users with at least 5 pre-interactions to ensure that they were indeed active in the original community.

A straightforward metric to compute is simply the ratio between the number of post-interactions and the number of pre-interactions for each user. However, this is problematic because we require users to have at least 5 pre-interactions but have no constraints on post-interactions. This causes our sample to be biased towards users with more pre-interactions than post-interactions.

To address this concern, for each exploring user u_e, we sample a *similarly*

active user u_{ne} in the original community who *never* interacts with the spinoff community. We call this user a "*nonexplorer*". The rough idea is demonstrated by the light user in Figure 5.5, who had a similar number of pre-interactions and made a post in the original community around t so that we know she was still active. The details of this sampling process are given in the appendix.

Metric: exploration effect. After we identify explorers and matching nonexplorers, we compute the fraction of explorers who have more post-interactions than pre-interactions in the older community, and then compute the same fraction for nonexplorers. We take the difference between these two fractions and call it the "exploration effect" (see Equation 5.2, in the appendix). Higher values of this quantity indicate that u_e was more active in the original community than the nonexplorer u_{ne}. We use the macro average to aggregate results from different spinoff pairs because the number of explorers varies between pairs.

The only parameter in our framework is w. Since our primary objective of interest in this work is the effect of the interaction with the spinoff community, we choose a relatively small window (30 days) to mitigate confounding factors that may affect user behavior over time and the dynamic nature of online communities (Danescu-Niculescu-Mizil et al., 2013; Backstrom et al., 2006; Ducheneaut et al., 2007; Kairam et al., 2012; Kumar et al., 2010). Our results are robust to reasonable changes in w (e.g., w = 20 days produces very similar results).

5.5.2 More active after exploring the spinoff community

We now apply this framework and examine how explorers behave in general. Surprisingly, we find that explorers are relatively more active in the original

community compared to nonexplorers, i.e., the exploration effect is generally positive. We then further split explorers based on their activity level and study how our observation differ depending on activity level.

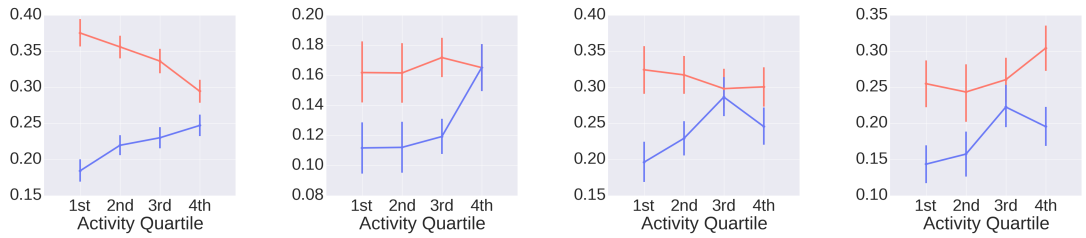
Comparisons across categories. Figure 5.6a presents exploration effect results for categories in our Table 5.3 taxonomy.[9] Somewhat counterintuitively, we find that for most spinoffs, users who explore become *more* active in the original subreddit after exploring, compared to similarly active users who never interacted in the new community (see Figure 5.6a).

Interestingly, the magnitude of this result varies based on the spinoff pair considered. We observe that "action" explorers are around 10% more likely to increase their activity after exploring, for example. "place" explorers, on the other hand, are roughly 2% less likely to increase their activity.

Our possible explanation for this observation is that users who explore "action" communities are often seeking to actively engage with a topic in a fashion above and beyond simple discussion. For example, the subreddit Bitcoin (which focuses on high-level discussions of the crypto-currency) and its spinoff pair BitcoinMining (which focuses on lower-level issues, e.g., hardware useful for mining Bitcoins) exhibit a difference in interaction ratio of roughly 10%. If a user explores BitcoinMining after interacting with Bitcoin, this is likely a strong indication of their interest in digging deeper into the topic itself. It's possible that viewing Bitcoin through the perspective of BitcoinMining increases overall engagement with the topic, at least in the short term.

In contrast, exploring "place" subreddits does not result in increased home

[9]Results are only reported for categories with more than 4 spinoff pairs.



(a) AskReddit vs TrueAskReddit; (n=8816 user pairs)

(b) Science vs AskScience; (n=5516 user pairs)

(c) Android vs androidapps; (n=2951 user pairs)

(d) apple vs applehelp; (n=2221 user pairs)

Figure 5.7: Several examples of **explorer** and **nonexplorer** activity levels (with 95% CIs) split into quartiles by pre-activity. The x-axis is pre-interaction quartile, and the y-axis is the proportion of users whose post-interactions exceeded their pre-interactions. In all cases, explorers tend to have greater post-interaction levels than nonexplorers, reflective of the results from the previous section. These plots are meant to highlight the complex relationships between activity level and activity rates. We observe many statistically significant differences, but note that each spinoff community pair's behavior in this regard appears to be unique. In the first three pairs, we do see that explorers with the highest pre-activity level present a smaller difference from nonexplorers.

activity nearly as often. For example, Bitcoin has another spinoff pair, BitcoinUK, that has an exploration effect of roughly zero. We have previously seen in Figure 5.4b that "place" spinoffs share relatively few early participants with their parent communities. Taken together, these observations suggest that users seeking place-specific communities are not necessarily interested in engaging more deeply with the topic, so much as in *who* they discuss the topic with or *how* the topic affects them.

A closer look at the pairs. Figure 5.6b presents the top and bottom 5 pairs in terms of exploration effect. It further demonstrates how our results may vary across different pairs. All 5 bottom pairs exhibit a significantly negative exploration effect, which shows that it is not always the case that explorers are more active.

The bottom 5 partly supports our above discussion regarding places. Indeed, in "place" related pairs, gtaonline pulled people from gta and so did canadaguns for guns. Among the top 5, there is an even spread among several categories including "learning" (gaybros vs. askgaybros), "action" (airsoft vs. airsoftmarket), and "medium" (cringe vs. cringepics). The surprising affix is *true*. Although it seems to suggest superiority and separation, explorers actually become more active in the original community in this case, too.

Discussion. Our findings resonate with the results in the previous chapter that users who continually explore new communities are, on average, more active than users who don't. However, no causal relationship can be established that explains this result: exploration does not necessarily *cause* increased activity, but may indicate a strong signal of interest level in our dataset.

5.5.3 Variations between explorers with different activity levels

We have established that users tend to be relatively more active in the "older" community after exploration, and have examined the variation across different categories and pairs. However, how does this effect differ for users with different pre-interaction levels? One could imagine that activity level prior to exploration affects whether or not users are more active after exploring. For example, upon discovering an alternative community, it's possible very active users might remain more attached to their home community, whereas relatively inactive users might not have the same level of commitment.

To address this question, we split users into pre-interaction quartile levels within their spinoff pair, so that the users with the least number of pre-interactions are put in bin one, and users with the greatest number of pre-interactions are put in bin four. We then compute the exploration effect for users in each quartile.[10]

Figure 5.7 presents the fraction of users who had more post-interactions than pre-interactions for, respectively, explorers and nonexplorers in several popular subreddit pairs. In general, the relative effects of exploration appear to be different based on how active users are, but there are complex and varied relationships between user activity level and how much defection matters; these relationships differ based on which spinoff pair is considered. Since we split users further into quartiles, the amount of data is not sufficient to reach conclusions for all pairs.

One relatively consistent pattern across pairs is that explorers with the highest pre-activity level usually have a smaller difference from the nonexplorers compared to explorers in the lowest quartile, as shown in the left three figures in Figure 5.7, although this is not true for Figure 5.7d.

The trend of how the fraction or the difference changes with different pre-activity quartile is even more complex. Consider the case of Figure 5.7a; this figure illustrates that for users with low activity levels (first/second quartiles) exploring is much more indicative of increased future activity than not exploring, and the difference is much less apparent for users with high activity levels – exploring and not exploring are associated with more similar levels of activity for users in the third/fourth quartiles.

[10]We have previously referred to Equation 5.2 as exploration effect, but plot p_e and p_{ne} separately in these plots under the same name.

Note that other pairs exhibit different patterns. For Science vs AskScience (Figure 5.7b) and Android vs androidapps (Figure 5.7c), the most active users (those in the 4^{th} quartile) appear to experience a slight "dip" in terms of the exploration effect.

5.6 Related Work

While there has been considerable interest in the topic in the social sciences (e.g., (Hurtado, 1997; Berry, 1997)), the study of situations wherein users engage with *multiple*, distinct communities represents a relatively new but increasingly relevant research area for computer scientists. Indeed, Kim (Kim, 2000) argues that a growing Web *needs* subdivisions, while Jones and Rafaeli (Jones and Rafaeli, 2010) also argue that an effective community splitting strategy is essential for virtual communities and online discourse to thrive. Furthermore, Birnholtz et al.'s (Birnholtz et al., 2015) study of confession groups on Facebook could be viewed in the context of "place" style affixes.

A number of studies have examined multi-community platforms in different contexts. Subcommunity survival (Turner et al., 2005; Iriberri and Leroy, 2009; Kraut and Resnick, 2012) is sometimes framed in the context of a meta-community. Also, Fisher et al. (Fisher et al., 2006) find that different newsgroups exhibit different conversation patterns, though they don't examine if the same users behave differently across platforms (as in Vasilescu et al. (2013b)). Finally, Adamic et al. (2008) examine the quality of user answers across different categories of Yahoo Answers.

Despite exhibiting some undesirable upvoting patterns (Gilbert, 2013), Red-

dit itself has been used as a data source in various contexts. For instance, the study of altruistic requests (Althoff et al., 2014), the study of domestic abuse discourse (Schrading et al., 2015), and work about post titles (Lakkaraju et al., 2013) demonstrate that useful information can be learned from Reddit comments and upvotes.

5.7 Conclusion

In this chapter, we use a dataset of all posts and comments from Reddit over an eight-year period to explore the space of naming affixes that lead to highly related communities on the platform. After building a taxonomy, we examine the early participants and other temporal aspects of the pairs, and introduce the idea of a spinoff community being "born out" of its unaffixed parent. Finally, we present the surprising result that users who explore in spinoff communities generally become relatively *more* active in their home communities instead of being "distracted". We also find that the magnitude of this effect (and sometimes its sign) depends on the type of community pair and how active a user was prior to exploration.

There are several directions for possible future work. First, it would be interesting to examine more closely the *origins* of highly related communities. If a community is created because of a disagreement (e.g., Zachary's Karate Club (Zachary, 1977)) one could potentially identify general characteristics of increasing unrest prior to a fission. Also, it would be interesting to delve deeper into differences between discourse on content in highly related communities pairs; how does discussion on TrueAtheism differ from discourse on Atheism, for ex-

ample. It would be useful for community organizers if we can detect when a spinoff community is necessary or beneficial. Furthermore, it is an important direction to understand the mechanism behind our observation that users who explore in spinoff communities generally become relatively *more* active in their home communities. This could be potentially useful for community organizers to identify complementary communities.

Finally, we note that our consideration has presupposed a pairwise framing, i.e., we always assumed a *pair* of communities. In some cases, we noted more complex phenomena underlying community creation. For example, a number of srs communities were all created in a short period of time. Also, the world of pokemon subreddits may consist of multiple affixes that lead to different subdivisions. In general, one could generalize pairwise interactions to explore more complex relationships between communities.

5.8 Appendix: Sampling Method for Control Users

The goal of this section is to describe how we sample a control user u_{ne} corresponding to each each exploring user u_e. Ultimately, to compute the exploration effect, we need to find someone who never posts in the new subreddit, but takes a similar number of actions in the same time period. To choose this similarly active, nonexplorer user, we sample u_{ne} as follows:

1. From the set of all nonexplorer users, find the subset who have an interaction in the original community within 24 hours of u_e's exploration time t. Let these interactions occur at time t'. If a nonexplorer user has more than one interaction between $t - 24$ hours and $t + 24$ hours, take the closest to t.

2. Find the user u_{ne} in this candidate set that minimizes the difference between their own number of pre-interactions (re-centered at their t') and u_e's. Specifically, if we let $p(u, t_a, t_b)$ be the number of interactions of user u in the original subreddit between t_a and t_b we find the loyal user

$$\operatorname*{argmin}_{u_{ne}} |p(u_e, t - w, t) - p(u_{ne}, t' - w, t')| .$$

3. If this difference is less than 5% of u_e's pre-interactions, a similarly active user u_{ne} has been successfully sampled.

Figure 5.5 demonstrates a pair of users that could be plausibly sampled in this manner. Both the light and dark users have the requisite 5 pre-interactions, and the light user makes a post within 24 hours of the dark user's first exploration.

After sampling k such user pairs $\{\langle u_{i,ne}, u_{i,e}\rangle\}|_{i=1}^{k}$ for a given pair of subreddits[11], we first compute the proportion of exploring/nonexplorer users whose activity increased, i.e., have more post-interactions than pre-interactions. For instance, this fraction for exploring users is computed as

$$p_e := \frac{1}{k} \sum_{i=1}^{k} \mathbb{1} \left[post(u_{i,e}) > pre(u_{i,e}) \right] . \tag{5.1}$$

Finally, for each spinoff pair of communities, the quantity we are interested in is

$$p_e - p_{ne} . \tag{5.2}$$

We generally call the quantity given in Equation 5.2 the "exploration effect". A larger exploration effect indicate that an explorer is more active in the *original* subreddit after posting to the splinter subreddit, when compared to a similarly active nonexplorer. In Figure 5.7, we plot p_e and p_{ne} separately, whereas in Figure 5.6 we plot $p_e - p_{ne}$.

[11] We discard the pair of communities if $k < 100$.

CHAPTER 6

FUTURE WORK

This thesis investigates the power of language and the space of multiple communities in a quantitative fashion. There are many future directions in the constantly evolving world.

The holy grail of my research to understand the effect of wording on social interaction is to create a writing assistant that helps every individual communicate better. The automated tools for writing nowadays only correct typos and simple grammatical errors. There is great room for potential tools that assist people to navigate the huge space of wordings and to compose a better wording tailored for their communicative goals. More generally, such tools may be also used in the reverse direction, e.g., to debias a piece of writing and make it neutral.

As these automated tools for individual users are being implemented and deployed, another fundamental question is the interplay between humans and machines. For instance, in Chapter 2, we have shown that machines outperform humans in predicting which tweet will be retweeted more. Does this mean that machines are going to overtake humans? One way to approach this question is to improve our understanding of the comparative advantages of machine intelligence and human intelligence. In many domains, I expect complementary abilities as a result of the different computational mechanisms between machines and humans. If machines and humans are complementary, we can design methods to integrate these strengths for better decision making. Even if machines may dominate humans in some domains, an interesting research question is then to understand whether it is possible to teach humans with the

"knowledge" of machines.

As for activities across multiple communities, the emergence of various communities remains to be further understood. In particular, many open questions arise when we examine social networks or communities in conflict. Does the deterioration of the relationship between two members slowly propagate and lead to two disconnected groups? Or does a small fight escalate to a community-level conflict that forces members to take stances? New models and datasets may be required to address these questions.

The final line of work is to explore policy implications in offline activities. An interesting observation is that to better run a website, service providers make many predictions at individual levels. For instance, service providers are actively trying to predict whether users will abandon the service or how long users will stay. Accurate predictions are vital for the service in resource allocation and user intervention. These predictions are ideal applications for machine learning because the right answer is not obvious and similar decisions are made many times. As pointed out in Kleinberg et al. (2015), there are similar problems in policymaking. An example is that predicting the current flu status of each person provides important information for the government to distribute resources and make emergency plans. This would be a great opportunity for machine learning researchers to extend impact to our offline life.

With real data to analyze and real systems that have been evolving and will be built, it is an exciting time to study human behavior and build socio-technical systems for people. This thesis contributes to a new paradigm of understanding human behavior and building such systems from computational perspectives.

BIBLIOGRAPHY

Lada A. Adamic, Jun Zhang, Eytan Bakshy, and Mark S. Ackerman. Knowledge Sharing and Yahoo Answers: Everyone Knows Something. In *Proceedings of WWW*, 2008.

Tim Althoff, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. How to ask for a favor: A case study on the success of altruistic requests. In *Proceedings of ICWSM*, 2014.

Shlomo Argamon and Shlomo Levitan. Measuring the usefulness of function words for authorship attribution. In *Proceedings of ACH/ALLC*, 2005.

Aristotle. *The Art of Rhetoric*. 350 BC.

Yoav Artzi, Patrick Pantel, and Michael Gamon. Predicting responses to microblog posts. In *Proceedings of NAACL (short paper)*, 2012.

Vikas Ganjigunte Ashok, Song Feng, and Yejin Choi. Success with style: Using writing style to predict the success of novels. In *Proceedings of EMNLP*, 2013.

Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group Formation in Large Social Networks: Membership, Growth, and Evolution. In *Proceedings of KDD*, 2006.

Michael Bailey, Daniel J. Hopkins, and Todd Rogers. Unresponsive and unpersuaded: The unintended consequences of voter persuasion efforts. In *Meeting of the Society for Political Methodology at the University of Virginia*, 2014.

Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an influencer: Quantifying influence on twitter. In *Proceedings of WSDM*, 2011.

Eytan Bakshy, Solomon Messing, and Lada A. Adamic. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239):1130–1132, 2015.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

John W. Berry. Immigration, Acculturation, and Adaptation. *Applied Psychology*, 46(1):5–34, 1997.

Jeremy Birnholtz, Nicholas Aaron Ross Merola, and Arindam Paul. "Is it weird to still be a virgin": Anonymous, locally targeted questions on Facebook confession boards. In *Proceedings of CHI*, 2015.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Robert M. Bond, Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415): 295, 2012.

Youmna Borghol, Sebastien Ardon, Niklas Carlsson, Derek Eager, and Anirban Mahanti. The untold story of the clones: Content-agnostic factors that impact YouTube video popularity. In *Proceedings of KDD*, 2012.

Christopher J. Bryan, Gregory M. Walton, Todd Rogers, and Carol S. Dweck. Motivating voter turnout by invoking the self. *Proceedings of the National Academy of Sciences*, 108(31):12653–12656, 2011.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911, 2014.

C. Bühler. The curve of life as studied in biographies. *Journal of Applied Psychology*, 19(4):405, 1935.

Michael Burgoon, Stephen B. Jones, and Diane Stewart. Toward a message-centered theory of persuasion: Three empirical investigations of language intensity. *Human Communication Research*, 1(3):240–256, 1975.

Amparo Elizabeth Cano-Basave and Yulan He. A Study of the Impact of Persuasive Argumentation in Political Debates. In *Proceedings of NAACL*, 2016.

Shelly Chaiken. The heuristic model of persuasion. In *Social influence: The Ontario Symposium*, 1987.

Marilyn J. Chambliss and Ruth Garner. Do adults change their minds after reading persuasive text? *Written Communication*, 13(3):291–313, 1996.

Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. How Community Feedback Shapes User Behavior. In *Proceedings of ICWSM*, 2014.

Dennis Chong and James N. Druckman. Framing theory. *Annual Review of Political Science*, 10:103–126, 2007.

Cindy Chung and James W. Pennebaker. The psychological functions of function words. *Social communication*, pages 343–359, 2007.

Robert B. Cialdini. *Influence: The Psychology of Persuasion*. HarperCollins, 1993.

Robert B. Cialdini, Wilhelmina Wosinska, Daniel W. Barrett, Jonathan Butner, and Malgorzata Gornik-Durose. Compliance with a request in two cultures:

The differential influence of social proof and commitment/consistency on collectivists and individualists. *Personality and Social Psychology Bulletin*, 25(10): 1242–1253, 1999.

Geoffrey L. Cohen, Joshua Aronson, and Claude M. Steele. When beliefs yield to evidence: Reducing biased evaluation by affirming the self. *Personality and Social Psychology Bulletin*, 26(9):1151–1164, 2000.

Joshua Correll, Steven J. Spencer, and Mark P. Zanna. An affirmed self and an open mind: Self-affirmation and sensitivity to argument strength. *Journal of Experimental Social Psychology*, 40(3):350–356, 2004.

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. Mark my words! linguistic style accommodation in social media. In *Proceedings of WWW*, 2011.

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

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of WWW*, 2012b.

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 *Proceedings of WWW*, 2013.

Koustuv Dasgupta, Rahul Singh, Balaji Viswanathan, Dipanjan Chakraborty, Sougata Mukherjea, Amit A. Nanavati, and Anupam Joshi. Social Ties and

Their Relevance to Churn in Mobile Telecom Networks. In *Proceedings of EDBT*, 2008.

Munmun De Choudhury, Winter A. Mason, Jake M. Hofman, and Duncan J. Watts. Inferring Relevant Social Networks from Interpersonal Communication. In *Proceedings of WWW*, 2010.

James Price Dillard and Lijiang Shen. *The Persuasion Handbook: Developments in Theory and Practice*. SAGE Publications, 2014.

John DiNardo. Natural experiments and quasi-natural experiments. In *The New Palgrave Dictionary of Economics*. Palgrave Macmillan, 2008.

M. Brent Donellan, Richard E. Lucas, and William Fleeson, editors. *Personality and Assessment at Age 40: Reflections on the Past Person-Situation Debate and Emerging Directions of Future Person-Situation Integration [Special Issue]*. Number 43(2) in *Journal of Research in Personality*. 2009.

Gideon Dror, Dan Pelleg, Oleg Rokhlenko, and Idan Szpektor. Churn Prediction in New Users of Yahoo! Answers. In *Proceedings of WWW (Companion)*, 2012.

Nicolas Ducheneaut, Nicholas Yee, Eric Nickell, and Robert J. Moore. The life and death of online gaming communities: A look at guilds in World of Warcraft. In *Proceedings of CHI*, 2007.

Olive Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961.

Amanda M. Durik, M. Anne Britt, Rebecca Reynolds, and Jennifer Storey. The effects of hedges in persuasive arguments: A nuanced analysis of language. *Journal of Language and Social Psychology*, 2008.

Alice H. Eagly and Shelly Chaiken. *The Psychology of Attitudes*. Harcourt Brace Jovanovich College Publishers, 1993.

Erik H. Erikson and Joan M. Erikson. *The life cycle completed (extended version)*. WW Norton & Company, 1998.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

Noura Farra, Swapna Somasundaran, and Jill Burstein. Scoring persuasive essays using opinions and their targets. In *Proceedings of the NAACL Workshop on Innovative Use of NLP for Building Educational Applications*, 2015.

Song Feng, Jun Seok Kang, Polina Kuznetsova, and Yejin Choi. Connotation lexicon: A dash of sentiment beneath the surface meaning. In *Proceedings of ACL*, 2013.

Danyel Fisher, Marc Smith, and Howard T. Welser. You are who you talk to: Detecting roles in usenet newsgroups. In *Proceedings of HICSS*, 2006.

Rudolph Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32 (3):221, 1948.

B. J. Fogg. Mass interpersonal persuasion: An early view of a new phenomenon. In *Persuasive Technology*, 2008.

ForWhatReason. Eli5 "the great digg migration". https://www.reddit.com/r/explainlikeimfive/comments/m0w30/eli5_the_great_digg_migration/. Accessed: 2016-06-02.

Eric Gilbert. Widespread underprovision on Reddit. In *Proceedings of CSCW*, 2013.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. Part-of-speech Tagging for Twitter: Annotation, Features, and Experiments. In *Proceedings of NAACL (short paper)*, 2011.

Corrado Gini. *Variabilità e mutabilità*. C. Cuppini, Bologna, 1912.

David Godes, Dina Mayzlin, Yubo Chen, Sanjiv Das, Chrysanthos Dellarocas, Bruce Pfeiffer, Barak Libai, Subrata Sen, Mengze Shi, and Peeter Verlegh. The firm's management of social interactions. *Marketing Letters*, 16(3-4):415–428, 2005.

Marco Guerini, Carlo Strapparava, and Oliviero Stock. Evaluation metrics for persuasive NLP with Google AdWords. In *Proceedings of LREC*, 2010.

Marco Guerini, Carlo Strapparava, and Gözde Özbal. Exploring text virality in social networks. In *Proceedings of ICWSM (poster)*, 2011.

Marco Guerini, Alberto Pepe, and Bruno Lepri. Do linguistic style and readability of scientific abstracts affect their virality? In *Proceedings of ICWSM (poster)*, 2012.

Marco Guerini, Gözde Özbal, and Carlo Strapparava. Echoes of persuasion: The effect of euphony in persuasive communication. In *Proceedings of NAACL*, 2015.

Anna Guimarães, Ana Paula Couto da Silva, and Jussara M Almeida. Temporal analysis of inter-community user flows in online knowledge-sharing net-

works. In *Proceedings of SIGIR 2015 Workshop on Temporal, Social and Spatially-aware Information Access*, 2015.

David A. Hanauer, Yang Liu, Qiaozhu Mei, Frank J. Manion, Ulysses J. Balis, and Kai Zheng. Hedging their mets: The use of uncertainty terms in clinical documents and its potential implications when sharing the documents with patients. In *Proceedings of the AMIA Annual Symposium*, 2012.

Lars Kai Hansen, Adam Arvidsson, Finn Årup Nielsen, Elanor Colleoni, and Michael Etter. Good friends, bad news-affect and virality in Twitter. *Communications in Computer and Information Science*, 185:34–43, 2011.

Kazi Saidul Hasan and Vincent Ng. Why are you taking this stance? Identifying and classifying reasons in ideological debates. In *Proceedings of EMNLP*, 2014.

Chip Heath, Chris Bell, and Emily Sternberg. Emotional selection in memes: The case of urban legends. *Journal of personality and social psychology*, 81(6): 1028, 2001.

Jack Hessel, Alexandra Schofield, Lillian Lee, and David Mimno. What do vegans do in their spare time? Latent interest detection in multi-community networks. 2015.

Jack Hessel, Chenhao Tan, and Lillian Lee. Science, AskScience, and BadScience: On the Coexistence of Highly Related Communities. In *Proceedings of ICWSM*, 2016.

George C. Homans. Social Behavior as Exchange. *American Journal of Sociology*, 63(6):597–606, 1958.

Liangjie Hong, Ovidiu Dan, and Brian D. Davison. Predicting popular messages in Twitter. In *Proceedings of WWW*, 2011.

Carl I. Hovland, Irving L. Janis, and Harold H. Kelley. *Communication and Persuasion: Psychological Studies of Opinion Change*, volume 19. Yale University Press, 1953.

Craig R. Hullett. The impact of mood on persuasion: A meta-analysis. *Communication Research*, 32(4):423–442, 2005.

Aída Hurtado. Understanding Multiple Group Identities: Inserting Women into Cultural Transformations. *Journal of Social Issues*, 53(2):299–327, 1997.

Ken Hyland. *Hedging in Scientific Research Articles*. John Benjamins Publishing, 1998.

Alicia Iriberri and Gondy Leroy. A Life-cycle Perspective on Online Community Success. *ACM Computing Surveys*, 41(2):11:1–11:29, 2009.

Aaron Jaech, Victoria Zayats, Hao Fang, Mari Ostendorf, and Hannaneh Hajishirzi. Talking to the crowd: What do people react to in online discussions? In *Proceedings of EMNLP*, 2015.

Quentin Jones and Sheizaf Rafaeli. Time to split, virtually: 'Discourse architecture' and 'community building' create vibrant virtual publics. *Electronic Markets*, 10(4):214–223, 2010.

Sanjay Ram Kairam, Dan J. Wang, and Jure Leskovec. The Life and Death of Online Groups: Predicting Group Growth and Longevity. In *Proceedings of WSDM*, 2012.

Douglas T. Kenrick and David C. Funder. Profiting from controversy: Lessons from the person-situation debate. *American Psychologist*, 43(1):23, 1988.

Amy Jo Kim. *Community Building on the Web: Secret Strategies for Successful Online Communities*. Addison-Wesley Longman Publishing Co., Inc., 1st edition, 2000.

J. Peter Kincaid, Robert P. Fishburne Jr, Richard L. Rogers, and Brad S. Chissom. Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for Navy enlisted personnel. *Chief of Naval Technical Training, Naval Air Station, Research Branch Report 8-75*, 1975.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. Prediction Policy Problems. *American Economic Review*, 105(5):491–495, 2015.

Robert E. Kraut and Paul Resnick. *Building Successful Online Communities: Evidence-Based Social Design*. The MIT Press, 2012.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, 2009.

Ravi Kumar, Jasmine Novak, and Andrew Tomkins. Structure and evolution of online social networks. In *Link mining: models, algorithms, and applications*, pages 337–357. Springer, 2010.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. From word embeddings to document distances. In *Proceedings of ICML*, 2015.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *Proceedings of WWW*, 2010.

Himabindu Lakkaraju, Julian McAuley, and Jure Leskovec. What's in a Name? Understanding the Interplay between Titles, Content, and Communities in Social Media. In *Proceedings of ICWSM*, 2013.

George Lakoff. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic*, 2, 1975.

David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Computational Social Science. *Science*, 323(5915):721–723, 2009.

Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *Proceedings of ACL*, 2014.

Marco Lippi and Paolo Torroni. Argumentation Mining: State of the Art and Emerging Trends. *ACM Transactions on Internet Technology*, 16(2):10:1–10:25, 2016.

Annie Louis and Ani Nenkova. What makes writing great? First experiments on article quality prediction in the science journalism domain. *Transactions of ACL*, 2013.

Pamela J. Ludford, Dan Cosley, Dan Frankowski, and Loren Terveen. Think Different: Increasing Online Community Participation Using Uniqueness and Group Dissimilarity. In *Proceedings of CHI*, 2004.

Zongyang Ma, Aixin Sun, and Gao Cong. Will this #hashtag be popular tomorrow? In *Proceedings of SIGIR*, 2012.

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

Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

Neil McIntyre and Mirella Lapata. Learning to tell tales: A data-driven approach to story generation. In *Proceedings of ACL-IJCNLP*, 2009.

David McRaney. The backfire effect. http://youarenotsosmart.com/2011/06/10/the-backfire-effect/, 2011. Accessed: 2015-10-15.

Katherine L. Milkman and Jonah Berger. What makes online content viral? *Journal of Marketing Research*, 49(2):192–205, 2012.

Tanushree Mitra and Eric Gilbert. The Language that Gets People to Give: Phrases that Predict Success on Kickstarter. In *Proceedings of CSCW*, 2014.

Raquel Mochales and Marie-Francine Moens. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22, 2011.

Natalie Wolchover. Why is everyone on the internet so angry? http://www.scientificamerican.com/article/why-is-everyone-on-the-internet-so-angry/. Accessed: 2016-06-02.

Kate G. Niederhoffer and James W. Pennebaker. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360, 2002.

Brendan Nyhan and Jason Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330, 2010.

Daniel J. O'Keefe. Standpoint explicitness and persuasive effect: A meta-analytic review of the effects of varying conclusion articulation in persuasive messages. *Argumentation and Advocacy*, 34(1):1–12, 1997.

Daniel J. O'Keefe. Justification explicitness and persuasive effect: A meta-analytic review of the effects of varying support articulation in persuasive messages. *Argumentation and Advocacy*, 35(2):61–75, 1998.

James W. Pennebaker and Laura A. King. Linguistic styles: language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6):1296–1312, 1999.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. Linguistic inquiry and word count: LIWC 2007. Technical report, 2007.

Isaac Persing and Vincent Ng. Modeling argument strength in student essays. In *Proceedings of ACL*, 2015.

Saša Petrović, Miles Osborne, and Victor Lavrenko. RT to win! Predicting message propagation in Twitter. In *Proceedings of ICWSM*, 2011.

Richard E. Petty and John T. Cacioppo. *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*. Springer Science & Business Media, 2012.

Richard E. Petty and Duane T. Wegener. Matching versus mismatching attitude functions: Implications for scrutiny of persuasive messages. *Personality and Social Psychology Bulletin*, 24(3):227–240, 1998.

Richard E. Petty, Duane T. Wegener, and Leandre R. Fabrigar. Attitudes and attitude change. *Annual Review of Psychology*, 48(1):609–647, 1997.

Pew Research Center. Internet gains on television as public's main news source.

Emily Pitler and Ani Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of EMNLP*, 2008.

Eva M. Pomerantz, Shelly Chaiken, and Rosalind S. Tordesillas. Attitude strength and resistance processes. *Journal of Personality and Social Psychology*, 69(3):408, 1995.

Samuel L. Popkin. *The Reasoning Voter: Communication and Persuasion in Presidential Campaigns*. University of Chicago Press, Chicago, 1994.

Kathleen Kelley Reardon. *Persuasion in Practice*. Sage, 1991.

Yuqing Ren, Robert Kraut, and Sara Kiesler. Applying Common Identity and Bond Theory to Design of Online Communities. *Organization Studies*, 28(3): 377–408, 2007.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of EMNLP*, 2013.

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

Sara Rosenthal and Kathleen McKeown. I couldn't agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *Proceedings of SIGdial*, 2015.

Matthew Rowe. Mining User Lifecycles from Online Community Platforms and their Application to Churn prediction. In *Proceedings of ICDM*, 2013.

Matthew J. Salganik, Peter Sheridan Dodds, and Duncan J. Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–856, 2006.

Nicolas Schrading, Cecilia O. Alm, Ray Ptucha, and Christopher Homan. An analysis of domestic abuse discourse on Reddit. 2015.

Claude E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

Marvin E. Shaw. *Group Dynamics: The Psychology of Small Group Behavior*. McGraw Hill, New York, 1971.

John C. Sherblom. Organization involvement expressed through pronoun use in computer mediated communication. *Communication Research Reports*, 7(1): 45–50, 2009.

Matthew P. Simmons, Lada A. Adamic, and Eytan Adar. Memes online: Extracted, subtracted, injected, and recollected. In *Proceedings of ICWSM*, 2011.

Edward H. Simpson. Measurement of diversity. *Nature*, 1949.

Philipp Singer, Fabian Flöck, Clemens Meinhart, Elias Zeitfogel, and Markus Strohmaier. Evolution of Reddit: From the Front Page of the Internet to a Self-referential Community? In *Proceedings of WWW (Companion)*, 2014.

Peter K. Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell, and Neil Tippett. Cyberbullying: its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry*, 49(4):376–385, 2008.

Swapna Somasundaran and Janyce Wiebe. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 2010.

Dhanya Sridhar, James Foulds, Bert Huang, Lise Getoor, and Marilyn Walker. Joint models of disagreement and stance in online debate. In *Proceedings of ACL*, 2015.

Christian Stab and Iryna Gurevych. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of EMNLP*, 2014.

Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H. Chi. Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In *Proceedings of SocialCom*, 2010.

Tao Sun, Ming Zhang, and Qiaozhu Mei. Unexpected relevance: An empirical study of serendipity in retweets. In *Proceedings of ICWSM*, 2013.

Latanya Sweeney. Discrimination in Online Ad Delivery. *Queue*, 11(3):10:10–10:29, 2013.

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

Chenhao Tan, Lillian Lee, and Bo Pang. The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter. In *Proceedings of ACL*, 2014.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions. In *Proceedings of WWW*, 2016.

Matt Thomas, Bo Pang, and Lillian Lee. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of EMNLP*, 2006.

Zakary L. Tormala and Richard E. Petty. What doesn't kill me makes me stronger: The effects of resisting persuasion on attitude certainty. *Journal of Personality and Social Psychology*, 83(6):1298, 2002.

Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. *Sociometry*, pages 425–443, 1969.

Oren Tsur and Ari Rappoport. What's in a hashtag?: Content based prediction of the spread of ideas in microblogging communities. In *Proceedings of WSDM*, 2012.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. Metaphor detection with cross-lingual model transfer. In *Proceedings of ACL*, 2014.

Tammara Combs Turner, Marc A. Smith, Danyel Fisher, and Howard T. Welser. Picturing Usenet: Mapping computer-mediated collective action. *Journal of Computer-Mediated Communication*, 10(4):00–00, 2005.

Orit Tykocinski, E. Tory Higgins, and Shelly Chaiken. Message framing, self-discrepancies, and yielding to persuasive messages: The motivational significance of psychological situations. *Personality and Social Psychology Bulletin*, 20 (1):107–115, 1994.

Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*, 2011.

Bogdan Vasilescu, Vladimir Filkov, and Alexander Serebrenik. StackOverflow and GitHub: Associations between Software Development and Crowdsourced Knowledge. In *Proceedings of SocialCom*, 2013a.

Bogdan Vasilescu, Vladimir Filkov, and Alexander Serebrenik. StackOverflow and GitHub: Associations between software development and crowdsourced knowledge. In *Proceedings of SocialCom*, 2013b.

Bogdan Vasilescu, Alexander Serebrenik, Prem Devanbu, and Vladimir Filkov. How Social Q&A Sites Are Changing Knowledge Sharing in Open Source Software Communities. In *Proceedings of CSCW*, 2014.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207, 2013.

Duane T. Wegener and Richard E. Petty. Effects of mood on persuasion processes: Enhancing, reducing, and biasing scrutiny of attitude-relevant information. In Leonard L. Martin and Abraham Tesser, editors, *Striving and Feeling: Interactions Among Goals, Affect, and Self-regulation*. Psychology Press, 1996.

Wikipedia. Digg. https://en.wikipedia.org/wiki/Digg. Accessed: 2016-06-02.

Frank Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

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

Jiang Yang, Xiao Wei, Mark S. Ackerman, and Lada A. Adamic. Activity Lifespan: An Analysis of User Survival Patterns in Online Knowledge Sharing Communities. In *Proceedings of ICWSM*, 2010.

Wayne W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33(4):452–473, 1977.

Fan Zhang and Diane Litman. Using Context to Predict the Purpose of Argumentative Writing Revisions. In *Proceedings of NAACL*, 2016.

Haiyi Zhu, Jilin Chen, Tara Matthews, Aditya Pal, Hernan Badenes, and Robert E. Kraut. Selecting an Effective Niche: An Ecological View of the Success of Online Communities. In *Proceedings of CHI*, 2014a.

Haiyi Zhu, Robert E. Kraut, and Aniket Kittur. The Impact of Membership Overlap on the Survival of Online Communities. In *Proceedings of CHI*, 2014b.

Yin Zhu, Erheng Zhong, Sinno Jialin Pan, Xiao Wang, Minzhe Zhou, and Qiang Yang. Predicting User Activity Level in Social Networks. In *Proceedings of CIKM*, 2013.

Julia R. Zuwerink and Patricia G. Devine. Attitude importance and resistance to persuasion: It's not just the thought that counts. *Journal of Personality and Social Psychology*, 70(5):931, 1996.