SOCIAL NEWS CONSUMPTION
IN SYSTEMS WITH CROWD-SOURCED CURATION

A Dissertation

Submitted to the Graduate School
of the University of Notre Dame
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

by

Maria Glenski

Tim Weninger, Director

Graduate Program in Computer Science and Engineering
Notre Dame, Indiana
April 2019
SOCIAL NEWS CONSUMPTION
IN SYSTEMS WITH CROWD-SOURCED CURATION

Abstract

by

Maria Glenski

People frequently supplement or have replaced their consumption of news from traditional print, radio, or television news sources with social news consumption from online social media platforms such as Facebook, Twitter, or Reddit. Reliance on social media sites as primary sources of news and information continues to grow and shows little sign of decreasing in the future. Tasked with curating an ever-increasing amount of content, providers leverage user interaction feedback to make decisions about which content to display, highlight, and hide. The sheer volume of new information being produced and consumed only increases the reliance that individuals place on anonymous others to curate and sort the massive amounts of information.

Here, I describe several analyses and predictive models of user-behavior in social news platforms such as: user-interactions that rely on or influence the aggregate, anonymous crowd-ratings used to identify news-worthy content and user-interactions with news sources of varied credibility in particular. The central focus of this work is to understand not only how individuals consume social news, but also how they contribute to the spread and reception of credible news and misinformation. Experimental results and predictive models demonstrate the influence of algorithmic biases on social news consumption patterns and the distinctions in the consumption of, response to, and propagation of information from news sources of varied credibility.
To my parents, Susan and Duane.
CONTENTS

List of Figures ................................................................. v

List of Tables ................................................................. xii

Acknowledgments .............................................................. xiii

Chapter 1: Introduction ......................................................... 1
  1.1 Overview and Contributions ........................................... 5
  1.2 Impact ................................................................. 6

Chapter 2: Rating Effects on Social News Posts and Comments ........... 8
  2.1 Introduction ............................................................ 8
  2.2 Methodology ........................................................... 15
    2.2.1 Post Experiment ................................................ 15
    2.2.2 Comment Experiment ......................................... 16
  2.3 Results ................................................................. 17
    2.3.1 Delay Effects .................................................... 22
    2.3.2 Reaching the Front Page ........................................ 25
    2.3.3 Subreddit Effects ................................................ 27
  2.4 Related Work .......................................................... 29
  2.5 Discussion and Conclusions .......................................... 31
    2.5.1 Voting and Viewing Mechanics ................................ 31
    2.5.2 Vote-based Manipulation ....................................... 32
    2.5.3 Conformity and Influence ...................................... 34
    2.5.4 Voting .......................................................... 35

Chapter 3: Consumption and Curation Habits in Social Rating Systems ... 36
  3.1 Introduction ............................................................ 36
  3.2 Information Consumers and Curators .................................. 39
    3.2.1 Headline Browsers .............................................. 42
    3.2.2 Vote Behavior ................................................... 44
    3.2.3 Position Bias ..................................................... 46
  3.3 Performance Deterioration in Browsing Sessions ...................... 48
    3.3.1 Community Variety ............................................. 50
    3.3.2 Performance Deterioration ...................................... 53
### Table of Contents

3.4 Discussion and Conclusions ................................................. 60

**Chapter 4: Social News Consumption from Sources of Varying Credibility** ................................................. 62

4.1 Introduction ................................................................. 62

4.2 Data Collection and Annotation .......................................... 66
  4.2.1 News Source Labels .................................................. 66
  4.2.2 Population Data ...................................................... 68
  4.2.3 Demographics Data .................................................. 70

4.3 Methodology ................................................................. 72

4.4 Population Level Deception Propagation ................................. 75

4.5 Deception Propagation By Demographics ................................. 83

4.6 Summary ...................................................................... 88

4.7 Limitations ................................................................. 89

4.8 Discussion and Conclusions ............................................... 89

**Chapter 5: User-Reactions to Social News Sources of Varying Credibility** ................................................. 92

5.1 Introduction ................................................................. 92

5.2 Classifying Reactions in Social Media Postings .......................... 93
  5.2.1 Data and Annotations ................................................ 94
  5.2.2 Model ................................................................. 94
  5.2.3 Model Performance .................................................. 95

5.3 Measuring Reactions to Trusted and Deceptive News Sources ......... 96
  5.3.1 Twitter and Reddit News Data ..................................... 97
  5.3.2 Methodology .......................................................... 98
  5.3.3 Results ............................................................... 99

5.4 Reactions from Human-like versus Bot-like Users ...................... 102
  5.4.1 Related Work ........................................................ 102
  5.4.2 Data Collection and Annotation ................................... 104
  5.4.3 Methodology ........................................................ 105
  5.4.4 Results ............................................................... 109
    5.4.4.1 Prevalence of Bots .......................................... 109
    5.4.4.2 Reaction Variety ............................................. 112
    5.4.4.3 Reaction Speed ............................................... 113
    5.4.4.4 Reaction Inequality ........................................... 114

5.5 Discussion and Conclusions ............................................... 118

**Chapter 6: Conclusions and Future Work** .................................. 120

6.1 Future Directions .......................................................... 122
  6.1.1 Enhancing User-Behavior Models .................................. 122
  6.1.2 Identification of Deceptive News Sources From Initial Behavior .......................................................... 123
  6.1.3 Extended Analysis of Propagation from News Sources of Varying Credibility .................................................. 123

Bibliography ........................................................................... 124
2.1 Composite, redacted screenshot of Reddit. (A) There are many possible ranking systems on Reddit; this image shows the first post when ordered by the “top”—scored posts within the past month. (B) Authenticated users may up-vote, or down-vote once on any post; the score of a post is congruent to the total number of up-votes minus the total number of down-votes. (C) Each post displays its rank on the far left corresponding to its position in the selected ranking system, a title text, the host (domain) of the linked content on the far right, and the number of comments below (this post also has 29,266 comments). (D) The top 200 comments are displayed in order as well corresponding to the chosen ranking system and number of points the comment has received; an orange-red arrow indicates that the current user up-voted this comment. (E) This comment has a score of 5,997, which is congruent to the number of up-votes minus the number of down-votes that the comment has received. (F) Comment threads are hierarchical such that each comment can have have children, siblings, etc. thus comment orderings are based on their vote score relative to the sibling comments in the thread hierarchy.

2.2 Final scores for artificially, randomly up-treated posts, down-treated posts, and scores for untreated posts in the control group are shown. Red inner error bars show the standard error of the mean; black outer error bars show the 95% confidence interval. At left, (a) shows the scores in the heavily skewed full distribution for posts. At right, (b) shows the scores in the heavily skewed full distribution for comments, with significant decreases for down-treated comments when compared to the control group.

2.3 Top 99% of final scores for artificially, randomly up-treated posts, down-treated posts, and scores for untreated posts in the control group are shown. Red inner error bars show the standard error of the mean; black outer error bars show the 95% confidence interval. When the highest 1% of post scores are removed, the score distribution becomes much less skewed resulting in tighter error bounds, which further result in significant increases for up-treated posts and significant decreases for down-treated posts when compared to the control group. Again, when the highest 1% of comment scores are removed, the score distribution becomes less skewed resulting in tighter error bounds, but with slight but not significant increases for up-treated comments when compared to the control group.
2.4 Final scores separated into their respective treatment delay intervals. At left (a) shows final scores for artificially, randomly up-treated posts, down-treated posts, and scores for untreated posts in the control group and, at right, (b) shows final scores for artificially, randomly up-treated comments, down-treated comments, and scores for untreated comments in the control group. Horizontal lines show the overall mean of each treatment group. The top 1% of scores were removed to un-skew the score distribution.

2.5 Results of Kolmogorov-Smirnov and Mann-Whitney U Tests (with top 1% included) for each treatment delay where ▲ represents tests comparing up-treatment with the control group and ▼ represents tests comparing down-treatment with the control group. † indicates results that are not statistically significant.

2.6 The middle 9 deciles of final scores for each treatment according to their delay intervals. These results show that most posts and comments receive a median score of 2 or less, and that treatment has the most effect in the higher deciles of the score distribution.

2.7 The probability of a post (at left) or a comment (at right) receiving a corresponding score by treatment type. The inset graphs show the complete probability distribution functions. The outer graphs show the probability of a post/comment receiving scores between 500 and 2000 – an approximation for *trending* or *frontpage* posts. Up-treated posts are 24% more likely to reach a score of 2000 than the control group.

2.8 Mean scores of down-treated, control group and up-treated posts in the top 8 most active subreddits at left (a) and the mean scores of down-treated, control group and up-treated comments in the top 9 subreddits from which we collected the most data at right (b), *i.e.*, subreddits that are most active and most often appearing as “rising” on Reddit. Black outer error bars show the 95% confidence interval and red inner error bars show error.

3.1 Number of interactions for the 309 registered Reddit users in the dataset within each category between August 2015 and August 2016.

3.2 Distribution by vote-type of votes cast by the 309 registered Reddit users between August 1, 2015 and July 31, 2016.

3.3 Headline browsing tendencies. At left is plotted a histogram of users as a function of their tendency to interact with each pageload. The typical user interacts with a loaded page about half the time. At right is plotted an un-grouped scatterplot of the same data. The scatterplot indicates that the number of pageloads is weakly correlated with the user’s tendency to interact the content ($R^2 = 0.09, p < 0.01$).
3.4 Voting tendencies. At left is plotted a histogram of the percentage of users as a function of their tendency to vote. In center, the scatterplot indicates a correlation between the number of a user’s interactions that were vote-interactions and the total number of a user’s interactions ($R^2 = 0.64, p < 0.01$). At right, the scatterplot indicates a weak correlation between the proportion of a user’s interactions that were vote-interactions and the total number of a user’s interactions ($R^2 = 0.04, p < 0.01$).

3.5 User behavior before voting on posts. Bar plots illustrate the likelihood that users vote on a post after viewing its comments section (top plot), content (middle plot), or either (bottom plot). The bottom plot shows that nearly one-third of voters rarely (i.e., < 20% of the time) view the content or comments of a post before casting a vote.

3.6 Effects of rank on tendencies to browse, upvote, or downvote posts from a user’s frontpage (or /r/all) or a specific subreddit.

3.7 Sessionizing Heuristic. A histogram of inactivity delays between consecutive interactions (left) shows the uptick at 1 hour that signals the boundary between intra- and inter-session delays indicating the use of a session inactivity ceiling of 1 hour. The distribution of resulting session lengths (right), shows where we partitioned the sessions into short, medium, and long sessions.

3.8 Distributions across categories based on their sessions by length (left) and by users (right).

3.9 Community Variety. Distinct subreddits loaded (left) and interacted with (right) during a session by session length. Despite being highly skewed, the mean-average number of subreddits loaded are 1.4, 2.5, and 4.8 for short-, medium-, and long- browsing users, respectively, and even lower 1.0, 1.3, and 1.5, respectively, for subreddits users actively interacted with.

3.10 Voting Variety. Number of subreddits by number of voting users (on left) exhibits a power law distribution. Number of subreddits by total number of votes (on right) also exhibits a power law distribution with a much longer tail. Blue circles and red squares indicate upvotes and downvotes respectively.

3.11 Examples of session partitioning for the original, delay-randomized, and event-randomized datasets. Circles represent interactions $I_j$, arrows represent inactivity delays between consecutive interactions $I_j$ and $I_{j+1}$, and boxes distinguish between sessions. The first row illustrates an example of how data is partitioned into sessions in the original dataset; the second row shows how sessions change when delays are randomly shuffled; and the third row shows how the sessions change when the events are shuffled.
3.12 Median post rank as a function of session decile for short-, medium-, and long-browsing users. Linear regression lines are plotted and the correlation of determination $R^2$ value is shown to the right of each regression line. † and ‡ symbols represent statistically significant effects with $p < 0.05$ and $p < 0.01$ respectively. The user sessions plot (at left) shows that the median ranks of browsed posts tends to increase for all types of users. As expected, baseline plots illustrating session shuffled by randomized delays (center) and randomized events (right) show little, if any, correlation.

3.13 Slope ($m$) and effect sizes ($R^2$) of various factors as a function of session progress. † and ‡ symbols represent statistically significant correlations with $p < 0.05$ and $p < 0.01$ respectively.

3.14 Median subreddit-normalized change in score for upvoted posts (SRN$\Delta^{\wedge}$) as a function of the fraction of session elapsed for short-, medium-, and long-browsing users. Linear regression lines are plotted, and the correlation of determination $R^2$ value is shown to the right of each regression line. † and ‡ symbols represent statistically significant effects with $p < 0.05$ and $p < 0.01$ respectively. The user sessions plot (on left) shows how the median SRN$\Delta^{\wedge}$ changes as a session progresses: the median SRN$\Delta$ of browsed posts tends to diminish for all types of users, but only long-browsing users show statistically significant correlation.

3.15 Effect sizes ($R^2$) for analysis of covariance (ANCOVA) tests that compare regressions on user sessions against delay-randomized and event-randomized sessions. Results of significance tests between the observed behavior and the delay-randomized (D) or event-randomized (E) sessions are represented with ◦, †, and ‡ symbols for $p \geq 0.05$, $p < 0.05$, and $p < 0.01$ respectively.

4.1 Tweet volume for the ten most frequently occurring news sources within each type as a function of their share of all tweets of the given type. For example, anonews.co is the 10th most frequently occurring clickbait source, but is responsible for only 0.03% of the clickbait tweets. In contrast, the most frequent clickbait source theblaze is responsible for 43%. The 10th (ZeeNews) and most frequent (el.pais) trusted sources have a smaller range in terms of shares of trusted tweets with ZeeNews responsible for 3.5% and el.pais for 11.7%.

4.2 Tweet volume by month. The percentage of retweets within each source-type is plotted as a function of the month retweeted.
4.3 Lorenz Curves and Gini Coefficients. As a graphical representation of income inequality within a population, Lorenz curves plot the share of income by the cumulative share of the population. Lorenz curves that measure the inequality in propagation plot the share of total propagation, i.e. the y% of tweets posted, by the share of the population who propagated, i.e. the cumulative x% of active users. The Gini coefficient is the proportion of the area under the line of perfect equality \((a_1 + a_2)\) that is captured between the line of perfect equality and the Lorenz curve \((a_1)\).

4.4 Lorenz curves of the propagation inequality for each source type, averaged across sources using user-source pairs instead of users as diffusion units. Propagation using direct retweets at left and @mention tweets at right. The lorenz curve representative of perfect equality is also included for reference.

4.5 Gini coefficients and Palma ratios for direct retweets (RT) from or tweets that @mention sources (@) for each news type, averaged across sources using source-user combinations as diffusion units. Higher values indicate more inequality, highest values are highlighted in bold.

4.6 Log-log plots of the distributions of the number of source tweets with a given number of users who retweeted for the five most frequently occurring sources in each type. More frequently occurring sources are plotted in darker shades. This figure is best viewed in color.

4.7 CDF plots of diffusion delays (in hours) by news type for direct retweets (top) and retweets with @mentions of the source account (bottom). The inset of each plot shows a closer view of the initial diffusion, highlighted with a box in the larger plot.

4.8 Gini coefficients and Palma ratios for each demographic attribute, averaged for each source type using source-user combinations as diffusion units. † denotes that a large majority (> 75%) of users were classified as the given demographic.

4.9 Interaction Volumes. Demographic where more users shared each individual source post from sources of a given type with the common language effect size (as % of comparisons) for each attribute. Statistical significance from MWU tests of \(p < 0.01\) for all results.

4.10 Interaction Delays. Demographics who directly retweet from or @mention news source types after a longer delay and the common language effect size (as %). A dash (—) is shown if no significant differences were found. All results are statistically significant (MWU tests of \(p < 0.01\)).

4.11 Interaction Tendencies. Demographics more likely to share a direct retweet from a source of a given type at least once. For each attribute, the demographic with the higher likelihood is listed if statistically significant (MWU \(p < 0.01\)) with the common language effect size (as %). A dash (—) is shown if no significant difference is found.
5.1 Architecture of neural network model used to predict reaction types.

5.2 Comparison of our model’s performance, measured using F1 score, trained only on content features, with the performance reported by Zhang et al. [139] trained on content, author, thread, structure, and community features.

5.3 Distributions of deceptive news sources and reactions to those sources (Reddit comments or tweets, respectively) for the Reddit and Twitter datasets across the four sub-categories of deceptive news sources.

5.4 Distributions of the five most frequently occurring reaction types within comments on Reddit and tweets on Twitter for each class of news sources.

5.5 CDF plots of the volumes of reactions by reaction delays for the frequently occurring reactions (i.e., reactions that occur in at least 5% of comments) for each source-type, using a step size of one hour. The CDF for Elaboration-reactions to Deceptive (no disinformation) Twitter news sources is occluded by the CDF for Deceptive Twitter news sources. This figure is best viewed in color.

5.6 As a graphical representation of income inequality within a population, Lorenz curves plot the share of income by the cumulative share of the population. The Gini coefficient is the proportion of the area under the line of perfect equality \( a_1 + a_2 \) that is captured between the line of perfect equality and the Lorenz curve \( a_1 \).

5.7 Prevalence of Bots. Distributions across bot accounts (bot score \( \geq 0.5 \)), human accounts (bot score \( < 0.5 \)), and unknown accounts (for which we could not collect a bot score) within the set of users who reacted (Users) and the set of reaction tweets (Tweets) for each class of news source. Highest proportions of each user type are highlighted in bold and lowest proportions are in italics.

5.8 Bot score distributions, using a bin width of 0.05, for users who reacted (left) and reaction-tweets (right). Mann Whitney U comparisons of raw distributions found that the average bot score of a user who posted a reaction-tweet is higher \( (p < 0.01) \) than the average bot score of a user who reacted for all source types except for Conspiracy-sources, where the average bot score of a user who posted a reaction-tweet is lower \( (p < 0.01) \).

5.9 Distributions of predicted reaction-types within tweets that directly responded to sources of each source-type.

5.10 Proportions of reactions posted by bot (B), human (H), or unknown (U) user-accounts for each source class and reaction type combination for the most frequent reaction types. Source class(es) with the lowest proportions for each account type are highlighted with bold for each of the reaction types.
5.11 Frequencies of most common reaction-types within reactions to news sources of each class posted by bot accounts (above) and human user accounts (below), as a percentage of reactions posted by accounts within each population.

5.12 Cumulative distribution function (CDF) plots of the volumes of reactions by reaction delays in hours (i.e. the delay between when a source posted content and when the reaction tweet was posted) for bot and human user accounts for the most frequently occurring reactions (occur in at least 10% of tweets) for each source-type, using a step size of one day.

5.13 Lorenz curves for each of the frequently occurring reactions (occurring in at least 5% of tweets) for each source-type. These Lorenz curves plot the share of reactions by the cumulative share of the population (bots, humans, or accounts without bot scores) as a graphical representation of inequality in reaction volume within each population. The gray dash-dotted line reflects the Lorenz curve that would result from a population wherein each user was responsible for an equal number of reactions. Gini coefficients for bot (B) and human (H) accounts and statistical significance results of Mann Whitney U (MWU) comparisons of Lorenz curves are listed in the top left corner of each subplot: ** if \( p < 0.01 \), * if \( p < 0.05 \), and − if \( p \geq 0.05 \). Lorenz curves and Gini coefficients are presented faded where there are no significant differences between bot and human users.

5.14 Relative Participation Inequality. The difference (\( \Delta \)), if statistically significant (MWU \( p < 0.01 \)), between Gini coefficients for bot (B) and human (H) user accounts and the relative increase (\( \% \Delta \)) from the human user to bot Gini coefficient, i.e. \( (B-H)/H \). A dash (−) is shown if no significant difference was found (\( p \geq 0.05 \)). The highest relative increases are highlighted in bold within source types and italicized within reaction types.
# LIST OF TABLES

1.1 Thesis Overview .................................................. 5
2.1 Distributions across treatments for the data collected ............. 17
3.1 The 10 most frequent subreddits from which interactions occurred .... 41
4.1 Summary of our Twitter dataset .................................. 68
4.2 Overlaps of user accounts who retweeted multiple news source types ..... 82
5.1 Summary of Annotated Datasets .................................. 95
5.2 Summary of Twitter and Reddit one-hop news datasets ............. 98
5.3 Summary of English-reactions from users who reacted frequently (> 4 reactions between Jan 2016 and Jan 2017) ......................... 106
ACKNOWLEDGMENTS

First, I would like to thank my advisor, Tim Weninger for his support, guidance, and advice. I could not have asked for a better advisor and mentor during my Ph.D study and am immensely grateful for the many productive and entertaining discussions throughout the years, all of the engaging research I have been able to pursue under his direction, and the opportunities I have been able to take advantage of as a member of his group.

Next, I would like to thank all my collaborators including Corey Pennycuff, Greg Stoddard, Paul Resnick, Svitlana Volkova, and Emily Saldanha. I would especially like to thank Dr. Svitlana Volkova. I am grateful to have had the opportunity to work with her over the past two years and her support, guidance, and mentorship have been invaluable.

I would also like to acknowledge and thank my committee, Dr. Meng Jiang, Dr. Huan Liu, and Dr. Dong Wang for their engagement and insightful comments and suggestions.

Special thanks also go to Joyce Yeats, Ginny Watterson, M. D. McNally, and Diane Wordinger in the Department of Computer Science and Engineering for all their help and support throughout the years. Joyce Yeats has helped me navigate innumerable aspects of our program’s or the graduate school’s requirements and processes which significantly reduced the amount of stress and confusion in my life over the past five years.

Additionally, I would like to thank all current and past members of the Weninger group for their support, interesting discussions, and the fun we have had in the lab but particularly Sal Aguinaga, Corey Pennycuff, and Baoxu (Dash) Shi for all their support, spending four years asking and answering questions that often prompted new ideas, and being my academic-siblings during the trials and triumphs of my first few years.
My sincere thanks also to all my friends and family for their support and encouragement over the years. In particular, my mother for encouraging my interest in mathematics and patterns from a young age and my sister Jennifer for being one of my biggest champions who was always just a phone call away and a great travel inspiration and companion who kept me grounded and refreshed throughout the years.
CHAPTER 1

INTRODUCTION

People use social media not only for entertainment and social networking but also as primary sources of news and information. Crowd-sourced creation, curation, and consumption of information has become the norm. An August 2017 survey from the Pew Research Center found that 67% of Americans report that they get at least some of their news from social media, an increase of 5% over the previous year [109], and this reliance has remained consistently high (at 68%) according to a subsequent survey conducted in August 2018 [86]. Pew Research Center also reported an increase in the rate of Americans who often get their news online in some way, increasing from 38% to 43% from 2016 to 2017 [13]. At the same time, reliance on television source (e.g. local, nightly, or cable news programs) decreased from 57% to 50%. While a higher proportion still rely on television sources, the gap has narrowed from a 19% difference to just 7%. The speed and convenience of social news are top advantages identified by social news consumers but inaccuracy is a large concern for many – 57% of social media users who consume news on one or more of the platforms they use expect the news they see to be “largely inaccurate” [86]. With such a reliance on social media as a source of news and information, the spread of misinformation is a significant concern.

Social news aggregators like Digg, HackerNews, and Reddit are systems that allow individuals to post news and information for others to view, vote, and comment on. These systems represent a stark departure from traditional media outlets where a news organization, i.e., a handful of television, radio, or newspaper producers, sets the topics and directs the narrative. These socio-digital communities increasingly set the news agenda,
cultural trends, and popular narrative of the day [70, 74, 90]. It has become commonplace to have news segments focused around what is “trending,” entire television shows devoted to covering happenings on social media, and even live presidential debates where topics are selected based on popular questions to social media Web site. Live commentaries (e.g., “live tweeting”), of events, crises, or performances for example, has become routine. As these trends continue and grow, it is important to understand not only how individuals consume information that is delivered to them but also how their online behavior shapes the news and information that countless others consume.

Tasked with curating an ever-increasing amount of content, providers leverage the collective ratings of the crowd to identify which content to show users. Users in turn rely on the anonymous, aggregate ratings of others to make important decisions about which products to buy, movies to watch, news to read, or even political candidates to support. These online, anonymous ratings are replacing the traditional word-of-mouth communication about an object’s, idea’s, or product's quality. Instead of primarily asking friends or family, users often read through reviews or trust star-ratings. The sheer volume of new information being produced and consumed only increases the reliance that individuals place on anonymous others to curate and sort massive amounts of information.

More than a century ago the experiments of Francis Galton determined that the median estimate of a group can be more accurate than estimates of experts [39]. Surowiecki’s book *The Wisdom of the Crowds* finds similar examples in stock markets, political elections, quiz shows, and a variety of other fields where large groups of people behave intelligently and perform better than an elite few [116]. However, other experiments have shown that when individuals’ perceptions of quality and value follow the behavior of a group, the resulting herd mentality can be suboptimal for both the individual and the group [14, 56, 77]. The rating mechanisms found in socio-digital platforms provide a type of Web-democracy that is open to anyone who can reach the platform. However, most users are passive browsers or “lurkers.” Only a handful of users actually contribute ratings, and even fewer still con-
tribute new content [42, 94, 120]. Given the widespread use and perceived value of these systems [42], it is important to consider whether they can successfully harness the wisdom of the crowd to accurately aggregate individual information or how they can be misused to propagate low quality or deceptive content.

Alongside the proliferation of “lurkers” compared to users who actively contribute or curate, social influence biases present an additional challenge to the reliance on social ratings. When users provide ratings to social posts (or comments), the system’s ranking algorithm reorders the content promoting highly rated content over others. Previous studies have found that this so-called ranking bias effect encourages herding behavior that results in suboptimal outcomes [44, 71, 91, 105]. This leads to asymmetric outcomes for posts of identical content [67, 104]. In social networks, these algorithmic biases may be responsible for viral information cascades [8, 30] and may play a significant role in the spread of misinformation in social media [37].

Typically, when researchers have studied social media, they study the collective actions of a community or population by looking at coarse-grained signals of posts, comments, and vote or rating totals. For example, the 90-9-1 rule, a representation of the proportion of Wikipedia browsers, editors, and creators [94], might be reflected in social news aggregators as the number of users who browse, comment, and submit content. Work in this area has found that user behavior (and subsequently, the content shown to other users) can be easily influenced and manipulated through injections of artificial ratings [43, 44, 91, 129]. This is especially relevant in light of the finding that 59% of bitly-URLs on Twitter are shared without the linked site ever having been visited [38]. In the context of social news consumption, especially when considering readers who are using social media as a primary source of information, the implications of this ease of influence are particularly concerning.

Previous work, especially within the area of social news, has focused on influence campaigns and the spread of (mis)information organically or through bots. Given a particular event, e.g., an election or a natural disaster, researchers typically follow information cas-
cades to tease out the diffusion processes and infer various characteristics about how social media users responded to the event [35, 117]. These studies have resulted in important findings about the effects of information contagion [70], influence campaigns [44], bots [37], and spam [50], etc., within specific newsworthy events. Misinformation spread in social networks has become a critical focus because of the substantive reliance on social rating platforms for news and the increasing consequences of the diffusion of misinformation.

Many studies focused on misinformation in social media have focused on rumor and misinformation detection with a primary focus on the network’s role in information diffusion models [65, 66, 101, 135]. Other studies compare the behavior of traditional and alternative media [112], classify media sources into sub-categories of misinformation [125], or attempt to detect rumor-spreading users [102]. A recent study by Vosoughi et al. [126] found that news fact-checked and found to be false spread faster and to more people than news items found to be true, and that although amplified by bots much of the diffusion is user-driven. These and others have found that the size and shape of (mis)information cascades within a social network depends heavily on the initial reactions of the users. Yet, we still lack an understanding of how users react to news sources of varying credibility and how their various initial responses contribute the the spread of (mis)information.

In this thesis, I investigate social news consumption in the presence of crowd-sourced curation and identify significant distinctions in the consumption and propagation of information from and the response to news sources of varied credibility. I show that informative patterns of behavior can be extracted from initial consumption or information propagation behavior. These patterns can be used to advance the understanding of social news consumption, develop informative models of user behavior in the presence of crowd-sourced curation, and motivate the development of robust models to efficiently identify questionable news sources without reliance on content-level annotations or user-reports.
TABLE 1.1

THESIS OVERVIEW

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Research Problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Rating Effects on Social News Posts and Comments</td>
<td>What effect does random vote perturbation have on content popularity in social news aggregator systems?</td>
</tr>
<tr>
<td>3</td>
<td>Consumption and Curation Habits in Social Rating Systems</td>
<td>How do users browse, vote, and engage with crowd-aggregated social news?</td>
</tr>
<tr>
<td>4</td>
<td>Social News Consumption from Sources of Varying Credibility</td>
<td>How do users, in general and by demographic, interact with news sources of varying credibility?</td>
</tr>
<tr>
<td>5</td>
<td>User-Reactions to Social News Sources of Varying Credibility</td>
<td>How do users react to news sources of varying credibility and are there significant differences between human and bot reactions?</td>
</tr>
</tbody>
</table>

1.1 Overview and Contributions

In this chapter, I have introduced the focus of this thesis, framed within the context of previous work and its relevance within our social media dependent society. The topics discussed in this thesis are presented in Table 1.1. Chapter 2 examines the effects of single random votes on the popularity outcomes of social news posts and comments. Chapter 3 quantifies the patterns of consumption and curation behavior and the effects of algorithmic biases such as rating, rank, and visibility biases via an in-vivo observational study of browsing and voting patterns on Reddit. Chapter 4 investigates social news propagation from news accounts identified as spreading verified versus deceptive content and quantify differences in behavior for a variety of classes of news sources and between demographic sub-populations of users. Chapter 5 analyzes the variety, volume, speed, and disparity of user-reactions to news sources of varying credibility across platforms (Twit-
ter and Reddit) and user-populations (Humans and Bots). Finally, Chapter 6 presents the conclusions and future working directions to develop enhanced models of user behavior on social news platforms, in general and specifically in terms of propagation of information from news sources of varied credibility, and the identification of deceptive news sources without crowd-sourced or expert-annotated labelling.

1.2 Impact

The work described in this dissertation will support the continued analysis and modelling of different types of online social behavior relevant to news source consumption and information spread from sources of varied credibility, e.g., information campaigns, coordinated effort, competing campaigns, recurrence, intimidation campaigns etc. — work that will not only be critical for national security considerations but could also lead to healthier interactions with and boosted levels of trust in social media news. The three main avenues of impact from this work are (1) novel patterns of behavior on social news platforms; (2) state-of-the-art models for social media content, behavior, and user demographics; and (3) novel datasets comprising social news platform interactions. These patterns, models, and datasets can be used by computer and computational social scientists alike in a myriad of analyses regarding social media users, social news consumption, or online social behavior.

Patterns Each chapter presents novel patterns of behavior in online rating systems, social news consumption in general, and consumption of social news from credible to deceptive news sources from a variety of user-populations. These patterns not only advance the understanding of how the mechanisms of socially curated news influence consumption but can also be used to motivate the development of new or enhance the performance of existing predictive models. For example, the predictive signals identified in Chapter 3 have been effectively utilized to predict future user-interactions from content, context, and a users’ interaction histories.
Models  Chapter 4 introduces state-of-the-art models to identify age, gender, income, and education levels of users from their social media postings, outperforming previous models trained and tested on the same dataset [47]. The linguistically-infused neural network model in Chapter 5 is able to identify the reaction type, i.e., the primary discourse act, of a social media posting from the content of the post and its parent text alone [46]. These models rely on content and derivative features of the content itself which allows future researchers and studies to apply them across multiple platforms.

Datasets  Chapter 3 introduces a novel dataset of browsing and voting interactions on Reddit. It comprises user-interactions with links, changes in sorting approaches, votes, and page-load interactions with web pages within the reddit.com domain for 309 distinct users over the course of the 12 months between August 2015 and August 2016. These private browsing and curating interactions are not publicly or easily available for collection. The release of this dataset enables future work and analyses to be conducted without the need to develop a tool to collect these signals of user behavior and recruitment of users willing to use such a tool.

Chapters 4 and 5 present datasets of one-hop propagation of information from news sources of varied credibility, labelled according to news source credibility with subsets that include additional annotations of inferred demographics and level of automation for the users interacting with these social news sources. These annotated interactions enable a more nuanced behavioral analysis and, paired with the findings highlighted alongside the introduction of these datasets, can be used to motivate more fine-grained future studies into direct interactions with varied classes of online news media.
CHAPTER 2

RATING EFFECTS ON SOCIAL NEWS POSTS AND COMMENTS

At a time when information seekers first turn to digital sources for news and opinion, it is critical that we understand the role that social media plays in human behavior. This is especially true when information consumers also act as information producers and editors through their online activity. In this chapter, in order to better understand the effects that editorial ratings have on online human behavior, we report the results of a two large-scale *in-vivo* experiments in social media. We find that small, random rating manipulations on social media posts and comments created significant changes in downstream ratings resulting in significantly different final outcomes.

2.1 Introduction

What is becoming known as collective intelligence bares the potential to enhance human capability and accomplish what is impossible individually [17, 132]. For example, more than a century ago the experiments of Francis Galton determined that the median estimate of a group can be more accurate than estimates of experts [39]. Surowiecki’s book *The Wisdom of the Crowds* finds similar examples in stock markets, political elections, quiz shows and a variety of other fields where large groups of people behave intelligently and perform better than an elite few [116]. However, other experiments have shown that when individuals’ perceptions of quality and value follow the behavior of a group, the resulting herd mentality can be suboptimal for both the individual and the group [14, 56, 77].

We often rely on online reviews contributed by anonymous users as an important source of information to make decisions about which products to buy, movies to watch, news to
read, or even political candidates to support. These online reviews replace traditional word-
of-mouth communication about an object’s or idea’s quality [21]. The sheer volume of new information being produced and consumed only increases the reliance that individuals place on anonymous others to curate and sort massive amounts of information. Because of the economic and intrinsic value involved, it is important to consider whether this approach can successfully harness the wisdom of crowd to accurately aggregate individual information.

By relying on the judgments of others, we may be susceptible to malicious ratings with some ulterior motive. Unfortunately, there is a gap in our knowledge and capabilities in this area, including untested and contradictory social theories. Fortunately, these gaps can be filled using new experimental methodologies on large, socio-digital data sets. The main idea is to determine if these socio-digital platforms produce useful, unbiased, aggregate outcomes, or (more likely) if, and how, opinion and behavior is influenced and manipulated. Work of our own and recent tangential experiments [67, 91, 105, 129] suggest that decisions and opinions can be significantly influenced by minor manipulations, yielding different social behavior. The main focus of this chapter is the determine what effect, if any, malicious voting behavior has on social news posts and comments.

Unfortunately, causal determinations are difficult to assess. In a closely related experiment, Wu and Huberman measured rating behavior in two different online platforms. The first allowed users to see prior ratings before they voted and the other platform hid the prior ratings until after the user voted. They found that when no information about previous ratings or page views are available, the ratings and user-opinions expressed tend to follow regular patterns. However, in cases where the previous ratings were made known, the user-opinions tended to be either neutral or form a polarized consensus. In the latter case, new opinions tend to reinforce previous opinions and thus become more extreme [134].

Because of the information overload caused by billions of daily shares, tweets, posts and comments, nearly all social media Web sites have sophisticated ranking algorithms
that attempt to identify the relatively few items that their users will find interesting. When new or better items are shared or posted, the ranking systems rely significantly upon user ratings to accelerate the discovery of new or interesting content. Information that is rated positively will be ranked higher, and will therefore be more visible to other users, which further increases the likelihood that it will receive further ratings [18, 97].

A recent experiment by Lerman and Hogg further studied the effects that presentation order has on the choices that users make. In this study, several users were asked to read and rate social media posts ranked by various ordering algorithms. Lerman and Hogg found that different ranking systems result in very different outcomes. Random orderings result in the most unbiased ratings, but may show a lot of uninteresting content resulting in poor user engagement. The “popularity ranking,” which rated posts by how many previous positive-votes it received led to highly inconsistent outcomes and showed that small early differences in ratings led to inconsistent rating scores [21].

Social news Web sites represent a stark departure from traditional media outlets in which a news organization, i.e., a handful of television, radio or newspaper producers, sets the topics and directs the narrative. Socio-digital communities increasingly set the news agenda, cultural trends, and popular narrative of the day [74]. News agencies frequently have segments on “what’s trending,” entire television shows are devoted to covering happenings on social media, and even live presidential debates select their topics based on popular questions posed to social media Web sites. As this trend continues and grows, the number of blogs, news outlets, and other sources of user generated content has outpaced the rate at which Web users can consume information. Social news Web sites are able to automatically curate, rank and provide commentary on the top content of the day by harnessing the wisdom of the crowd.

The recent popularity of social networks has led to the study of socio-digital influence and popularity cascades where models can be developed based on the adoption rate of friends (e.g., shares, retweets). Bakshy et al., find that friendship plays a significant
role in the sharing of content [7]. Similarly, Leskovec et al. were able to formulate a generative model that predicts the size and shape of information cascades in online social networks [72].

However, social media users seem to be unaware of the effects of social manipulation. A recent survey of Reddit users aimed to determine what the sampled community thought drove Reddit-users to up-vote or down-vote various posts. The surveyors expected that the leading indicators would be that users are more likely to up-vote or like 1) content that others have liked, indicating social influence; 2) content that was submitted or contributed by a well known user, indicating trust or model-based bias; or 3) content that is relevant to the user’s interests. Contrary to our scientific understanding of social influence, the surveyed users indicated that social influence had little effect on their voting likelihood [100]. Other than the need to raise awareness of the impact of social influence within social media communities, these results suggest that online social media aggregators are a viable testbed for theories of trust and social influence.

Like social networks, online social news platforms allow individuals to contribute to the wisdom of the crowd in new ways. These platforms are typically Web sites that contain very simple mechanics. In general, there are 4 operations that are shared among social news sites:

1. individuals generate content or submit links to content,
2. submissions are rated and ranked according to their rating scores,
3. individuals can comment on the submitted content,
4. comments are rated and ranked according to their rating scores.

Simply put, these platforms allow individuals to submit content and vote on the content they like or dislike.

The voting mechanism found in socio-digital platforms provides a type of Web-democracy that is open to everyone. Given the widespread use and perceived value of these voting systems [42], it is important to consider whether they can successfully harness the wisdom of
the crowd to accurately aggregate individual information. In our study, we determine what effect, if any, ranking and vote score has on rating behavior. This is accomplished via an *in vivo* experiment on the social media Web site, Reddit, by inserting random votes into the live rating system.

Reddit is a social news Web site where registered users can submit content, such as direct posts or links. Registered users can then up-vote submissions or down-vote submissions to organize the posts and determine the post’s position on the site; posts with a high vote score (*i.e.*, up-votes – down-votes) are ranked more highly than posts with a low vote score. Reddit is organized into many thousands of “subreddits,” according to topic or area of interest, *e.g.*, news, science, compsci, datamining, and theoryofreddit, and posts must be submitted to a subreddit. A user that subscribes to a particular subreddit will see highly ranked posts from that subreddit on their frontpage, which Reddit describes as ‘the front page of the Internet’ and is unique for each user. Figure 2.1 illustrates an example post and a small piece of its comment section.

As in most social media Web sites, users are free to comment on the posts. Reddit has a unique commenting framework that ranks comments based on their scores relative to their sibling comments. For instance, all root comments, *i.e.*, comments with no parent, are ranked together, and all of the children-comments of some single parent-comment are ranked together. It is possible, even frequent, that a comment deep within the comment thread-tree will have a higher score than its parent or ancestor-comments [128]. By default, Reddit only displays the top 200 comments, even though it is common for popular posts to receive thousands of comments. Therefore, many comments in popular threads are never viewed, which likely exacerbates the rich-get-richer effect that is already seen in certain ranking systems.

It is important to note that, unlike other online social spaces, Reddit is not a social *network*. the notion of friendship and friend-links, like on Facebook, is mostly absent on Reddit. Although usernames are associated with posts and comments, the true identity
Figure 2.1. Composite, redacted screenshot of Reddit. (A) There are many possible ranking systems on Reddit; this image shows the first post when ordered by the “top”-scored posts within the past month. (B) Authenticated users may up-vote, or down-vote once on any post; the score of a post is congruent to the total number of up-votes minus the total number of down-votes. (C) Each post displays its rank on the far left corresponding to its position in the selected ranking system, a title text, the host (domain) of the linked content on the far right, and the number of comments below (this post also has 29,266 comments). (D) The top 200 comments are displayed in order as well corresponding to the chosen ranking system and number of points the comment has received; an orange-red arrow indicates that the current user up-voted this comment. (E) This comment has a score of 5,997, which is congruent to the number of up-votes minus the number of down-votes that the comment has received. (F) Comment threads are hierarchical such that each comment can have have children, siblings, etc. thus comment orderings are based on their vote score relative to the sibling comments in the thread hierarchy.
of registered users is generally unknown and in many cases fiercely guarded. In fact, we attempted to find friendship by looking at user-pairs that frequently reply to each other in comments; unfortunately, more than 99.9% of the comments were in reply to a user that they had never previously replied to. Thus, we typically refer to Reddit a social non-network, and the vast amount of previous social network literature does not apply.

In this chapter, we report the results of two large in-vivo experiments on Reddit; the first \( (N = 93,019) \) up-voted or down-voted posts at random and the second \( (N = 128,316) \) up-voted or down-voted comments at random. Based on these experimental treatments we observe the effects that votes have on the final score of a post or comment as a proxy for observing herding effects in social news. Unlike the experimental study performed by Muchnik et al., and other behavioral studies our experiments: 1) manipulate votes of posts and comments rather than just comments, 2) leverages Reddit’s dynamic, score-based ranking system rather than a time-only ranking system, 3) does not involve friendship or the use of social networks, and 4) randomly delays the vote treatment rather than always performing the treatment immediately upon creation. These differences are significant in that this is the first ever vote manipulation experiment on a global scale, live, working system. The use of randomized trials eliminates concerns about various confounding factors, and we have made our data and analysis scripts available to the community for replication and further research.

We find that small, random rating manipulations on social media posts and comments created significant changes in downstream ratings resulting in significantly different final outcomes. We found positive herding effects for positive treatments on posts, increasing the final rating by 11.02% on average, but not for positive treatments on comments. Contrary to the results of related work, we found negative herding effects for negative treatments on posts and comments, decreasing the final ratings on average, of posts by 5.15% and of comments by 37.4%. Compared to the control group, the probability of reaching a high rating \((\geq 2000)\) for posts is increased by 24.6% when posts receive the positive
treatment and for comments is decreased by 46.6% when comments receive the negative treatment.

2.2 Methodology

2.2.1 Post Experiment

During the 6 months between September 1, 2013 and January 31, 2014 a computer program was executed every 2 minutes that collected post data from Reddit through an automated two-step process. First, the most recent post on Reddit was identified and assigned to one of three treatment groups: up-treated, down-treated, or control. Up-treated posts were artificially given an up-vote (a +1 rating) and down-treated posts were given a down-vote (a -1 rating). Up-treatment, down-treatment and the control have an equal likelihood of being selected. Vote treated posts are assigned a random delay ranging from no delay up to an hour delay in intervals of 0, .5, 1, 5, 10, 30 and 60 minutes. Second, each post was re-sampled 4 days later and final vote totals were recorded.

These treatments created a small, random manipulation signalling positive or negative judgment that is perceived by other voters as having the same expected quality as all other votes thereby enabling estimates of the effects of a single vote while holding all other factors constant. This data collection resulted in 93,019 sampled posts, of which 30,998 were up-treated and 30,796 were down-treated; each treatment type was randomly assigned a delay interval with equal likelihood. Treatments were removed from the vote scores before data analysis was performed, i.e., up-treated post-scores were decremented by 1 and down-treated post-scores were incremented by 1.

During the experimental time period, Reddit reported that their up-vote and down-vote totals were “fuzzed” as an anti-spam measure; fortunately, they certified that a post’s score (i.e., up-votes minus down-votes) was always accurate. In July of 2014, after the data gathering phase of this experiment had ended, Reddit removed the vote totals from
their Web site and replaced it with a semi-accurate points system; Reddit administrators currently assert that the rankings are always accurate, even though their reported scores may not be.

2.2.2 Comment Experiment

During the 6 months between September 1, 2013 and January 31, 2014 a computer program, separate from the post experiment, was executed every 2 minutes that collected comment data from Reddit through an automated two-step process. First, the most recent comment on the top ranked post ordered by the ”rising” ranking algorithm on the Reddit frontpage was identified and assigned to one of three treatment groups: up-treated, down-treated, or control. Up-treated comments were artificially given an up-vote (a +1 rating) and down-treated comments were given a down-vote (a -1 rating). Up-treatment, down-treatment and the control have an equal likelihood of being selected. Vote treated comments are assigned a random delay ranging from no delay up to an hour delay in intervals of 0, .5, 1, 5, 10, 30 and 60 minutes. Second, each comment was re-sampled 4 days later and final vote totals were recorded.

These treatments produced a score manipulation similar to that of the post experiment, wherein all other factors were held constant enabling a clear causal signal to be measured. This data collection resulted in 96,486 sampled comments, of which 35,704 were up-treated and 31,830 were down-treated; each treatment type was randomly assigned a delay interval with equal likelihood. Treatments were removed from the vote scores before data analysis was performed, i.e., final up-treated comment-scores were decremented by 1 and final down-treated comment-scores were incremented by 1. To our knowledge, comment scores were not fuzzed in the same way that post scores are fuzzed, so absolute point scores reported here should be accurate.

The voting agents used were periodically checked to ensure that they had not been blocked or their votes ignored. The voting agent did not target any one type of content.
or subreddit or content provider, which are among the most common types of vote-spam, therefore, we are certain that all of our votes were counted. Distributions across treatments for the post and comment experiments are summarized in Table 2.1.

2.3 Results

We first compared the final vote totals of each treatment. These findings measure the overall effect that up-treatments and down-treatments have on the overall life of a post or comment. Figure 2.2 shows the full distribution of the final post scores and comment scores for each treatment group. Black outer error bars show the 95% confidence interval and red inner error bars show the standard error of the mean. The full distribution of post scores in Figure 2.2 (a) is extremely positively skewed with a skewness of 11.2 and a kurtosis of 149.8. If we remove the top 1% highest scoring posts from the data set the skewness and kurtosis values drop to 6.5 and 54.9 respectively giving a better, although still skewed, view of the treatment effects. Figure 2.3 (a) shows the distribution of the final post scores with the top 1% of posts removed. In this case, the up-treated posts have a significantly higher final score, and the down-treated posts have a significantly lower final score.

The distribution of comment scores in Fig. 2.2 (b) is even more positively skewed than the distribution of post scores with a skewness of 16.4 and a kurtosis of 339.7 but
Figure 2.2. Final scores for artificially, randomly up-treated posts, down-treated posts, and scores for untreated posts in the control group are shown. Red inner error bars show the standard error of the mean; black outer error bars show the 95% confidence interval. At left, (a) shows the scores in the heavily skewed full distribution for posts. At right, (b) shows the scores in the heavily skewed full distribution for comments, with significant decreases for down-treated comments when compared to the control group.

when the top 1% highest scoring comments are removed, the skewness and kurtosis values dropped to 6.7 and 58.1, similar to the skewness and kurtosis for the distribution of post scores when the top 1% highest scoring posts are removed. In this case, the down-treated comments have a significantly lower final score but the up-treated comments do not have a significantly higher final score.

Tests of statistical significance, e.g., T-test, are known to improperly reject the null hypothesis when the data distribution is non-normal or highly skewed. This is indeed the case in our result set as is indicated by the abnormally high skewness and kurtosis scores.

Removal of the top 1% of scores is one way to unskew the data, hence the tightening of error bars and narrowing of confidence internals in Fig. 2.3 as compared to the full results in Fig. 2.2. Another way to unskew data is to take the log of each value in the distribution, which unfortunately removes negative scores from the analysis, a significant limitation for
Figure 2.3. Top 99% of final scores for artificially, randomly up-treated posts, down-treated posts, and scores for untreated posts in the control group are shown. Red inner error bars show the standard error of the mean; black outer error bars show the 95% confidence interval. When the highest 1% of post scores are removed, the score distribution becomes much less skewed resulting in tighter error bounds, which further result in significant increases for up-treated posts and significant decreases for down-treated posts when compared to the control group. Again, when the highest 1% of comment scores are removed, the score distribution becomes less skewed resulting in tighter error bounds, but with slight but not significant increases for up-treated comments when compared to the control group.

Student’s T-Test on the full set (i.e., 100%) of log-scores for posts also showed that the up-treated posts were significantly higher than the control group ($p = 1.69 \times 10^{-20}$), and that the down-treated posts were significantly lower than the control group ($p = 1.69 \times 10^{-09}$), although scores less than or equal to 0 were removed to calculate the log of the final scores. For comments we find that Student’s T-Test on log-scores demonstrated that up-treated posts were significantly higher than the control group ($p = 2.69 \times 10^{-24}$), and down-treated posts were significantly lower than the control group ($p = 9.18 \times 10^{-05}$). Unfortunately, the distribution of log-scores was still far from normal, so the T-Test is
likely to give improper results.

With this in mind, we used the non-parametric, 1-dimensional Kolmogorov-Smirnov (K-S) test as well as the Mann-Whitney U (M-W) Test to determine the significance between treatments and control. Both the M-W and the K-S tests are nonparametric tests to compare two unpaired groups of data. They each compute p-values that test the null hypothesis that the two groups have the same distribution. They do have some important differences though. The M-W test operates by ranking all the values from low to high, and then computes a p-value that depends on the differences between the mean ranks of the two groups. The K-S test compares the cumulative distribution of the two data sets, and computes a p-value that depends on the largest difference between the two distributions. The differences between the two tests are important, but they both compute p-values that can be used to judge the statistical significance of the treatment effects. Thus, we will display the results of both tests.

The K-S Tests showed that the final score distribution of all up-treated posts were more positively skewed than posts in the control group (K-S test statistic: 0.08; \( p < 2.2 \times 10^{-16} \)), which were more positively skewed than down-treated posts (K-S test statistic: 0.11; \( p < 2.2 \times 10^{-16} \)). The same K-S test on comment scores shows significantly higher final scores for up-treated comments and down-treated comments \( (p < 2.2 \times 10^{-16}) \). The reason that the p-value of the K-S statistic is reported as being less than \( 2.2 \times 10^{-16} \) is because floating point underflow error prevents a more precise calculation in the R-based K-S test calculator.

Finally, we performed the independent 2-group M-W Test comparing treatments (up-treated and down-treated) with the control. We again find significant differences comparing the up-treated post scores to the control \( (p = 5.9 \times 10^{-53}) \) and the down-treated post scores to the control \( (p = 7.8 \times 10^{-73}) \). The same M-W Test on comments also showed significant differences in the final scores of up-treated comments compared to the control group \( (p = 5.57 \times 10^{-15}) \), and significantly different final scores in down-treated comments.
Figure 2.4. Final scores separated into their respective treatment delay intervals.

At left (a) shows final scores for artificially, randomly up-treated posts, down-treated posts, and scores for untreated posts in the control group and, at right, (b) shows final scores for artificially, randomly up-treated comments, down-treated comments, and scores for untreated comments in the control group. Horizontal lines show the overall mean of each treatment group. The top 1% of scores were removed to un-skew the score distribution.

compared to the control group ($p = 7.52 \times 10^{-8}$).

In general, an up-vote increases a post’s score on the site which increases its visibility according to the default ranking algorithms. The increased visibility of the post makes it more likely to be viewed by others. However, making a post more visible does not necessarily mean that it will receive more up-votes and continue to increase or even maintain its visibility; it may instead receive down-votes, thereby decreasing the posts visibility. That is, until we consider that the vast majority of votes cast on Reddit are up-votes and down-voting is actually discouraged unless the post is spam, off-topic, or otherwise improper. Thus, we are confident that the increase in the final post score after positive vote manipulation in the presence of popularity ranking mechanics is largely due to the increase in visibility due to the treatment up-vote.

Comments, in contrast, have a vastly different visibility mechanism than posts. Reddit comment threads are hierarchical, wherein the default “best” (highest up-vote to down-vote ratio) ordering mechanism sorts comments among its siblings only. The visibility
of a comment in the hierarchy depends not only on its ordering among its siblings but also the rank of any parent or ancestor comments it has. Because our voting mechanism selected the most recent comment, it may be the case that the selected comment was a child or other descendant of a highly visible comment. As such, it may be the case that the treated comment was already highly visible by its relative position in the comment hierarchy. Unfortunately, we did not record the relative position of each treated comment and are unable to find correlation between relative visibility and treatment effects.

Another difference in the comments experiment is that, by default, only the top 200 comments are visible. By selecting “rising” posts, our collection methodology makes it highly likely that the comment that we select is within the first 200, and is therefore at least initially visible. Unfortunately, for large comment threads a single down-treatment may be enough to make the comment no longer visible under default orderings. This is probably why down-treated comments have such a low overall score compared to up-treated or control groups.

2.3.1 Delay Effects

Up-votes and down-votes for post receiving treatments were performed after a 0, 0.5, 1, 5, 10, 30 or 60 minute delay chosen at random, and Figures 2.2 and 2.3 does not distinguish between the effects of vote-treatments performed after the various delay periods. Figure 2.4 separates the results for posts from Figure 2.3 (a) and Figure 2.3 (b) into their respective treatment delay groups in Figure 2.4 (a) and Figure 2.4 (b), respectively. We expected that immediate votes would have a larger effect than votes performed after a long delay. However, these results show, surprisingly, that a delay in treatment generally did not have a significant effect on the mean outcome of a post’s final score.

Unfortunately, displayed error bounds and confidence intervals, which are computed from Student’s T-Test, have little meaning when the data is so highly skewed; K-S test results shown in Figure 2.5 again showed that all up-treated posts were more positively
Figure 2.5. Results of Kolmogorov-Smirnov and Mann-Whitney $U$ Tests (with top 1% included) for each treatment delay where ▲ represents tests comparing up-treatment with the control group and ▼ represents tests comparing down-treatment with the control group. † indicates results that are not statistically significant.
skewed than posts in the control group and that the effects generally diminished as the delay interval increased. Similarly, the control group was more positively skewed than the down-treated posts, but the effects were mixed as the delay interval increased.

As for comments, the K-S test results were more mixed in Figure 2.5 but still mostly statistically significant. The up-treated comments were significantly more positively skewed than the control group comments, and the down-treated comments resulted in a significantly lower score. Interestingly, the p-values of the comment scores diminished as the delay grows longer, meaning that the vote treatment on comments are not effective a half-hour or an hour after the comment has been made. In short, timely voting on a comment is more important than timely voting on a post on average.

M-W tests of statistical significance, also shown in Figure 2.5 demonstrate that post treatments have a significant effect across all delay periods, and that this effect only slightly diminishes (if at all) when the delay approaches 1 hour. As for comment treatments, the M-W tests showed significance results similar to those from the K-S tests. Namely, the effect of up-treatment, as measured by the p-value scores, diminished as the delay grew bigger and led to an insignificant effect when the delay was 1 hour. The effect of down-treatment was significant for short delay periods, but was not significant for delays of 10 minutes and 30 minutes, and was only barely significant for delays of 1 hour.

The results from the statistical tests on the comment treatments from Figure 2.5 and Figure 2.4(b) appear to be in conflict. Figure 2.4(b) seems to show that negative treatments have a big effect on the final outcome of the comment for all delay levels, while positive-treatments have a little effect. However, proper statistical tests show that the truth is more nuanced. Ultimately, with this type of data, the best way to show aggregate results is through n-tile plots. With this in mind, Figure 2.6 shows the inner-deciles of the results as a function of their treatment delay. Taken together these results show graphically what the tests of statistical significance imply: that up-treated posts tend to score more highly than the control group, and that down-treated posts tend not to not score as highly as the control
group. The decile plots also show that the majority of posts (deciles $\leq 50\%$) receive at most a final score of 2, and that most comments never receive any votes at all.

2.3.2 Reaching the Front Page

Overall, the results suggest that an up-treatment increases the probability that a post will result in a high score relative to the control group, and that down-treatments decrease that probability relative to the control group. However, on Reddit and other social news sites only a handful of posts become extremely popular. On Twitter and Facebook this is generally referred to as a trending topic, but on Reddit the most popular posts are the ones that reach the front page. Unfortunately, reaching the front page is a difficult thing to discern because each user’s homepage is different, based on the topical subreddits to which the user has subscribed. Nevertheless, we crudely define a post as having become popular, i.e., is trending, on the frontpage, etc., if it has a score of more than 500. Using this definition, Figure 2.7 (a) shows the probability that post reaches a given final score under the two treatment conditions. These probability distribution functions are monotonically decreasing, positively skewed, and show that up-treatment results in a large departure from the control group for posts and down-treatment results in a large departure from the control group for comments. However, despite our earlier claims of up-treatment and down-treatment symmetry on post results, these results show that, in the upper limits of the distribution, down-treatments do not effect the final score results. These results mean that, compared to the control group, an up-treated post is 7.9\% more likely to have a final score of at least 1000, and an up-treated post is 24.6\% more likely to have a final score of at least 2000.

The probability that a comment reaches a high score is generally lower than the probability of a post reaching the same high score because posts are generally more viewed and voted on than comments. Indeed, in order to even view the comments, a user must first view, or at least click-on, the post. Also, lower rated comments or comments with
Figure 2.6. The middle 9 deciles of final scores for each treatment according to their delay intervals. These results show that most posts and comments receive a median score of 2 or less, and that treatment has the most effect in the higher deciles of the score distribution.
Figure 2.7. The probability of a post (at left) or a comment (at right) receiving a corresponding score by treatment type. The inset graphs show the complete probability distribution functions. The outer graphs show the probability of a post/comment receiving scores between 500 and 2000 – an approximation for *trending* or *frontpage* posts. Up-treated posts are 24% more likely to reach a score of 2000 than the control group.

Multiple levels of ancestor comments above them are often hidden until a user chooses to reveal them. Figure 2.7 (b) shows the probability that a comment reaches a given final score under the two treatment options as in Figure 2.7 (a). Interestingly, we find that an up-treatment has very little effect on the probability of a comment reaching a high score; yet, a down-treatment has a dramatic negative effect on that probability.

2.3.3 Subreddit Effects

We finally investigated treatment effects in the top 10 most frequent subreddits. These do not necessarily correspond to the top 10 most popular subreddits or the subreddits with the most comments. Rather, they are the subreddits to which posts are most frequently submitted or whose posts are most frequently ranked first on Reddit’s “rising” ranking system due to our data collection methodology.

From the top 10 subreddits for posts, we removed *politic* and *r/friendsafari* and from the top 10 subreddits for comments, we removed *r/friendsafari*. These subreddits were
removed from our analysis because posts in \texttt{politic} are automatically submitted by a computer program, and because posts and comments in \texttt{friendsafari} cannot be down-voted according to the subreddit rules. Thus, only 8 subreddits for posts and 9 subreddits for comments are shown in Figure 2.8 which illustrates the effects of treatment on post and comment scores on average within top 10 subreddits.

Figure 2.8 (a) and Mann-Whitney test results show significant positive effects on post scores in \texttt{r/AdviceAnimals}, \texttt{r/AskReddit} and \texttt{r/videos}, and significant negative effects on post scores in \texttt{r/AskReddit} and \texttt{r/pics}. These results illustrate similar symmetric effects that we found on posts overall. Voting effects on comment scores within subreddits are shown in Figure 2.8 (b). While we find that down-treatments typically result in significantly lower final comments scores compared to the control, up-treatments rarely result in significantly higher final comment scores as shown by Figure 2.8 (b).

Within the top 500 subreddits for posts, we found that 22\% had significant up-treatment effects, 21.6\% had significant down-treatments, and 5.4\% of subreddits had significant differences in both the up-treatment and down-treatment results when compared to the control group. There was also no correlation in the top 500 between up-treatment significance and
number of submissions the subreddit received ($r^2 = 0.014$; p-value = 0.007), nor down-
treatment significance and number of submissions the subreddit received ($r^2 = 0.010$; p-value = 0.026).

2.4 Related Work

Although this is the first in-vivo Reddit experiment, our work is motivated and informed by multiple overlapping streams of literature and build on substantial prior work from multiple fields such as: herding behavior from theoretical and empirical viewpoints [105, 128]; social influence [7]; collective intelligence [2, 56]; and online rating systems [78]. A recent study by Muchnik et al on a small social news Web site, similar to Reddit, found that a single up-vote/like on an online comment significantly increased the final vote count of the treated comment; interestingly, the same experiment also found that a single negative rating had little effect on the final vote count [91].

In a separate line of work, Sorensen used mistaken omissions of books from the NY Times bestsellers list to identify the boost in sales that accompany the perceived popularity of a book’s appearance on the list [111]. Similarly, when the download counters for different software labels were randomly increased, Hanson and Putler found that users are significantly more likely to download software that had the largest counter increase [53]. Salganik and Watts performed a study to determine the extent to which perception of quality becomes a “self-fulfilling prophecy.” In their experiment they inverted the true popularity of songs in an online music marketplace, and found that the perceived-but-false-popularity became real over time [104].

These experiments aim to determine the causal effect of social influence on rating behavior, as well as the mechanisms driving socio-digital influence. Although these experiments are first-of-a-kind, they are motivated and informed by multiple overlapping streams of literature and build on substantial prior work from multiple fields such as: herding behavior from theoretical [10, 14, 52] and empirical viewpoints [3, 19, 73, 105, 134]; social
influence in networks \([4, 7, 72, 84, 92]\); collective intelligence \([14, 17, 56, 132]\); and online rating systems \([20, 21, 28, 31, 32, 59, 76, 78, 82, 92, 134, 142]\). Interestingly, most of the previous work is geared towards marketing science because of the close relationship between business and consumer opinion.

The dynamics of online reviews, ratings and votes have received a lot of recent attention in the computing and marketing literature because the dynamics of online reviews for books, restaurants, hotels, etc. have become a vital business interest \([28, 31, 32, 59, 76, 82, 142]\). Recent work in text mining is able to automatically determine the positivity and negativity of user-opinion \([78–80]\) even among different aspects of a certain product (e.g., large can be a good thing when talking about portion size, but bad when talking about camera size) \([81]\). These papers attempt to codify ratings from plain, user-generated text and then determine relationships between the ratings and popularity.

Nonetheless, studies that aim to demonstrate the ease of online manipulation of ratings or voting tend to be limited. The biggest limitation is that these studies assume that the manipulators have full knowledge of the voting preferences of every user – a valid assumption in theoretical work, but a meaningless assumption in real-world applications \([11, 12, 41, 106]\). There is some work that considers manipulators who have a limited \([24]\) or probabilistic \([9, 83]\) knowledge of the voting preferences, but these assumptions are still too limiting for our purposes.

On the practical side, one obvious case of online manipulation is spam, particularly a new type of spam called social spam. Social spam is on the rise, with Nexgate Research reporting a tripling of social spam activity every six months \([93]\) and BusinessWeek magazine reporting that only 40% of online social media users are real people \([63]\). There has been some practical work on detecting social spam in online social networking Web sites \([119]\) like Facebook and Twitter, but not in social news platforms like Reddit and HackerNews. The largest and perhaps most effective type of social spam relies on social networks to broadcast and propagate the advertisement or message \([6, 40, 85, 140]\). These
social network spammers are also the easiest to detect and shut down. However, social news platforms are purposefully not social networks.

2.5 Discussion and Conclusions

In general, we find that the positive treatment of a single, random “up-vote” on a Reddit post has a corresponding positive herding effect that increases post scores on average and in the top limits of the heavily skewed score distribution but that a single, random “up-vote” on a Reddit comment had no significant positive herding effects. We further found that the negative treatment of a single, random “down-vote” on a post or comment has a corresponding negative herding effect that significantly decreased the post or comment scores on average, in contrast to the asymmetric findings of Muchnik et al. [91], who found no significant effects of a negative treatment. However, our results begin to resemble asymmetry in the top limits of the post score distribution meaning that a negative treatment does not decrease the probability that a post will receive a high score in the way that it does for comments.

Separating treatments by their delay intervals did not yield a significant difference in effect overall. K-S and M-W tests found that up- and down-treatments for most delay intervals had significant effects compared to the control. In general, the time that a vote is placed did not change the overall effect for post scores, but longer delays did diminish the effects that votes had on comment scores.

2.5.1 Voting and Viewing Mechanics

Research in social news manipulation has been shown a great deal of interest in recent years because of its centrality in shaping the news and opinion of society. There are several conflicting reports that now need to be teased apart. The work by Muchnik et al. showed positive herding effects but not negative herding effects [91] on non-Reddit social media comments ranked by recency, rather than popularity, and in the presence of a friendship
social network. The voting and visibility mechanics of Reddit, which govern the data collection in this chapter, are vastly different then the small or contrived experiments studied in earlier work.

The post experiment and results are actually more in line with past research on ranking and visibility bias [71] because of how the ranking mechanics of posts impact visibility. The results of our analysis as described above and the behavior of voting on Reddit, with an overwhelming majority of votes being upvotes and the discouragement of down-voting posts that are appropriate, support an increase in popularity from an increase in visibility. Thus, we are confident that the increase in the final post score and the probability of reaching a high post score after positive vote manipulation in the presence of popularity ranking mechanics is largely due to the increase in visibility due to the treatment up-vote.

2.5.2 Vote-based Manipulation

Collectively, this work, in the presence of other work on this subject [45, 71, 91, 114], shows that votes determine visibility which, in turn, drives more votes. The 1% rule, or its variants like the 90-9-1 rule, the 80/20 rule or the pareto principle, when applied to social media indicates that about 90% of users only view content, 9% of users edit content (including voting), and 1% of users actively contribute new content [54, 120]. On all manner of vote-based social media platforms, the 10% of users who actually vote are the ones who determine the kind of content that becomes widely visible and circulated among the remaining 90% of the viewing public. Therefore, that active 10% determines the ideas and opinions that the public is exposed to and influenced by.

Clearly, there is a huge incentive for opinion-pushers to manipulate the visibility of certain ideas and opinions on social media Web sites. There are several types of vote-based manipulation techniques that exist. We discuss a few of them here.
**Vote Brigading**  Vote brigading is when a large group of people all conspire to up-vote or down-vote a particular post or idea. This is not unique to Reddit, as Twitter has its Retweet armies that attempt to manipulate the velocity of some discussion in order to artificially force a topic to become a “trending” topic. The social media Web site Digg was particularly susceptible to vote brigading, wherein only users with many friends could ever hope to have a post reach the frontpage because the poster’s friends would initially vote on the post in order to raise its visibility enough so that the wider community to see it.

Fortunately, most forms of this type of vote manipulation can be easily detected and stopped with spam detection and prevention techniques [33, 69]. As part of a larger strategy, Reddit now encourages hyperlinks between subreddits to be tagged with a no-participation URL, which restricts access for non-subscribers of the subreddit to read-only, in order to prevent “cross-subreddit contamination”[1]. For example, a no-participation link from /r/yankees to a post in /r/redsox (a historical baseball rivalry) would prevent Yankees fans from downvoting posts that favor Red Sox fans.

**Vote Nudging**  Vote nudging is the type of vote manipulation that is studied in this chapter and is the easiest, and most common, type of vote manipulation on social media. A post or comment is most susceptible to being ignored when it is young. Vote nudging is when someone asks a few friends to up-vote the post or comment in order to give it a positive boost during its initial appearance. After the initial boost, the post is left to grow normally.

As we have shown in this study, vote nudging can be extremely successful because the default ranking system gives higher visibility to posts with more, timely votes. Vote nudging also prevents instances when an unrelated user down-votes, and effectively kills a posts’ changes of becoming visible, because a post with three or four up-votes may be able to withstand the effects of a down-vote better than a post with no up-votes.

It is difficult to say how much vote nudging happens in social media. It is common for

[http://www.reddit.com/r/NoParticipation](http://www.reddit.com/r/NoParticipation)
users to have multiple accounts for this reason, but multiple votes from the same IP address is easy for spam prevention systems to catch.

**Reverse Vote Nudging**  Reverse vote nudging is when a user down-votes all of the posts or comments that are similar to their post or comment in order to make a relative gain on competing content. For example, if a user contributes a post about the winner of a baseball game to /r/yankees, several other users may also have contributed posts about the same baseball game at about the same time to /r/yankees. In order the increase the relative ranking of their own post, the user may down-vote all of the other posts by the other users thereby increasing the relative ranking of their submission.

Similarly, a user may wish to down-vote all posts or comments that are ranked just above the user’s submission in order to increase the relative ranking of their submission. Using the same baseball example as earlier, if there are posts dealing with other Yankees content that are ranked just above the user’s post, then the user may increase the ranking of their own post, and therefore increase its visibility, by downvoting the other content.

### 2.5.3 Conformity and Influence

Comment threads on Reddit are a unique supplement to the posted content. In fact, it is widely thought that most social media users, across all types of platforms, read the title of the post and skip directly to the comments section – although this has not been empirically researched. Also, Reddit, Youtube, Twitter and many other social media platforms, to some extent, show the current score of each comment in the comment thread. Thus, the opinions and ideas expressed in each comment are given an explicit rating from the voting user base that is often viewed as the prevailing opinion of the overall population.

The Asch conformity experiments in the 1950’s and onward showed that perceptions of popular opinion can have profound effects on individual perceptions of the truth [5]. Social comment threads frequently have instances where the highest scored comments
represent an incorrect fact or are contrary to be the prevailing public opinion, perhaps due to comment manipulation discussed above. However, it is sometimes uncomfortable for many comment readers to hold opinions contrary to what they perceive to the the prevailing opinion. This disillusionment sometimes leads to a position change, but can also lead to a retreat inwards due to confirmation bias, which, in the worst case, leads to radicalization.

2.5.4 Voting

The nature of the manner in which social platforms rank items for viewing typically utilizes the ratings, in this case the post or comment scores, of the items being ranked. The results of our experiments show that random vote perturbations through vote treatments impact the scores of posts and comments on Reddit. These results underscore the need for counter measures against vote chaining and social engineering strategies as multiple artificial votes are likely to increase the herding effect.

Finally, we bring attention back to what Eric Gilbert calls, the ‘widespread underprovision of votes’ in social media like Reddit [42]. Although our data does not draw these figures explicitly, we estimate that a very small number of the daily visitors to social media Web sites actually vote on the items they view. This seems to be an even further skewed anecdote of the 1-9-90 rule of social networking [120], and may be an underestimated reason behind the results presented in this chapter.
CHAPTER 3

CONSUMPTION AND CURATION HABITS IN SOCIAL RATING SYSTEMS

As crowd-sourced curation of news and information have become the norm, it is important to understand not only how individuals consume information through social news Web sites, but also how they contribute to their ranking systems. In this chapter, we present findings that highlight the browsing and voting behavior of the study’s participants. We find that most users do not read the article that they vote on, and that, in total, 73% of posts were rated (i.e., upvoted or downvoted) without first viewing the content. We also show evidence of cognitive fatigue in the browsing sessions of users that are most likely to vote.

3.1 Introduction

As crowd-sourced curation of news and information have become the norm, it is important to understand not only how individuals consume information through social news Web sites, but also how they contribute to their ranking systems. The sheer volume of new content that is produced and consumed only increases the reliance that individuals place on the ability of anonymous others to curate and sort massive amounts of information. In this chapter, we introduce a new dataset containing the activity logs that recorded all activity for 309 Reddit users for one year and an analysis of the browsing and voting behavior of the study participants’ 1,801,405 total interactions August 1, 2015 through July 31, 2016.

Social news aggregators like Digg, HackerNews, and Reddit are social rating systems that allow individuals to post news and information for others to view, vote, and comment on. Unlike online social networks such as Facebook and Twitter, social news aggregators
typically lack a strong user identity, \textit{i.e.}, users are mostly anonymous and social relationships like friendship or leader-follower do not exist. In the absence of social relationships, social news aggregators rely heavily on user-votes when deciding which content to deliver to the user. In this way, news aggregators allow \textit{information consumers} to also act as \textit{information curators} through their votes.

Social news aggregators represent a stark departure from traditional media outlets in which a news organization, \textit{i.e.}, a handful of television, radio, or newspaper producers, sets the topics and directs the narrative. Socio-digital communities increasingly set the news agenda, cultural trends, and popular narrative of the day [70, 74, 90]. News agencies frequently have segments on “what’s trending,” entire television shows are devoted to covering happenings on social media, and even live presidential debates select their topics based on popular questions posed to social media Web sites. As this trend continues and grows, it is important to understand not only how individuals consume information that is delivered to them but also how their online behavior shapes the news and information that countless others consume.

When researchers study social media, they study the collective actions of a community by looking at coarse-grained signals of posts, comments, and vote totals. For example, the 90-9-1 rule, representing the proportion of Wikipedia browsers, editors, and creators [94], might be reflected in social news aggregators as the number of browsers, commenters, and posters. Other work in this area has found that user behavior can be easily influenced and manipulated through injections of artificial ratings [43, 44, 91, 129]. This is especially relevant in light of the finding that 59\% of bitly-URLs on Twitter are shared without ever being read [38]. Votes, artificial or not, influence the score of a post, which is the primary means of ranking the content. This leads to a “rich get richer” effect where a positive rating directly results in a higher score and therefore a higher ranking; the increased ranking raises the visibility of the post, which provides a better opportunity for other users to view and vote on the post.
Nearly all studies involving the analysis of social media rely exclusively on the information provided by the social media platform itself, typically via API calls or by crawling the Web site. Unfortunately, this means that many questions about browsing behavior, voting habits, and commenting remain unanswered.

**Complete activity logs of Reddit usage.** In contrast to previous studies, our analysis of social media behavior is based on complete activity logs for 309 users of the popular social news aggregator Reddit from August 1, 2015 to July 31, 2016. Study participants were asked to download a browser extension (available for the Firefox and Chrome browsers) that reported usage data of all clicks, pageloads, and votes made within the `reddit.com` domain. Upon installation, the browser extension asked the user to opt-in to the data collection via an informed consent form; the user could modify their consent or uninstall the extension by clicking a small icon that was embedded next to their username in the top-right of the Reddit Web page. This experiment was reviewed and approved by ethics review boards at the University of Notre Dame and the Air Force Office of Scientific Research.

**Recruitment and Sampling Bias.** Study participants were recruited through posts to various subreddits. Participants were required to be a registered Reddit user. Their account must have been created at least a week before installing the browser extension in order to remove the potential for malicious users. We only recorded activity that occurred while the participant was signed in to their Reddit account. Activity that occurred within private or “incognito” browsing modes was not collected. All data was anonymized on the client-side. We did not collect IP addresses of the user, nor did we collect posts or comments made by the user as these could be used to easily de-anonymize the data.

It is important to be cognizant of potential sampling bias. The only way to collect this data is for users to self-select, *i.e.*, opt-in, to the project. Self-selection bias, especially in online polling, for example, will often skew responses in favor of the most motivated group. On the contrary, our system does not ask any questions; it merely observes user behavior. Our recruitment strategy may be affected by undercoverage bias, where certain groups are
not included in the sample. Because of the potential for undercoverage bias, we do not intend for this data to be representative of broad or group-based opinion or popularity. Indeed, the opinions and views expressed on Reddit itself are not broadly representative of the general public. Therefore, in this study we only consider how registered users browse and vote on Reddit, not what they browse and vote on.

We offer a first look at this newly collected data through two complementary topics:

1. **Information Consumers and Curators.** Because of the critical role that news and information play in everyday life, it is important to understand the dynamics of modern media delivery. The present work is a first look at the private behavior of users of a large social news aggregation Web site. In the first section, we investigate the duality of users in social media, wherein information consumers, *i.e.*, browsers or lurkers, also act as information curators through their votes. With the new data we can begin to answer questions about the variety of information that people browse, and their behavior before and after a vote.

2. **Performance Deterioration in Browsing Sessions.** It is also important to consider the acuity of a user during curation. Human performance is known to diminish after periods of sustained mental effort. The same is probably true in online browsing sessions. Recent work has found that a user’s online effort, measured by comment length and writing level, typically declines as a browsing session progresses [110]. However, relatively little is known about how this phenomenon affects voting behavior, which is critical to understand because votes directly determine the content that other people view.

We found that most users do not read the article that they vote on, and that, in total, 73% of posts were rated (*i.e.*, upvoted or downvoted) without first viewing the content. We also found evidence of cognitive fatigue in the browsing sessions of users that are most likely to vote.

3.2 Information Consumers and Curators

As a way of introducing the dataset, we begin with some simple aggregate statistics. We have a total of 1,801,405 total interactions from 309 registered Reddit users. Figure 3.1 breaks down these interactions into their various types and Table 3.1 illustrates a few of
the most common topics or communities within each interaction type, listing the 10 most frequently occurring subreddits for each interaction type.

*Link-interactions* occur whenever a user clicks on a content-hyperlink of any kind, including a post’s title, a post’s comment section, content expanders, etc. Simply put, a link-interaction occurs whenever a user views the content or comments of a post. We find that 89.2% of all link interactions occur from within a ranked list, *i.e.*, a link that is ranked amidst other content such as on a user’s frontpage or a subreddit. Most of the remainder occur within a comments section, which are threaded rather than ranked. Regardless of view, 21.1% of all clicks were on post titles; 20.9% and 16.1% were on image expanders and collapsers respectively; 10.6% of clicks opened the comments section; 5.4% and 3.3% were on text expanders and collapsers respectively; and 4.6% and 3.2% were on video expanders and collapsers respectively.

*Sort-interactions* are those clicks that reorder the ranking of information displayed to the user. The default ordering is best, which ranks posts by a time-normalized score function. Re-ordering is relatively rare, occurring in only 3% of all pageloads. When reordered, users choose to sort by new 68.1% of the time followed by top, hot, rising, and controversial at rates of 14.0%, 9.9%, 6.0%, and 2.0% respectively.

*Pageload-interactions* occur anytime a Web page within the Reddit domain is loaded or refreshed. Thus, a pageload event will duplicate certain types of sort- or link-interactions.
TABLE 3.1

THE 10 MOST FREQUENT SUBREDDITS FROM WHICH INTERACTIONS OCCURRED

<table>
<thead>
<tr>
<th>Subreddit</th>
<th># Link-interactions</th>
<th>Subreddit</th>
<th># Sort-interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frontpage</td>
<td>333155</td>
<td>r/stevenuniverse</td>
<td>1962</td>
</tr>
<tr>
<td>r/all</td>
<td>68817</td>
<td>r/sandersforpresident</td>
<td>1397</td>
</tr>
<tr>
<td>r/videos</td>
<td>20190</td>
<td>r/nba</td>
<td>1299</td>
</tr>
<tr>
<td>r/funny</td>
<td>15878</td>
<td>r/politics</td>
<td>739</td>
</tr>
<tr>
<td>r/trees</td>
<td>9760</td>
<td>Frontpage</td>
<td>733</td>
</tr>
<tr>
<td>r anime</td>
<td>9498</td>
<td>r/games</td>
<td>343</td>
</tr>
<tr>
<td>r/sandersforpresident</td>
<td>8947</td>
<td>r/halo</td>
<td>306</td>
</tr>
<tr>
<td>r/leagueoflegends</td>
<td>8930</td>
<td>r/frugal</td>
<td>275</td>
</tr>
<tr>
<td>r/youtubehaiku</td>
<td>8194</td>
<td>r/personalfinance</td>
<td>252</td>
</tr>
<tr>
<td>r/wtf</td>
<td>7511</td>
<td>r/santaslittlehelpers</td>
<td>231</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subreddit</th>
<th># Pageload-interactions</th>
<th>Subreddit</th>
<th># Vote-interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frontpage</td>
<td>174805</td>
<td>r/kotakuinaction</td>
<td>26024</td>
</tr>
<tr>
<td>r/all</td>
<td>22565</td>
<td>r/sandersforpresident</td>
<td>20691</td>
</tr>
<tr>
<td>r/sandersforpresident</td>
<td>17913</td>
<td>r/askreddit</td>
<td>18499</td>
</tr>
<tr>
<td>/nba</td>
<td>15347</td>
<td>r/fatlogic</td>
<td>16105</td>
</tr>
<tr>
<td>r/funny</td>
<td>12127</td>
<td>r/politics</td>
<td>15264</td>
</tr>
<tr>
<td>r/stevenuniverse</td>
<td>10942</td>
<td>Frontpage</td>
<td>14007</td>
</tr>
<tr>
<td>r/askreddit</td>
<td>10837</td>
<td>r/trollxchromosomes</td>
<td>7985</td>
</tr>
<tr>
<td>r/politics</td>
<td>8904</td>
<td>r/thathappened</td>
<td>6124</td>
</tr>
<tr>
<td>r/kotakuinaction</td>
<td>7420</td>
<td>r/all</td>
<td>5289</td>
</tr>
<tr>
<td>r/videos</td>
<td>6977</td>
<td>r/news</td>
<td>4725</td>
</tr>
</tbody>
</table>
For example, a click into a comment section will be recorded as a link-interaction as well as a pageload. Content expanders do not count as a pageload, nor do outbound clicks to non-Reddit Web sites (but outbound clicks do count as link-clicks). We find that the Reddit frontpage was loaded in 23.6% of all pageloads, followed by r/all with 4.8% of all pageloads, further followed by a litany of subreddits.

*Vote-interactions* are those clicks that upvote, downvote, un-upvote or un-downvote a post or a comment. In order to un-vote, a user must have previously cast an initial vote; we do not substract un-votes from the vote totals. Figure 3.2 shows the breakdown of votes in our dataset. Interestingly, we find that there are 357% more votes on comments than there are on posts, perhaps due to the relative abundance of comments.

It is important to note that we only intend for these aggregate statistics to describe the size and scope of our data collection. It is likely that the interactions of a handful of power users are able to skew these results. A better way to analyze consumer and curation behavior on Reddit is to investigate interactions from a user perspective. Thus, throughout the remainder of this section we investigate social media consumption and curation interactions by plotting the distributions of users within each task.

3.2.1 Headline Browsers

Previous studies frequently show that the vast majority of Web users do not contribute content [54, 120]. However, these studies all tend to assume the non-contributors still
Figure 3.3. Headline browsing tendencies. At left is plotted a histogram of users as a function of their tendency to interact with each pageload. The typical user interacts with a loaded page about half the time. At right is plotted an un-grouped scatterplot of the same data. The scatterplot indicates that the number of pageloads is weakly correlated with the user’s tendency to interact the content ($R^2 = 0.09, p < 0.01$).

generally browse the content of a post. The problem with this assumption is that social media aggregators do not show the content initially; instead, they present the user with a ranked list of short headlines, which are usually carefully crafted to encourage the reader to click to see more. For our first task, we ask: How often do users browse headlines without clicking through to see the content?

For this task, we want to find the number of users who loaded a page on Reddit, but did not interact with the content on the page at all, i.e., the user only browsed the headlines. We define a content-interaction as any interaction that opens the post’s content either via content-expander, click-to-comments, click-to-external-URL, etc. Although we cannot determine what a user read but did not click, it is reasonable to assume that nearly all pageloads were purposeful, and that the user read at least part of the loaded Web page.

Figure 3.3 shows the interaction rates as a histogram (left) and as a scatterplot (right). We find that 25% of participants had content-interactions with less than 10% of their pageloads. The data shows that most study participants were headline browsers. Specifically, 84% of participants interacted with content in less than 50% of their pageloads, and the vast majority (94%) of participants in less than 60% of their pageloads. Because of the
cognitive effort required to interact with each post, we expected to see a negative correla-
tion between the number of pageloads and interaction percentage. However, the scatterplot
in Figure 3.3 (right) shows that a user’s propensity to interact with a page is only slightly
correlated with the number of pageloads ($R^2 = 0.09, p < 0.01$).

3.2.2 Vote Behavior

Unlike browsing interactions, i.e., viewing the post content or comments section, de-
scribed above, voting interactions are critical to the functioning of social media aggregators
like Reddit because they provide explicit feedback to the ranking system, which uses the
ratings to reorder the content. Here we look specifically at these vote-interactions to better
understand voting behavior.

Like previous studies which have found that most users do not vote on the majority
of posts [42], we show in Figure 3.4 (left) that many of our participants do not vote of-
ten or at all. About 50% of users vote in less than 1% of their interactions. While the
middle scatterplot in Figure 3.4 shows that a user’s number of vote-interactions is posi-
tively correlated with the user’s total number of interactions ($R^2 = 0.64, p < 0.01$) when
fitting a power regression, the scatterplot on right shows that a user’s tendency to vote
(the proportion of their interactions that were votes, denoted % Vote-Interactions) is only
very weakly positively correlated with the user’s total number of interactions ($R^2 = 0.04,
p < 0.01$). As a user’s total number of interact increases, they tend to have a larger number
of vote-interactions but not an increased tendency to vote.

When users do vote, we ask: How often do users actually read the content of the post
before they vote on it? To answer this question, we examined all of the posts that were
voted on and for each user that cast a vote, we counted the number of times that the user
viewed the content or comments section of the post before voting. The distribution of these
counts, as a proportion of the user’s total number of vote-interactions (denoted % of User’s
Vote-Interactions) for all participants with vote-interactions is plotted in Figure 3.5. The
Figure 3.4. Voting tendencies. At left is plotted a histogram of the percentage of users as a function of their tendency to vote. In center, the scatterplot indicates a correlation between the number of a user’s interactions that were vote-interactions and the total number of a user’s interactions \((R^2 = 0.64, p < 0.01)\). At right, the scatterplot indicates a weak correlation between the proportion of a user’s interactions that were vote-interactions and the total number of a user’s interactions \((R^2 = 0.04, p < 0.01)\).

The bottom bar plot shows that almost one-third (31%) of all participants that voted only rarely \((i.e., < 20\% \text{ of the time})\) viewed a post’s content and/or comments before voting and thus make their decision based on the headline alone. Conversely, about 17% of all participants regularly \((i.e., \geq 80\% \text{ of the time})\) viewed a post’s content or comments before voting. The remaining users are spread relatively equally between the two extremes of rarely and regularly browsing before voting.

Similarly, the top plot in Figure 3.5 shows the distribution of users that view the comments section of a post before they vote, and the middle plot shows the distribution of users that view the content of a post before they vote. These plots are not mutually exclusive, \textit{e.g.}, a user may view both the content and comments section before voting. We find that viewing the comments rarely precedes a vote, but the likelihood that a user views the content of a post before voting has a more even distribution.

Browsing the content of a post is certainly a more demanding task than not browsing. So we expected to see that frequent voters are more likely to vote without viewing the post’s
Figure 3.5. User behavior before voting on posts. Bar plots illustrate the likelihood that users vote on a post after viewing its comments section (top plot), content (middle plot), or either (bottom plot). The bottom plot shows that nearly one-third of voters rarely (i.e., < 20% of the time) view the content or comments of a post before casting a vote.

content. However, contrary to our expectations, we found that there was no correlation ($R^2 = 0.06$) between a user’s total number of votes and a user’s tendency to vote without browsing (not illustrated). In total, out of 41,540 posts that were voted on, 30,421 (i.e., 73%) were rated without first viewing the content or comments of the post.

3.2.3 Position Bias

Another well studied phenomenon found in social news aggregators is position bias, wherein people tend to pay more attention to items at the top of a list than those at the bottom [18, 57]. As a consequence of position bias, the presentation order has a strong effect on the choices that people make [71, 137]. This is especially true for Web search results [26], and has even been found to affect the answers that people select when answering multiple choice questions [15].

Here we describe position bias on Reddit and also show how position affects voting
behavior. Before we begin, it is important to note that each Reddit post must be submitted to a subreddit. A subreddit is a topic-specific community (e.g., /r/news, /r/ama, /r/machinelearning) with its own set of relevancy rules, guidelines and mores. If a registered user subscribes to some subreddit, then the posts submitted to that subreddit will appear on the user’s frontpage in an order computed by one of Reddit’s ranking systems (e.g., best, top, rising, new). A user can also view the posts submitted to a specific subreddit by navigating, searching, or typing the subreddit’s name. Alternatively, a user can view the posts from all subreddits, regardless of subscription status, by visiting the special /r/all subreddit. The choice of view, either the frontpage/all or a specific subreddit, drastically changes the scope of the posts that the user sees, and may also change how the user interacts with the content.

Figure 3.6 shows the effect that rank has on the likelihood that study participants browsed (i.e., viewed the content or comments section), cast an upvote, or cast a downvote on a post. The interaction percentages are divided into interactions that occurred from a mixed-subreddit view (i.e., the frontpage or /r/all) or from a specific subreddit. As we observed earlier, the number of browse interactions, upvotes, and downvotes differ by an
order of magnitude. In Figure 3.6 we plot the percentage of each interaction at each rank as a percentage of the total interactions within each view. For example, the top-left point shows that 7.9% of all clicks in a subreddit were to browse the content or comments of the top-ranked post.

Position bias is clearly evident in our results: a user is about 4 times more likely to interact with the top post than the 10th post. Line plots of browsing behavior show an uptick in interaction likelihood just before rank 25. This increase has been previously attributed to ‘contrarians’, who navigate the list backwards [71, 104]. A large decrease in browsing likelihood is found between rank 25 and 26, which corresponds to Reddit’s default pagination break.

Across all interaction types, posts are more likely to be browsed, upvoted, and downvoted when viewed from within their specific subreddit, than when viewed from the front page. After the pagebreak, the gap between views narrows and even flips in the case of browsing interactions. This indicates that users are more willing to explore further down the ranked list when presented with a more diverse set of posts.

3.3 Performance Deterioration in Browsing Sessions

In this section, we present several experiments that highlight the effects of user behavior over the course of a browsing session. For these tasks, we first need to identify the browsing sessions for each user. In the collected data, a timestamp is recorded for each interaction and a browsing session is loosely defined as a period of regular activity. In a recent study, Halfaker et al presented a methodology for identifying clusters of user activity by plotting the histogram of the time elapsed between interactions. They argue that the appearance of regularity within delays implies a good sessionizing heuristic. The results of Halfaker et al as well as sessionizing results for delays between comments by Singer et al found that a delay threshold of about 1 hour is an appropriate threshold [51,110]. In other words, these studies suggest that if a user is inactive for more than 1 hour, then that user’s session has
ended and the next activity will begin a new browsing session.

Using the same methodology, the inter-activity delays of our data is plotted in Figure 3.7 (left). We can clearly see an uptick in the number of delays that occur in the time-bin prior to one hour. These results are in alignment with the previous studies, so we confidently define a browsing session for a particular user as a contiguous block of activity with delays of at most one hour. This created 41,385 unique browsing sessions across all 309 users for a mean-average of 133 sessions per user over the year.

Not all users are equally active, and the activity of each user in a session varies significantly. For instance, although the mean-average session length is about 53 minutes (median: 21 minutes), Figure 3.7 (on right) shows that many browsing sessions last for only one-click, while the longest session lasted for several hours. In our analysis, we partitioned the sessions into equal-sized bins of short, medium, and long, resulting in session length groupings of $<3 \text{ min}$, $3\text{min}-53\text{min}$, and $>53 \text{ min}$ respectively. We then classified users as being short-, medium-, or long-browsing users based on the session length of the majority of the user’s sessions. Figure 3.8 summarizes the distribution of users and sessions across types. We find that most users in our dataset had primarily short-browsing sessions, i.e. , the plurality of their sessions lasted less than three minutes, and nearly equal
numbers of medium-browsers and long-browsers.

3.3.1 Community Variety

Despite the potential for social news and social networking sites to expose users to diverse content and new perspectives, many recent studies show the opposite effect. Through their friendships, filters, subscriptions, etc., users frequently filter out those ideas and opinions that clash with their own [25, 88]. This has led to the rise of the “filter bubble” and the “echo chamber” wherein users only view what they agree with [95], and has been associated with the adoption of more extreme views over time [115] and even misconception of factual information [29]. We expect to find a similar echo chamber on Reddit because a user’s frontpage is populated from only those subreddits, which are topic-based communities, to which the user subscribes. Here, we take a first look at the variety of the communities browsed on Reddit within browsing sessions.

Unfortunately, it is difficult to computationally discern the ideological leanings for each of the 5,235 distinct subreddits in our data set to perform content-based analysis of variety. Instead, we look at the number of distinct subreddits that a user loads as a measure of the variety of communities that a user views content from. While the content within a single community may cover a narrow or diverse range of topics, content is contributed and curated by the user-base of that community. Since we have seen above, in Figure 3.3,
Figure 3.9. Community Variety. Distinct subreddits loaded (left) and interacted with (right) during a session by session length. Despite being highly skewed, the mean-average number of subreddits loaded are 1.4, 2.5, and 4.8 for short-, medium-, and long- browsing users, respectively, and even lower 1.0, 1.3, and 1.5, respectively, for subreddits users actively interacted with.

that typical users have a low tendency to interact with each pageload, we use the number of distinct subreddits interacted as another estimate at the community variety of a browsing session, in this case as a proxy of the variety of communities the user more actively engages with. Figure 3.9 shows, as expected, that each session-type and user-type exhibits different browsing behavior. It is reasonable to expect that the number of subreddits found within a short-browsing session will be smaller than the number of subreddits found within long browsing sessions simply because longer sessions have a greater opportunity to view more content or communities.

Despite the distributions illustrated in Figure 3.9 being highly skewed left, we find a surprising lack of variety when we look at the mean-averages. We find that users navigate to a mean-average of 1.4, 2.5, and 4.8 distinct subreddits for short-, medium-, and long-browsing user-types respectively. Despite the number of subreddit pageloads, we further find that users only interact with posts from 1.0, 1.3, and 1.5 subreddits from short-, medium-, and long-browsing users on average. In other words, most users only click or vote on a very specific set of posts, even in long-sessions. The discrepancy between subreddit-pageloads and interactions can be explained as headline browsing behavior; that
is, even if a user views a variety of subreddits, it is unlikely that the user will actually click past the headline. If plotted as a function of session-lengths rather than user-type, Figure 3.9 illustrates the same lack of community variety with nearly identical the mean-averages.

The dearth of community variety also occurs across browsing sessions. Despite collecting 1,018 subscribe-events and 437 unsubscribe-events over the year-long data collection period, the median user subscribed to 0 new subreddits. In fact, all of the subscription events can be attributed to only 109 unique users (104 subscribed, 44 unsubscribed), and from those 104 subscribers only 26 subscribed to 10 or more new subreddits. In summary, although many of our study participants spent a lot of time interacting with Reddit, we find that most users, even highly active users, perform most of their activity within a handful of subreddits.

Because of the demonstrated lack of browsing variety, we also suspect a lack of voting variety. Figure 3.10 (on left), which illustrates the distribution of voters to subreddits, shows that the majority of subreddits in our collection have only one voting user and that the distribution resembles a power law distribution. Figure 3.10 (on right) illustrates the distribution of votes to subreddits. The distribution of subreddits to votes (on right) has an
even longer tail than the distribution of subreddits to voters (on left) because a single user may interact with many posts and/or comments within a single subreddit thereby severely skewing the results. In both plots the finding is the same: most subreddits receive very few votes from very few voters; only a handful of subreddits receive the majority of the votes.

3.3.2 Performance Deterioration

It is widely recognized that human performance tends to deteriorate during periods of sustained mental or physical effort. Physical fatigue is most often associated with prolonged exercise or exertion where the physical performance of an individual diminishes over time. Similarly, researchers have found evidence of cognitive fatigue where an individual’s ability to complete cognitive tasks diminishes over time. Understanding the processes behind performance deterioration is critical in many enterprises where individuals must maintain high performance over long periods. Many complicated and delicate surgeries, for example, may take hours to complete, yet still require high cognitive performance [55]. Many other high-stakes tasks, such as flying and driving, are equally prone to performance deterioration [16]. Cognitive fatigue has also been studied in test taking, where individuals must have consistent performance on cognitively demanding tasks that take long periods [1]. As society increasingly relies on the ratings of users to curate online news and information, it is important that we understand whether and how cognitive fatigue affects online user behavior. In this section, we explore this topic by looking at user behavior over the course of a browsing session.

To understand how user behavior changes over the course of a session, we must first develop comparative baselines. To that end, we generated two randomized session datasets from the observed data that we label as delay-randomized and event-randomized sessions [110]. In the delay-randomized dataset, we randomly shuffle the ordering of the delays between interactions while holding the ordering of interaction events constant. For the event-randomized dataset, we randomly shuffle the ordering of the interaction events.
while holding the delay between interactions constant. We illustrate how sessions differ between the original, delay-randomized, and event-randomized datasets in Figure 3.11.

The interactions of each browsing session are placed into decile bins, where the first 10% of interactions [0,10] are placed in the same bin, the second 10% (10,20] are placed in the same bin, and so on. Decile binning provides an opportunity to analyze aggregate statistics about the behavior of participants as they progress through their session. As in previous analysis, we separate participants by their typical session lengths into short-browsing, medium-browsing, and long-browsing users using the same criteria as in the previous section. Using this methodology, we can begin to ask questions about the activities performed by users over the course of their browsing session.

We first look at the ranks of posts with which a user interacts. Reddit orders content in various ways. Generally, and by default, posts are ranked by the number of upvotes minus the number of downvotes normalized by the time since submission. Based on the interaction percentages by post rank illustrated in Figure 3.6, we expect that readers begin

---

1Figure adapted from [110] with permission.
Figure 3.12. Median post rank as a function of session decile for short-, medium-, and long-browsing users. Linear regression lines are plotted and the correlation of determination $R^2$ value is shown to the right of each regression line. † and ‡ symbols represent statistically significant effects with $p < 0.05$ and $p < 0.01$ respectively. The user sessions plot (at left) shows that the median ranks of browsed posts tends to increase for all types of users. As expected, baseline plots illustrating session shuffled by randomized delays (center) and randomized events (right) show little, if any, correlation.

Based on these initial findings, we expect that the median rank of posts that are viewed or voted on will rise as the user session progresses. To test this hypothesis, Figure 3.12 (a) shows that the median post rank is statistically correlated with the session decile for all user-types. As expected, short-browsing users have a higher correlation than medium- and long-browsing users because users who spend more time browsing are more likely to view more subreddits, each of which start their rank numbering at 1. These results are in contrast to the randomized baselines drawn in Figure 3.12 (b) and Figure 3.12 (c), which mostly show non-significant correlations.

We performed the same regression analysis on several other performance factors. Rather than visually illustrating each graph, we show a summary of the regression results in Fig-
Figure 3.13. Slope \((m)\) and effect sizes \((R^2)\) of various factors as a function of session progress. \(^\dagger\) and \(^\ddagger\) symbols represent statistically significant correlations with \(p < 0.05\) and \(p < 0.01\) respectively.

The coefficient of determination \((R^2)\) values are tabulated along with signs of their slopes for each user browsing type. The top row repeats the \(R^2\) values and tests of statistical significance results from Figure 3.12 (on left). The rank of upvote and downvote interactions is not highly correlated with session progress, meaning that although users are likely to browse lower-ranked items as their session progresses, they are not more likely to vote on them.

Unlike a post’s rank, which is typically a value between 1 (the top ranked post) and 25 (the bottom of the first page), the score of a post is highly skewed and varies significantly depending on the subreddit to which it was submitted. We ignore these confounding factors, for now, and tabulate their \(R^2\) values in the middle row of Figure 3.13. A post’s score at the time it was viewed is significantly correlated with session progress in short-browsing users, but not for medium- and long-browsing users. Generally, these results have the same interpretation as the post rank results: Short-browsers typically view posts with high scores.
at the beginning of their session and posts with progressively lower scores as their session continues. We believe that non-correlation in the case of medium- or long-browsing users is because medium and long-browsing users are more likely to visit multiple subreddits (as indicated in Figure 3.9), which rank their posts separately.

Two weeks after the end of the data collection period, we crawled each of the recorded posts to determine their “final” score. Like the score at time of view, Figure 3.13 shows that the final score of each post that a user viewed is negatively correlated with session progress in short- and long-browsing users.

Although informative as to the nature of social media browsing, these results say little about whether users have any cognitive fatigue. In the next experiment we ask: Is rating performance correlated with session progress?

To answer this question we first need to define rating performance in terms of social media sessions. The change in a post’s score from the time of view to the final score $\Delta \text{Score}_{I,F}$ is a good starting point because this metric gives an indication of how much the post changed after it was interacted with (i.e., upvoted or downvoted). Positive values for this metric will indicate that a post’s score increased after the user’s vote; the higher the increase, the more predictive power that the user displayed. For example, a user who interacts with a popular post after it was identified as popular by others (i.e., received a large proportion of its total votes) would have a lower $\Delta \text{Score}_{I,F}$ than a user that interacted with the same post earlier when there was a weaker signal of popularity, thus indicating more predictive power.

Unfortunately, this metric alone is highly dependent on the size of the subreddit to which the post was submitted. For example, in smaller subreddits a score of 10 or 20 is considered to be a high score whereas in large subreddits a score of 10 or 20 is actually a very low score. To adjust for subreddit size effects, we normalize $\Delta \text{Score}_{I,F}$ by the absolute value of the median final score of all collected posts within a subreddit. To ensure a representative sample size, we only consider subreddits that have 50 or more posts col-
lected in our dataset. We abbreviate this function as $\text{avg}_{SR}(\text{Score})$, and the final upvote performance metric is:

$$\frac{\Delta \text{Score}_{I,F} + 0.5}{|\text{avg}_{SR}(\text{Score})| + 0.5},$$

where $+0.5$ is added as Laplace smoothing to avoid division by zero. We call this metric the subreddit normalized change in score (SRN$\Delta$), which can be in response to an upvote (SRN$\Delta^\wedge$) or a downvote (SRN$\Delta^\vee$).

SRN$\Delta$ has a couple of important properties. First, by normalizing by the average of all posts’ final scores within each subreddit, this metric partially represents the power of a single vote. Second, SRN$\Delta$ indicates the change as a proportion of the average score of posts within the subreddit, where positive values indicate an increase in the score of the post between the time of a user’s interaction and the end of voting.

Figure 3.14 shows the results of the upvote performance analysis. We do not find any correlation among the variables in the randomized session plots. Within the original user sessions, we find statistically significant negative correlations between rating performance and session progress in the long-browsing user sessions, that is, scores of posts that long-browsing users vote on early in their session tend to increase more than posts that they vote on later in their session.

Because these experiments are not randomized, we cannot say for certain whether the diminishing rating performance indicates a decay in the user’s influence, predictability, or some other characteristic. Earlier work in this area has found significant influence effects [44, 91, 129] on a user’s vote due to ranking bias [57], but we make no such claim.

There may exist other confounding factors that explain the deterioration in rating performance. To address this potential, we performed analysis of covariance (ANCOVA) tests for each user-type (i.e., short-, medium-, long-browsers). The ANCOVA test is like the standard analysis of variance (ANOVA) test except that ANCOVA compares the response variable (e.g., SRN$\Delta$), with an independent variable (session progress) and a factor (session ordering) [130]. By comparing the regressions of the original session data against the
Figure 3.14. Median subreddit-normalized change in score for upvoted posts (SRN∆∧) as a function of the fraction of session elapsed for short-, medium-, and long-browsing users. Linear regression lines are plotted, and the correlation of determination $R^2$ value is shown to the right of each regression line. † and ‡ symbols represent statistically significant effects with $p < 0.05$ and $p < 0.01$ respectively. The user sessions plot (on left) shows how the median SRN∆∧ changes as a session progresses: the median SRN∆ of browsed posts tends to diminish for all types of users, but only long-browsing users show statistically significant correlation.

respective regressions of the delay-randomized and event-randomized sessions, the ANCOVA test can determine if the differences in regressions are statistically different from the random baselines. Results for each experiment are tabulated in Figure 3.15. Short- and medium-browsing users are not significantly different from the random baselines, but the long-browsing users have a significant negative decay.

We experimented with several alternatives to the normalization used in SRNΔ including: (1) using the mean score of a subreddit, rather than the median; and (2) normalizing by the average change in score for each subreddit $\text{avg}(\Delta\text{Score}_{I,F})$, rather than the average final score. Various combinations of these alternatives created small changes in the regression and ANCOVA analysis, but did not alter the results of significance tests.
<table>
<thead>
<tr>
<th>User-Type</th>
<th>Short</th>
<th></th>
<th>Medium</th>
<th></th>
<th>Long</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R^2$</td>
<td>D</td>
<td>E</td>
<td>$R^2$</td>
<td>D</td>
<td>E</td>
</tr>
<tr>
<td><strong>Post Rank</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>View</td>
<td>.42</td>
<td>◦ ‡</td>
<td>.25</td>
<td>○</td>
<td>.37</td>
<td>‡ ‡</td>
</tr>
<tr>
<td>Upvote</td>
<td>.15</td>
<td>○</td>
<td>.20</td>
<td>†</td>
<td>.37</td>
<td>‡ ‡</td>
</tr>
<tr>
<td>Downvote</td>
<td>.12</td>
<td>○</td>
<td>.22</td>
<td>†</td>
<td>.20</td>
<td>○</td>
</tr>
<tr>
<td><strong>Post Score</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time of View</td>
<td>.15</td>
<td>○ ○</td>
<td>.05</td>
<td>○ ○</td>
<td>.11</td>
<td>○ ○</td>
</tr>
<tr>
<td>Final Score</td>
<td>.13</td>
<td>○ ○</td>
<td>.13</td>
<td>○ ○</td>
<td>.17</td>
<td>○ ○</td>
</tr>
<tr>
<td><strong>SRNΔ</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Upvote</td>
<td>.18</td>
<td>○ ○</td>
<td>.05</td>
<td>○ ○</td>
<td>.35</td>
<td>○ ‡</td>
</tr>
<tr>
<td>Downvote</td>
<td>.10</td>
<td>○ ○</td>
<td>.17</td>
<td>○ ○</td>
<td>.15</td>
<td>○ ○</td>
</tr>
</tbody>
</table>

Figure 3.15. Effect sizes ($R^2$) for analysis of covariance (ANCOVA) tests that compare regressions on user sessions against delay-randomized and event-randomized sessions. Results of significance tests between the observed behavior and the delay-randomized (D) or event-randomized (E) sessions are represented with ◦, †, and ‡ symbols for $p \geq 0.05$, $p < 0.05$, and $p < 0.01$ respectively.

3.4 Discussion and Conclusions

This chapter presented a user study of browsing and voting behavior on the social news aggregator Reddit. We find that the vast majority of participants are headline browsers, who only view the summary headlines without any further interaction to view the content or read the comments. This shallow browsing behavior is also evident in voting interactions where 73% of posts were voted on without first viewing the post’s content and nearly a third of voters almost always (i.e., $\geq 80\%$ of the time) vote without browsing a post’s content or its comments. Voting dictates the visibility of individual posts and directly determines the content that is presented to others. Our results indicate that voting is probably heavily dependent on the quality of a post’s title or its headline rather than the content itself [67]. When combined with the clear evidence of position bias in our dataset, this dependence raises concerns about the manner by which platforms rank content through voting using an assumption that votes correspond with the quality of the content itself.
Interestingly we did not find evidence to link browsing behavior before voting or the effect of position bias with cognitive fatigue, with no statistically significant correlation between the number of votes a user casts and the effort expended before the vote, *i.e.* whether a user browsed the content or comments of a post before voting. In fact, when we investigated performance deterioration within sessions, we found more evidence of deterioration for long-browsing users than short- or medium-browsing users. While there were some statistically significant increases in the rank of posts interacted with by short- or medium-browsing users when compared to our events-randomized baseline, there was no statistically significant decline in rating performance. Long-browsing users, however, had a statistically significant decline in the predictive power of their upvotes, which we approximated with our SRN$\Delta^\wedge$ metric, as their session progressed. This decline could be indicative of a loss of predictive power or accuracy because of cognitive fatigue in long-browsing users.

We also found a severe lack of browsing and voting variety. Participants seem more inclined to exert more effort when interacting with a more diverse range of content, indicated by a greater probability of browsing further down and onto the second page when viewing a frontpage over a specific subreddit. However, the scope within a user’s browsing session remains relatively narrow despite having access to a potentially wide range of diverse content and communities through the numerous subreddits available. Despite an intuitive increase in the number of subreddits as session length increases, users are typically engaging with a limited number of communities regardless of session length. There is a similar lack of variety in voting by users; most subreddits received very few votes and most voters voted in very few subreddits. This shows that while a diverse range of communities are available, users tend to self-select a narrow scope to interact with.
CHAPTER 4

SOCIAL NEWS CONSUMPTION FROM SOURCES OF VARYING CREDIBILITY

As people rely on social media as their primary sources of news, the spread of misinformation has become a significant concern. In this chapter, we present a large-scale study of news in social media where we analyze eleven million posts and investigate propagation behavior of users that directly interact with news accounts identified as spreading trusted versus malicious content. Unlike previous work, which looks at specific rumors, topics, or events, we consider all content propagated by various news sources. Moreover, we analyze and contrast population versus sub-population behavior (by demographics) when spreading misinformation, and distinguish between two types of propagation, *i.e.* direct retweets and mentions. Our evaluation examines how evenly, how many, how quickly, and which users propagate content from various types of news sources on Twitter.

4.1 Introduction

People use social media not only for entertainment and social networking but also as their primary source of news and information. An August 2017 survey from the Pew Research Center found that 67% of Americans report that they get at least some of their news from social media, an increase of 5% over the previous year [109]. Of those who use Twitter, 74% said they received news from the platform. With such a reliance on social media as a source of news and information, the spread of misinformation is a significant concern.

Previous work in social media analysis, especially within the area of social news, has focused on influence campaigns and the spread of (mis)information either organically or
through bots. Given a particular event like an election or a natural disaster, researchers typically follow information cascades to tease out diffusion processes and infer various characteristics about how social media responded to the event [35, 117]. These studies have resulted in important findings about the effect of such items as information contagion [70], influence campaigns [44], bots [37], and spam [50], etc., within specific newsworthy events.

In this chapter we take a different view. Rather than studying information propagation one newsworthy event at a time, we seek to quantify and compare socio-digital phenomena according to the source of the information. This is especially prescient in the current climate where the reliability of traditional sources of news and information are contested. For this we rely on previous work by Volkova et al. [125] that aimed to classify information sources according to their quality (i.e. their accuracy according to fact-checking organizations) and their intent (i.e. whether the author intends to deceive the reader or not). Along these two axes we focus on news sources that fall into one of the following five categories: trusted, clickbait, conspiracy theories, propaganda, and disinformation.

The goal of this chapter is to quantify and compare the immediate propagation (i.e. the direct consumption) of information from the different types of news sources. In service of this goal we identified 282 news sources on Twitter and collected 11 million direct interactions (i.e. retweets and mentions) with those source accounts from two million unique users. Unlike previous work, we report our findings on information propagation behaviour from sources of varying levels of credibility at the population level as well as for various user-demographics. With this data and the news source type classification, which is described in more detail in the next section, we can quantify how social media users interact with different types of news sources. This is the focus of the four research questions outlined below.
RQ1: How evenly do users share content from news sources of varying credibility?

Several previous studies have investigated the makeup of users that retweet content from specific news sources as a way to identify sources that spread rumors or disinformation. In the context of social media, the 1% rule and its variants indicate that most users only browse content while a mere 1% of users contribute new content [54, 120]. Within the subset of those who actively contribute new content, Kumar and Geethakumari [64] found larger disparity among users who retweeted news from sources identified as spreading disinformation. That is, a small number of highly active users were responsible for the vast majority of retweets of disinformation. However, fitting the template of most social network research, the study focused only on keywords related to the events in Egypt and Syria in 2013. To answer this research question more generally, the present work quantifies and compares the disparity in sharing behavior of millions of users across the various categories of news sources. Specifically, for each type of news source — clickbait, propaganda, etc., we ask whether information sharing is equally distributed across users, or instead if there are a small group of vocal users responsible for the majority of the information propagation.

RQ2: How many users share content from different types of news sources?

To identify rumor-spreading users, Rath et al. [102] proposed an RNN model with believability scores to weight edges in a user-retweet network. Believability scores for pairs of users were calculated from the users’ scores of trustingness and trustworthiness. They used the propensity of other users to retweet a source as a proxy for the trustworthiness of the source. Users were considered to be more trustworthy if more users retweeted them, and were considered more trusting if they retweeted content from a larger number of other users. Using a similar proxy for trustworthiness, we consider whether deceptive sources can be identified by how trustworthy they are, i.e. how many users retweeted their posts.
RQ3: How quickly do users share from news sources of each type?

Information diffusion studies have often used epidemiological models to understand the diffusion of information, both suspicious and trusted, among social media users [60, 118, 127, 136]. These models were originally formulated to model the spread of disease within a population. In the social media context, users are considered to be “infected” when they propagate information to other users. Jin et al. [60] modeled diffusion using the SEIZ (Susceptible, Exposed, Infected, and Skeptics) model and compared ratios of the transition rates into and out-of their “exposed” category, i.e. whether people are exposed to misinformation faster than they spread it. Authors found that users tend to share information about factual events more quickly than misinformation or rumors (both verified false or ambiguous in veracity). A recent study by Vosoughi et al. [126] found that news fact-checked and found to be false spread faster and to more people than news items found to be true. Because our methodology considers all content directly shared from various sources (rather than content about specific events), we are able to determine whether deceptive or trusted sources have slower immediate share-times overall.

RQ4: Who is sharing from different news sources?

Existing work focuses on user responses to rumor diffusion as belief exchange that is caused by influence from friends, e.g. the Tipping Model [107]. Wu et al. [135] combined content analysis of rumor-tweets to detect false rumors on Weibo as early as one day after the initial broadcast with 90% confidence. Among their most important features was the type of user performing the sharing. Ferrara [34] used a similar classification and found those with high followings generated highly-infectious cascades. Studies have also identified that someone who believes in one conspiracy theory is also likely to believe in others [48, 75]. Goertzel [48] found that “young people were slightly more likely to believe in conspiracy theories” but belief was not significantly correlated with gender or the level of education of the participants. Through our analysis of a sample of users with inferred demographics, we can identify whether there are different patterns in how
users interact with conspiracy sources. In this chapter, we focus on a broad question about user sharing behavior to discover informative patterns within sub-populations not only in the propagation of rumor versus non-rumor or a single category of deceptive content but within interactions spreading information from varying types of news-providers.

In order to tackle mis- and disinformation spread in social networks, it is important to address motivations about why people share deceptive news. Motivating factors can be psychological (clickbait), political (propaganda, conspiracy, disinformation), financial (clickbait, disinformation), and social (conspiracy, propaganda) among others. By analyzing how information from different types of deceptive news sources is propagated across a social network, this study quantifies how people share from news sources who spread misleading, manipulated, or fabricated information; who these disinformation propagators are; and how much deceptive information is being shared.

4.2 Data Collection and Annotation

In this section we describe how the news sources, the population data, and the demographics data used in this study were collected and annotated with source type labels.

4.2.1 News Source Labels

As previously discussed, we focus on news sources that fall into one of five classes along a spectrum of varying credibility levels. We define trusted news sources and each of the sub-categories of deceptive news sources as follows:

**Trusted** news sources provide factual information with no intent to deceive the audience, e.g. “Umberto Eco, Italian semiotician and best-selling author, dies at 84 [URL] [URL].”

**Clickbait** news sources use attention-grabbing, misleading, or vague headlines such as “That’s about as tone deaf as it gets right there. [URL]” to attract an audience.
Conspiracy theory news sources provide uncorroborated or unreliable information to explain events or circumstances, e.g., “Video: Hoboken train wreck planned? [URL].”

Propaganda news sources provide intentionally misleading information to advance a social or political agenda. For example, “The evidence clearly shows that building new nuclear power plants will make global warming worse. [URL].”

Disinformation news sources share fabricated and factually incorrect information meant to deceive an audience. For example, “The great cholesterol and statins con finally unravels: [URL] #statins #cholesterol #heartdisease [URL].”

Lists of news sources and their labels were previously aggregated by Volkova et al. [125] through a combination of crowd-sourcing and public resources. Authors manually constructed a list of trusted news sources that were confirmed to have Twitter-verified accounts that posted content in English. Deceptive news sources, i.e., clickbait, conspiracy, and propaganda, were collected from several public resources that annotate suspicious news accounts and their associated websites. Labels for each of the sub-categories of deceptive news sources were also manually verified to ensure quality.

Disinformation labels were collected from a unique source of public data that comprises confirmed cases of disinformation campaigns: https://euvsdisinfo.eu/, which is also available on Twitter through the @EUvsDisinfo account. Weekly reports contain disinformation summaries with the countries and languages targeted, as well as the URLs of sources of disinformation, people and organizations who reported disinformation, and manually generated disproofs (when applicable). We limit our analysis to disinformation news accounts collected by the European Union’s East Strategic Communications Task Force between 2015 and 2016. As of November 2016, EUvsDisinfo reports included almost 1,992 confirmed disinformation campaigns found in news reports from around Europe and beyond.

1Deceptive news lists used by Volkova et al. [125] include: http://www.fakenewswatch.com/ http://www.propornot.com/p/the-list.html
### TABLE 4.1

**SUMMARY OF OUR TWITTER DATASET**

<table>
<thead>
<tr>
<th></th>
<th># Sources</th>
<th># Tweets</th>
<th># Retweeting users</th>
<th># @-mentioned users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trusted (T)</td>
<td>182</td>
<td>6,567,002</td>
<td>1,423,227</td>
<td>390,164</td>
</tr>
<tr>
<td>Clickbait (CB)</td>
<td>11</td>
<td>40,347</td>
<td>19,361</td>
<td>6,002</td>
</tr>
<tr>
<td>Conspiracy (CS)</td>
<td>13</td>
<td>126,246</td>
<td>35,799</td>
<td>9,171</td>
</tr>
<tr>
<td>Propaganda (P)</td>
<td>26</td>
<td>609,251</td>
<td>233,799</td>
<td>34,532</td>
</tr>
<tr>
<td>Disinformation (D)</td>
<td>50</td>
<td>3,487,732</td>
<td>292,437</td>
<td>82,638</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>282</strong></td>
<td><strong>10,819,357</strong></td>
<td><strong>1,784,655</strong></td>
<td><strong>471,967</strong></td>
</tr>
</tbody>
</table>

#### 4.2.2 Population Data

Our dataset, summarized in Table 4.1, includes approximately 11 million tweets that either retweeted or @mentioned news sources of varying degrees of credibility. We collected all direct mentions of 282 sources over 13 months between January 2016 and January 2017. Then, we assigned the category-label from each news source mentioned or retweeted to each individual tweet. Figure 4.1 shows the relative size and frequency of the ten most frequently occurring sources for each news type in the dataset.

Distributions of tweet volume over time are illustrated in Figure 4.2. Trusted and disinformation-spreading sources are referenced and retweeted consistently across the entire year. The other three deceptive types have concentrated peaks in activity sharing or mentioning source accounts. Clickbait- and Conspiracy-spreading sources both have a single peak in June 2016 (26%) and July 2016 (21%), respectively. Tweets that reference or retweet propaganda-spreading sources were most heavily posted in October and November 2016 (49%), with another spike in July 2016 (17%).

Recent work has found that accounts spreading disinformation are significantly more likely to be automated accounts [108]. However, in this chapter we are interested in the
Figure 4.1. Tweet volume for the ten most frequently occurring news sources within each type as a function of their share of all tweets of the given type. For example, *anonews.co* is the 10th most frequently occurring clickbait source, but is responsible for only 0.03% of the clickbait tweets. In contrast, the most frequent clickbait source *theblaze* is responsible for 43%. The 10th (*ZeeNews*) and most frequent (*el_pais*) trusted sources have a smaller range in terms of shares of trusted tweets with *ZeeNews* responsible for 3.5% and *el_pais* for 11.7%.

impact that news sources have in the system as a whole, capturing the publicly visible responses to news sources by all accounts — whether activity from that account is manual, automated, or some mixture of both manual and automated behavior. Therefore, we did not remove bots or automated accounts from the population dataset. For the dataset we used in our demographics-based analysis, described in the next subsection, we focus specifically on personal accounts.

We did compare the news accounts and tweets captured in our dataset with the list of Twitter accounts connected to Russia’s “Internet Research Agency” recently released by the US House Intelligence Committee.² However, because we are focused primarily on

sources of news (rather than specific rumors or events), our dataset only contained 180 of the 2752 flagged-accounts, and only 4763 retweets; so we did not focus on these accounts in particular in this chapter.

4.2.3 Demographics Data

In addition to the population-level aggregate statistics, we study information propagation and influence across various user demographics. To accurately obtain user demographics we curated a subset of non-organization user accounts that generated enough information to render a demographic description with high confidence. To identify this
sample, we first restrict the dataset to users who 1) retweeted or @mentioned deceptive news sources in at least 5 posts during the data collection period, and 2) posted in English. This resulted in a subset of 106,849 users. From this list we collected the 200 most recent tweets from Twitter’s public API. We used these tweets and the Humanizr classifier to identify person-accounts, which they defined as ”a personal account is one controlled by an individual” [87]. This resulted in a sub-sample of 66,171 person accounts.

To infer user demographics, i.e. gender, age, income, and education, we employed a demographic classifier trained on a large, previously annotated Twitter dataset [123]. Specifically, our demographic classifier relied on a Convolutional Neural Network (CNN) architecture initialized with 200-dimensional GLOVE embedding vectors pre-trained on Twitter tokens [99]. Following previous methodology, each demographic-attribute was assigned one of two mutually exclusive classes [123]. For example, we classified gender as either male (M) or female (F), age as either younger than 25 (Y) or 25 and older (O), income as below (B) or at and above (A) $35,000 a year, and education as having only a high school education (H) or at least some college education (C). The area under the ROC curve (AUC) for 10-fold cross-validation experiments were 0.89 for gender, 0.72 for age, 0.72 for income, and 0.76 for education. These are the state-of-the-art results for user demographics prediction on Twitter, an improvement on performance of previous models that used the same dataset [122].

Users in this dataset were primarily predicted to be male (96%), older (95%), with higher incomes (81%), college-educated (82%), and classified as regular users (59%). It is important to note that our user sample is representative of those users who frequently interact with deceptive source accounts. It is not a balanced sample of global demography or a representative sample of Twitter itself. In fact, a survey by the Pew Research Center in 2016 found that only 17% of Twitter users had a high school education or less, 38% were between 18 and 29 years old, and 47% were male [49]. Each reported category was less skewed towards the majority class in our demographic attributes than we found in the
sample of users who frequently interacted with deceptive source accounts.

We also include each user’s role in their network based on the friend and follower counts collected from user metadata. For this analysis we borrow the leader/follower heuristic to assign a user as an opinion leader (L) if they are followed by more users than they follow or a regular user (R) if they follow more users than they have followers [135].

4.3 Methodology

Here, we describe the methodology used to analyze propagation behavior of news content and misinformation across and within the five types of sources identified. Again, we focus on propagation at the source level rather than the content or individual tweet level. We consider propagation of all content, deceptive or not, from sources of each type.

RQ1: How evenly do users share content from news sources of varying credibility?

To answer this question we compare the distributions of how users share information using three measures commonly used to measure income inequality: Lorenz curves, Gini coefficients, and Palma ratios. Rather than measure how much of the total population’s income each individual is responsible for, we repurpose these metrics to measure how much of the total tweet volume each user is responsible for. This allows us to compare propagation inequality across source types the way economists compare income inequality across countries.

Lorenz curves have traditionally been used to illustrate the distribution of income or wealth graphically [61]. In those domains, the curves plot the cumulative percentage of wealth, income, or some other variable to be compared against the cumulative (in increasing shares) percentage of a corresponding population. The degree to which the curve deviates from the straight diagonal ($y = x$) representative of perfect equality represents the inequality present in the distribution. In our case, the Lorenz curve is adapted to illustrate the cumulative percentage of propagation (tweets shared) as a function of the cumulative
Figure 4.3. Lorenz Curves and Gini Coefficients. As a graphical representation of income inequality within a population, Lorenz curves plot the share of income by the cumulative share of the population. Lorenz curves that measure the inequality in propagation plot the share of total propagation, i.e. the y% of tweets posted, by the share of the population who propagated, i.e. the cumulative x% of active users. The Gini coefficient is the proportion of the area under the line of perfect equality ($a_1 + a_2$) that is captured between the line of perfect equality and the Lorenz curve ($a_1$).

The Gini coefficient is defined as the proportion of the area under the line of perfect equality that is captured above the Lorenz curve, i.e. $\frac{a_1}{a_1 + a_2}$ in Figure 4.3. The Gini coefficients reported in subsequent sections are calculated using the formula in Eq. (4.1), which is an approximation of the points of the Lorenz curves observed in the collected data. Gini coefficients can grow to be greater than 1 but only if individuals within a population can be responsible for negative proportions, e.g. if individuals can have negative incomes if using our income example. In our data, users must be responsible for at least 1 share in order to be considered part of the dataset so the Gini coefficients have an upper-bound of 1.

$$\hat{G} = 1 - \sum_{k=1}^{n} (X_k - X_{k-1})(Y_k + Y_{k-1})$$  

(4.1)
The third measure we consider is the Palma ratio. It is defined as the ratio of the share of the top 10% to the bottom 40% of users in the population. Again, using income as our example, in perfect equality each individual in the population is responsible for an equal amount, e.g. an equal share of income, resulting in a Palma ratio of $1/4$. The Palma ratio was formulated as another measure of inequality because the Gini coefficient is most sensitive to changes in the middle, which is relatively stable [23]. The Palma ratio, on the other hand, is sensitive to changes at the extremes.

We use the Gini coefficient, Palma ratio, and the Lorenz curves to provide a balanced understanding of the inequalities in the distributions of how information is spread online across the five types of news sources.

**RQ2: How many users share content from different types of news sources?**

To answer this question we measure the size of the Twitter audience that directly retweets content as a measure of source influence. We compare the distributions of and the average number of users who propagate content posted by sources for each source type. This allows us to measures how large of a direct response each source post causes across source types. When we consider the results at the user level, we compare the behavior of each sub-population of users.

**RQ3: How quickly do users share from news sources of each type?**

To answer this question we measure the speed by which users share content with their followers. Specifically, we measure the delay from the original tweet from the news source to the time of the retweet. It is important to note that it is not the goal of the present work to measure the entire cascade of information propagation; rather, we are interested in direct retweets of news source accounts and, therefore, only collect these specific share events. To compensate for these methodological decisions, we borrow similar statistics from recent work that measured all share events [36]. We can then draw conclusions by comparing our measurements of direct shares against the global aggregate.
RQ4: Who is sharing from different news sources?

To answer this question at an aggregate level we look at user overlap and user-network similarities across the five types of news sources. We hypothesize that there will be large overlaps with trusted sources for the sets of users interacting with sources identified as spreading deceptive content, but that this overlap will probably not be symmetric. That is, that users who spread content from suspicious sources may also spread content from trusted sources. However, users who spread content from trusted sources may be less likely to also spread content from suspicious sources.

Then, we compare source types by certain social network statistics including density, edges to nodes ratios, or average indegree, outdegree, shortest path length, etc. Each source is represented as its own social graph (sub-network). Nodes in each graph represent users who retweeted or @mentioned the news source, or who were @mentioned (using @user) in a tweet connected to the source (through an @mention or a retweet of the source by another user). Edges represent the links between these users on a per-source basis. Specifically, we draw an edge between users $x$ and $y$ if $x$ retweets $y$, $x$ includes $@y$ in a tweet, or $z$ mentions both $@x$ and $@y$ in the same tweet.

We also measure shares across various user demographics and five news source types. We present the results of Mann Whitney U (MWU) tests to identify statistically significant differences in who is sharing information for each type of news, along with common language effect sizes to illustrate the magnitude of those differences.

4.4 Population Level Deception Propagation

We first look at the behavior of users spreading information from each type of news source at the population level. Here we present the results of experiments that use the large dataset of almost 11 million tweets over the course of 2016. We compare and contrast how evenly, how much, how quickly, and who shares content within and across user sub-populations interacting with different types of news sources — trusted, clickbait, con-
RQ1: How evenly do users share content from news sources of varying credibility?

Across all types, including trusted sources, we see large diffusion inequality. We find that a relatively small subset of users are responsible for a large proportion of each sources’ shares. However, these inequalities are not equal across news source types. To look at information propagation inequality for each source type overall, we use each source-user pair as a single propagation unit in the propagation frequency distributions used to generate Lorenz curves and calculate Gini coefficients and Palma ratios for each source type.

We illustrate the inequalities in the propagation of each source type with their respective Lorenz curves in Figure 4.4. The curve for perfect equality is also included for easy comparison not only across news types but within the context of best and worst cases. We see that, for direct retweets, the Lorenz curve for disinformation sources is the furthest from the line of perfect equality. There is a significant gap ($p < 0.01$) between it and
Figure 4.5. Gini coefficients and Palma ratios for direct retweets (RT) from or tweets that @mention sources (@) for each news type, averaged across sources using source-user combinations as diffusion units. Higher values indicate more inequality, highest values are highlighted in bold.

<table>
<thead>
<tr>
<th>Source-Type</th>
<th>Gini Coefficient</th>
<th>Palma Ratio</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RT @</td>
<td>RT @</td>
<td></td>
</tr>
<tr>
<td>Trusted</td>
<td>0.57</td>
<td>0.56</td>
<td>3.68</td>
</tr>
<tr>
<td>Clickbait</td>
<td>0.40</td>
<td>0.46</td>
<td>1.91</td>
</tr>
<tr>
<td>Conspiracy</td>
<td>0.67</td>
<td>0.49</td>
<td>5.95</td>
</tr>
<tr>
<td>Propaganda</td>
<td>0.48</td>
<td>0.49</td>
<td>2.61</td>
</tr>
<tr>
<td>Disinformation</td>
<td><strong>0.83</strong></td>
<td><strong>0.68</strong></td>
<td><strong>20.13</strong></td>
</tr>
</tbody>
</table>

the next closest Lorenz curve (conspiracy sources). Except for conspiracy theory sources which are more equally diffused, the arrangement of Lorenz curves for each source type from closest to furthest from perfect equality is the same for @mentions as for retweets. While Lorenz curves for @mention tweets are more closely plotted around the curve for propaganda sources, there is still significantly more inequality in the propagation of disinformation sources \(p < 0.01\).

Figure 4.5 presents the Gini coefficients and Palma ratios for each news type. Again, we see the greatest inequality in retweet diffusion for disinformation sources, although much lower for @mentions of those sources than direct retweets. The 10% most active users who directly retweet disinformation-spreading sources share 20.13 times as many tweets as the least active 40%, compared to around 2 - 6 times as much for each of the other source types. This ratio drops to 6.49 for @mentions of disinformation sources.

Shares of @mention propagation are more equally distributed among users than direct retweets for trusted, conspiracy, and disinformation. Direct retweet of clickbait and propaganda sources are more evenly shared by users than @mentions, but only slightly. The Gini coefficients also illustrate the larger inequality gap for disinformation sources that we saw with the Lorenz curves, reaching 0.83 for direct retweets. All other source types except for direct retweets of conspiracy sources have Gini coefficients below the minimum.
coefficient for a Pareto 80:20 distribution (0.60). Interestingly, we see that clickbait and propaganda sources are more evenly propagated than trusted sources. The more-even distribution of clickbait articles is not surprising — the whole point of clickbait articles is to motivate many people to click and share the articles.

We also compared the Gini coefficients and Palma ratios of individual sources. The only statistically significant findings \((p < 0.01)\) were in the differences between disinformation-spreading sources and all other types of news sources. In particular, disinformation sources had higher Gini coefficients than trusted sources in 63% of comparisons, and propaganda sources in 66% of comparisons. Using Palma ratios, disinformation sources had higher ratios than trusted sources and propaganda sources in 72% and 75% of comparisons, respectively. These results show that the volume of retweets for disinformation sources are more unevenly distributed than trusted or propaganda sources; these results also demonstrate that the unevenness of distribution is more heavily evident in the extremes of the distributions — among the 10% most and 40% least active users.

**RQ2: How many users share content from different types of news sources?**

We compared the mean number of users who retweet each source post and found that the 95% confidence intervals of those means overlap for all source type comparisons except between conspiracy and disinformation sources. This is not unexpected because the retweet distributions are so heavily skewed. Long tails may heavily influence the means. However, we found that the distributions of the number of users who retweet each source post for trusted and disinformation sources differ significantly \((p < 0.05)\). Further, when we compare the distributions of the mean number of users who retweeted each source tweet for each of the sources, we also see statistically significant differences \((p = 0.03)\) where the mean number of users who retweeted a disinformation source is greater than that of a trusted source in 60% of comparisons. Disinformation sources have, on average, more users retweet each source post.

Figure 4.6 illustrates the distributions of the number of users who retweet each source post.
Figure 4.6. Log-log plots of the distributions of the number of source tweets with a given number of users who retweeted for the five most frequently occurring sources in each type. More frequently occurring sources are plotted in darker shades. This figure is best viewed in color.
tweet for the five most frequently retweeted sources of each type. We see a clear difference in the behavior of trusted sources and disinformation sources. Disinformation sources show a significant shift in the bulk of source tweets compared to the other source types. The same shift is seen in the most popular propaganda source. These sources have a greater proportion of their tweets retweeted by a larger number of users than the other types of sources, including trusted news. As expected, the number of users who retweet is closely correlated with the number of source tweets (Pearson = 0.76, \( p < 0.01 \)).

**RQ3: How quickly do users share from news sources of each type?**

Next, we study how quickly users share direct retweets compared to tweets that @mention source accounts. We find that, as expected, *the majority of retweets occur within 24 hours of the original tweet being posted*, regardless of whether the share was a direct retweet or an @mention of a source. Although previous work found longer delays when deceptive content like rumor is propagated compared to verified news [60], these trends did not appear for all types of deceptive sources when we considered all content posted, *i.e.* deceptive sources may post both deceptive and non-deceptive content.

Delays of retweets from suspicious sources are, on average, longer than for trusted sources; however, their 95% confidence intervals overlap. When we examine cumulative distribution functions (CDFs) of delays for each source type, shown in Figure 4.7, we observe some statistically significant differences (MWU \( p < 0.01 \)) in how long users wait before they retweet from a specific type of deceptive source. Users retweet from trusted, conspiracy, and disinformation sources after similarly short delays (soon after content is posted). However, users who retweet from clickbait sources wait significantly longer \( (p < 0.01) \) after the sources post content. Those who retweet propaganda sources wait the longest after original postings \( (p < 0.01) \).

There are noticeable differences between @mention tweets and direct retweets. A much larger percentage of @mention tweets are shared within the first hour after the original posting occurs than the content retweeted directly from a source for all source types.
This would also include content originally posted that was then retweeted by at least one other user before being retweeted again with the source account mentioned within the retweet, *e.g.* through the use of RT@source. Diffusion delays for tweets that @mention disinformation sources in the body of the tweet appear to skew toward shorter delays in the bottom plot of Figure 4.7 than those mentioning trusted sources. However, MWU tests found the distributions did not differ significantly.

**RQ4: Who is sharing from different news sources?**

Finally, we look at *who* shared content from sources at an aggregate level. For this, we compared the overlap of users across all types of news sources. Table 4.2 shows these

---

Figure 4.7. CDF plots of diffusion delays (in hours) by news type for direct retweets (top) and retweets with @mentions of the source account (bottom). The inset of each plot shows a closer view of the initial diffusion, highlighted with a box in the larger plot.
### TABLE 4.2

OVERLAPS OF USER ACCOUNTS WHO RETWEETED MULTIPLE NEWS SOURCE TYPES

<table>
<thead>
<tr>
<th>proportion of users who retweet column-type sources</th>
<th>who also retweet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trusted</td>
<td>0.009190</td>
</tr>
<tr>
<td>Clickbait</td>
<td>0.014988</td>
</tr>
<tr>
<td>Conspir.</td>
<td>0.067495</td>
</tr>
<tr>
<td>Propag.</td>
<td>0.055091</td>
</tr>
<tr>
<td>Disinfo.</td>
<td>0.675585</td>
</tr>
</tbody>
</table>

Intersections as a percentage of the set of users who shared content from the column’s type of sources. This measurement approximates the likelihood that users who interacted with sources of the column’s type also interacted with sources of the row’s type.

For example, say we are interested in users who interact with both trusted and clickbait news sources. If we are interested in how many of the users who retweet from trusted news sources also retweet from clickbait news sources, $P(CB|T) = \frac{CB\cap T}{T} = \frac{13,080}{1,423,227} = 0.009190$, we look to the second cell of the first column. However, we may also be interested in how many of the users who retweet from clickbait news sources also retweet from trusted news sources, calculated as $P(T|CB) = \frac{CB\cap T}{CB} = \frac{13,080}{19,361} = 0.675585$ and reported in the second cell of the first row. The ratio of $P(T|CB)$ to $P(CB|T)$, $\frac{P(T|CB)}{P(CB|T)} = \frac{0.675585}{0.009190} = 73.51$, corresponds to the ratio of the population of users who retweet from trusted news sources to the population of users who retweet from clickbait news sources, i.e. $\frac{P(T|CB)}{P(CB|T)} = \frac{CB\cap T}{TB\cap T} = \frac{CB\cap T}{CB} \frac{T}{TB} = \frac{T}{CB} = \frac{1,423,227}{19,361} = 73.51$.

We see the highest intersections with trusted sources for all types of deceptive sources. That is, 68%, 60%, 41%, and 27% of users who retweeted information from clickbait,
conspiracy, propaganda, and disinformation sources, respectively, also shared information from trusted sources. Intuitively, this makes sense because the mainstream trusted sources are likely to post a more general or broad range of content than the other types of sources which may be more targeted toward a niche set of users or viewpoints.

We see an interesting overlap between propaganda, clickbait, and conspiracy sources. There is a high proportion of users who shared from clickbait-sources (39%) or conspiracy-sources (43%) who also propagated information from propaganda-sources. However, relatively few users who shared from propaganda-sources also shared from clickbait- or conspiracy-sources — only 3% and 7%, respectively. Users who retweeted clickbait- and conspiracy-sources are fairly likely to have retweeted propaganda-sources, but not the other way around. In fact, users who retweeted clickbait- and conspiracy-sources are the most likely to have also retweeted other source types.

We then studied social graphs of users who directly interacted with each type of news sources. We compared network statistics of each set of graphs and report the key novel findings below:

- The density (i.e. edge to node ratio) of propaganda networks are significantly different from trusted networks ($p < 0.01$).

- The density, average in-degree, and average out-degree of trusted networks differ from conspiracy theory networks ($p < 0.05$) and propaganda networks ($p < 0.01$).

- Average shortest path lengths of disinformation networks differ significantly from trusted ($p < 0.01$), conspiracy theory ($p < 0.05$), and propaganda networks ($p < 0.05$).

4.5 Deception Propagation By Demographics

Next we report how evenly, how much, how quickly, and who shares content from news sources across and within each source type in context of the predicted user demographics of our sample of 66,171 users who actively engage with deceptive news sources.
Figure 4.8. Gini coefficients and Palma ratios for each demographic attribute, averaged for each source type using source-user combinations as diffusion units. † denotes that a large majority (>75%) of users were classified as the given demographic.

**RQ1: How evenly do users share content from news sources of varying credibility?**

As we found at the population level, there is a relatively small group of users who share more than others. This is also reflected in the Gini coefficients and Palma ratios for all source type and demographic combinations in Figure 4.8. When we compare content shares from trusted (T) sources, we see the greatest differences between users in different age brackets. The biggest differences between sub-populations, however, occur between users who share from conspiracy and disinformation sources. Men who shared from conspiracy sources did so much more unevenly than women who shared from these sources, with an 84% higher Palma ratio. A similar pattern occurs between users of different incomes; the Palma ratio is 2.95 times higher for users with higher incomes than users with incomes below $35,000.

As highlighted in Figure 4.8, the Palma ratios for disinformation sources are consis-
Figure 4.9. Interaction Volumes. Demographic where more users shared each individual source post from sources of a given type with the common language effect size (as % of comparisons) for each attribute. Statistical significance from MWU tests of \( p < 0.01 \) for all results.

Tentatively higher than all other source types. *The greatest differences in equality at the extremes of the distribution of active users is seen in younger users where the most active 10% propagate 41.34 times as much as the least active 40%.* Disinformation sources are heavily retweeted by a slight proportion of the populations of younger users, users with lower incomes, and users with only a high school education who retweet a disinformation source at least once.

**RQ2: How many users share content from different types of news sources?**

Next, we look at how many users within each demographic sub-population shared individual posts from sources of each type. In Figure 4.9 we see that there are statistically significant differences (MWU \( p < 0.01 \)) in how many users retweeted each source post or post that @mentioned sources of each type. The demographic that had more users share each post is almost always the demographic which the majority of users were predicted to have (male, older, income above $32,000, or college-educated). In one exception we find that users with income below $35,000 (in 81% of comparisons) or with only a high school education (in 75% of comparisons) comprise the dominant share of users who shared individual posts from disinformation sources, despite there being fewer users predicted to have these attributes.
a. Direct retweets from news sources

<table>
<thead>
<tr>
<th>Source-Type</th>
<th>Gender</th>
<th>Age</th>
<th>Income</th>
<th>Education</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trusted</td>
<td>F</td>
<td>53</td>
<td>Y</td>
<td>56</td>
<td>52</td>
</tr>
<tr>
<td>Clickbait</td>
<td>——</td>
<td>——</td>
<td>A</td>
<td>54</td>
<td>52</td>
</tr>
<tr>
<td>Conspiracy</td>
<td>F</td>
<td>58</td>
<td>O</td>
<td>55</td>
<td>62</td>
</tr>
<tr>
<td>Propaganda</td>
<td>F</td>
<td>52</td>
<td>O</td>
<td>57</td>
<td>56</td>
</tr>
<tr>
<td>Disinformation</td>
<td>F</td>
<td>51</td>
<td>O</td>
<td>50</td>
<td>56</td>
</tr>
</tbody>
</table>

b. Retweets with @mentions of news sources

<table>
<thead>
<tr>
<th>Source-Type</th>
<th>Gender</th>
<th>Age</th>
<th>Income</th>
<th>Education</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trusted</td>
<td>F</td>
<td>54</td>
<td>Y</td>
<td>53</td>
<td>56</td>
</tr>
<tr>
<td>Clickbait</td>
<td>F</td>
<td>54</td>
<td>Y</td>
<td>58</td>
<td>59</td>
</tr>
<tr>
<td>Conspiracy</td>
<td>——</td>
<td>——</td>
<td>——</td>
<td>——</td>
<td>53</td>
</tr>
<tr>
<td>Propaganda</td>
<td>F</td>
<td>52</td>
<td>O</td>
<td>56</td>
<td>56</td>
</tr>
<tr>
<td>Disinformation</td>
<td>F</td>
<td>54</td>
<td>O</td>
<td>56</td>
<td>56</td>
</tr>
</tbody>
</table>

Figure 4.10. Interaction Delays. Demographics who directly retweet from or @mention news source types after a longer delay and the common language effect size (as %). A dash (—) is shown if no significant differences were found. All results are statistically significant (MWU tests of $p < 0.01$).

**RQ3: How quickly do users share from news sources of each type?**

We compare the speed with which each demographic shares content posted by each type of source. CDFs of diffusion delays for each demographic (not illustrated) resulted in plots similar to those found in Figure 4.7. Figure 4.10 illustrates which demographic takes a longer time to propagate content.

Except for comparisons between predicted gender or age brackets for users who retweet clickbait sources, we find significant differences in diffusion delays for direct retweets for all other news source types. Users predicted to have only high school education directly retweet news sources of all types faster than those with a college education. Men retweet from sources more quickly than women for all source types except clickbait. Users with predicted income below $35,000, predicted to have only a high school education, or who are “regular” users share content from clickbait sources faster than their counterparts. Interestingly, we see that while older users will retweet trusted sources more quickly than younger users, there is a greater delay when they retweet the most suspicious sources —
conspiracy, propaganda, or disinformation, relative to the delays of younger users.

On the other hand, we see fewer occurrences of significant differences between sub-populations when users retweet content that only @mention the source rather than directly retweet it. Users with different predicted genders and age brackets now share tweets that @mention clickbait sources after different delays. Users predicted to be men, older, college-educated, or to have incomes below $35,000 will share tweets that @mention clickbait sources faster than their counterparts, but only slightly (effect size of 54-59%).

**RQ4: Who is sharing from different news sources?**

Finally, we study who is sharing in terms of the predicted user demographics. Figure 4.11 presents the demographic that is more likely to directly retweet each source type at least once. We see that older users are less likely than younger users to share content from a disinformation source, but more likely for any other type of news. Similar patterns occur in users predicted to have higher versus lower income or education levels (who may also be older). When we consider the predicted gender, we see a similar trend where women are more likely to share content at least once from disinformation sources and less likely than men for all other types of suspicious sources. However, women are no less or more likely than men to share content from trusted sources. We did not find any significant differences in demographics for @mentions.
4.6 Summary

Our extensive large-scale population-level and demographics analysis of the propagation behavior of users who directly interact with different types of news sources identified several key differences. Some characteristics, like diffusion inequality and the number of users who retweet per post, show large differences between trusted and disinformation news sources. Other results highlight key differences between propaganda, clickbait, and conspiracy news sources. Together, these novel results may be used to differentiate news sources of varying degrees of credibility without the need for expensive content-level annotation.

Recall that this chapter explores four research questions at the population-level and across various demographics. A summary of our novel findings is presented below.

**RQ1: How evenly do users share content from news sources of varying credibility?**

- **Population:** Direct retweets of disinformation sources are most highly retweeted from a small group of users that actively engage with those sources regularly. Propaganda is the next most unevenly shared news, followed by trusted news, conspiracy, and clickbait.

- **By demographics:** We find striking differences in sharing behavior across different user demographics. The largest imbalance is in the most active 10% of young users, which retweet disinformation sources 41.34 as much as the least active 40%.

**RQ2: How many users share content from different types of news sources?**

- **Population:** We did not find statistically significant differences between the number of users that retweet suspicious news across source types. The exception to this finding was disinformation sources, which have a higher number of users who retweet each post compared to trusted sources.

- **By demographics:** Users, on average, with an annual income below $35,000 (in 81% of comparisons) or high school-educated users (in 75% of comparisons) share content from disinformation sources more often than their counterparts.
RQ3: How quickly do users share from news sources of each type?

- Population: Trusted, conspiracy, and disinformation sources have similarly short delays between the time a source posts content and the time that users share it. Delays are significantly longer for clickbait and propaganda sources ($p < 0.01$).

- By demographics: Older users retweet trusted sources more quickly than younger users. Younger users share the most suspicious sources (conspiracy, propaganda, and disinformation) more quickly than older users.

RQ4: Who is sharing from different news sources?

- Population: Users who share information from clickbait and conspiracy news sources are also likely to also share from propaganda sources, but not the other way around.

- By demographics: Users who are within the majority predicted demographics are more likely to share at least once from all types of sources except disinformation, where the minority demographics is more likely to share. We found no significant results for which demographics are more likely to @mention sources.

4.7 Limitations

It is important to highlight some of the limitations in our study. First, the data samples used were not random samples nor representative of the overall Twitter population. It is important to note that we do not make claims about the behavior of all Twitter users. We instead focus on the behavior of users who share information from deceptive news sources identified as conspiracy, propaganda, clickbait and disinformation. Second, the demographic labels were assigned by an imperfect classifier (even though state-of-the-art performance has been achieved). Some of the demographic classes had AUROC rates of 0.72. It is unclear if there is a classification bias in one direction or another. Nevertheless, we caution the reader against making strong claims for individual demographic classes.

4.8 Discussion and Conclusions

To summarize, this is the first study that reports novel observable differences of information spread at the news account level used to understand re-sharing behavior from
sources of varying degrees of credibility in social media. More specifically, this work quantifies how people share misleading, manipulated, or potentially fabricated information from social media news sources; who these propagators are; and how much and how evenly deceptive information is being shared. The properties we highlight in the previous section are differences in the way users directly interact with news sources that could be used to differentiate sources of varying credibility or trustworthiness without the need for tweet level annotation of deceptive versus trusted content or third-party source annotations.

The results of our findings can be used to inform many practical applications including but not limited to: informing models and simulations of deceptive content spread across languages, geolocations, specific groups of users with different demographics and interests e.g., gatekeepers or persistent groups, and types of content e.g., deceptive posts during natural disasters or health messaging campaigns. These models can in turn be used to identify sources of varying credibility, or sources which require further investigation into credibility. For example, such a model that tags trusted versus deceptive, or potentially deceptive, sources could be used to guide not only the general public when they consume information from social news sources but journalists and fact-checkers who focus on verifying news sources and information being spread.

Our analysis focuses on first-hop spreaders of deceptive news content in one social media platform and could naturally be extended to measure how information spreads from deceptive news sources beyond the first hop across many social environments. One could construct information cascades initiated by immediate-hop spreaders and measure how deceptive news propagate. For example, how deep deceptive news travel, how broad they go, how many unique users they reach, how many total users they affect, how long does it take for them to reach the audience of a certain size, how deceptive news evolves while being re-shared, or what the mechanisms of re-sharing are e.g., retweets, quotes, comments etc. Another interesting application could be to evaluate how the same deceptive news content, that is potentially seeded by adversaries, propagates across different social platforms e.g.,
Twitter, Reddit vs. Facebook. Moreover, understanding the intent behind misleading and fabricated news spread is another practical application of our analysis, that goes beyond rumor propagation work. In general, analyzing different types of online social behavior relevant to information spread e.g., information campaigns, coordinated effort, competing campaigns, recurrence, intimidation campaigns etc. is not only critical for national security but would also ensure healthier interactions and boost the level of trust in social media.
CHAPTER 5  
USER-REACTIONS TO SOCIAL NEWS SOURCES OF VARYING CREDIBILITY

In the age of social news, it is important to understand the types of reactions that are evoked from news sources with various levels of credibility. In this chapter, we seek to better understand how users react to trusted and deceptive news sources across two popular, and very different, social media platforms. To that end, we (1) develop a model to classify user reactions into one of nine types, such as answer, elaboration, and question, etc; (2) measure the speed and the type of reaction for trusted and deceptive news sources for 10.8M Twitter posts and 6.2M Reddit comments; and (3) present key findings of our analysis into the prevalence of bots, the variety and speed of bot and human reactions, and the disparity in authorship of reaction tweets between these two sub-populations.

5.1 Introduction

Society’s reliance on social media as a primary source of news has spawned a renewed focus on the spread of misinformation [68]. Most studies that investigate misinformation spread in social media focus on individual events and the role of the network structure in the spread [65, 66, 101, 135] or detection of false information [102]. These and other studies have found that the size and shape of (mis)information cascades within a social network depends heavily on the initial reactions of the users. Yet, we still lack an understanding of how users (human and automated alike) react to news sources of varying credibility and how their various response types contribute to the spread of (mis)information.

Recent studies have found that 59% of bitly-URLs on Twitter are shared without ever being read [38] and 73% of Reddit posts were voted on without reading the linked arti-
Instead, users tend to rely on the commentary added to retweets or comment sections of Reddit-posts for information about the content and its credibility. Faced with this reality, we ask: what kind of reactions do users find when they browse sources of varying credibility? As reliance on social media as a source of news increases and the reliability of news sources is increasingly debated, it is important to understand how users (human and automated) react to news sources of varying levels of credibility and what kinds of reactions are presented to other users as they browse content from the various classes of news sources.

Another direction of recent work has been comparative analyses of the speed of diffusion between news stories of varied veracity. A recent study by Vosoughi et al. [126] found that news stories that were fact-checked and found to be false spread faster and to more people than news items found to be true. Zeng et al. [138] found that tweets which deny rumors had shorter delays than tweets of support. One of the goals of this chapter is to determine if these trends are maintained for various classes of news sources on Twitter and Reddit and among sub-populations of users. As we noted above, most studies that examine misinformation spread focus on individual events such as natural disasters [117], political elections [35], or crises [113] and examine the response to the event on social media. In contrast, our methodology considers immediate reactions to news sources of varying credibility.

5.2 Classifying Reactions in Social Media Postings

Discourse acts, or speech acts, can be used to identify the use of language within a conversation, e.g. agreement, question, or answer. In this section, we describe our approach to classify user reactions into one of eight types of discourse: agreement, answer, appreciation, disagreement, elaboration, humor, negative reaction, or question, or as none of the given labels, which we call “other”, using linguistically-infused neural network models.
5.2.1 Data and Annotations

We use a manually annotated Reddit dataset from Zhang et al. [139] to train our reaction classification model. Annotations from 25 crowd-workers labelled the primary discourse act for 101,525 comments within 9,131 comment threads on Reddit. The Reddit IDs, but not the text content of the comments themselves, were released with the annotations. So we collected the content of Reddit posts and comments from a public archive of Reddit posts and comments. Some content was deleted prior to archival, so the dataset shown in Table 5.1 is a subset of the original content. Despite the inability to capture all of the original dataset, a summary of the original and collected datasets as distributions of comments across reaction types (i.e. primary discourse acts of comments) in Table 5.1 shows that our dataset has a similar distribution to the original.

5.2.2 Model

We develop a neural network architecture that relies on content and other linguistic signals extracted from reactions and parent posts and takes advantage of a “late fusion” approach previously used effectively in vision tasks [62, 96]. More specifically, we combine a text sequence sub-network with a vector representation sub-network as shown in Figure 5.1. The text sequence sub-network consists of an embedding layer initialized with 200-dimensional GloVe embeddings [99] followed by two 1-dimensional convolution layers then a max-pooling layer followed by a dense layer. The vector representation sub-network consists of two dense layers. We incorporate information from both sub-networks through concatenated padded text sequences and vector representations of normalized Linguistic Inquiry and Word Count (LIWC) features [98] for the text of each post and its parent.

---

[1] bigquery.cloud.google.com/dataset/fh-bigquery:reddit_posts|reddit_comments
### TABLE 5.1

**SUMMARY OF ANNOTATED DATASETS**

<table>
<thead>
<tr>
<th>Reaction Type</th>
<th>Zhang <em>et al.</em></th>
<th>Present work</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
<td>%</td>
</tr>
<tr>
<td>agreement</td>
<td>5,054</td>
<td>4.73</td>
</tr>
<tr>
<td>answer</td>
<td>41,281</td>
<td>38.63</td>
</tr>
<tr>
<td>appreciation</td>
<td>8,821</td>
<td>8.25</td>
</tr>
<tr>
<td>disagreement</td>
<td>3,430</td>
<td>3.21</td>
</tr>
<tr>
<td>elaboration</td>
<td>19,315</td>
<td>18.07</td>
</tr>
<tr>
<td>humor</td>
<td>2,358</td>
<td>2.21</td>
</tr>
<tr>
<td>negative reaction</td>
<td>1,901</td>
<td>1.78</td>
</tr>
<tr>
<td>other</td>
<td>1,979</td>
<td>1.85</td>
</tr>
<tr>
<td>question</td>
<td>10,568</td>
<td>9.89</td>
</tr>
<tr>
<td>no majority label</td>
<td>12,162</td>
<td>11.38</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>106,869</strong></td>
<td><strong>100</strong></td>
</tr>
</tbody>
</table>

5.2.3 Model Performance

As shown in Figure 5.2, our linguistically-infused neural network model that relies solely on the content of the reaction and its parent has comparable performance to the more-complex CRF model by Zhang *et al.* [139], which relies on content as well as additional metadata like the author, thread (*e.g.* the size of the the thread, the number of branches), structure (*e.g.* the position within the thread), and community (*i.e.* the subreddit in which the comment is posted).
5.3 Measuring Reactions to Trusted and Deceptive News Sources

In this section, we present key results of our analysis of how often and how quickly users react to content from sources of varying credibility using the reaction types predicted by our linguistically-infused neural network model. That is, we investigate the variety and speed of immediate reactions to news sources of varying credibility so we can determine whether trusted or deceptive news sources evoke more reactions of a given type or faster responses from social media users.
5.3.1 Twitter and Reddit News Data

We focus on trusted news sources that provide factual information with no intent to deceive and deceptive news sources. Deceptive sources are ranked by their intent to deceive as follows: clickbait (attention-grabbing, misleading, or vague headlines to attract an audience), conspiracy theory (uncorroborated or unreliable information to explain events or circumstances), propaganda (intentionally misleading information to advance a social or political agenda), and disinformation (fabricated or factually incorrect information meant to intentionally deceive readers).

Trusted, clickbait, conspiracy, and propaganda sources were previously compiled by Volkova et al. [125] through a combination of crowd-sourcing and public resources. Trusted news sources with Twitter-verified accounts were manually labeled and clickbait, conspiracy, and propaganda news sources were collected from several public resources that annotate suspicious news accounts. We collected news sources identified as spreading disinformation by the European Union’s East Strategic Communications Task Force from euvsdisinfo.eu. In total, there were 467 news sources: 251 trusted and 216 deceptive.

We collected reaction data for two popular platforms, Reddit and Twitter, using public APIs over the 13 month period from January 2016 through January 2017. For our Reddit dataset, we collected all Reddit posts submitted during the 13 month period that linked to domains associated with one of our labelled news sources. Then we collected all comments that directly responded to those posts. For our Twitter dataset, we collected all tweets posted in the 13 month period that explicitly @mentioned or directly retweeted content from a source and then assigned a label to each tweet based on the class of the source @mentioned or retweeted. A breakdown of each dataset by source type is shown in Table 5.2. Figure 5.3 illustrates the distribution of deceptive news sources and reactions across the four sub-categories of deceptive news sources. In our analysis, we consider the

---

Example resources used by Volkova et al [125] to compile deceptive news sources: http://www.fakenewswatch.com/ and others.
**TABLE 5.2**

**SUMMARY OF TWITTER AND REDDIT ONE-HOP NEWS DATASETS**

<table>
<thead>
<tr>
<th>Source Type</th>
<th>Reddit Dataset</th>
<th>Twitter Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># Sources</td>
<td># Comments</td>
</tr>
<tr>
<td>Trusted</td>
<td>169</td>
<td>5,429,694</td>
</tr>
<tr>
<td>Deceptive (no disinfo)</td>
<td>128</td>
<td>664,670</td>
</tr>
<tr>
<td>Deceptive</td>
<td>179</td>
<td>795,591</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>348</strong></td>
<td><strong>6,225,285</strong></td>
</tr>
</tbody>
</table>

![Histogram](image)

**Figure 5.3.** Distributions of deceptive news sources and reactions to those sources (Reddit comments or tweets, respectively) for the Reddit and Twitter datasets across the four sub-categories of deceptive news sources.

set of all deceptive sources and the set excluding the most extreme (disinformation).

**5.3.2 Methodology**

We use the linguistically-infused neural network model from Section 5.2 to label the reaction type of each tweet or comment. Using these labels, we examine how often response types occur when users react to each type of news source. For clarity, we report the five
Figure 5.4. Distributions of the five most frequently occurring reaction types within comments on Reddit and tweets on Twitter for each class of news sources.

most frequently occurring reaction types (expressed in at least 5% of reactions within each source type) and compare the distributions of reaction types for each type of news source.

To examine whether users react to content from trusted sources differently than from deceptive sources in terms of speed, we measure the reaction delay, which we define as the time elapsed between the moment the link or content was posted/tweeted and the moment that the reaction comment or tweet occurred. We report the cumulative distribution functions (CDFs) for each source type and use Mann Whitney U (MWU) tests to compare whether users respond with a given reaction type with significantly different delays to news sources of different levels of credibility.

5.3.3 Results

For both Twitter and Reddit datasets, we found that the primary reaction types were answer, appreciation, elaboration, question, or “other” (no label was predicted). Figure 5.4 illustrates the distribution of reaction types among Reddit comments (top plot) or tweets (bottom plot) responding to each type of source, as a percentage of all comments/tweets reacting to sources of the given type (i.e. trusted, all deceptive, and deceptive excluding disinformation sources).
Figure 5.5. CDF plots of the volumes of reactions by reaction delays for the frequently occurring reactions (i.e., reactions that occur in at least 5% of comments) for each source-type, using a step size of one hour. The CDF for Elaboration-reactions to Deceptive (no disinformation) Twitter news sources is occluded by the CDF for Deceptive Twitter news sources. This figure is best viewed in color.
For Twitter, we report clear differences in user reactions to trusted vs. deceptive sources. Deceptive (including disinformation) sources have a much higher rate of appreciation reactions and a lower rate of elaboration responses, compared to trusted news sources. Differences are still significant \((p < 0.01)\) but the trends reverse if we do not include disinformation sources. We also see an increase in the rate of question-reactions compared to trusted news sources if we exclude disinformation sources.

For Reddit, there appears to be a very similar distribution across reaction types for trusted and deceptive sources. However, MWU tests still found that the differences between trusted and deceptive news sources were statistically significant \((p < 0.01)\) — regardless of whether we include or exclude disinformation sources in the deceptive class. Posts that link to deceptive sources have higher rates of question, appreciation, and answering reactions, while posts that link to trusted sources have higher rates of elaboration, agreement, and disagreement.

Next, we compared the speed with which users reacted to posts of sources of varying credibility. Our original hypothesis was that users react to posts of trusted sources faster than posts of deceptive sources. The CDFs for each source type and platform (solid and dashed lines represent Reddit and Twitter respectively) are shown in Figure 5.5. We observe that the lifetime of direct reactions to news sources on Twitter is often more extended than for sources on Reddit. One exception is answer reactions which almost always occur within the first hour after the Twitter new source originally posted the tweet being answered. This may be due to the different ways that users consume content on the two platforms. Users follow accounts on Twitter, whereas on Reddit users “follow” topics through their subscriptions to various subreddits. Users can view the news feeds of individual sources on Twitter and view all of the sources’ posts. Reddit, on the other hand, is not designed to highlight individual users or news sources; instead new posts (regardless of the source) are viewed based on their hotness score within each subreddit.

In addition, we observe that reactions to posts linked to trusted sources are less heavily
concentrated within the first 12 to 15 hours of the post’s lifetime on Reddit. The opposite is found on Twitter. Twitter sources may have a larger range of reaction delays, but they are also more heavily concentrated in the lower end of that range ($p < 0.01$).

5.4 Reactions from Human-like versus Bot-like Users

In this section, we present a comparison of human-like and bot-like users’ reactions to news sources of varying credibility. We focus on how behavior of bot and human users differ in four specific areas: 1) concentration of reactions to news sources of each level of credibility, i.e. are bots responsible for a larger proportion of the reactions for one class of news sources over another? (prevalence of bots), 2) the variety of reactions each class of news sources evoke, (reaction variety), 3) the speed with which reactions are posted, (reaction speed), and 4) how equally the volume of reactions are spread across the set of users who reacted (reaction inequality).

5.4.1 Related Work

**Prevalence of Bots.** Previous studies have identified the widespread presence of automated accounts or “bots” on social media. A 2014 filing from Twitter acknowledged that 8.5 percent of its active monthly users were automated accounts\(^3\) and subsequent studies found this to be a low estimate of the actual prevalence of bot accounts\(^{[22, 121]}\). Recent work has found that accounts spreading disinformation are significantly more likely to be automated accounts\(^{[108]}\). Other studies highlight evidence of bot participation in political discussion\(^{[58, 89, 133]}\) and astroturf campaigns that present the appearance of widespread support of a candidate, opinion, or topic artificially\(^{[103]}\). A 2018 Pew Research center study found that the majority (66%) of links tweeted to popular news sites are posted by accounts that are likely to be bots, i.e. whose behavior is more similar to bot accounts than

\(^3\)https://www.sec.gov/Archives/edgar/data/1418091/000156459014003474/twtr-10q_20140630.htm?ga=1.155500795.1900968760.1407851022

102
to humans [131]. We seek to answer whether similar trends hold among reactions to news sources.

**Reaction Variety.** Linguistic markers have been found to be effective for early detection of rumors in social networks. For example, Kwon et al. [66] demonstrated better detection performance of rumors on Twitter by using user and linguistic features rather than structural or temporal network features. Similarly, Zhao et al. [141] identified clusters of tweets that contain disputed claims by searching for fact-checking language. Recently, Zhang et al. [139] classified Reddit comments into eight types including agreement, answer, appreciation, disagreement, elaboration, humor, negative reaction, and question, and analyzed patterns from these discussions arranged by various subreddits. Our work goes one step further and employs information credibility classifiers like those mentioned above in order to better understand how (and how fast) human users and bots react to information posted by news sources of varying credibility.

**Reaction Speed.** Information diffusion studies have often used epidemiological models, originally formulated to model the spread of disease within a population, in the context of social media [60, 118, 136]. In this context, users are *infected* when they spread information to other users. A recent study by Vosoughi et al. [126] found that news that was fact-checked (post-hoc) and found to be false had spread faster and to more people than news items that were fact-checked and found to be true. In this work, we examine the speed at which users react to content posted by news sources of varying credibility and compare the delays of different types of responses. By contrasting the speed of reactions of different types, from different types of users (bot and human), and in response to sources of varying credibility, we are able to determine whether deceptive or trusted sources have slower immediate share-times overall and within each combination of user, reaction, and news source types.

**Reaction Inequality.** In the context of social media, the 1% rule and its variants indicate that most users only browse content while a mere 1% of users contribute new con-
tent [54, 120]. Within the subset of those who actively contribute new content, Kumar and Geethakumari [64] found a larger disparity among users who retweeted news from sources that were identified as spreading disinformation. That is, a small number of highly active users were responsible for the vast majority of retweets of disinformation. This study focused only on keywords related to the events in Egypt and Syria in 2013. To answer this research question more generally, the present work quantifies and compares the disparity in sharing behavior of users who frequently reacted to news sources across the various categories of sources, in particular the disparity within each of the reaction types. Specifically, for each type of reaction and each type of news source, we examine whether reactions from bots and human users who frequently reacted are equally distributed across the population of users or if there are a small group of vocal users responsible for the majority of the reaction-tweets.

5.4.2 Data Collection and Annotation

Deceptive news sources that primarily share clickbait, conspiracy theories, or propaganda were previously collected by Volkova et al. [125] from several public resources that annotate suspicious news accounts. The authors also compiled a set of trusted news sources that tweet in English with Twitter-verified accounts which were manually labeled. We collected a set of news sources from https://euvsdisinfo.eu/ that were identified as a source of disinformation by the European Union’s East Strategic Communications Task Force. As of November 2016, EUvsDisinfo reports include almost 1,992 confirmed disinformation campaigns found in news reports from around Europe and beyond. We limited our set to news sources identified between 2015 and 2016 [124].

In total, we focused on 282 news sources which were identified as sources who spread:

- **trusted news (T):** factual information with no intent to deceive the audience;

---

• **clickbait (CB)**: attention-grabbing, misleading, or vague headlines to attract an audience;

• **conspiracy theories (CS)**: uncorroborated or unreliable information to explain events or circumstances;

• **propaganda (P)**: intentionally misleading information to advance a social or political agenda; or

• **disinformation (D)**: fabricated and factually incorrect information meant to intentionally deceive the audience.

We collected tweets posted between January 2016 and January 2017 that explicitly @mentioned or directly retweeted content from one of our 282 sources via the public Twitter API and assigned a label to each tweet based on the class of the source @mentioned or retweeted. Then, we focused on the subset of 4,613,517 tweets identified as English-content in the Twitter metadata. We further focused on users who frequently interacted (at least five times) with the news sources we considered, using tweets posted in any language, which resulted in 431,771 English-tweets for 255 news sources from 184,248 distinct, frequently interacting users. We then classified each of the reaction-tweets as an agreement, answer, appreciation, disagreement, elaboration, humor, negative reaction, question, or other. To do so, we used linguistically-infused neural network models from Section 5.2.

Finally, we gathered botometer scores [27] for each user who posted a reaction-tweet and partitioned the data into bot reactions and human-user reactions using a bot-score threshold of 0.5. That is, human-user reactions were posted by users with a bot score under the threshold of 0.5 and the bot reactions dataset comprises tweets posted by users with bot scores at or above the threshold. A summary of the dataset across source types is presented in Table 5.3.

5.4.3 Methodology

In this subsection, we describe the methodology we used to examine the behavior of bot and human user accounts across varying reactions and in reaction to news sources of
TABLE 5.3

SUMMARY OF ENGLISH-REACTIONS FROM USERS WHO REACTED FREQUENTLY (> 4 REACTIONS BETWEEN JAN 2016 AND JAN 2017)

<table>
<thead>
<tr>
<th>Source-Type</th>
<th>Sources # Accounts</th>
<th>Sources # Tweets</th>
<th>Reactions # Users</th>
<th>Reactions # Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trusted</td>
<td>173</td>
<td>1,633,996</td>
<td>173,098</td>
<td>2,875,120</td>
</tr>
<tr>
<td>Clickbait</td>
<td>10</td>
<td>13,764</td>
<td>8,088</td>
<td>22,352</td>
</tr>
<tr>
<td>Conspiracy</td>
<td>13</td>
<td>31,584</td>
<td>14,047</td>
<td>80,025</td>
</tr>
<tr>
<td>Propaganda</td>
<td>25</td>
<td>81,305</td>
<td>51,160</td>
<td>295,070</td>
</tr>
<tr>
<td>Disinformation</td>
<td>34</td>
<td>68,319</td>
<td>26,131</td>
<td>164,040</td>
</tr>
</tbody>
</table>

varying levels of credibility. As previously discussed, we focus on four types of behavior: prevalence of bots, reaction variety, reaction speed, and the inequality of reaction volume.

First, we examine the prevalence of bots, *i.e.* the relative presence of bots in reactions to news sources of each type. We consider the following two distributions: 1) bot scores of users who reacted to news sources of a given type and 2) bot scores associated with reaction-tweets (the bot scores of users who posted the reaction). The distribution of reaction-users focuses on the distribution of bot scores over the set of unique users who reacted, each user is represented once and only once. On the other hand, users may be represented multiple times in the distribution of bot scores associated with reaction-tweets, if they reacted to a news source of a given class multiple times. With these two distributions of bot scores, we are able to examine the prevalence of bots within the population of reacting users and within the population of reactions broadcast.

As a result of our bot classification methodology, we are able to examine user types using coarse and fine-grained classifications. We first examine the distributions of bots and humans users at a coarse granularity with a binary classification of users as either a bot
or human user account. Then, we consider a fine-grained distinction using the bot scores of users and compare the distributions of bot scores for users who react and of bot scores associated with reaction-tweets (i.e. the bot score of the user who posted). Mann Whitney U (MWU) tests that compare distributions across types of sources and types of users are used to identify statistically significant differences in these fine-grained distributions.

The next characteristic that we evaluate is the variety of reactions each class of news source elicits from bots and from human users. We compare distributions across reaction types overall and separated them into each category of user. Comparisons of reaction variety within each user type allows us to identify certain reactions, classes of news sources, or reactions to a class of news source that have higher concentrations of bot (or human) reactions. Then we consider the tendency of each user type by comparing the frequencies of each reaction type across all classes of sources between bot and human users.

Next we examine the speed of reactions. To answer whether how quickly bots or human users react differs or whether users react to content from trusted sources faster than from deceptive sources, we look at reaction delays for each user type, reaction type, and response to each class of news sources. We define the reaction delay as the time elapsed between the source tweet and when the reaction occurred. We compare the cumulative distribution functions (CDFs) of each user type within and across each type of source to analyze the delay patterns.

Finally, we compare the inequality in reactions among bots and human users. That is, how evenly the volume of reaction-tweets is spread across users of each type; Does each user post an equal number of reactions? We do so using two measures that have been commonly used to measure income inequality: Lorenz curves and Gini coefficients. Rather than measure how much of the total population’s income each individual is responsible for, we repurpose these metrics to measure how much of the total reaction-tweet volume each user is responsible for. We adapt Lorenz curves to measure the inequality in reaction volume by plotting the share of the total reaction volume, i.e. the y% of reaction-tweets
posted, by the share of the population who reacted, \textit{i.e.} the cumulative $x\%$ of users ordered by least to most reaction-tweets posted. This allows us to compare \textit{reaction inequality} across source types the way that economists compare income inequality across countries or regions.

Lorenz curves have traditionally been used to illustrate the distribution of income or wealth graphically \cite{61}. In those domains, the curves plot the cumulative percentage of wealth or income compared against the cumulative (in increasing shares) percentage of a corresponding population. The degree to which a Lorenz curve deviates from the straight diagonal line ($y = x$) representative of perfect equality represents the inequality present in the distribution. In our case, the Lorenz curve is adapted to illustrate the cumulative percentage of propagation (tweets shared) as a function of the cumulative percentage of
users posting, as shown in Figure 5.6.

\[
\hat{G} = 1 - \sum_{k=1}^{n} (X_k - X_{k-1})(Y_k + Y_{k-1})
\]  

(5.1)

The Gini coefficient is defined as the proportion of the area under the line of perfect equality that is captured above the Lorenz curve, \emph{i.e.} \( \frac{a_1}{a_1 + a_2} \) in Figure 5.6. The Gini coefficients reported in subsequent sections are calculated using the formula in Eq. 5.1, which is an approximation of the points of the Lorenz curves observed in the collected data. Using income as an example, Gini coefficients can grow larger than 1 but only if individuals within the population can be responsible for negative shares, that is, if individuals can have negative incomes. In our data, users must be responsible for at least 1 reaction-tweet in order to be considered part of the dataset, so Gini coefficients in our analysis have an upper-bound of 1.

5.4.4 Results

Here we present the key results of our analysis of the behavior of bots and human users in reaction to news sources of varying credibility: the prevalence of reactions from bots and the variety, speed, and the inequality in volume of reaction tweets evoked by each class of news source.

5.4.4.1 Prevalence of Bots

In this subsection, we consider the prevalence of bot users among the audience and reactions broadcast to the community. The distributions of users across bot, human, and unknown (accounts for which we could not collect bot scores) within each class of news source are presented in Figure 5.7.

As shown in Figure 5.7, bots are responsible for approximately 9-15% of the reactions to sources of any given type but only comprise around 7-10% of users responsible for
<table>
<thead>
<tr>
<th>Source-Type</th>
<th>% Bot Users</th>
<th>% Bot Tweets</th>
<th>% Human Users</th>
<th>% Human Tweets</th>
<th>% Unknown Users</th>
<th>% Unknown Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trusted</td>
<td>7.47</td>
<td>12.57</td>
<td>77.32</td>
<td>74.46</td>
<td>15.22</td>
<td>13.03</td>
</tr>
<tr>
<td>Clickbait</td>
<td><strong>10.17</strong></td>
<td><strong>15.06</strong></td>
<td>74.10</td>
<td>72.62</td>
<td>15.73</td>
<td>12.35</td>
</tr>
<tr>
<td>Conspiracy</td>
<td>7.90</td>
<td>8.90</td>
<td>72.79</td>
<td>76.50</td>
<td><strong>19.31</strong></td>
<td>14.62</td>
</tr>
<tr>
<td>Propaganda</td>
<td>6.80</td>
<td>11.54</td>
<td>75.00</td>
<td>70.56</td>
<td>18.20</td>
<td><strong>17.94</strong></td>
</tr>
<tr>
<td>Disinformation</td>
<td>9.64</td>
<td>13.29</td>
<td>73.11</td>
<td>70.18</td>
<td>17.25</td>
<td>16.65</td>
</tr>
</tbody>
</table>

Figure 5.7. Prevalence of Bots. Distributions across bot accounts (bot score ≥ 0.5), human accounts (bot score < 0.5), and unknown accounts (for which we could not collect a bot score) within the set of users who reacted (Users) and the set of reaction tweets (Tweets) for each class of news source. Highest proportions of each user type are highlighted in bold and lowest proportions are in italics.

reaction-tweets. We see that although conspiracy sources have the lowest presence of human users within the population of users who react, they have the highest proportion of reactions authored by human-users. *Trusted news sources have the highest relative presence of human users.* Interestingly, disinformation news sources have only the second highest proportions of bots for users who reacted as well as reaction tweets posted. Instead, clickbait news sources have the highest presence of bots with 10.17% of users who were responsible for 15.06% of the reaction-tweets for clickbait sources identified as bots.

Figure 5.8 illustrates the distributions of bot scores of users who reacted (left) and the scores associated with reaction-tweets, *i.e.* the bot score of the user who posted the tweet, (right). When we compare distributions of users’ bot scores across classes of news sources, we find statistically significant differences. Mann Whitney U comparisons identified significant ($p < 0.01$) differences between distributions for clickbait and trusted or propaganda news sources, where reactions and users who post reactions to clickbait sources have higher bot scores, on average, than trusted or propaganda news sources. Although the distributions of bot scores of unique users and scores associated with reaction tweets are not statistically significant, the slight changes in the shape of the distributions, *e.g.* between the two distributions for Conspiracy sources, paired with the discrepancies
Figure 5.8. Bot score distributions, using a bin width of 0.05, for users who reacted (left) and reaction-tweets (right). Mann Whitney U comparisons of raw distributions found that the average bot score of a user who posted a reaction-tweet is higher ($p < 0.01$) than the average bot score of a user who reacted for all source types except for Conspiracy-sources, where the average bot score of a user who posted a reaction-tweet is lower ($p < 0.01$).
in Figure 5.7 hint at the inequality of reaction tweet volume. That is, they indicate that reactions are not evenly spread across users. We investigate this further in our analysis of reaction inequality.

5.4.4.2 Reaction Variety

We plot the distributions of reaction-types for each of the five classes of news sources in Figure 5.9 and the distribution across bot, human, and unknown users for each source class and reaction type combination for the most frequent reaction types in Figure 5.10. When we compare the distributions of reaction types, we see that the most common reaction types (i.e. present in $\geq 10\%$ of reactions) are answer, elaboration, question, and “other” across all classes of media. In Figure 5.11 we present the relative frequencies of the most common reactions within the reaction-tweets posted by a given user type in response to news sources of a given class. These plots focus more closely on how reaction frequencies differ within a single user-type population.

When we examine the distributions of each class, we find several key differences in the variety of reactions elicited. Conspiracy news sources have the highest relative rate of elaboration responses, i.e. “On the next day, radiation level has gone up. [url]”, with a more pronounced difference within the bot population. Conspiracy news sources also have the lowest relative rate of answer reactions within the bot population, but not within human users. Clickbait news sources, on the other hand, have the highest relative rate of answer reactions and the lowest rate of question reactions across both populations of user types.

Conspiracy and propaganda news sources have higher rates of human question-reactions than they do human answer-reactions; human users who react to these types of news sources question content from the source more often than they respond with an answer. When we examine bot reactions, we see a similar trend for conspiracy sources but a higher relative rate of answer reactions than question reactions to propaganda sources.
Figure 5.9. Distributions of predicted reaction-types within tweets that directly responded to sources of each source-type.

<table>
<thead>
<tr>
<th></th>
<th>Answer</th>
<th>Elaboration</th>
<th>Question</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>H</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td>Trusted</td>
<td>0.16</td>
<td>0.69</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>Clickbait</td>
<td>0.24</td>
<td>0.66</td>
<td><strong>0.11</strong></td>
<td></td>
</tr>
<tr>
<td>Conspiracy</td>
<td><strong>0.09</strong></td>
<td>0.75</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>Propaganda</td>
<td>0.25</td>
<td><strong>0.57</strong></td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>Disinformation</td>
<td>0.15</td>
<td>0.68</td>
<td>0.17</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5.10. Proportions of reactions posted by bot (B), human (H), or unknown (U) user-accounts for each source class and reaction type combination for the most frequent reaction types. Source class(es) with the lowest proportions for each account type are highlighted with bold for each of the reaction types.

5.4.4.3 Reaction Speed

Next, we study the speed with which bot and human users react to news sources. CDF plots for reaction delays of the most frequently occurring reactions are shown in Figure 5.12. These plots illustrate the percentage of reactions that occur within the first $x$ hours after a source posted the original content users reacted to. As expected, a large proportion of the reaction activity occurs soon after a news source posts across all reaction and source type combinations.

Mann Whitney U tests that compared distributions of reaction delays found that humans elaborate on and question content from clickbait sources faster than bots do ($p < 0.01$).
Figure 5.11. Frequencies of most common reaction-types within reactions to news sources of each class posted by bot accounts (above) and human user accounts (below), as a percentage of reactions posted by accounts within each population.

This is reflected in Figure 5.12 where we see the CDF curve for humans pulls above the curve for bots due to the heavier concentration (at least 80%) of reactions with very short ($\leq 6$ hours) delays, compared to bot users with approximately 60-70% of reactions that occurred within the first 6 hours. We see similar trends for all the other combinations of reaction and source types but a few notable exceptions. In the case of answer-reactions in response to content from propaganda news sources, bots respond with significantly shorter delays than human users do ($p < 0.01$). MWU tests comparing bot and human answer-reactions to clickbait and disinformation sources were not found to differ with statistical significance.

5.4.4.4 Reaction Inequality

Finally, we investigate reaction inequality to answer the question: does each user share an equal number of reactions, or are some user or users responsible for a disproportionate number of the reaction tweets for each of the most common reaction types (answer, elaboration, question, and “other”) ? In Figure 5.13 we present the Lorenz curves for bots and
Figure 5.12. Cumulative distribution function (CDF) plots of the volumes of reactions by reaction delays in hours (i.e. the delay between when a source posted content and when the reaction tweet was posted) for bot and human user accounts for the most frequently occurring reactions (occur in at least 10% of tweets) for each source-type, using a step size of one day.
human users when we consider populations with reaction tweets for each combination of reaction type and class of source.

There are significant differences (MWU $p < 0.01$) between the Lorenz curves for bot and human users for all combinations of reaction and source types except for elaboration reactions to clickbait news sources and elaboration, question, and “other” reactions to conspiracy sources. In these cases, human users are also unevenly responsible for reaction tweets, i.e. a subset of the human users are responsible for a disproportionate number of the human-reactions, and the disparity between users who react infrequently and those who post a substantial number of reactions is similar to those within the corresponding populations of bot users.

When users reacted to conspiracy sources, the volume of reaction tweets are similarly unequally distributed across users within the populations of bots and human users except for answer-reactions. Answer-reactions posted in response to conspiracy sources have a smaller prolific subset of bot users responsible for an unexpectedly large volume of the reaction tweets. Human users also respond unevenly with a subset of users who post a disproportionate amount of the reactions, but to a lesser extent than the population of bot users who posted reactions. We see similar patterns across all significant comparisons. That is, *bot-like populations, if significantly different from the corresponding human-like user population, always have a higher level of disparity in reaction volumes than the corresponding human-like users.*

Figure 5.14 presents the increases in Gini coefficient from the human user to bot populations. For clarity, we present only the significant increases ($p < 0.05$) with dashes (−) in place of results without significance. Increases are presented in both absolute terms and relative to the Gini coefficient of the human user population. The most extreme difference is seen in answer-reaction to propaganda sources, with the bot population having a Gini coefficient 58.6% (+0.34) larger than human users do. We find that the highest relative increases for the more deceptive news source classes (conspiracy, propaganda, and
Figure 5.13. Lorenz curves for each of the frequently occurring reactions (occurring in at least 5% of tweets) for each source-type. These Lorenz curves plot the share of reactions by the cumulative share of the population (bots, humans, or accounts without bot scores) as a graphical representation of inequality in reaction volume within each population. The gray dash-dotted line reflects the Lorenz curve that would result from a population wherein each user was responsible for an equal number of reactions. Gini coefficients for bot (B) and human (H) accounts and statistical significance results of Mann Whitney U (MWU) comparisons of Lorenz curves are listed in the top left corner of each subplot; ** if \( p < 0.01 \), * if \( p < 0.05 \), and – if \( p \geq 0.05 \). Lorenz curves and Gini coefficients are presented faded where there are no significant differences between bot an human users.
Table 5.14. Relative Participation Inequality. The difference ($\Delta$), if statistically significant (MWU $p < 0.01$), between Gini coefficients for bot (B) and human (H) user accounts and the relative increase ($\%\Delta$) from the human user to bot Gini coefficient, *i.e.* $(B-H)/H$. A dash (−) is shown if no significant difference was found ($p \geq 0.05$). The highest relative increases are highlighted in bold within source types and italicized within reaction types.

<table>
<thead>
<tr>
<th>Reaction</th>
<th>Trusted</th>
<th>Clickbait</th>
<th>Conspiracy</th>
<th>Propaganda</th>
<th>Disinfo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answer</td>
<td>0.15</td>
<td>21.74</td>
<td>0.21</td>
<td>38.89</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>19.05</td>
<td></td>
<td>58.62</td>
<td>0.11</td>
</tr>
<tr>
<td>Elaboration</td>
<td>0.09</td>
<td>18.00</td>
<td></td>
<td></td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12.96</td>
<td></td>
<td>16.42</td>
<td>5.63</td>
</tr>
<tr>
<td>Question</td>
<td>0.12</td>
<td>25.53</td>
<td>0.13</td>
<td>40.63</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>34.78</td>
<td></td>
<td>9.80</td>
<td>5.63</td>
</tr>
<tr>
<td>Other</td>
<td>0.11</td>
<td>28.95</td>
<td>0.09</td>
<td>31.03</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25.64</td>
<td></td>
<td>9.62</td>
<td>9.80</td>
</tr>
</tbody>
</table>

Disinformation) occur when we compared answer-reactions. The highest relative increase for elaboration reactions occurs within elaboration-reactions to trusted news sources. The highest relative increase in inequality for reactions to trusted news source, however, occurs within the class of “other” reactions, *i.e.* reactions that our annotation model did not predict to be one of the eight reaction types. In contrast, we see the lowest significant relative differences between human and bot users in reactions to disinformation sources. We see that the Gini coefficients for bots are only 5.6% higher than humans for elaboration-reactions, and approximately 10% higher for both question-reactions and other-reactions.

5.5 Discussion and Conclusions

In this chapter, we have presented a content-based model that classifies user reactions into one of nine types, such as answer, elaboration, and question, etc.; a large-scale analysis of Twitter posts and Reddit comments in response to content from news sources of varying credibility, and a novel analysis of bot and human-user reactions to sources of varying levels of credibility using fine-grained reaction labels.

Our analysis of user reactions to trusted and deceptive sources on Twitter and Reddit shows significant differences in the distribution of reaction types for trusted versus de-
ceptive news. However, due to differences in the user interface, algorithmic design, or user-base, we find that Twitter users react to trusted and deceptive sources very differently than Reddit users. For instance, Twitter users questioned disinformation sources less often and more slowly than they did trusted news sources; Twitter users also expressed appreciation towards disinformation sources more often and faster than towards trusted sources. Results from Reddit show similar, but far less pronounced, reaction results.

Examining the reactions of human-like and bot-like users, we identified several key differences in the prevalence of bots within reactions and populations of users who reacted, the variety of reactions each news source evokes, the speed with which different reactions occurred and the inequality of participation in the set of reactions. Users interacting with trusted sources seek verification of the information or claims being spread more often than those spreading information from deceptive sources. We found the highest rate of correction in tweets spreading information from or referencing disinformation-spreading sources. Intuitively, these sources are more likely to be sharing incorrect information and thus, a higher occurrence of users responding with language correcting the information makes sense.
CHAPTER 6

CONCLUSIONS AND FUTURE WORK

In previous chapters we have discussed several analyses of user behaviors on social media — user curation, consumption and reactions to social news sources of varied degrees of credibility. Each chapter has introduced quantifiable patterns of consumption and/or curation behaviors or state-of-the-art predictive models related to such user interactions. Several of these patterns have been used \[43, 46\], and can be used in future work, as informative signals for predictive models and analyses of user interactions with social media or news content. A summary of the key contributions of each chapter is as follows.

**Rating Effects on Social News Posts and Comments:** In Chapter 2, we highlighted how small, random rating manipulations on social media posts and comments create significant changes in downstream ratings for significantly different final outcomes. We found positive herding effects for positive treatments on posts, increasing the final rating by 11.02% on average, but not for positive treatments on comments. Contrary to the results of related work, we found negative herding effects for negative treatments on posts and comments, decreasing the final ratings on average, of posts by 5.15% and of comments by 37.4%. Compared to the control group, the probability of reaching a high rating (≥2000) for posts is increased by 24.6% when they receive the positive treatment and for comments is decreased by 46.6% when they receive the negative treatment.

**Consumption and Curation Habits in Social Rating Systems:** Chapter 3 examined activity logs that recorded all activity for 309 Reddit users for one year and presented findings that highlight the browsing and voting behavior of the study’s participants. We found that
most users do not read the article that they vote on, and that, in total, 73% of posts were rated (i.e. upvoted or downvoted) without first viewing the content. We also show evidence of cognitive fatigue in the browsing sessions of users that are most likely to vote and severe lacks of browsing and voting variety by all users. Despite an intuitive increase in the number of subreddits as session length increases, users are typically engaging with a limited number of communities regardless of how long they browse. While a diverse range of communities are available, users tend to self-select a narrow scope to interact with.

Social News Consumption from Sources of Varying Credibility: Chapter 4 identified key differences in propagation behavior from trusted versus suspicious news sources and introduced state-of-the-art models to predict users’ demographics (age, gender, income, and education). One of the key differences identified was a high inequity in the diffusion rate wherein a small group of highly active users were responsible for the majority of information spread from disinformation news sources overall and within each demographic considered. Another was that news content is shared significantly more quickly from trusted, conspiracy, and disinformation sources compared to clickbait and propaganda. Finally, users who interact with clickbait and conspiracy sources are likely to share from propaganda accounts, but not the other way around. Analysis by demographics showed that users with lower annual income and education share more from disinformation sources compared to their counterparts. Older users propagate news from trusted sources more quickly than younger users, but they share from suspicious sources after longer delays.

User-Reactions to Social News Sources of Varying Credibility: Chapter 5 introduces the state-of-the-art linguistically-infused neural network models that predict the reactions present in users’ social media responses. We show that there are significant differences between trusted and deceptive news sources on Twitter in the speed and the type of reactions posted in response to source content, but far smaller differences on Reddit. Bots are responsible for 9-15% of the reactions to sources of any given type but comprise only 7-10% of accounts responsible for reaction-tweets; trusted news sources have the highest propor-
tion of humans who reacted; bots respond with significantly shorter delays than humans when posting answer-reactions in response to sources identified as propaganda. Finally, we reported significantly different inequality levels in reaction rates for accounts identified as bots vs not.

6.1 Future Directions

The exploration of consumption and curation behavior of users interacting with, propagating information from, and reacting to social news sources summarized above open many directions for future work. The results of these findings can be used to inform many practical applications including but not limited to: informing models and simulations of deceptive content spread across languages, geolocations, specific groups of users with different demographics and interests e.g., gatekeepers or persistent groups, and types of content e.g., deceptive posts during natural disasters or health messaging campaigns.

6.1.1 Enhancing User-Behavior Models

Intuitively, the first avenue of future work follows from the studies presented in Chapters 3 and 4 — modelling users’ social news interactions in the presence of algorithmic biases using predictive signals informed by the behavior patterns identified in these studies. Initial models were developed and presented at the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining in 2017 [43]. Ultimately, the performance results of these models illustrated that the history of user-interactions on a social media platform contains enough information to accurately predict a user’s future interactions. As these models were developed prior to the study described in Chapter 4, future work expanding these initial models to incorporate additional signals identified can be used for improved or more generalizable models of social news interactions.
6.1.2 Identification of Deceptive News Sources From Initial Behavior

The patterns of initial behavior of news sources of varying credibility and user-reactions delineated in this dissertation can also be used to identify sources of varying credibility without relying on time-consuming and labor-intensive manual annotations. Similarly, models that rely on these signals can also be used to identify sources that may require further investigation into credibility. For example, such a model that tags trusted versus deceptive, or potentially deceptive, sources could be used to guide not only the general public as they consume information from social news sources but journalists and fact-checkers who focus on verifying sources and the information being spread in real or close to real time. In this case, the timely nature of models that rely on these initial interaction patterns over post hoc labelling or manual verification is a significant asset.

6.1.3 Extended Analysis of Propagation from News Sources of Varying Credibility

The analysis in Chapter 4 focused on first-hop spreaders of deceptive news content in one social media platform and could naturally be extended beyond the first hop across many social environments. For example, how deep or broad posts from news sources of varying credibility travel, how many unique users they reach or affect, how long does it take for them to reach the audience of a certain size, or how deceptive news evolves when re-shared, etc. Another interesting application is to evaluate how the same deceptive news content propagates across different social platforms e.g., Reddit vs. Facebook and when seeded by different types of users (e.g., automated “bots” versus manual accounts or users of varying demographics). We found that users of varying demographics respond and share content differently, do they influence the overall spread of information they seed? Understanding the intent behind the spread of misleading and fabricated versus verified news is also a practical application of the analyses described in this dissertation.


