Partisanship, Friendship, and Censorship in Online Social Networks

by

Qi Yang

Submitted to the Institute for Data, Systems, and Society
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Social and Engineering Systems

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2021

Author .....
Institute for Data, Systems, and Society
August 16, 2021

Certified by .....
Tauhid Zaman
Associate Professor of Operations Management, Yale University
Thesis Supervisor

Certified by .....
David Rand
Associate Professor of Management Science and Brain and Cognitive Sciences, MIT
Thesis Committee Member

Certified by .....
Devavrat Shah
Professor of Electrical Engineering and Computer Science, MIT
Thesis Committee Member

Accepted by .....
Fotini Christia
Chair, Social and Engineering Systems Doctoral Program

Partisanship, Friendship, and Censorship in Online Social Networks

by

Qi Yang

Submitted to the Institute for Data, Systems, and Society
on August 16, 2021, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Social and Engineering Systems

Abstract

With the widespread adoption of social media in today's society, it has become increasingly important to understand users' behaviors and the underlying factors driving these behaviors. This thesis considers different aspects of this problem, from tie formation to influence campaigns, examining factors of partisanship, friendship, and censorship in today's online social networks.

We begin in the first chapter with an overview of social media and the key behaviors of the users of these platforms and the platforms themselves that we will be studying. In the second chapter, we introduce the follow back problem, and examine how different following strategies and political ideologies can influence the follow back rate. After obtaining followers, one can then begin posting content to influence them. In the third chapter, we consider this type of influence campaign. Recent studies have shown that exposure to opposing opinions causes a backfire effect, where people become more steadfast in their original beliefs. We demonstrate a technique known as pacing and leading which can mitigate this backfire effect over time.

In the fourth chapter, we consider the challenge of inferring political bias in a hyper-partisan media ecosystem. From empirical studies in the previous chapters, we discovered that Twitter exhibited an anti-conservative bias when suspending users. However, many studies find that conservatives are more likely to share misinformation on social media. Therefore, it is possible that the suspensions are due to enforcing an unbiased policy aimed at limiting the spread of misinformation. Here, we evaluate the two possible hypotheses empirically by examining the suspension of Twitter users. We found that the observation that Republicans were more likely to be suspended than Democrats provides no evidence that Twitter was biased against conservatives. Instead, this asymmetry can be explained entirely by the tendency of the Republicans to share more misinformation.

Lastly, the COVID-19 pandemic created large shifts in how people stay connected with each other due to social distancing and isolation measures. In the fifth chapter, we study research questions around the impact of COVID-19 on online public and private sharing propensity, its influence on online communication homophily, and cor-

relations between online communication and offline case severity in the United States. To do so, we study the usage patterns of 79 million US-based users on Snapchat, a large, leading mobile multimedia-driven social sharing platform. Our findings suggest that COVID-19 has increased private communication, while decreased publicly shared content when users are out-and-about, decreased homophily across locations, ages and genders, and has a positive correlation with widening gaps between across-state and within-state communication increases after the onset of COVID-19.

Thesis Supervisor: Tauhid Zaman
Title: Associate Professor of Operations Management, Yale University

Thesis Committee Member: David Rand
Title: Associate Professor of Management Science and Brain and Cognitive Sciences, MIT

Thesis Committee Member: Devavrat Shah
Title: Professor of Electrical Engineering and Computer Science, MIT

Acknowledgments

First and foremost, I would like to thank my advisor, my friend, my academic 'mom' Tauhid Zaman, without whom I could not have finished my PhD, and could not have done it so happily. I'm so lucky to be his student, and he is always funny, supportive, and caring. I enjoyed all the projects we have worked on together, and all the places we have worked at, either in the meeting room, or in the casino (crazy, right?).

Next, I would like to thank my thesis committee members, David Rand and Devavrat Shah. It has been a great pleasure to work with Dave, and I'm amazed by his sharp mind and exceptional writing during the co-authorship. I have always admired Devavrat, and although we have not had the chance to collaborate, he has supported me throughout my PhD.

I would also like to appreciate the support from IDSS. I would like to thank Munther Dahleh, Ali Jadbabaie, and Fotini Christia for their great job creating a collaborative and intellectually exhilarating environment. I am also indebted to many IDSS staff, especially Elizabeth Miles. Moreover, I would like to thank a few other MIT faculty members, Patrick Jaillet and John N. Tsitsiklis for their guidance and support.

This thesis is a result of collaboration with some of the greatest minds in our field. I would like to thank my lab squad Nico and Hunter, my co-authors Mohsen Mosleh, Khizar Qureshi, Neil Shah, Rajan Vaish, Xiaolin Shi, and others from Snapchat Research. I appreciate their time for the valuable discussions and their candid comments on improving my work. All of these people are talented researchers and good friends, and I am very fortunate to work with them.

I also want to say thank you to all my friends at MIT: Jinglong, Penny & Yuanhan, Yi & Lei, Yan & Hui, Yuan, Hanwei, Minghao, Kenny, Amir, Paolo, Manon, Manxi, Dan, Neil, Yiqun, Sharon, Jackie and many others. Thank you all for making my PhD life amazing and colorful.

Finally, I must express my deepest gratitude to my family. I would like to thank my parents, my grandparents, and my BF (boyfriend and best friend) Jinglong who

have been providing me with their endless support and unconditional love.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Technology has changed the ways people interact with others in their daily lives. Over the past decades, there has been a rapid growth of online social network usage. The rise in the number of individuals using Facebook, Instagram, Twitter, Snapchat, and other social media platforms — as well as the amount of time they spend on them — has gathered academics' interest in the effects of online social networks on our lives. There are three elements that are crucial to online social networks research: friendship, partisanship, and censorship.

Friendship

Friendship defines the structure of social networks. People can communicate with others and belong to different networks via virtual communities on social media. By interacting with others online, people utilize social media to obtain information and learn about other ideas and perspectives on problems, topics, and events. Most significantly, new social media is used for socializing; it is a type of media that allows individuals to participate in online discussions and discourse without having to meet face-to-face.

Friendships on social media can be traditional friendships, but they can also be longtime acquaintances (for example, from high school) or very informal interactions between individuals who have never met in person. Some researchers believe that

social networking platforms have not changed the nature of friendship, but rather that the term friend (as in social media friends) is being confused with conflated with weaker social ties such as acquaintances (Amichai-Hamburger, Kingsbury, and Schneider, 2013, Boyd and Ellison, 2007).

Similarity between friends, or between individuals sharing a social link, is known as homophily in the social networks literature, and has long been observed and studied (McPherson, Smith-Lovin, and Cook, 2001). Homophily is the principle that contact between similar people occurs at a higher rate than among dissimilar people. This principle has implications in information diffusion, grouping and community formation, online exposure and more (McPherson, Smith-Lovin, and Cook, 2001). Catanzaro, Caldarelli, and Pietronero (2004), Krivitsky et al. (2009), Shah (2020) study incorporation of homophily as a first-class citizen in network and graph modeling. Guacho et al. (2018), Akoglu, Chandy, and Faloutsos (2013), Pandit et al. (2007) exploit homophily principles in networks to detect misbehaving users and misinformative articles. Recently, several works (Bessi et al., 2015, Gillani et al., 2018, Kumar and Shah, 2018) discuss homophily's role in echo-chamber formation, opinion polarization and misinformation spread in social media. Homophily can also lead to between-group segregation of interpersonal relations in teams (Lau and Murnighan, 1998), and can occur due to formative effects and preferential selection (Currarini, Jackson, and Pin, 2009).

Political campaigns

Social media began as a means to connect with friends and family, but it was quickly adopted by companies looking to reach out to customers through a popular new communication channel. The capacity to connect and exchange information with anybody on the planet, or with a large number of people at once, is the power of social media (Dwivedi, Kapoor, and Chen, 2015). Companies are not the only ones who realized the advertising and networking advantages of these sites. Political parties and politicians increasingly use social media to communicate interactively with citizens. Social media also permits them to have more personal interactions with their audiences, in-

creasing their engagement in politics. This online political campaigning has received increased academic attention in recent years (Kruikemeier et al., 2013, Boulianne, 2015). During the 2004 election cycle, political campaigns began to explore the benefits of these sites, and by 2008 the campaigns in the U.S. presidential election began to fully understand the power of these sites.

In recent years, online political campaigning via social media has received increased academic attention (Boulianne, 2015). Many studies have shown the benefits of social media in enhancing political campaigns, civic engagement and political participation (Jungherr, 2016, Kalsnes, 2016, Conway, Kenski, and Wang, 2015). Other researchers have established that internet use provides an opportunity to voter-consumers to access political information and has positive effects on citizens' involvement in politics, and consequently, contributes to the quality of democracy (Carpini, 2004, Dimitrova et al., 2014, Wang et al., 2012).

Social media also provides a platform for one to persuade a potentially large audience (Perrin, 2015). However, the structure of these networks present their own obstacles to persuasion. Because users can choose from whom they receive information, these networks exhibit a great deal of homophily, where neighbors have similar opinions (Bakshy, Messing, and Adamic, 2015). This creates echo-chambers where users are not frequently exposed to arguments contrary to their own positions and existing opinions are often reinforced. Moreover, even if users are exposed to opposing views, empirical research has shown that when opinions differ greatly, making an argument can actually cause the opinions of the audience to shift away from the argument (Lord, Ross, and Lepper, 1979, Nyhan and Reifler, 2010, Bail et al., 2018).

Partisanship

Partisanship and political polarization in the United States has become a central focus of social scientists in recent decades (DiMaggio, Evans, and Bryson, 1996, Baldassarri and Bearman, 2007, Baldassarri and Gelman, 2008). Americans are deeply divided on controversial issues such as inequality, gun control, and immigration — and divisions about such issues have become increasingly aligned with partisan identities in

recent years (Mason, 2018). Observational studies find that Americans are substantially more likely to have face-to-face social interactions with copartisans (Gentzkow and Shapiro, 2011) and to be connected to copartisans on social media networks (Colleoni, Rozza, and Arvidsson, 2014). These partisan divides are often attributed to "echo chambers," or patterns of information sharing that reinforce pre-existing political beliefs by limiting exposure to opposing political views (Flaxman, Goel, and Rao, 2016). Scholars argue that personalized recommendation algorithms on social media exacerbate political polarization as users receive news only through like-minded sources (Bakshy, Messing, and Adamic, 2015). Researchers have also found that people evaluate counter-attitudinal information more critically than attitude-congruent information (Taber and Lodge, 2006). Under some circumstances, this exposure to counter-attitudinal information may cause people to actively counter-argue, resulting in even stronger attitudes and increased political polarisation, the so-called *backfire effect* (Bail et al., 2018).

Censorship

Social media companies are also taking an active role in shaping the uses of their platforms and intervening to mitigate the effect of objectionable content. Nonetheless, it has been widely claimed by those on the political right that social media platforms are biased against conservatives. Many prominent conservatives, including former President Donald Trump, have filed lawsuits against social media companies, accusing them of censoring users with an anti-conservative bias (Bond, 2021). This view is shared by many Republican voters - for example, in an August 2020 poll, roughly seven-in-ten Republicans said major social media sites tend to favor liberals over conservatives (Vogels, Perrin, and Anderson, 2020). Yet these charges are based on anecdotal instances of particular platform actions (e.g. Twitter's permanent suspension of Trump's account), rather than any systematic comparison of enforcement on conservatives versus liberals.

More importantly, charges of anti-conservative bias must contend with alternative explanations for the preferential suspension of conservatives - most notably, attempts

by platforms to combat misinformation. There is widespread bi-partisan concern about misinformation and "fake news" on social media, and widespread bi-partisan agreement that technology companies should take action against such content. For example, a July 2020 poll found that a majority of supporters of both parties believed that technology companies are responsible for preventing misuse of their platforms aimed at influencing elections (van Green, 2020), and a November 2020 poll found that a majority of supporters of both parties believed that social media companies should be held responsible for false or inaccurate content posted on their platforms (Koopman, 2021).

However, while the desire to combat misinformation is bi-partisan, there are substantial partisan asymmetries in the spreading of misinformation. Numerous studies of the 2016 U.S. presidential election, for example, have found that conservatives shared dramatically more fake news on social media than liberals (Guess, Nagler, and Tucker, 2019, Grinberg et al., 2019). Thus, even if platforms are more likely to suspend conservatives than liberals, this asymmetry in suspension could simply arise from politically neutral enforcement aimed at satisfying the bi-partisan demand for a reduction in online misinformation, rather than anti-conservative bias.

COVID-19's impact on social networks

In 2020, the sudden onset of a new, highly contagious coronavirus has created large shifts in how people stay connected with each other due to isolation measures. With the ongoing COVID-19 pandemic, people have increased their social media usage to seek information about the pandemic according to surveys (Wiederhold, 2020). The widespread effects of COVID-19 have led to several recent initiatives in studying its interplay with social media use: Kim (2020) collected comments from Korean social media to analyze negative emotions and societal problems during COVID-19. Lin, Liu, and Chiu (2020) used Google Keyword Search frequency to predict the speed of COVID-19 spread in 21 countries/regions. Singh et al. (2020) characterizes Twitter conversation around COVID-19, and indicates that online conversation about the virus leads to new cases geographically. Several prior works have also studied misin-

formation around COVID-19: Depoux et al. (2020) remarks upon the rapidity of the panic and spread of misinformation. Huynh and others (2020) studies how COVID-19's risk perception in Vietnam is heavily mediated by baseline and geographical social media use. Pennycook et al. (2020) found that many people disseminated false information related to the virus because they failed to reason appropriately if content was true or false before sharing, and that propensity to share was misaligned with people's ability to judge accuracy. Brennen et al. (2020) indexes many common false claims about COVID-19 circulating on social media, and notes that the majority are misinformative (improper context, misleading) rather than disinformative (fabricated or imposter content).

Thesis Outline

Social media networks have been a central feature of modern social life. Despite having fewer total users than other major platforms (e.g. Facebook), Twitter has become a major conduit of information and news, with traffic on the platform having the ability to influence politicians and have substantial economic impacts on companies. A key feature of social media platforms in general, and Twitter in particular, is that users can influence their followers - and thus that a key goal of many users is to convert those whom they want to influence into followers. This challenge of getting a particular user to follow you has been deemed the *follow back problem* and we will examine different factors that affect the follow back rate in our work. In Chapter 2, we introduce the follow back problem, and examine how different following strategies and political ideologies can influence the follow back rate.

Then in Chapter 3, after establishing ties with social media users, we examine influence campaigns. Online social networks create echo-chambers where people are infrequently exposed to opposing opinions. Even if such exposure occurs, the persuasive effect may be minimal or nonexistent. Recent studies have shown that exposure to opposing opinions causes a backfire effect, where people become more steadfast in their original beliefs. We demonstrate a technique known as pacing and leading

which can mitigate this backfire effect over time.

In Chapter 4, we study the issue of inferring political bias in online social networks. We find that the observation that Republicans are more likely to be suspended than Democrats on Twitter provides no evidence that Twitter is biased against conservatives. Instead, this asymmetry can be explained entirely by the tendency of the Republicans to share more misinformation

Lastly, during the past year, the Covid-19 pandemic has created large shifts in how people stay connected with each other in lieu of social distancing and isolation measures. In Chapter 5, we study the usage patterns of 79 million US-based users on Snapchat. Our findings suggest that COVID-19 has increased private communication, while decreased publicly share content when users are out-and-about, decreased homophily across locations, ages and gender, and has a positive correlation with widening gaps between across-state and within-state communication increases after the onset of COVID-19.

Chapter 2

The Follow Back Problem in a Hyper-Partisan Environment

Online social networks provide users with a platform to persuade and influence a potentially large audience. One can increase the audience size by increasing the number of social ties. The nature of these ties vary across social media platforms. In Facebook, these ties are undirected relationships. A pair of users sharing a tie in Facebook are referred to as friends. In other platforms, such as Twitter and Instagram, the ties are directed. A user can choose to follow another user. The pair of users in a directed tie are referred to as the follower and the following. Followers are able to see content posted by their following. Therefore, to gain influence, a social media user would strive to obtain many followers in a directed social network. This challenge of obtaining followers is referred to as the *follow back problem*. In this chapter we will study different strategies for this problem. In particular, we will focus on the follow back problem in the context of politically polarized social networks and study how political ideologies affect social media users' propensity to follow each other.

2.1 Previous Literature

Prior work has suggested that requesting to follow others with similar interests increases the probability of reciprocation (Smith and Giraud-Carrier, 2010). The impact

of homophily in obtaining followers is particularly strong with respect to political affiliations. An observational study found that Twitter users are more likely to be connected to co-partisans on social networks (Colleoni, Rozza, and Arvidsson, 2014). A recent field experiment demonstrated that Democrats and Republicans were much more likely to reciprocate follows from co-partisans Mosleh et al. (2021b).

A key element to the follow back problem is the idea of mutual ties. When two social media users are connected to a third, there is a tendency for the two users to close this tried by forming ties with each other. This phenomenon is referred to as *triadic closure* (Granovetter, 1977). The effect of triadic closure on influence was demonstrated empirically in Ugander et al. (2012) and F. Nagle (2009). In Boshmaf et al. (2013) the authors found that triadic closure had an effect on an user's ability to form connections with new users in Facebook. While triadic closure was initially defined for undirected graphs, there have been extensions developed for directed graphs (D.M. Romero, 2010). Cheng et al. (2011) found that the number of two-step directed paths between two nodes has strong predictive power for the follow back rate. Their analysis showed that five or six two-step paths served as a threshold, above which there was a significantly higher follow back rate.

2.2 Study One: Impact of Interaction Types and Political Partisanship

We first conduct an experiment on Twitter to study the effect of two factors on the follow back rate: interaction type and political partisanship. There are two interactions we consider. These actions are designed to build some form of rapport with the user. First is liking a user's tweet. This is a very basic action and lets the user feel that their content is popular. Second is following the user. This signifies that one wants to be in the user's audience for their content. In addition to interaction types, we also consider political ideology. We measure follow back rates between Democrats and Republicans in both co-partisan an counter-partisan pairings.

2.2.1 Experiment Design

For the experiment we created multiple Twitter bot accounts that varied in their political partisanship. Our bot accounts were designed to appear human with identical descriptions, except for their political identification. In total we created ten accounts, half of which were Republicans and half of which were Democrats. Examples of these bots are shown in Figure 2-1. More details about bot creation and bot automation can be found in Appendix A. We had these bots interact with subjects who identified as either Democrats or Republicans.

Figure 2-1: Design of bot accounts. We created ten human-like identical looking bot accounts (five Republican and five Democrat). The bot accounts followed a set of elite accounts according to their political partisanship, and retweeted randomly from them every day. Their political stances were also revealed in their description with partisan hastags.

To identify experimental subjects, we collected a list of Twitter users who tweeted or retweeted the hashtags #Tump2020 or #VoteBidenHarris2020. These hashtags signaled support for one of the main candidates in the 2020 U.S. presidential election. We further confirmed these users' political ideologies by examining their media consumption (Eady et al., 2019a).

From this full list, we constructed a politically balanced set of users to form the subject pool for our experiment. We removed users from this set with more than 15,000 followers as these accounts would have a very low probability of following back due to their popularity. We also removed users with zero friends or zero followers

as these accounts may not be active at all. Finally, we removed users for whom our partisanship estimator (Grinberg et al., 2019) was unable to return a score, usually due to the account having no media consumption, being banned, having strict privacy settings, or being restricted.

We used randomized assignment by blocking to balance subjects over the experimental conditions and boost our causal inference precision (Higgins, Sävje, and Sekhon, 2016). We created homogeneous blocks of users for the experiment based on (1) user partisanship, (2) logarithm of the number of their followers (we used this transformation since follower counts are highly skewed), (3) number of days with at least one tweet in the past 14 days (to measure recent activity on the platform), and (4) follow back rate, which was measured by number of mutual friendships divided by total number of followers. Using this blocking, users were randomly assigned to one of six conditions, in which the user was followed by a bot account that was either a co-partisan or counter-partisan, with one of the following three interaction strategies: directly follow, like the subject's tweet, or a combination of both following and liking. In total, our bot accounts successfully followed 8,104 users (3,952 Republicans and 4,152 Democrats).

2.2.2 Results

Figure 2-2 shows the follow back rate of Democratic and Republican users that reciprocated the bot accounts' interaction with a follow back in each experimental condition. We interpret the outcomes using a linear probability model predicting whether the user followed the bot based on co-partisanship with the bot, interaction strategy of the bot, political partisanship of the user, and all other measured features. We also report exact p-values (pFRI) calculated via Fisherian randomization inference based on 10,000 permutations. We found counter-partisan pairings had a significantly lower follow back rate than co-partisan pairings (b=-0.046, SE=0.007, $p<.001$, pFRI$<.001$). Overall, users were twice as likely to follow back a co-partisan bot compared to a counter-partisan bot. Moreover, we see that in terms of interaction type, liking a users' tweets was not effective. The follow back rate was significantly lower

than following or liking and following (b=-0.076, SE=0.007, p<.001 , pFRI<.001). Adding a like to a follow did not make a significant difference in the follow back rate. There is no difference in the follow back behaviors between Republicans and Democrats.

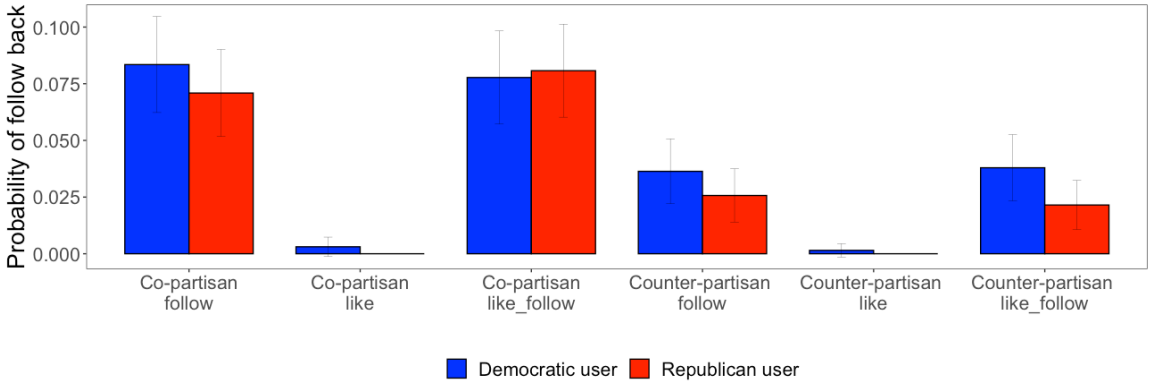


Figure 2-2: Plot of the mean follow back probability for Democratic and Republican users following the bot accounts in each experimental condition. Error bars indicate 95% confidence intervals.

2.3 Study Two: Impact of Mutual Ties

In our second study, we study the effect of mutual ties on follow back rates. Study one provided each bot with a set of followers. The followers of these followers have a mutual tie with the bot: they each follow someone who follows the bot. The subjects for study two are the followers of the bot followers. Because these subjects all have a mutual tie with the bot and span different political ideologies, we are able to use them to test the combined impact of political partisanship and mutual ties on follow back rates.

2.3.1 Experiment Design

We began by gathering 5,856 users who are followers of subjects who followed the bots in study one. For each bot we randomly assigned the followers of their followers into two groups. Then each bot would interact with its subjects in two phases. In phase

one, the bot followed the users in group one. In phase two, which is a week after phase one, we blocked every user who followed the bot, including both phase one subjects as well as subjects from study one. Now the users in group two followed no one who followed the bot. The bot then followed the group two users. The experiment design is illustrated in Figure 2-3. To control for temporal effects, we used a few idle bot accounts to follow random users on Twitter in phases one and two. We found that the follow back rates were not significantly different under a two-sample binomial proportions test for the two phases.

Since not all followers of the bot followers have media consumption, we defined their partisanship as the partisanship of who they followed in study one. We also controlled for covariates including the number of followers (log scaled), number of friends (log scaled), as well as the number of mutual friends within the study two user set, which accounts for possible spillover effects.

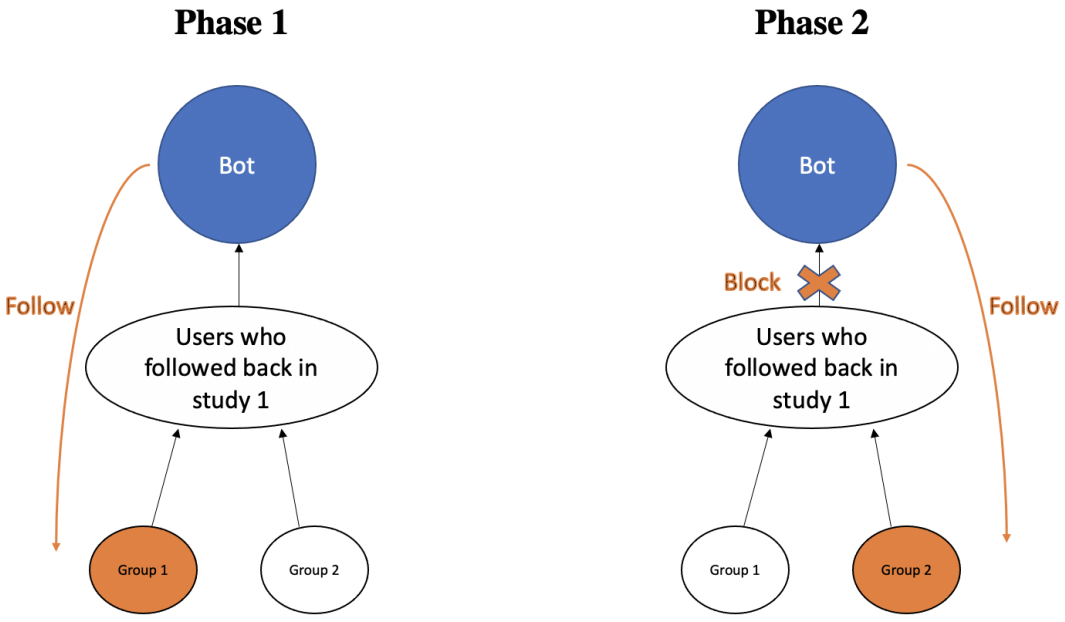


Figure 2-3: Design of study two. In phase one, the bot account follows users in group one who follow a user who followed the bot from study one. In phase two, the bot account blocks all of its followers and follows the users in group two.

2.3.2 Results

Figure 2-4 shows the follow back rate of Democratic and Republican users in each experimental condition. Recall that in phase one, there is a mutual tie between the bot and the subject, whereas in phase two, there is none. We interpret the outcomes using a linear probability model predicting whether the subject followed the bot based on the existence of a mutual tie, the political partisanship of the following of the subject from study 1, co-partisanship with the bot, and all other covariates. We also report exact p-values (pFRI) calculated via Fisherian randomization inference based on 10,000 permutations.

First, we found that the presence of a mutual tie does not make a difference in follow back rate. In phase one where mutual ties exist, the average follow back rate is 25%, whereas in phase two, the average follow back rate is almost identical at 26.7%. Moreover, we examined the effect of shared partisanship. Here shared partisanship is inferred by the partisanship of the following of the subject from study one. We found that counter-partisan users have a significantly lower follow back rate (b=-1.038, SE=0.319, p<.001 , pFRI<.001). Users were still twice more likely to follow back a co-partisan bot compared to a counter-partisan bot, similar to study one.

2.3.3 Discussion

In these two studies, we tested the causal effect of partisanship, intertwined with interaction type and the existence of mutual ties on the follow back rate. Users were roughly twice as likely to follow back bots whose partisanship matched their own. Moreover, the follow back rate when the bot only liked the users' content without following as well was significantly lower than following or liking and following. Given that the bot followed the user, also liking their tweet did not make a significant difference in the follow back rate. Lastly and interestingly, we did not find evidence for a higher follow back rate when a mutual tie existed between the bot and the user. This is in contrast with other extant work that found a positive impact of mutual ties. There are many possible hypotheses for this. One could be that a single mutual

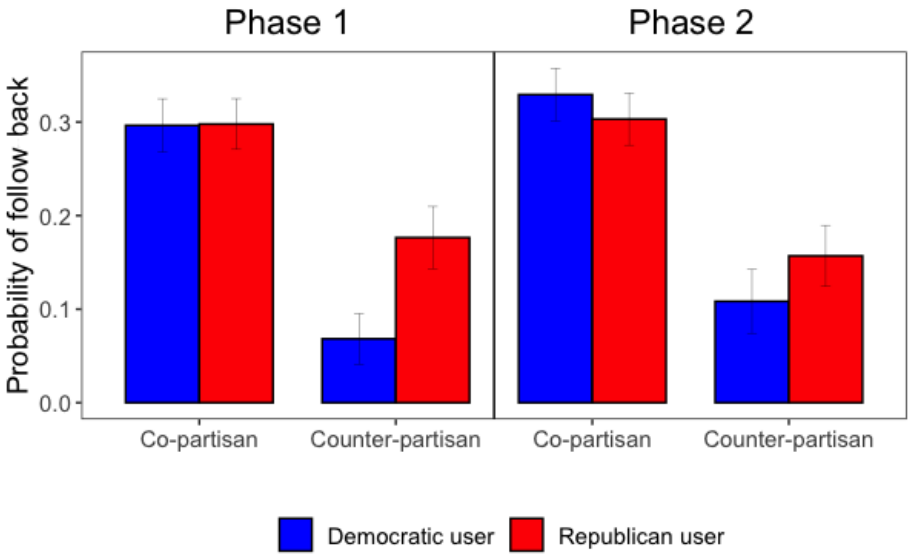


Figure 2-4: Plot of the mean follow back probability for study two in phases one (mutual tie present) and two (no mutual tie). Error bars indicate 95% confidence intervals.

tie was not sufficient to affect the follow back rate. This aligns with the findings of Cheng et al. (2011), where five to six mutual ties were needed to increase the rate. Another hypothesis is that the impact of partisanship is so strong that it overrides any sort of social impact of a mutual tie. Further studies are required to identify the precise reason mutual ties do not affect follow back rates in partisan settings.

Chapter 3

Mitigating the Backfire Effect Using Pacing and Leading

Online social networks create echo-chambers where people are infrequently exposed to opposing opinions. Even if such exposure occurs, the persuasive effect may be minimal or nonexistent. Recent studies have shown that exposure to opposing opinions causes a backfire effect, where people become more steadfast in their original beliefs. We conducted a longitudinal field experiment on Twitter to test methods that mitigate the backfire effect while exposing people to opposing opinions. Our subjects were Twitter users with anti-immigration sentiment. The backfire effect was defined as an increase in the usage frequency of extreme anti-immigration language in the subjects' posts. We used automated Twitter accounts, or bots, to apply different treatments to the subjects. One bot posted only pro-immigration content, which we refer to as arguing. Another bot initially posted anti-immigration content, then gradually posted more pro-immigration content, which we refer to as pacing and leading. We also applied a contact treatment in conjunction with the messaging based methods, where the bots liked the subjects' posts. We found that the most effective treatment was a combination of pacing and leading with contact. The least effective treatment was arguing with contact. In fact, arguing with contact consistently showed a backfire effect relative to a control group. These findings have many limitations, but they still have important implications for the study of political polarization, the backfire effect,

and persuasion in online social networks.

3.1 Introduction

Today online social networks provide a platform for one to persuade a potentially large audience (Perrin, 2015). However, the structure of these networks present their own obstacles to persuasion. Because users can choose from whom they receive information, these networks exhibit a great deal of homophily, where neighbors have similar opinions (Bakshy, Messing, and Adamic, 2015). This creates echo-chambers where users are not frequently exposed to arguments contrary to their own positions and existing opinions are often reinforced. Moreover, even if users are exposed to opposing views, empirical research has shown that when opinions differ greatly, making an argument can actually cause the opinions of the audience to shift away from the argument (Lord, Ross, and Lepper, 1979, Nyhan and Reifler, 2010, Bail et al., 2018). This *backfire effect* poses a major challenge when trying to persuade or influence individuals.

Within such online settings it has been found that the use of uncivil or extreme language can spread in such online settings (Cheng et al., 2017). Such language can create animosity among social media users and prevent constructive discussions. Given the scale and importance of online social networks, it is important to develop methods to persuade in these environments. However, the combination of the backfire effect and echo-chambers present major obstacles to persuasion. The structure of echo chambers prevent one from seeing contrary opinions, but if one does, the backfire effect limits their persuasion ability. It would be useful to have a method that allows one to present arguments in online social networks in a manner that mitigates the backfire effect and the usage of extreme language.

In this work we conduct a field experiment to test persuasion methods in an online social network. Our standard method, which we refer to as *arguing*, simply has one present arguments for the target position without any other interaction with the audience. Arguing can be viewed as a messaging based persuasion method because

it only involves content posted by the arguer. We test another messaging method we refer to as *pacing and leading* which is based on the idea that persuasion is more effective if there is some sort of bond or connection between arguer and audience. This method begins by having the arguer emotionally pace the audience by agreeing with their opinion on the persuasion topic. This is done to form a bond with the audience. Then over time, the arguer shifts its own opinion towards the target position which will lead the audience to this position. In addition to messaging based methods, we also test a persuasion method based on interaction with the audience that we refer to as *contact*. This method has the arguer like the social media posts of its audience. This interaction can serve as a form of social contact in an online setting and potentially lead to more effective persuasion when combined with messaging based methods.

Our experiment tests two primary hypotheses. The first hypothesis is that pacing and leading will mitigate the backfire effect more than standard arguing through the effect of in-group membership, which means that the arguer and audience belong to a common social group. Theories of inter-group conflict suggest that persuasion is more effective when the arguer and audience are in-group (Tajfel et al., 1979). In (Munger, 2017) race was used as an in-group feature to persuade users in the online social network Twitter to not use extreme language. It was found that in-group persuasion (arguer and audience have the same race) was more effective than out-group persuasion (arguer and audience have different races). This study demonstrated that race was an effective in-group feature for persuasion. We expect a similar finding when in-group membership is based the opinion towards the persuasion topic.

Our second hypothesis is that contact between the arguer and audience will mitigate the backfire effect. By having contact with the audience, the arguer can form a rapport with the audience and shift them to a more positive affective state. Persuasion strength may be enhanced by these psychological effects. Researchers have found that affective states impact the efficacy of persuasion (Rind, 1997, Rind and Strohmetz, 2001). The social influence literature is rife with evidence that social rapport and a positive relationship enhance persuasion and influence (Cialdini and Trost, 1998). Moreover, it has been found that a person's persuasive ability is strengthened

if the audience likes this person (Burger et al., 2001).

3.2 Experiment Design

The persuasion topic used in our study is immigration. Events such as the European refugee crisis have made immigration a charged political issue and it is an active topic of discussion on social networks. Several studies have measured population level sentiment on this topic in Twitter (Öztürk and Ayvaz, 2018, Backfried and Shalunts, 2016, Coletto et al., 2016). It was found in (Öztürk and Ayvaz, 2018) that English posts about the refugee crisis were more likely to have a negative opinion on the topic. A similar result was found for Twitter users in the United Kingdom (Coletto et al., 2016). Given the level of interest in the topic and its geo-political importance, immigration is an ideal topic to test persuasion methods. In our experiment we try to persuade individuals to have a more positive opinion of immigration.

We employ automated Twitter accounts, which we refer to as bots, to test different persuasion methods. Our experiment subjects are Twitter users who actively discuss immigration issues and have anti-immigration sentiment. Each bot implements a different persuasion method. One bot is a control which posts no content and does not interact with the subjects. One bot applies the arguing method by posting content which is pro-immigration. The third bot applies pacing and leading by posting content that is initially anti-immigration and then gradually become more pro-immigration. To test the contact treatment, we randomly selected half of the subjects from each bot and have the bots like the posts of these subjects. To assess the effectiveness of the different persuasion methods, we analyze the sentiment of content posted by these subjects over the course of the experiment. We now present details of our experiment design, which is illustrated in Figure 3-1.

The subjects for our experiment were Twitter users who have an anti-immigration sentiment. To find potential subjects we began by constructing a list of phrases that conveyed strong anti-immigration sentiment, such as #CloseThePorts, #BanMuslim, and #RefugeesNotWelcome. We used the Twitter Search API to find posts, known as

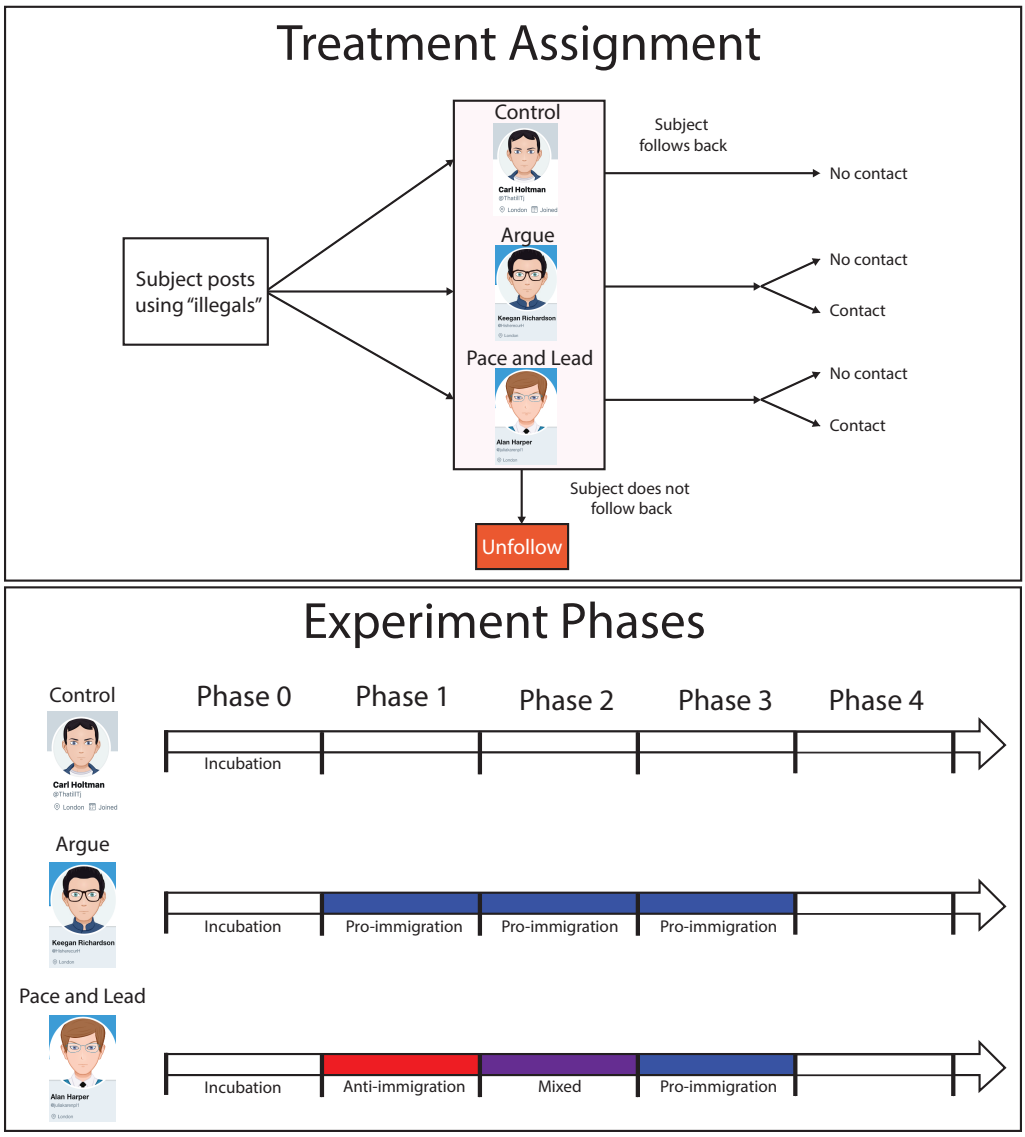


Figure 3-1: (top) Diagram illustrating the subject acquisition procedure for the experiment. (bottom) Timeline of experiment phases.

tweets in Twitter, that contained at least one of these keywords. More details about how to use Twitter API can be found in Appendix A. We then collected the screen names of the users who posted these tweets.

Our search procedure has the potential to find users who do not have anti-immigration sentiment. For instance, to convey support for immigrants, a user could post a tweet critical of an anti-immigration phrase. To make sure that there were not many users who fall in this category, we manually investigated 100 random users collected by our search procedure. We found that none of the users was pro-immigration, giving us confidence that the overwhelming majority of our potential subjects were anti-immigration.

We further narrowed our subject pool by requiring each user to satisfy the following criteria. First, their tweet must be in English and must not contain only punctuation or emojis. Second, the user should not be an automated bot account. The text conditions on the tweet were checked using simple pattern matching. Bot accounts were identified using the machine learning based Botometer algorithm (Davis et al., 2016). Users who Botometer identified as being the most bot-like were manually reviewed and eliminated if they are indeed bots.

We created Twitter accounts for the control, argue, and pace and lead treatments. One of the goals of our experiment was to test persuasion strategies in a realistic setting. Therefore, we wanted the bots to resemble human Twitter users, in contrast to the study in (Bail et al., 2018) where the subjects were told in advance the Twitter account they were following was a bot. To accomplish this, we had the bots be active on Twitter for a two month incubation period before we started the experiment. Each of the bots location was set to London, and they followed a number of popular British Twitter accounts. The bots were designed to look like white males with traditional European names. We used cartoon avatars for the profile pictures, similar to what was done in (Munger, 2017). We show the profile images for the bots and list their treatment type in Figure 3-1. During the incubation period, once or twice a day the bots posted tweets about generic, non-immigration topics and shared tweets about trending topics on Twitter, an act known as retweeting. They also tweeted articles or

videos talking about immigration, but not yet taking a stance on the issue. This was done to show that the bots had some interest in immigration before the experiment began. We provide examples of the incubation period tweets and retweets in Appendix B.

One month into the incubation period, we began obtaining subjects for the experiment. To participate in the experiment, the potential subjects needed to follow the bots so that the bots' tweets would be visible in their Twitter timelines. We randomly assigned each of the users in the subject pool to the bots. The bots then liked a recent tweet of their assigned users and followed them. The liking of the tweet and following were done to increase the follow-back rate of the potential subjects. To avoid bias before the experiment, all tweets the bots liked were manually verified to not be immigration related. After liking and following their assigned subjects' tweets, the bots were able to achieve an average follow back rate of 19.3%. In total we were able to obtain 1,336 subjects who followed the bots. To make the bots appear more human, we tried to keep their ratio of followers to following greater than one. To do this, the bots would wait one to seven days before unfollowing a user who did not follow-back. The actual wait time depended on the user activity level, with a longer wait time given for less active users. Details are provided in Appendix B.

The experiment had four different phases. We denote the incubation period as phase zero. Phases one, two, and three are the main active phases of the experiment. The control bot does nothing for these phases. The argue bot would post a pro-immigration tweet once a day in these phases. The pace and lead bot also posted tweets once a day in these phases, but the tweet opinion varied. In phase one the tweets were anti-immigration. In phase two the tweets expressed uncertainty about immigration or potential validity of pro-immigration arguments. In phase three the tweets were pro-immigration, similar to the argue bot. We constructed the tweets based on what we deemed a proper representation of the opinion for each phase. We show example tweets for the argue and pace and lead bots in the different phases in Appendix B. In phase four of the experiment the bots tweeted nothing. We used this phase to measure any persistent effect of the treatments. Each phase lasted

approximately one month, except for the incubation phase which lasted two months. The incubation phase began on September 27th, 2018 and the fourth phase was completed on March 1st, 2019. The experiment timeline is shown in Figure 3-1.

In addition to the tweeting based treatments, we also tested the contact treatment on the subjects. We randomly assigned 50% of the subjects of the argue and pace and lead bots to this treatment group. During phases one, two, and three, the bots liked the tweets of the subjects assigned this treatment. When the bot liked a subject's tweet, the subject is notified. Liking tweets would make the bot more visible to the subject and potentially give the subject a greater trust or affinity for the bot. The control bot did not apply the contact treatment to any of its subjects.

All subjects voluntarily chose to follow the bots, which may lead to a selection bias in our subjects. Therefore, our conclusions are limited to Twitter users willing to follow the bots and do not necessarily generalize to all Twitter users. However, since a follow-back is required for a Twitter account to implement a tweet based treatment, this is not a strong limitation of our conclusions. This experiment was approved by the Institutional Review Board (IRB) for the authors' institution and performed in accordance with relevant guidelines and regulations.

3.3 Results

We used the frequency of extreme anti-immigration language in the subjects' tweets to measure any persuasion effect the bots had. In particular, we counted how many of the subjects' tweets contained the word "illegals" in each phase. The term illegals is a pejorative term used by people with anti-immigration sentiment. For instance, there are tweets such as *I want a refund on all the tax money spent on illegals!!!* which show strong anti-immigration sentiment. The usage frequency of such extreme language can be used to gauge sentiment, as was done in (Munger, 2017). We chose the word illegals because it is consistently used by anti-immigration Twitter users, unlike hashtags that gain temporary popularity. We plot the illegals usage frequency in each phase and treatment group in Figure 3-2. This frequency is defined as the

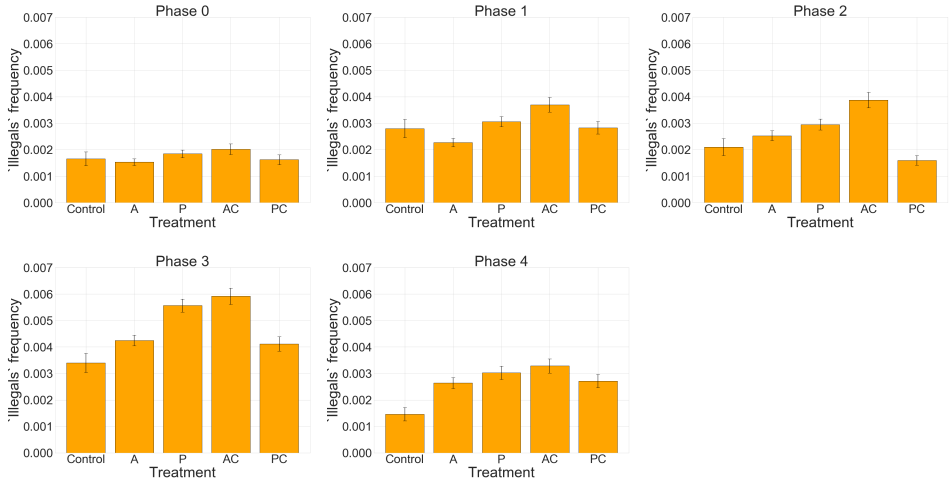


Figure 3-2: Plot of the frequency and standard error of usage of the word "illegals" in tweets for each phase and treatment group. The treatments are labeled as follows: A is argue without contact, AC is argue with contact, P is pace and lead without contact, and PC is pace and lead with contact.

number of tweets containing illegals divided by the total number of tweets for all subjects in each phase and treatment group. We note that the overall frequency is very low, but shows aggregate differences between phases. For instance, phase three has a higher frequency than the other phases for all treatments. This suggests that there are exogenous factors affecting the behavior of the subjects. Another interesting observation is in phase two, where we see that the pace and lead with contact treatment has a much lower frequency than the other treatments, while argue with contact has the highest frequency. Recall that in phase two pacing and leading has tweets that are slightly pro-immigration. We next perform a more quantitative statistical analysis to assess the different treatments.

We treat each tweet as a binary outcome that equals one if the tweet contains the word illegals. The probability of such an outcome is modeled using logistic regression. For a tweet i the probability is

$$\log\left(\frac{p_i}{1-p_i}\right) = \sum_{t=0}^{4} \beta_t x_{t,i} + \sum_{t=0}^{4} \beta_{a,t} x_{a,i} + \sum_{t=0}^{4} \beta_{p,t} x_{p,i} + \sum_{t=0}^{4} \beta_{ac,t} x_{ac,i} + \sum_{t=0}^{4} \beta_{pc,t} x_{pc,i} + \epsilon_i.$$

The coefficients β_t for $t = 0, 1, ..4$ model exogenous factors that may impact the probability during each phase. For instance, news stories related to immigration may increase the probability. We use separate treatment coefficients for each phase because the pace and lead treatment varies by phase. Recall that this treatment shifts the opinion of its tweets from anti- to pro-immigration over phases one to three. The treatment coefficients are indexed by subscripts indicating the treatment and phase. We use the subscript t for the phase, a for argue, and p for pace and lead. The subscript c indicates the contact treatment where the bots like the subjects' tweets. The x variables are binary indicators for the treatment group of the subject posting the tweet and in which phase the tweet occurred. User heterogeneity and other unobserved factors are modeled using a zero mean normally distributed random effect ϵ_i.

By regressing out the phase effect we can isolate the different treatments. We plot the resulting treatment coefficients separated by tweet group (argue or pace and lead) and contact group in Figure 3-3. This grouping makes differences in each individual treatment over the phases more visible. We also indicate on the plots which differences are statistically significant at a 1% level.

We first look at the effect of the contact treatment. In the top left plot of Figure 3-3 we see that the argue with contact coefficient is greater than argue without contact, and the difference does not vary much over the phases. The difference is significant for phases one, two, and three. In phases zero and four, where the bots do not tweet about immigration, there is no significant difference. The contact treatment may be making the bots' pro-immigration tweets more visible to the subject, resulting in a backfire effect where the subject uses the word illegals more frequently.

For pacing and leading in the top right plot of Figure 3-3, we see that the non-contact coefficient is greater than contact. In phases two and three the difference is significant. Contact appears to enhance the effectiveness of pro-immigration tweets in the later stages of the pacing and leading treatment. This is in contrast to arguing, where contact degrades the effectiveness of pro-immigration tweets.

We next look more closely at arguing versus pacing and leading when the contact treatment is fixed. In the bottom left plot of Figure 3-3 we see that without contact, the tweet treatment coefficients have a small difference which does vary appreciably across phases. Argue has a smaller coefficient, but the difference is statistically significant only for phases one and three.

For the contact group in the bottom right plot of Figure 3-3, the difference changes sign. Argue has the larger coefficient and the difference varies across the phases. Phase two shows a large significant difference. The difference is smaller in phase three, but still significant. The moderately pro-immigration tweets of the phase two pace and lead treatment seem to be more effective than the argue tweets when the bot has contact with the subject. The same can be said of fully pro-immigration tweets in phase three, but the advantage of pacing and leading over arguing is less than in phase two.

3.4 Discussion

Our results show that when the bots make contact with the subjects, pacing and leading was more effective than arguing in phases two and three. If the bots were arguing, then contact had the opposite effect and made the treatment less effective. We see a novel interaction effect, where combining pacing and leading with contact is the most effective treatment, especially in phase two.

Our findings suggest strategies one can use to overcome the challenges posed by echo-chambers and the backfire effect. We were able to penetrate echo-chambers by using bots which followed and liked the posts of the users. Penetrating an echo-chamber allows one to present arguments to the user. To overcome the backfire effect, we found that the bots should continuously like the posts of the users, and present arguments that are more nuanced and moderate in their language (phase two of the pacing and leading treatment). This softer approach proved more effective than standard arguing.

There are several interesting questions raised by our findings. One question con-

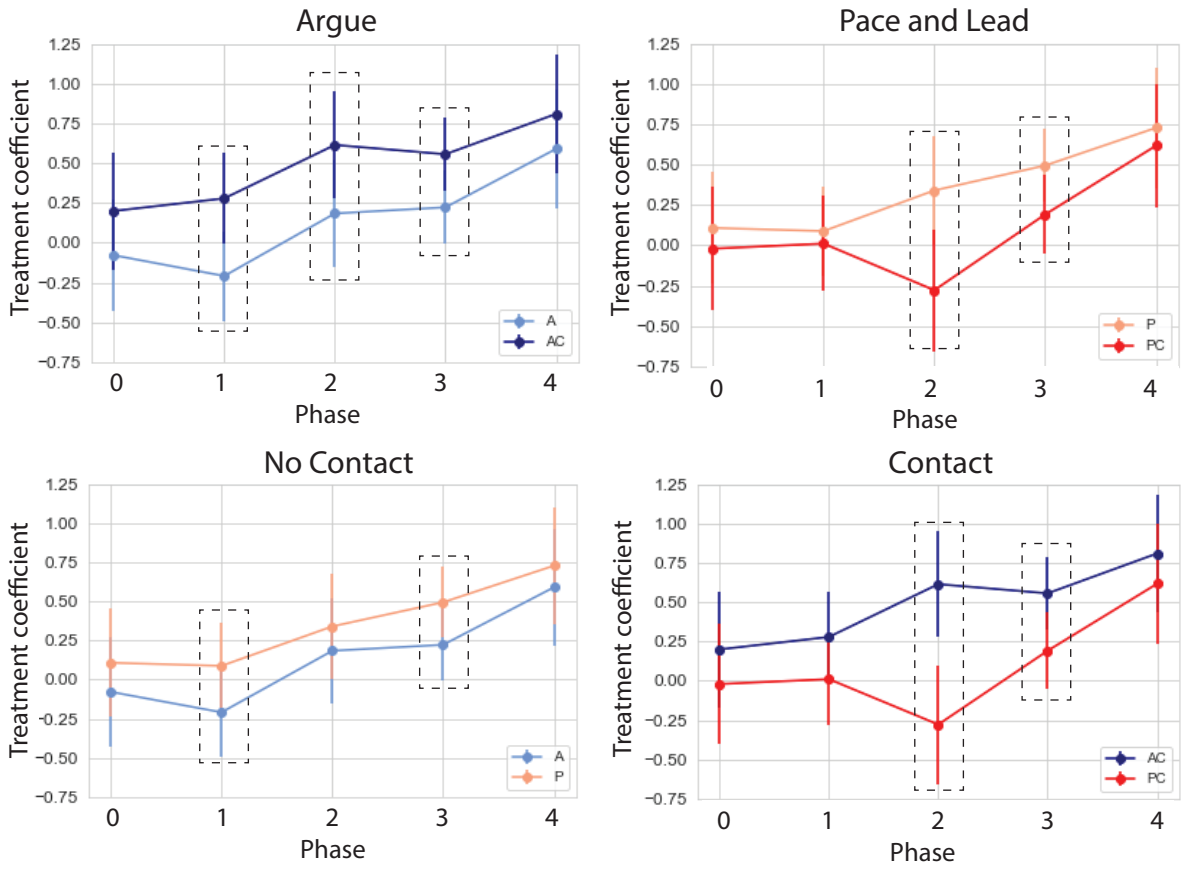


Figure 3-3: Plots of the regression coefficients (with standard errors) for the treatments in each phase. The title of each plot indicates the treatment component that is held fixed. The treatments are labeled as follows: A is argue without contact, AC is argue with contact, P is pace and lead without contact, and PC is pace and lead with contact. The dashed boxes indicate which coefficients have a difference that is statistically significant at a 1% level.

cerns the phases for pacing and leading. We found that the moderate posts were most effective. It is not clear if this treatment would work in isolation or if the phase one pacing treatment is necessary. We hypothesize that this period allows greater trust to be built between subject and bot, but our experiment does not confirm this. Another question is whether the phase three pace and lead treatment where the posts strongly advocate the target position is necessary. It may be that the moderate posts are sufficient to mitigate the backfire effect and potentially persuade the subject.

Finally, we note that care should be taken when trying to apply our results to more general settings. This study focused on the topic of immigration, which is an important political and policy issue. Discussion on this topic has split along traditional conservative liberal fault lines. We expect our findings to extend to similar political issues, but further study is needed. However, our subjects were Twitter users with anti-immigration sentiment who were willing to follow our bots. This represents a limited population in a very specific social setting. More work is needed to determine whether our findings replicate in different populations or within varied social settings.

Chapter 4

Inferring Political Bias in a Hyper-partisan Media Ecosystem

Many Republican politicians and voters have accused technology companies of anti-conservative bias when deciding what users to suspend. However, many studies find that conservatives are more likely to share misinformation on social media. Thus, it is possible that preferential suspension of conservatives may simply be the result of platforms attempting to reduce the spread of misinformation - a goal that has strong bi-partisan support. Here, we evaluate this possibility empirically by examining the suspension of Twitter users. We identified a set of 9,000 politically-engaged Twitter users, half Democratic and half Republican, in October 2020, and determined how many were suspended in the six months following the U.S. 2020 election. While only 4.9% of Democrats were suspended, 19.1% of Republicans were suspended. Yet the Republicans shared much more information from unreliable news sites than the Democrats, and suspension was predicted as well by the tendency to share information from unreliable sites - as judged either by professional fact-checkers or a politically-balanced crowd of laypeople - as it was predicted by partisanship. Thus, the observation that Republicans were more likely to be suspended than Democrats provides no evidence that Twitter was biased against conservatives. Instead, this asymmetry can be explained entirely by the tendency of the Republicans to share more misinformation.

4.1 Introduction

It has been widely claimed by those on the political right that social media platforms are biased against conservatives. Many conservative activists, including former President Donald Trump, have filed lawsuits against social media companies, accusing them of censoring users with an anti-conservative bias (Bond, 2021). This view is shared by many Republican voters - for example, in an August 2020 poll, roughly seven-in-ten Republicans said major social media sites tend to favor liberals over conservatives (Vogels, Perrin, and Anderson, 2020). Yet these charges are based on anecdotal instances of particular platform actions (e.g. Twitter's permanent suspension of Trump's account), rather than any systematic comparison of enforcement on conservatives versus liberals.

Even more importantly, charges of anti-conservative bias must contend with alternative explanations for the preferential suspension of conservatives - most notably, attempts by platforms to combat misinformation. There is widespread bi-partisan concern about misinformation and "fake news" on social media, and widespread bi-partisan agreement that technology companies should take action against such content. For example, a July 2020 poll found that a majority of supporters of both parties believed that technology companies are responsible for preventing misuse of their platforms aimed at influencing elections (van Green, 2020), and a November 2020 poll found that a majority of supporters of both parties believed that social media companies should be held responsible for false or inaccurate content posted on their platforms (Koopman, 2021).

However, while the desire to combat misinformation is bi-partisan, there are substantial partisan asymmetries in the spreading of misinformation. Numerous studies of the 2016 election, for example, have found that conservatives shared dramatically more fake news on social media than liberals (Guess, Nagler, and Tucker, 2019, Grinberg et al., 2019). Thus, even if platforms are more likely to suspend conservatives than liberals, this asymmetry in suspension could simply arise from politically neutral enforcement aimed at satisfying the bi-partisan demand for a reduction in online

misinformation, rather than anti-conservative bias.

4.2 Data Collection

In this study, we evaluate these issues empirically by examining the suspension of users on Twitter. We used the same user set from Chapter 2. Again, we collected a list of Twitter users who tweeted or retweeted either of the election hashtags #Trump2020 and #VoteBidenHarris2020 on October 6, 2020. We also collected the most recent 3,200 tweets sent by each of those accounts. We processed tweets and extracted tweeted domains from 34,920 randomly selected users (15,714 shared #Trump2020 and 19,206 shared #VoteBidenHarris2020), and filtered down to 12,238 users who shared at least 5 links to domains used by the ideology classifier of (Eady et al., 2019b). We also excluded 426 'elite' users with more than 15,000 followers who are likely unrepresentative of Twitter users more generally. We then constructed a politically balanced set by randomly selecting 4,500 users each from the remaining 4,756 users who shared #Trump2020 and 7,056 users who shared #VoteBidenHarris2020. We also counted which hashtags users have tweeted more often by collecting all of the hashtags they have tweeted in their most recent 3,200 tweets, and 94% of the 9,000 users have matched ideology. After nine months, on July 30, 2021, we checked the status of the 9,000 users and assessed suspension. To evaluate evidence for anti-conservative bias, we ask how well suspension probability can be predicted by users' political partisanship versus their tendency to share misinformation.

4.3 Results

We found dramatic differences in suspension rates between Republican and Democratic users. Accounts that had shared the #Trump2020 hashtag during the election were 3.9 times more likely to have been subsequently suspended than those that shared the #VoteBidenHarris2020 hashtag: While only 4.9% of the Democratic users had been suspended as of July 2021, 19.1% of the Republican users had been suspended

(Figure 1c).

On first inspection, this seems to indicate that Twitter was exhibiting strong anti-conservative bias in its suspension actions. However, consistent with past work (Guess, Nagler, and Tucker, 2019, Grinberg et al., 2019), Republican users in our dataset were much more likely to share news from untrustworthy news sites than Democratic users. To quantify the quality of news shared by each user, we leveraged a previously published set of 60 news sites (20 mainstream, 20 hyperpartisan, 20 fake news, with liberal and conservative leaning sites in each category) whose trustworthiness had been rated by eight professional fact-checkers (Pennycook and Rand, 2019). We followed the approach used in prior work (Pennycook et al., 2021, Mosleh et al., 2021a) and calculated a quality score for each user by averaging the trustworthiness ratings of any of their last 3,200 tweets as of October 2020 that contained links to any of those 60 sites. (Like most other researchers in this space (Guess, Nagler, and Tucker, 2019, Grinberg et al., 2019, Pennycook et al., 2021, Mosleh et al., 2021a), we use source trustworthiness as a proxy for article accuracy, because it is not feasible to rate the accuracy of every shared link.) The average quality of news shared in our dataset was much lower for Republican users compared to Democratic users ($t(8943)=119.75$, $p < .001$; $r(8943)=-0.78$, $p < .001$; see Figure 1a). And when considering suspension likelihood, the pattern mirrors that found for partisanship: Users who shared more news from untrustworthy news sites were much more likely to get suspended. For example, while only 4.7% of users with higher news sharing quality scores were suspended, 19.5% of users with lower news sharing quality scores were suspended (median split, Figure 1d).

We then ask how well a user's probability of being suspended can be predicted using their partisanship versus information sharing quality. To do so, we use the area-under-the-curve (AUC), which measures accuracy while accounting for differences in base rates and is a standard measure of model performance in fields such as machine learning. We find that the AUC when predicting suspension using partisanship is 0.67; using other classification approaches that produce continuous ideology ratings (Barberá et al., 2015, Grinberg et al., 2019, Eady et al., 2019b) produces similar results

(AUC are between 0.70 and 0.71). Critically, however, the AUC when predicting suspension using the fact-checker-based news quality ratings was just as high (AUC = 0.70). Thus, the preferential suspension of Republicans can also be explained by preferential suspension of users who shared information for untrustworthy news sites. Policies aimed at fighting misinformation in a nonpartisan way could have easily led to the observed difference in suspension of Republicans versus Democrats.

Some might argue that these findings are the result of professional fact checkers having a liberal bias. To address this potential concern, we next run the same analysis but instead of ratings of professional fact-checkers, we use trustworthiness ratings generated by politically-balanced crowds of demographically representative (quota-sampled) American laypeople recruited via Lucid (Barberá et al., 2015). We continue to find that the average quality of news sites - as assessed by politically-balanced layperson crowds - was significantly lower for Republican users compared to Democratic users ($t(8943)=102.01$, $p < .001$; $r(8943)=-0.73$, $p < .001$; Figure 1b). Replicating the fact-checker results, only 5.0% of users with higher crowd-based news sharing quality scores were suspended while 19.1% of users with lower crowd-based news sharing quality scores were suspended (median split, Figure 1e), and crowd-based news sharing quality was highly predictive of suspension (AUC = 0.69; Figure 1e). Thus, our findings cannot be attributed to liberal bias among professional fact-checkers.

Next, we note that these results are specific to measures of the quality of information shared, rather than other features of sharing. For example, suspension was not well predicted by the toxicity of language in users' posts (Per) (AUC = 0.56), their use of hate speech (AUC = 0.48) (Davidson et al., 2017), or their use of offensive language (AUC = 0.51) (Davidson et al., 2017).

Additionally, Twitter has anounced in January that they suspended more than 70,000 accounts linked to the far-right movement QAnon (Catherine Herridge). We also measured how many of our users' suspension was due to QAnon. We counted the occurance of all QAnon related words including '#qanon', '#wwg', '#wga', '#thegreatawakening', '#q', '#qarmy','#wwg1wga','#trusttheplan' and computed

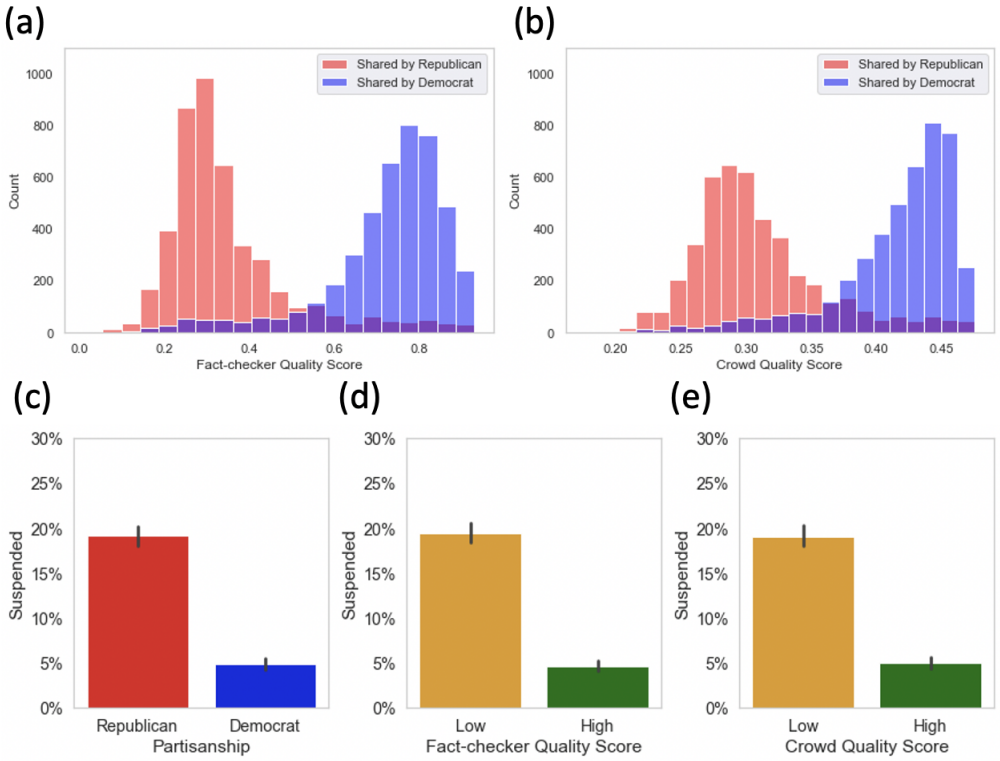


Figure 4-1: Top row: Distribution of news quality scores for links shared by Democrats versus Republicans, based on (A) professional fact-checker trustworthiness ratings and (B) politically-balanced layperson crowd trustworthiness. Bottom row: Percent of users suspended based on (C) partisanship, (D) median split of professional fact-checker trustworthiness ratings of shared news, and (E) median split of politically-balanced layperson crowd trustworthiness ratings of shared news. Error bars indicate 95% confidence intervals.

Features	AUC
hate speech	0.477268
offensive language	0.515389
toxic score	0.564129
q_binary	0.605647
q_freq	0.627015
q_log	0.628995
ideology-hashtag	0.668081
crowd source	0.685703
fact checker	0.698439
ideology-media score	0.700531
fact checker grinberg	0.704819
ideology-follower	0.715642

Table 4.1: AUC when predicting suspension using different features, ranked from low to high. Hate speech, offensive language, and toxicity of language in users' posts have low AUCs, whereas quality scores as well as users' ideology have high AUCs.

three QAnon features: the frequency of tweets with QAnaon related hashtags, the log of the number of QAnon tweets, as well as the binary indicator of whether the user has tweeted QAnon hashtags. 2,972 users have at least one tweet with QAnon hashtags. The suspension rate for these users is 18.5%, similar to the overall suspension rate of Republicans. The AUC of the three QAnon features are between 0.6 and 0.62. In short, it seems that suspenion cannot be well predicted by QAnon usage for our user set.

4.4 Discussion

Consistent with anecdotes cited as evidence for political bias on the part of technology companies, we found that Republican accounts were much more likely to have been suspended by Twitter following the 2020 election. However, the tendency to share links from misinformation sites pre-election was just as predictive of post-election suspension as was partisanship. This includes information quality measures generated by politically balanced crowds of laypeople, who - by virtue of being politically balanced - inherently cannot be accused of having an anti-conservative bias.

Thus, the fact that conservatives were much more likely to get suspended does

not provide evidence that conservatives were specifically targeted by Twitter due to their ideology. In the current hyperpartisan media ecosystem, partisanship is strongly confounded with the quality of information users share. Thus, Twitter's preferential suspension of Republicans could be entirely explained by a good-faith, politically-neutral effort to enact the bi-partisan desire for reduced misinformation online.

Chapter 5

Online Communication Shifts in the Midst of the COVID-19 Pandemic

The COVID-19 pandemic has created large shifts in how people stay connected with each other in lieu of social distancing and isolation measures. More and more, individuals have turned to online communications as a necessary replacement for in-person interaction. Despite this, the research community has little understanding of how online communications have been influenced by the offline impacts of COVID-19. Our work touches upon this topic. Specifically, we study research questions around the impact of COVID-19 on online public and private sharing propensity, its influence on online communication homophily, and correlations between online communication and offline case severity in the United States. To do so, we study the usage patterns of 79 million US-based users on Snapchat, a large, leading mobile multimedia-driven social sharing platform. Our findings suggest that COVID-19 has increased propensity to privately communicate with friends, while decreasing propensity to publicly share content when users are out-and-about. Moreover, online communications have observed a marked decrease in baseline homophily across locations, ages and genders, with relative increases in cross-group communications. Finally, we observe that increased offline positive COVID-19 case severity in US states is associated with widening gaps between across-state and within-state communication increases after the onset of COVID-19, as well as marked declines in public sharing.

5.1 Introduction

The COVID-19 pandemic has created seismic shifts in people's lives, with profound economic, social, family, work, and school disruptions. To flatten the curve of COVID-19 cases and mitigate negative social and economic impacts, governments have put in place numerous restrictions on constituents regarding limits to in-person interaction, enforced self-quarantine, and social distancing measures. With these practices in place, friends, families, and colleagues have been forced to suddenly adopt or augment new or existing communication modalities, with most work, school and social communications happening exclusively online in many parts of the world to this day (Koeze and Popper, 2020).

Despite the massive and wide-reaching shift in interpersonal interactions being moved to online ones, we as a community still have little knowledge about how people's online communication habits via social platforms have changed as a result of a sudden and critical externality, like COVID-19. For example, who are people talking to, and where? How has the severity of the pandemic influenced the interaction behavior? One might expect that the manner, content, and intent of communications might have changed substantially as well, due to concern for friends and family, reaching out in periods of isolation, dealing with mental health struggles, and more. Research on communication behaviors during pandemics to date, including early work on COVID-19, mainly focus on risk assessment and misinformation (Strekalova, 2017, Larson, 2018, Bursztyn et al., 2020); several prior works also studied communication on social media during natural disasters (Metaxa-Kakavouli, Maas, and Aldrich, 2018, Palen and Anderson, 2016a) like hurricanes and tropical storms, with attention to the utility of social platforms for information spreading and relief effort organization in the extreme short-term. However, none of the prior works touch upon interpersonal online communication shifts due to such a jarring externality.

Given the size and scope of the impacted population, building understanding of this topic is crucial: many individuals around the world have felt the apparent impacts of in-person restrictions and their implications for human interaction. How-

ever, there are numerous challenges in facilitating study of pandemic impact on online communications, making it challenging or impossible to study in the past: Firstly, pandemics are naturally rare events, severely limiting the timeframe of their investigation and would-be investigators. Secondly, previous pandemics have not occurred in the modern heyday of online and social platform-driven communications. Thirdly, investigation of online behaviors is challenging without appropriate data access and large enough scope of study.

In this work, we take advantage of a confluence of factors which allow us to overcome these issues, and enable our investigation of COVID-19's impact on private and public online communication patterns in the United States via Snapchat. Snapchat is a highly popular multimedia ephemeral messaging platform which launched in 2011, and has 238 million daily active users across the world (Snap Inc., 2020). It offers functionality for both public and private sharing of *Snaps* (short images or videos), through *Direct Snaps* (private, one-to-one communications) and *Story Snaps* (broader audience, either to all a users' friends or to all users on the platform). In our study, we examine online communication habits as proxied by both modalities.

Our work aims to address three key research questions to better understand shifts in online communication patterns before and after the onset of COVID-19:

1. **How has COVID-19 impacted online private and public sharing propensity?** We find that post-onset COVID-19 engagement is higher for online direct communication and lower for online geo-based public communications. Temporal analysis shows the change is sudden for most states in the US, and the difference is statistically significant for all states.

2. **How has COVID-19 influenced homophily in online communications between users?** We find that social distancing measures have reduced effects of homophily and induce increases in inter-state/gender/age online geolocation-based public communications.

3. **Are changes in online communication patterns correlated to the severity of offline COVID-19 impact?** We find that the number of COVID-19

cases in different states is not correlated to increases in communication frequency, but is positively correlated to differences in within-state and across-state communication metrics.

We took a quantitative approach to study these questions. To this end, we analyzed the communication patterns of over 79 million US-based Snapchat users from February to May 2020. In the remainder of the paper, we investigate these questions and provide detailed answers and discuss implications based on our analyses. We hope our work helps elucidate how COVID-19 has influenced changes in online communications as a function of in-person restrictions, and can help inform design improvements for social platforms as a result.

5.2 Background and Related Work

We discuss prior work in four areas: social media's role in times of crisis, communication during COVID-19, homophily in social media, and background about the Snapchat platform and prior work on Snapchat. Our work interfaces with each of these aspects.

5.2.1 Social Media's Role in Times of Crisis

Due to its unpredictable and negative nature, an ongoing crisis often produces a high amount of uncertainty and anxiety among the public during a short period, potentially resulting in large scale damages (Coombs, 2014). Prior work suggests that during these crises, social media usage increases (Ulvi et al., 2019), and the interplay between social media and crises has led to numerous prior works on the broad theme of crisis informatics (Palen and Anderson, 2016b). Several works (Ulvi et al., 2019, Goolsby, 2010, Hiltz, Diaz, and Mark, 2011) investigate the role of social media in information dissemination, coordination and public awareness about crises and natural disasters. Imran et al. (2015) also surveys mechanisms for information extraction and distillation from social media in times of crisis. Other works touch

on cultural comparisons of communication in critical times: Ding and Zhang (2010) studies compares institutional communications during the 2009-2010 H1N1 flu outbreak, while Welhausen (2015) discusses intercultural risk communications through data visualization during the 2014 Ebola outbreak in West Africa. Metaxa-Kakavouli, Maas, and Aldrich (2018) finds that online social ties play a critical and previously underestimated role in natural disaster preparedness. Higher levels of bridging and linking social ties correlate strongly with evacuation propensity. While these prior works mainly focus on public or health authority responses, and information distillation and distribution via social media in times of crisis, none study the fundamental, characteristic changes in underlying communication patterns on social media brought about by large-scale social distancing and new communication norms, as our work aims to.

5.2.2 Communication during COVID-19

With the ongoing COVID-19 pandemic, people have increased their social media usage to seek information about the pandemic according to surveys (Wiederhold, 2020). The widespread effects of COVID-19 have led to several recent initiatives in studying its interplay with social media use: Kim (2020) collected comments from Korean social media to analyze negative emotions and societal problems during COVID-19. Lin, Liu, and Chiu (2020) used Google Keyword Search frequency to predict speed of COVID-19 spread in 21 countries/regions. Singh et al. (2020) characterizes Twitter conversation around COVID-19, and indicates that online conversation about the virus leads new cases geographically. Several prior works have also studied misinformation around COVID-19: Depoux et al. (2020) remarks upon the rapidity of the panic and spread of misinformation. Huynh and others (2020) studies how COVID-19's risk perception in Vietnam is heavily mediated by baseline and geographical social media use. Pennycook et al. (2020) found that many people disseminated false information related to the virus because they failed to reason appropriately if content was true or false before sharing, and that propensity to share was misaligned with people's ability to judge accuracy. Brennen et al. (2020) indexes many common false

claims about COVID-19 circulating on social media, and notes that the majority are misinformative (improper context, misleading) rather than disinformative (fabricated or imposter content).

5.2.3 Homophily in Social Media

Homophily is the principle that contact between similar people occurs at a higher rate than among dissimilar people. This principle has implications in information diffusion, grouping and community formation, online exposure and more (McPherson, Smith-Lovin, and Cook, 2001). Catanzaro, Caldarelli, and Pietronero (2004), Krivitsky et al. (2009), Shah (2020) study incorporation of homophily as a first-class citizen in network and graph modeling. Guacho et al. (2018), Akoglu, Chandy, and Faloutsos (2013), Pandit et al. (2007) exploit homophilic principles in networks to detect misbehaving users and misinformative articles. Recently, several works (Bessi et al., 2015, Gillani et al., 2018, Kumar and Shah, 2018) discuss homophily's role in echo-chamber formation, opinion polarization and misinformation spread in social media. Homophily can also lead to between-group segregation of interpersonal relations in teams (Lau and Murnighan, 1998), and can occur due to formative effects and preferential selection (Currarini, Jackson, and Pin, 2009). While our work does not directly model homophily, it empirically studies variation in homophilic effects pre and post onset of COVID-19.

5.2.4 The Snapchat Platform

Snapchat is a popular, mobile multimedia-driven social messaging platform, introduced in September 2011. As of July 2020, Snapchat has roughly 238 million daily active users and enjoys widespread use (Snap Inc., 2020). Snapchat enables users to create short image or video snippets, called *Snaps* which can be both narrowcast (directly shared privately with friends) as *Direct Snaps*, or broadcast (made publicly visible either to all friends or all other users on the platform) as *Story Snaps*. Juhász and Hochmair (2018) found that Snapchat users are more likely to share Snaps to

everyone on the platform that are taken in highly trafficked areas, such as tourist hotspots or urban centers. Snaps can be further modified with geolocation-based filters, augmented reality (AR) lenses, stickers and more, adding metadata and context to the content (Verstraete, 2016, Rios, Ketterer, and Wohn, 2018). Snaps are ephemeral: Direct Snaps persist only until the recipient views them, after which they are deleted. Story Snaps are appended to a user's *Story* timeline, and automatically deleted 24 hours after posting. Several prior works study these features and associated user engagement: Bayer et al. (2016) notes that Snapchat users associate Snapchat communications with increased trust in the audience, and reduced self-curation due to ephemerality; this is unlike other platforms where content is pervasive and retained indefinitely, and promotes full curation of a singular external online profile (Uski and Lampinen, 2016). Katz and Crocker (2015), Habib, Shah, and Vaish (2019), Juhász and Hochmair (2018) note that users' individual sharing decisions on Snapchat in private versus public spheres are influenced by various contextual factors associated with identity, activity, location and time of sharing. Lamba and Shah (2019), Kaghazgaran et al. (2020) characterize and statistically model consumption behaviors of Direct Snaps and Story Snaps, respectively. Several works (Tang et al., 2020, Saha et al., 2021) also propose approaches to model ad response and user churn phenomena on Snapchat.

5.3 Data

We utilized rich engagement data spanning the time between February 15 to May 13 from Snapchat. We designate March 11 as the date threshold for partitioning our study period into pre and post COVID-19 timeframes, given that March 11th was the day on which the World Health Organization (WHO) declared the outbreak to be a pandemic[1]. Notably, March 11th also corresponds to the timeframe that many states started adopting stay-at-home orders. We do not consider dates prior to Feb 13th or post May 13th due to access limits (limited availability prior, and limited

[1]https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen

access post). We gathered user engagement metrics for 79 million Snapchat users.

We focused the study population to those whose primary/home location is in the United States, and have stayed in their home location before the pandemic (as observable from Snapchat usage logs). To achieve this goal, we first filtered our candidate population to users who were active around January 15, on which day the Center for Disease Control (CDC) reported the first COVID-19 case in the United States. We further selected users whose geolocation was in the U.S. (considering 51 "states," including the 50 conventional states, and DC) for all of our analyses. Lastly, we limited selection to users who had consistently reported locations in the same state in the past 1, 7, 30, and 90 days, thereby removing effects from "visitors" who were only active in the United States for a short period of time but active in another country for the majority. As a result, we obtained a representative set of residents for each state. Importantly, we held this population consistent across pre and post COVID-19 periods to limit exogenous effects from different samples (new or resurrected users).

Additionally, for Direct Snaps, we specifically gathered de-identified sender and recipient user ID, to evaluate the flow of conversations. Note that one Snap could be (privately) sent to multiple recipients, and create several parallel information flows. We also collect user attributes like (self-reported) gender and age for both users in the pair to evaluate homophilic tendencies, and evaluate location based on IP address. For Story Snaps, we specifically gathered de-identified poster user IDs, and attributes indicating whether the Story Snap met our location-based criteria.

Privacy, Ethics and Disclosure.

Our work uses sensitive data from Snapchat. It is conducted within Snapchat, and reflects our commitment to user privacy. Our analysis relies on de-identified data, and throughout the work, we discuss only aggregated metrics across a large cohort of users.

5.4 Methods

To quantify the impact of COVID-19 on online engagement habits, we conducted several empirical comparisons of engagement before and after the onset of COVID-19; we call these "pre COVID-19" and "post COVID-19," informally.

To answer RQ1, we analyzed several key daily metrics on public and private sharing to test if social distancing affected private and public communications differently on Snapchat. For public communications, we considered metrics based on Story Snaps (SS), due to their broadcast functionality:

1. **Total SS from the state.** We considered SS from senders whose home location was in the state.

2. **SS per poster.** We considered the ratio of total SS posted to posters to control for fluctuations in the number of posters.

Notably, we limited our focus to those SS which utilize geolocation-based overlay filters, AR lenses and other modifiers indicating that the user is out-and-about, and publicly sharing at a location of interest, further contrasting with private sharing norms. This is consistent with (Juhász and Hochmair, 2018)'s findings on sharing while in public being characteristically different from that in private. Enforcing the above condition allows us to focus on the set of SS which are not taken when the user is "at home." For private sharing, we considered metrics based on Direct Snaps (DS), due to their one-to-one narrowcast functionality:

1. **Total DS from the state.** We considered all DS from senders whose home location was in the state. This was the most common and straightforward way to track online engagement.

2. **DS per sender.** We divided total DS by number of users to derive each user's DS change on average.

3. **Recipients per DS sender.** We considered the number of recipient users engaged with per DS senders. Number of recipients demonstrated the size of social network.

4. **DS tie strength.** We define DS tie strength as the ratio between DS sent and the number of recipients. Higher DS tie strength indicates indicates tighter, or more concentrated engagement. Note that we define this notion as a node-level descriptor, rather than an edge-level one.

The above metrics are key indicators for user engagement, reflecting both raw and normalized quantities, as well as engagement concentration. We calculated these metrics on a daily basis, and then aggregated them based on pre and post COVID-19 timeframes. We used two-sample t-tests to determine whether measured quantities in the pre and post COVID-19 timeframes differed, and if those differences were statistically significant. We choose t-tests assuming normality in the means of the pre and post daily metric quantities, and roughly equal variances. We further corrected the p-values with the Benjamini-Hochberg (BH) procedure for controlling False Discovery Rate (FDR) in multiple testing (Benjamini and Hochberg, 1995). Moreover, to analyze the temporality of these changes, we also conducted change-point detection on the time series to test if engagement changed abruptly or gradually. For the purpose, we used the Pruned Exact Linear Time (PELT) search method (Killick, Fearnhead, and Eckley, 2012) to determine the existence and location of the change point.

To answer RQ2, we gathered user demographic descriptors like location, gender and age for both source and destination users across many communication pairs, to gauge shifts in the pre and post COVID-19 settings. We hypothesized that in-person distancing measures would qualitatively impact the types of online communications rather than just the quantities, and use distributions of these user properties across pairs to evaluate the variation in homophily (communications between alike-users) in the pre and post periods. Specifically, for each of the primary factors we studied, including location (within state versus across state), age (same age group versus different age group), and gender (same gender and other gender), we used two-sample t-tests on the daily metric quantities to measure the difference and significance of the pre and post COVID-19 periods, and draw conclusions reflecting whether lockdowns encourage users to interact more with others that are similar or dissimilar.

For RQ3, we aimed to evaluate the relationship between offline severity of COVID-

19, and online communication differences in public and private settings. Specifically, we collected statistics on COVID-19 cases in different states from Miller K (2020), and used linear regression to evaluate the correlations between the two.

5.5 Results

Below, we discuss our findings for each of the 3 RQs.

5.5.1 Public and Private Sharing Propensity (RQ1)

First, we analyzed how private and public sharing propensities shifted with the onset of COVID-19 and associated distancing measures.

Private Sharing.

We compared pre and post COVID-19 user engagement in several metrics based on Direct Snaps (DS), which has the one-to-one narrowcast functionality as discussed in details in *Methods*.

First, we calculated the mean of total DS for each state in the US, post COVID-19, as shown in Figure 5-1. The figure clearly shows that post COVID-19 private (DS) engagement is substantially higher for all states, ranging from a 9.7 to 25.5 percentage increase state-wise. Two-sample t-tests for each state also demonstrate $p < .05$ after BH correction, further confirming significant inequalities in the sample means in the pre and post COVID-19 periods. Recall that since our pre and post metrics are evaluated over a fixed user population, the normalized quantity (DS per sender) is also significant across pre and post periods. The geography of these changes is shown in Figure 5-2a.

Moreover, DS per sender and recipient per sender also have a significant increase after the onset of COVID-19. Post COVID-19 means are 8.1 to 24 percent higher for DS per sender, and 1.5 to 6.0 percent higher for recipient per sender, with more geographical detail in Figure 5-2b-5-2c. Two-sample t-tests for each state also demon-

strate $p < .05$ after BH correction, which suggests that on average, the private communication volume and the social communication network size increases for each user.

To further investigate if the increase in DS is attributed to all friends, or just top contacted friend, we considered tie strength (DS per recipient). We found that post COVID-19 means are 5.3 to 18.6 percent higher, and the increment is significant for all states, with $p < .05$ after BH correction. Figure 5-2d illustrates the geographical change on the map. This result demonstrated that on average, users are deepening their friendships with all friends. In short, in-person distancing measures led to substantial online private communication increases.

Geographical snapshots in Figure 5-2 show similarities across Total DS, DS per sender, and tie strength in highest increment states, with KS, CA and NM consistently showing highest % changes in these quantities, and ME with consistently low % changes. We observe some differing trends in Recipients per sender, which conveys more about communication breadth rather than depth like the other metrics; here, HI has the highest increase, perhaps owing to its disconnected status from the mainland.

Lastly, we performed change-point detection on each metric over the joint pre and post time period to test if engagement changed abruptly or gradually, as shown in Figure 5-3. We found that overwhelmingly, 49 states experienced a sudden surge in total DS and DS per sender, 47 in recipients per DS sender, as well as 48 in tie strength.

Public Sharing.

We also compared pre and post COVID-19 user engagement as measured by location-based Story Snaps (SS). Figure 5-4 shows the relative percentage change in total SS for each state in the US, post COVID-19. Clearly, these metrics drop consistently across states, ranging from a -78.98 to -35.31 percentage decrease, indicating the limited mobility of users and desire to share content publicly due to distancing and isolation measures. Two-sample t-tests for each state again confirm the significance of these effects, with all $p < .05$ after BH correction. These effects were also apparent in normalized per-poster quantities (not shown due to space constraints). Moreover,

a sudden change point was found in all 51 states: 45 states on March 16, and 6 states on March 21, indicating an abrupt and significant change in the public engagement.

5.5.2 Variation in Homophilic Tendencies (RQ2)

Next, we consider how homophilic communication tendencies (baseline rates of within-group communications) shifted post COVID-19. We analyzed communication pattern changes in three aspects: location (within state vs. across states), age (within age-group vs. across age-groups) and gender (within gender vs. across genders). We use variations in the DS tie strength to reason about change in homophily, by comparing changes in the difference in within-group and across-group DS tie strengths (we offer further discussion on the use of this metric below, for *Location*).

Location.

First, we considered within-state and across-state private communications. Before considering the DS tie strengths, we considered the two metrics contributing to the ratio: private communication volume (DS sent) and social network size (recipients per sender). As previously discussed, Figure 5-1 shows that private (user-user) communication volume increased for all states, but does not indicate the manner of this increase.

Thus, we evaluated the total DS by state, pre and post COVID-19, broken down by across-state and within-state communications. Figure 5-5 shows the absolute increase in total private sharing (DS) of within-state (red) and across-state (blue). In general, private communication quantities (DS) increased both within-state (red) and across-state (blue), and for most states, raw within-state communication volume increases outpaced across-state volume increases, due to their large baseline propensity (the majority of communications pre COVID-19 were within-state). However, upon considering the relative social network sizes of within-state and across-state groups, we see a different portrayal of the effect: Figure 5-6 shows that the across-state recipients per sender (blue) increased substantially for all states, while the within-state

quantity (red) increased only for some states but decreased for others. So, while the volume of DS sent increased more within-state, the recipients per sender increased more across-state.

To study these contrasting indications more carefully, we consider the DS tie strength ratio, which more concisely summarizes the depth of the relationship that a sender has with a set of recipients. In this case, we considered both the across-state and within-state DS tie strengths (adjusting the numerator and denominator to account only for across-state and within-state interactions, respectively). Interestingly, 44 states have higher increase in across-state than within-state tie strength, of which 22 of the increases are significantly higher ($p < .05$) when evaluate with a two-sample t-test with BH correction. This suggests that while both within and across-state tie strengths increased post COVID-19 (suggesting deepening relationships), users actually deepened across-state relationships *moreso* than within-state ones. We conclude that although relationships within-state also deepened post COVID-19, larger deepening of across-state relationships suggests a relative reduction of location-based homophily from the pre-COVID baseline.

Age.

Next, we considered communications between users within and across age-groups. We consider users partitioned into 5 age groups: 13-17, 18-20, 21-24, 25-34 and 35 plus. We use the DS tie strength metric to measure the communication intensity between the groups.

Figure 5-7 shows absolute increases in DS tie strength post COVID-19 between associated sender and correspondent (receiver) users, computed by subtracting the pre COVID-19 from the post COVID-19 metric. Darker shades of blue denote larger increases. Note that conditioning on each sender age group (vertical), the darker cells are those corresponding to different age-grouped users. For example, considering 13-17 aged senders, the communication increase was 0.534 within age-group, compared to the much more substantial 2.975 increase to 25-34 correspondents. Likewise, the communication increase from 35 plus age-group senders was 0.072 within-group, com-

	13-17	18-20	21-24	25-34	35+
13-17		✓	✓	✓	✓
18-20				✓	✓
21-24	✓			✓	✓
25-34	✓		✓		✓
35+	✓	✓	✓	✓	

Table 5.1: Two-sample t-test significance results on the difference of difference in DS tie strength between "within age-group" and "across age-group" categories. Most results indicate across age-group increases are significantly different (✓ indicates $p < .05$) larger than within age-group ones.

pared to 0.482 to 13-17 correspondents. Note that these quantities are deltas in the normalized DS tie-strength ratio; a 1 unit change is extremely large, indicating that on average, users send 1 more DS to all of their friends. Two-sample t-tests with BH correction confirmed that tie strength means were significantly different across all age groups pre and post COVID-19 (i.e. quantities in all cells of Figure 5-7 are significant). Moreover, we also conducted two-sample t-tests with BH correction to evaluate the difference-in-difference measurements (Lechner and others, 2011). Specifically, we considered the difference across age groups, in difference in DS tie strength pre and post COVID-19 between one's own age-group and other age-groups (i.e. cells on the diagonal, compared to cells off the diagonal), to evaluate e.g. whether the increases in tie strength from 13-17 senders to 13-17 correspondents are indeed significantly different to 18-20 correspondents (in other words, to evaluate whether the across-group effect is larger than the within-group effect). Table 5.1 shows the significance results (✓ indicates $p < .05$) for these difference-in-difference experiments, clearly indicating the effect is present and observable for the majority of age-group pairs. These results altogether suggest a reduction of age-group homophily from the pre COVID-19 baseline.

Gender.

Thus far, we observed that the post COVID-19 period marks an observed reduction of homophily from both the location and age lens. We next consider whether this

effect holds across gender as well, using within and across-gender DS tie strengths.

We found that in general, users have deeper DS tie strengths with the opposite gender pre COVID-19 (i.e. males have larger tie strengths to females on Snapchat, and vice versa). Post COVID-19, all 4 groups (MM, MF, FM, FF) saw statistically significant ($p < .05$ after BH correction) increases in tie strength (i.e. quantities in all cells of Figure 5-8 are significant). Moreover, communication with the opposite gender increased even more than with the same genders for both female and male senders (comparing cells in the same vertical). We conducted two-sample t-tests to evaluate the difference-in-difference (difference across gender, in difference in DS tie strength pre and post COVID-19) between one's same gender and the opposite gender. We found that the increases in tie strength for MF is indeed significantly larger than that for MM, and likewise increase in FM is significantly larger than that of FF (both $p < .05$).

Specifically, MF increased significantly more than MM, and FM significantly more than FF, indicating a further reduction of gender-based homophily in an already heterophilic regime.

5.5.3 Correlation with COVID-19 Case Severity (RQ3)

Private Sharing.

Although COVID-19 is an international emergency, its impacts have been disparate across locations. In the US, while some states saw more severe outbreaks and announced early lockdown orders, some have not observed the same and are more "under control" from an offline (on-ground) standpoint. Those affected by more severe distancing measures may be communicating more or less online than others: thus, we study the relationship between offline severity and online impact.

We consider DS tie strength as the target metric for evaluating engagement depth. If the tie strength increases, it indicates that users send more DS to their social networks, and are thus communicating more closely with their friends. We use the positive case count on May 16, 2020 as a measure of offline/on-ground COVID-19

severity, i.e. higher positive case count implies higher severity. Technically, we use the logarithm of the measure (monotonically increasing with respect to the actual case count), due to its large scale.

Firstly, we evaluated whether the offline severity was correlated with increase or decrease in tie strength between pre and post COVID-19 periods. Figure 5-9 shows that two quantities are not significantly correlated, and that tie strength increases in all states occur despite (or without regard to) the case severity.

Next, we considered the difference-in-difference of tie strength for within-state communication and across-state communication measurements. Figure 5-10 shows a positive relationship between case severity and difference-in-difference measurement (across-state minus within-state), which indicates that increased case severity is correlated with increased communication across-state. A significant regression equation was found ($p < 0.001$) with $R^2 = 0.2$. The prediction of difference-in-difference of "tie strength" is equal to $-1.994 + 0.03(\log \text{ of COVID}-19 \text{ cases})$. In other words, for every unit increase in log of COVID-19 cases, there is a 0.03 increase in the difference-in-difference measurement in Snaps per recipient. This significant positive correlation not only evidences the association between case severity and online communication, but it also further substantiates that distancing effects contribute to a reduction of location-based homophily (as in our result for RQ2).

Public Sharing.

While social distancing is positively correlated with the increase in private sharing, we also ask whether it correlates with the magnitude of the drop in location-based public sharing. Figure 5-11 shows a positive relationship between case severity and the drop in public story posting (SS), which indicates that increased COVID-19 case severity is associated with a larger drop in public sharing. A significant regression equation was found ($p < 0.01$) with $R^2 = 0.13$. The prediction of percentage drop in story posting is equal to $53.3 + 5.5e{-}5(\log \text{ of COVID}-19 \text{ cases})$. In other words, for every unit increase in log of COVID-19 cases, the drop in story posting grows by $5.5e{-}5\%$.

5.6 Discussion and Conclusion

In this work, we quantified the impact of COVID-19 on online engagement habits through various angles. First, we found that post-onset COVID-19 engagement is higher for private sharing and lower for location-based public sharing. This finding reflected that after the onset of COVID-19, due to stay-at-home orders and other quarantine policies, people reduced their time outside and increased their communications on social media platforms. In particular, as Juhász and Hochmair (2018) found that Snapchat users share public Snaps from highly trafficked areas, such as tourist hotspots or urban centers. We postulate that when these centers temporarily closed, it affected the foot traffic negatively, and further intensified people's discomfort with being in popular public areas, therefore decreasing location-based public sharing. This crisis highlights the particular strengths of social media especially when in-person interactions are limited. Since many people cannot connect with their friends and family in person, for the time being (and potentially longer), social media has become an even more dominant means of maintaining valued connections.

Second, we found that lockdowns temporarily reduced effects of homophily and induced increases in across-state/gender/age-group online communications. As social distancing measures were put into place, most relationships became effectively the same "online distance" (just a Snap) away. Moreover, people realized how important it is to stay in contact with their friends and family. Both of the reasons presumably led to increased diversity in online communications. For instance, due to the rise in COVID-19 cases, students returned home from colleges and stayed connected with friends across the country. Due to stay-at-home orders, people couldn't meet family members and relatives (of all age groups and gender) regularly, so they relied on online communication to check on each other. There are many such instances that got people to step out of their usual social circle and bond with those outside it. Presumably, our observations capture both a flattening of multiple social circles (that many previously have been a mixture of in-person and online interactions) for preservation of routine communication and deepening, as well as a resurrection of previously less accessible

relationships (far-away friends, colleagues and relatives).

Third, we concluded that the number of COVID-19 cases in different states is not correlated to the increase in private sharing frequency, but is significantly positively correlated to the difference in private sharing metrics of within-state and across-state communications. Summarily, COVID-19 cases do not affect increment in private sharing directly, but rather, the more severe pandemic is in a state, the more the reduction in location-based homophily is. Moreover, COVID-19 cases is also positively correlated with the reduction in location-based public sharing. These are surprising and interesting results. This is likely due to the effectiveness of stay-at-home orders and other social distancing measures. Overall, since the number of cases, testing access, response measures and enforcements varied from state to state in the US, the effects of the COVID-19's severity on communication patterns were not uniformly consistent. Generally, we believe that higher case-severity likely corresponds to higher public panic and more stringent distancing and isolation restrictions, leading to stronger effects where observed.

Limitations. The conclusions of this work are limited to US population, and may not generalize to other countries who had differing pandemic responses and lesser degrees of distancing. Moreover, our work was conducted using data from Snapchat, which offers a significant, but not comprehensive view of online communications – Snapchat's user population skews younger, and female, for example. Additionally, our study time period is subject to limitations of platform data access, integrity (i.e. user-misreported information) and availability, suggesting the value of more longitudinal work on the persistence, increase or decrease of the observed effects with the evolving response to COVID-19. Furthermore, our analysis with respect to on-ground case severity is impacted by the inconsistencies and challenges in measurement, detection and response imposed by external factors. Lastly, our analysis does not differentiate results across states. Future work can build on ours by carefully positioning findings with respect to diverse policy responses; the lack of to-date standardized data to quantify such complex policies and their public adherence makes this task non-trivial, and ripe for future work.

Impact. Overall, our work contributes to a more profound understanding of how COVID-19 has, and continues to influence online human behaviors, and moreover how online platforms provide an alternative foundation to support human connection during the pandemic. Notably, humans are inherently social. In times of physical distancing, they are inclined to compensate their social needs via online measures. Our findings shed light on how COVID-19's on-ground impact and associated in-person distancing and isolation measures influenced communication volumes and propensity differences in private and public sharing behaviors, variation and reduction in homophilic baselines, and correlate to the magnitude of such shifts. We hope that our study provides valuable and timely insights for other researchers, and inspires further work on longer-term impacts and communication changes as a result of a post COVID-19 world.

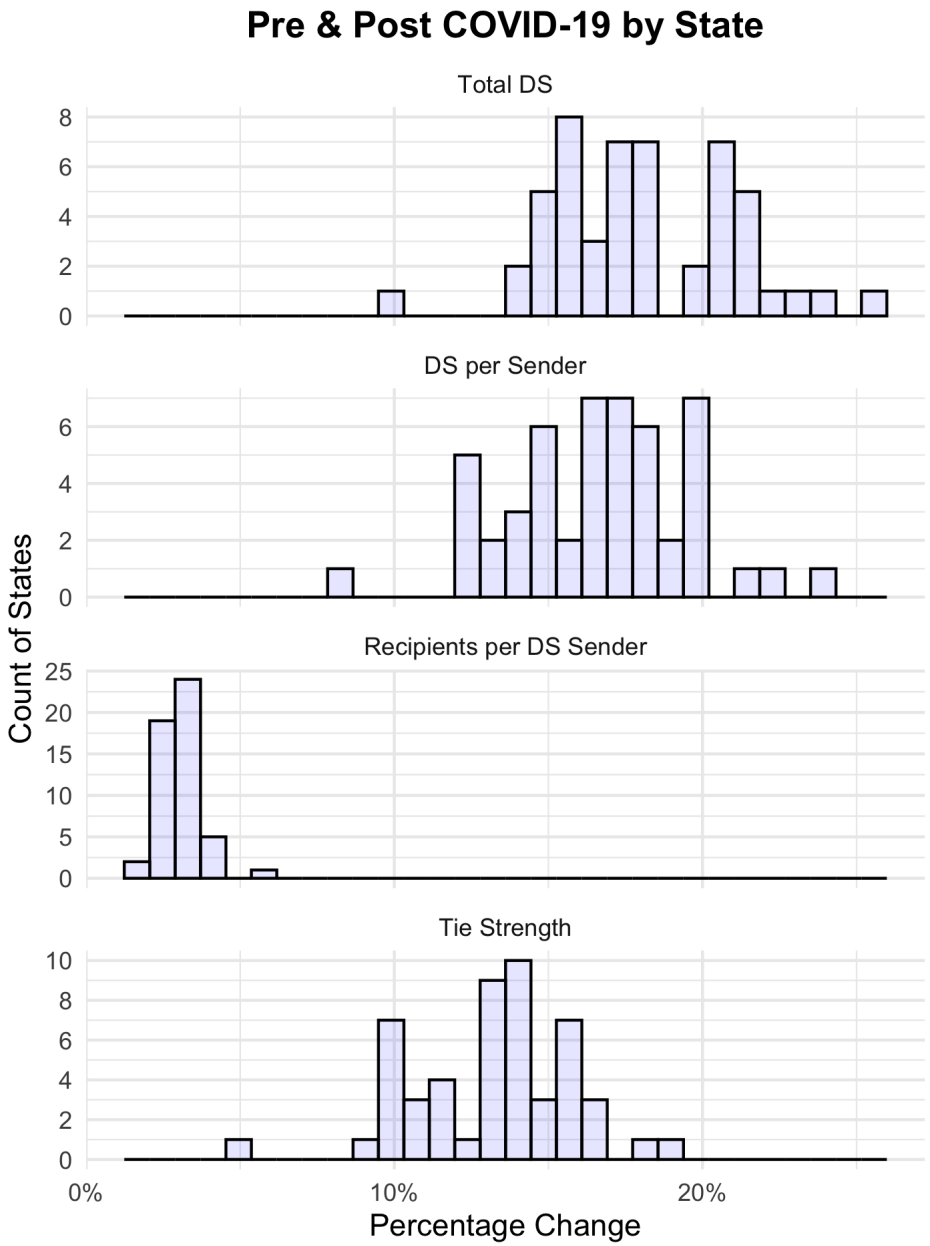
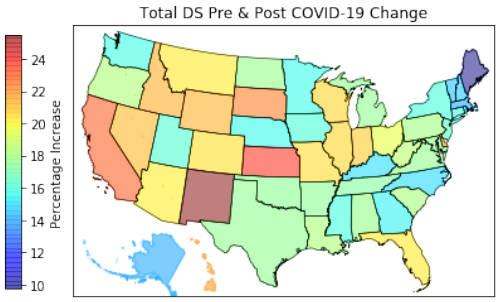
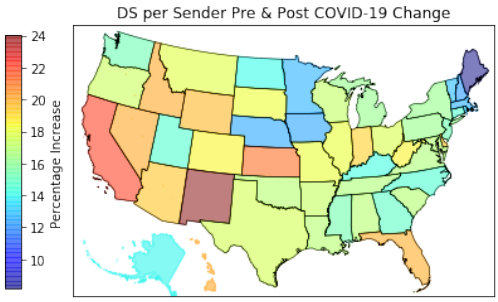


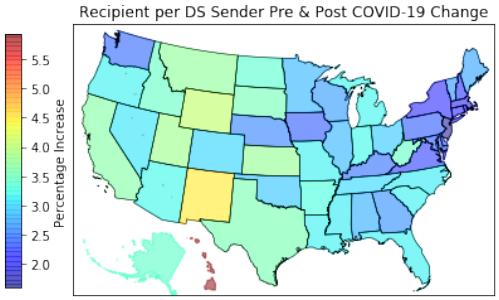
Figure 5-1: Percentage changes in private sharing (DS) across all the US states for several metrics indicate that online private sharing substantially increases (all $p < .05$).



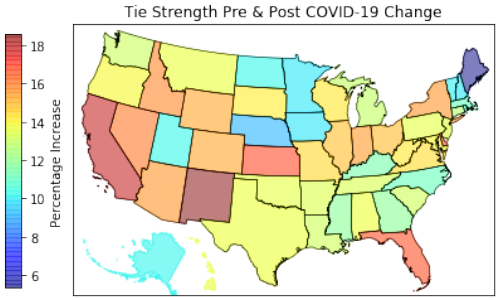
(a) ME, MA, and AK have the lowest, whereas CA, KS, and NM experience the highest increment.



(b) ME, NH, IA have the lowest, whereas KS, CA, and NM experience the highest increment.



(c) NJ, MA, and CT have the lowest, whereas WY, NM, and HI experience the highest increment.



(d) ME, NE, and IA have the lowest, whereas KS, CA, and NM experience the highest increment.

Figure 5-2: % changes in private sharing (DS) on the US map.

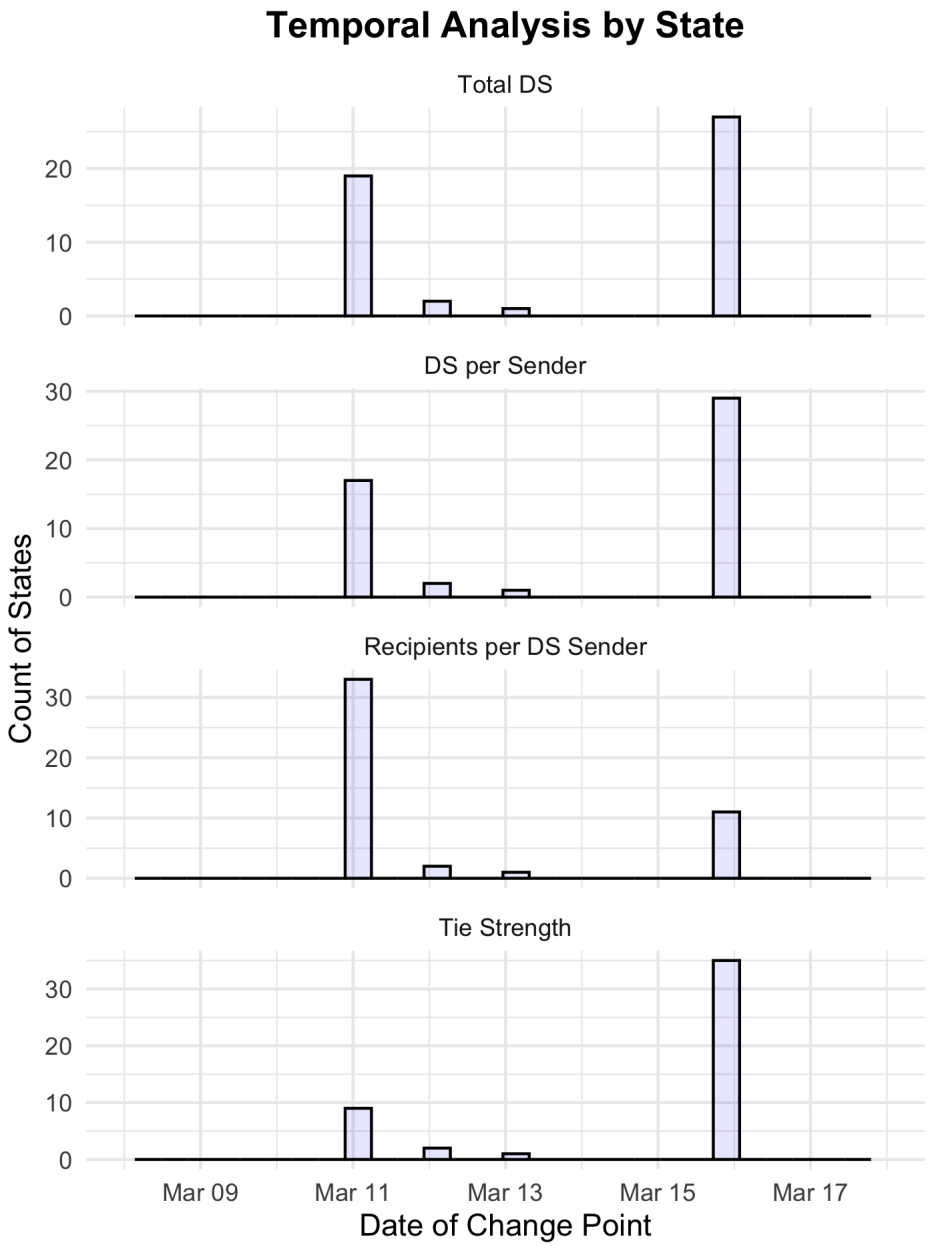


Figure 5-3: Change point detection in private sharing (DS) across all US states for several metrics indicate that online private sharing experienced a surge for most states.

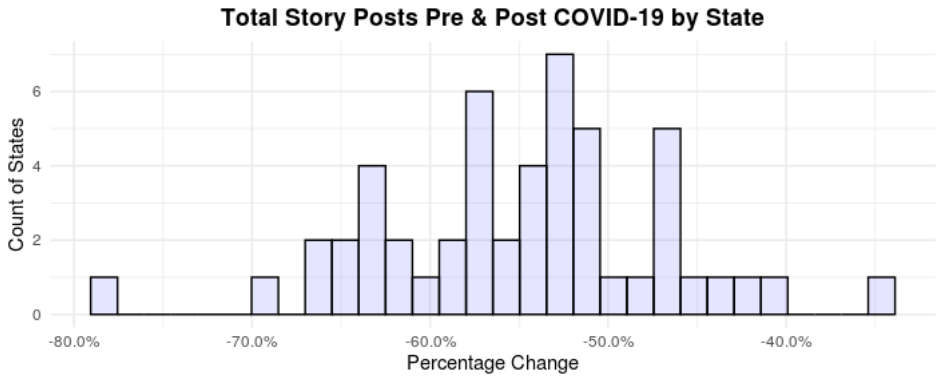


Figure 5-4: Percentage changes in public posting (SS) across all US states indicate that online location-based public sharing substantially decreases (all $p < .05$); post COVID-19 means are -78.98 to -35.31 percent lower.

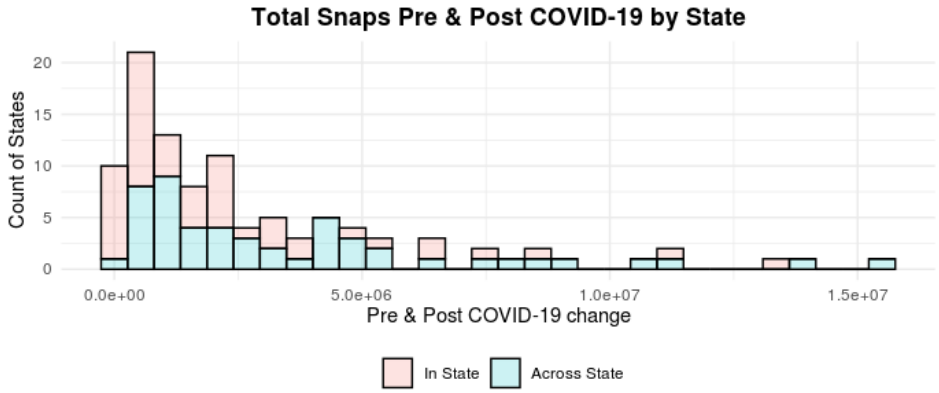


Figure 5-5: Raw increase in total private sharing (DS) of within-state (red) and across-state (blue). Within-state DS increases outsizes across-state DS increase for most states.

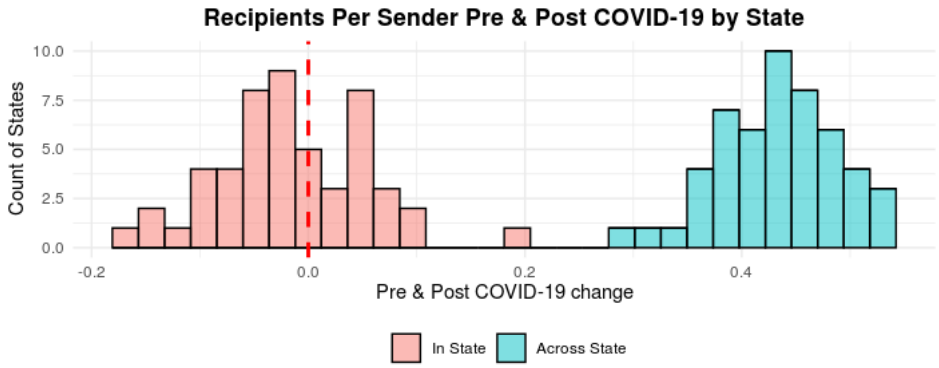


Figure 5-6: Social network size (measured by recipients per sender) consistently grows for across-state communications (blue), compared to mixed effects for in-state communications (red), indicating a reduction of location-based homophily and promotion of cross-location diversity.

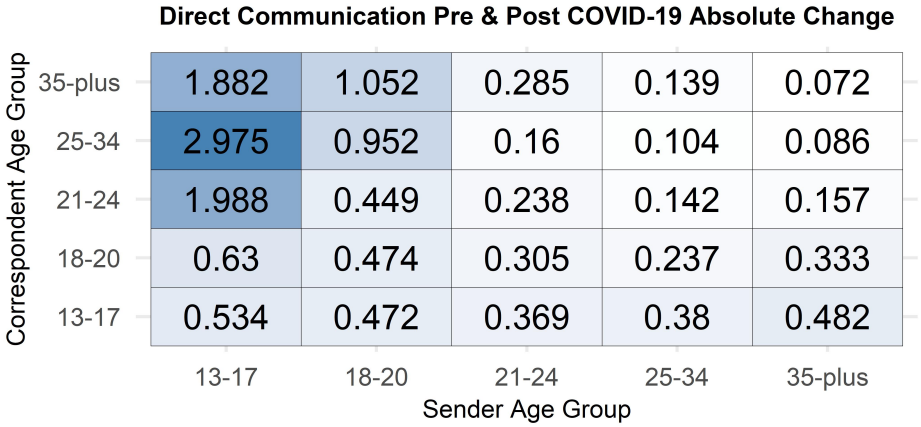


Figure 5-7: Absolute increases in private sharing (DS tie strength) between different age groups pre and post COVID-19 indicate reduction in age-group homophily. Users deepen communications both within and across age-groups, and seemingly moreso in the latter setting.

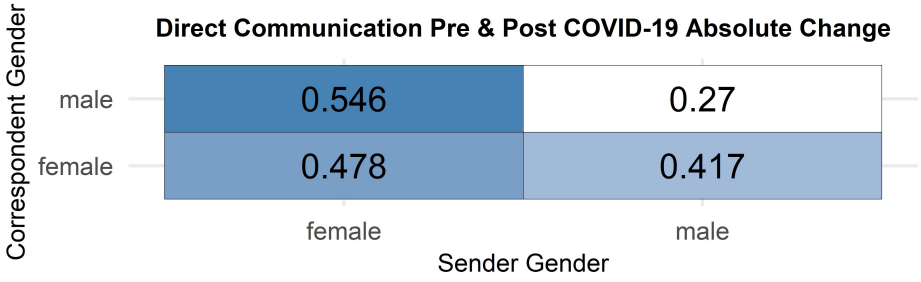


Figure 5-8: Absolute increases in private sharing (DS tie strength) between different gender groups pre and post COVID-19 indicate reduction in gender-group homophily. Users deepen communications both genders, and seemingly moreso with the opposite gender.

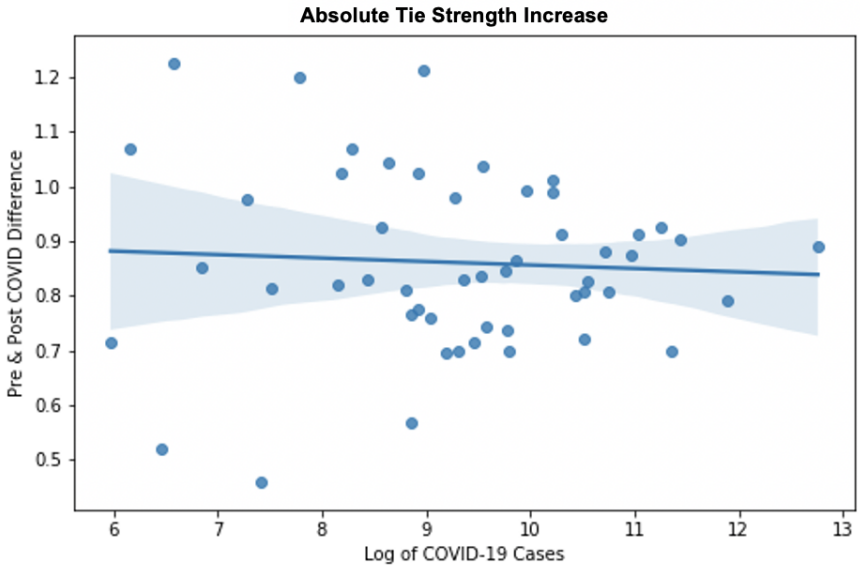


Figure 5-9: Offline COVID-19 case severity is not significantly correlated with online private sharing (DS) tie strength changes across states pre and post COVID-19.

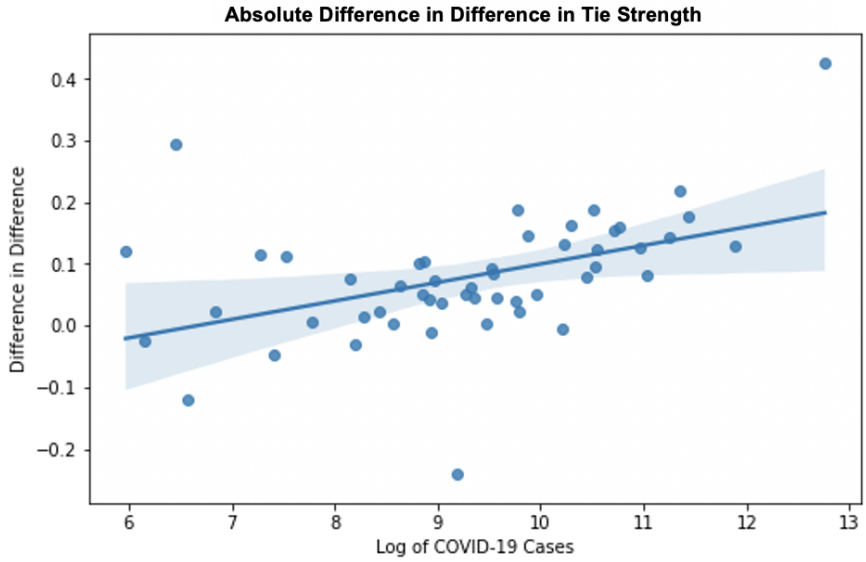


Figure 5-10: Offline COVID-19 case severity is significantly positively correlated with difference-in-difference (across-state minus within-state) measurements of online private sharing (DS) tie strength changes pre and post COVID-19. More COVID-19 cases is associated with larger margins between across-state and within-state tie strengths.

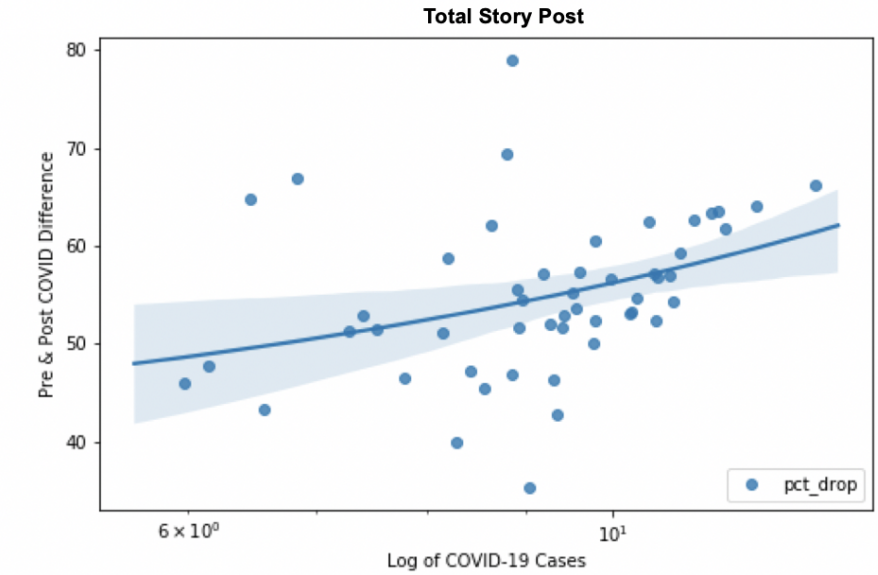


Figure 5-11: Offline COVID-19 case severity is significantly positively correlated with drops in online public sharing (SS) pre and post COVID-19. More COVID-19 cases is associated with larger reduction in public sharing activity.

Chapter 6

Conclusion

We conclude this thesis with a summary of our contributions. The overarching goal of this thesis was to understand users' behaviors and the underlying factors driving these behavior. We pursued this objective by considering three different aspects: partisanship, friendship, and censorship.

In Chapter 2, we introduced the follow back problem, and examined how different following strategies and political ideologies can influence the follow back rate. After obtaining followers, one can then begin posting content to influence them. We found that co-partisanship significantly increased the follow back rate. Moreover, following before friending was also crucial for users to follow back; liking their content only was not enough.

In Chapter 3, we considered influence campaigns. Ongoing examinations have shown that exposure to opposing opinions causes a backfire effect,, where individuals become more undaunted in their unique convictions. We showed a method known as pacing and driving which can mitigate this backfire effect over time.

In Chapter 4, we look at the difficulty of inferring political bias in a hyper-partisan media environment. We observed that Twitter exhibited an anti-conservative bias when suspending users, based on empirical investigations in the preceding chapters. On the other hand, many studies find that conservatives are more likely to share misinformation on social media. As a result, the bans might be the result of implementing an unbiased policy aimed at preventing the spread of disinformation. We found no

indication that Twitter was biased towards conservatives based on the fact that Republicans were more likely to be suspended than Democrats. Instead, this asymmetry could be explained by the Republicans' proclivity for spreading more misinformation.

In Chapter 5, we studied the impact of COVID-19 on online public and private sharing propensity, its influence on online communication homophily, and correlations between online communication and offline case severity in the United States. By tracking the usage patterns of 79 million US-based users on Snapchat, we found that COVID-19 has increased private communication, while decreased publicly share content when users are out-and-about, decreased homophily across locations, ages and genders, and has a positive correlation with widening gaps between across-state and within-state communication increases after the onset of COVID-19.

Appendix A

How to Make a Twitter Bot

A Twitter bot is a Twitter account that can automatically perform actions, like send tweets at a scheduled time or follow or unfollow accounts, and much more. These bots are created and managed via the Twitter API: `https://developer.twitter.com/en/products/twitter-api`.

A.1 Twitter API

A.1.1 Apply for an API Account

To start, you need to make a Twitter account. Notice that in order to apply for the Twitter API, you would need to register with a phone number. Google voice number would not be accepted more than twice, if you want to create multiple bots.

Then, go to `developer.twitter.com` and log in with the existing account and apply for a developer account. In the application, you need to explain your intended use of the Twitter API. If the application is not approved for the first time, especially when you apply for multiple APIs, you can email Twitter to follow up.

A.1.2 User API

On the Twitter developer site, navigate to the app dashboard and select your app, then select keys and tokens (it should be between App details and Permissions).

Here, you'll have the option to generate or regenerate Consumer API Keys, as well as your Access token and access token secret. These keys will allow you to access and control your account, so make sure to keep them to yourself.

A.2 Bot Profile

You can give your bot a name, as well as a short description. You can also change the profile picture. To make them seem real, use AI generated pictures: `https://generated.photos`. All images can be used for any purpose without worrying about copyrights, distribution rights, infringement claims, or royalties. We show an example in Figure A-1:

Figure A-1: Example of bot accounts.

A.3 Bot Activities

A.3.1 Incubation

After the bot account is created, you might need to train the bot for some time before you start interacting with experiment subjects, since an account with no followers and no tweets would appear fake, and might influence the experiment result. During the incubation period the bots would follow random users to gain some followers. The

bots could also post some tweets about generic topics, or share tweets about trending topics on Twitter.

A.3.2 Unfollow Timing

Usually, the bot would also unfollow the users after some given time to prevent the following count from being inflated, and to keep a better ratio of followers to following which is desirable for appearing human and gaining followers. The "unfollow time" depends on user tweet frequency and can be calculated in the following way. Let W_u denote how long the bot waits between following and unfollowing user u. The wait time should reflect how often a user checks Twitter and it should be shorter for more active users because we want to give the user time to log in and see that the bot had interacted with and followed them and then make the choice on whether or not to follow it. Also, we can set a limit for the bot to wait at least one day before unfollowing and at most seven days to ensure that it would not wait too long or unfollow too soon. Let μ_u and σ_u be the mean and standard deviation of the inter-tweet time for user u. Small values for μ_u indicate that u is a active Twitter user and checks the app often. Then W_u is given by

$$W_u = \min(7 \text{ days}, \max(1 \text{ day}, \mu_u + 4\sigma_u)) \tag{A.1}$$

The bot would unfollow the user if the user is not following the bot when the wait time had elapsed.

A.4 Sample Code

A.4.1 Running the Bot

Below is the sample Python code to set up your bot and perform basic operation such as making posts, following, searching, and checking the friendship between two users (if one follows another).

```
1  import tweepy
2  import time
3  import csv
4
5  def OAuth(consumer_key, consumer_secret, access_token, access_secret):
6      try:
7          auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
8          auth.set_access_token(access_token, access_secret)
9          return auth
10     except Exception as e:
11         return None
12
13
14 def like_retweet_follow(api, hashtag, number_of_tweets):
15     user_ids = []
16     for tweet in tweepy.Cursor(api.search, hashtag).items(
           number_of_tweets):
17         try:
18             tweet.favorite()  # like the tweet
19             tweet.retweet()  # retweet
20             api.create_friendship(tweet.user.id)# follow the author of
                   the tweet
21             user_ids.append(tweet.user.id)
22             print("Action Completed: Tweet liked, retweeted, and author
                   followed")
23             time.sleep(5)  # introduces time between each action
24         except tweepy.TweepError as err:
25             print(err.reason)
26     print(user_ids)
27
28 def make_post(api, tweet_to_post):
29     try:
30         api.update_status(tweet_to_post)
31         print("Action completed, posted: " + tweet_to_post)
32     except tweepy.TweepError as err:
```

```
33          print(err.reason)
34
35  def search(query,qt, api, user_follow):
36      '''
37      Given a query represented as a string (to search for) and an int qt,
            return a list of users who posted something
38      related to the query.
39      '''
40      users = []
41      for twt in tweepy.Cursor(api.search,q=query, lang = 'en').items(qt):
42          try:
43              screen_name = twt.user.screen_name
44              user_followers= twt.user.followers_count
45              tweet_text = twt.text
46              user = api.get_user(screen_name)
47              user_friends = twt.user.friends_count
48              ID = user.id_str
49              following_person = []
50              for person in user_follow:
51                  source_screen_name = screen_name
52                  target_screen_name = person
53                  friendship = api.show_friendship(source_screen_name=
                        source_screen_name, target_screen_name=
                        target_screen_name)
54                  following_person.append(friendship[0].following)
55              basic_info = [ID, screen_name, tweet_text,  query,
                    user_followers, user_friends]
56              basic_info.extend(following_person)
57              users.append(basic_info)
58
59          except tweepy.TweepError as error:
60              print(error.reason)
61          except StopIteration:
62              break
63      return users
64
```

```
65 def data_to_csv(data, name_of_file):
66     table_head = ['user_id', 'screen_name', 'tweet', 'query', '
           num_followers', 'num_friends', 'Donald Trump', 'Mike Pence', '
           Ted Cruz'
67                     'Sarah Palin', 'Newt Gingrich', 'Joe Biden', 'Barack
                          Obama', 'Alexandria Ocasio-Cortez', 'Bernie
                          Sanders', 'Kamala Harris']
68     data.insert(0, table_head)
69     with open(name_of_file, 'w', newline='') as fp:
70         a = csv.writer(fp, delimiter=',')
71         a.writerows(data)
72
73 bot_list = []
74 your_api_key = OAuth(your_api_key)
75 your_api = tweepy.API(your_api_key, wait_on_rate_limit = True,
       wait_on_rate_limit_notify = True)
76 bot_list.append(your_api)
77 #hashtag = "Python -filter:retweets" #keyword to search for, filtered
       for retweets to allow for user's to be followed properly
78 number_of_tweets = 2
79 post_to_tweet = "I hope you all are having a good day!"
80 #for bot in bot_list:
81     #like_retweet_follow(bot, hashtag, number_of_tweets)
82     #make_post(bot, post_to_tweet)
83
84
85 data = your_api.rate_limit_status()
86
87 #print(data['resources']['statuses']['/statuses/home_timeline'])
88 #print(data['resources']['users']['/users/lookup'])
89 #print(data['resources']['search']['/search/tweets'])
90
91
92
93
94 user_follow = ['realDonaldTrump', 'Mike_Pence', 'tedcruz', '
```

```
    SarahPalinUSA', 'newtgingrich',
95                  'JoeBiden', 'BarackObama', 'AOC', 'BernieSanders', '
                        KamalaHarris']
96 users_1 = search("#keepamericagreat", 150, your_api, user_follow)
97 print('got users')
98 data_to_csv(users_1, 'KAG_retweet.csv')
```

A.4.2 Extracting URL

Below is the sample code to extract URLs from all tweets, in order calculate media scores, and other quality scores that requires media consumption.

```
 1 #import tweepy #https://github.com/tweepy/tweepy
 2 import csv
 3 import json
 4 import re
 5 import urlexpander
 6 #from urllib.request import urlopen,build_opener,install_opener,
       HTTPHandler,HTTPSHandler
 7 import http.client
 8 import csv
 9 import pandas as pd
10 from urllib.parse import urlparse
11 from random import shuffle
12 import socket
13 import sys
14 #from utils import user2tweet
15 import datetime
16 import time
17 import os.path
18 from os import listdir
19 from os.path import isfile, join
20 from joblib import Parallel, delayed
21 import unshortenit
22 #from urlextract import URLExtract
23 from joblib import Parallel, delayed
```

```
24 import sys
25 from tqdm import tqdm
26
27 import argparse
28 import logging
29 import socket
30 import requests
31 dns_cache = {}
32 # Capture a dict of hostname and their IPs to override with
33 from pprint import pprint as pp
34 import urllib3
35
36 import requests
37 #import dns.resolver # NOTE: dnspython package
38 import tldextract
39
40 #from urllib3.util import connection
41
42
43
44 #http = urllib3.PoolManager()
45
46 #url = 'https://bit.ly/2SGzWMp'
47 #response = http.request('GET', url)
48
49 #print(response.geturl())
50
51
52
53 parser = argparse.ArgumentParser()
54 parser.add_argument("—input_path", type=str, help="input directory",
       default="")
55 parser.add_argument("—log_file", type=str, help="location of log file",
        default="")
56 parser.add_argument("—hist_depth", type=int, help="how far in hist
       since the pull date", default=365)
```

```
57
58
59
60 args = parser.parse_args()
61
62 mypath=args.input_path
63
64
65 hist_depth=args.hist_depth
66
67 if not os.path.exists('{}_processed'.format(mypath)):
68     os.makedirs('{}_processed'.format(mypath))
69 else:
70   print ('diretory {}_processed already exists!'.format(mypath))
71   logging.info('diretory {}_processed already exists!'.format(mypath))
72
73
74 if args.log_file=='':
75     log_file=mypath+"_processed/tweet_process.log"
76 else:
77     log_file=args.log_file
78
79 print ('log file',log_file)
80
81 logging.basicConfig(filename=log_file,
82                                 filemode='a',
83                                 format='%(asctime)s,%(msecs)d %(name)s %(
                                        levelname)s %(message)s',
84                                 datefmt='%H:%M:%S',
85                                 level=logging.INFO)
86 logging.info('start loging')
87 num_cores=50
88
89 #mypath='test'
90
91
```

```
92
93 if not os.path.exists('{}_processed'.format(mypath)):
94     os.makedirs('{}_processed'.format(mypath))
95 else:
96   print ('diretory {}_processed already exists!'.format(mypath))
97   logging.info('diretory {}_processed already exists!'.format(mypath))
98
99
100
101 files = [f for f in listdir(mypath) if isfile(join(mypath, f)) and f.
      endswith('.csv') ]
102 shuffle(files)
103
104
105 for file in files:
106   user=file.replace('.csv','')
107   fname="{}_processed/{}.txt".format(mypath,user)
108   if os.path.isfile(fname):
109   # print ('user {} already exists!'.format(user))
110     continue
111   print("processing user {}".format(user))
112   logging.info("processing user {}".format(user))
113   start_time = time.time()
114   with open("{}_processed/{}.txt".format(mypath,user), 'w') as outfile:
115     pass
116   try:
117     df = pd.read_csv("{}/{}.csv".format(mypath,user),dtype={'text':
          object})
118   except Exception as e:
119     logging.info(e)
120     print(e)
121
122 # for i,tweet in enumerate(tweet_data['tweet']):
123 #   #print (tweet)
124   # (sites ,dates ,tweetIDs,all_tweet_dates)=process_tweet_urls(tweet)
125 #
```

```
126   # print (sites ,dates ,tweetIDs,all_tweet_dates)
127   tweets=[]
128   #pulled_date=datetime.datetime.strptime(tweet_data['pulled_date'],'%Y
          -%m-%d %H:%M:%S')
129   tweet_data={}
130   tweet_data['tweet']=[]
131   for idx,row in df.iterrows():
132     tweet_data['tweet'].append({
133       'text':row['text'],
134       'created_at':row['date'],
135       'id':-99
136       }
137       )
138
139   for tweet_ in tweet_data['tweet']:
140     #if datetime.datetime.strptime(tweet_['created_at'],'%Y-%m-%d %H:%M
          :%S')> pulled_date- datetime.timedelta(days=hist_depth):
141     tweets.append(tweet_)
142
143   #print ('keys',tweet_data.keys())
144   tweets=[tweet_ for tweet_ in tweets if tweet_ is not None]
145   sites_short=[]
146   sites=[]
147   dates=[]
148   RTs=[]
149   quoteds=[]
150   tweetIDs=[]
151 # print(tweets[0])
152   #urls=[]
153   for tweet in tweets:
154     try:
155       urls = re.findall('http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[!*\(\)
              ,]|(?:%[0-9a-fA-F][0-9a-fA-F]))+', str(tweet['text']))
156     except Exception as e:
157       print(e)
158       print(tweet['text'])
```

```
159         raise
160       if 'RT' in tweet.keys():
161         RT=tweet['RT']
162       else:
163         try:
164           RT=int(str(tweet['text']).lower().startswith('rt @'))
165         except Exception as e:
166           print (e)
167           logging.info(e)
168           RT=int(str(tweet['text']).lower().startswith('rt @'))
169       quoted=None
170       if 'quoted' in tweet.keys():
171         quoted=tweet['quoted']
172       for url in urls:
173         sites_short.append(url)
174         RTs.append(RT)
175         quoteds.append(quoted)
176         tweetIDs.append(tweet['id'])
177         dates.append(tweet['created_at'])
178
179     sites=urlexpander.expand(sites_short,
180       n_workers=50,
181       chunksize=1280,
182       #cache_file='temp.json',
183       verbose=1)
184
185     sites_final=[]
186     dates_final=[]
187     RTs_final=[]
188     quoteds_final=[]
189     tweetIDs_final=[]
190     for site_,date_,RT_,quoted_,tweetID_ in zip(sites,dates,RTs,quoteds,
          tweetIDs):
191
192       site=urlexpander.get_domain(site_)
193       if site=='twitter.com':
```

```
194             continue
195         sites_final.append(site)
196         dates_final.append(date_)
197         RTs_final.append(RT_)
198         quoteds_final.append(quoted_)
199         tweetIDs_final.append(tweetID_)
200     #print(sites)
201 # print(urls,sites)
202     #results = Parallel(n_jobs=num_cores)(delayed(process_tweet_urls)(
            tweet_) for tweet_ in tqdm(tweets)  )
203 #       if cut_off>50:break
204 #    ts = time.strftime('%Y-%m-%d %H:%M:%S', time.strptime(tweet['
        created_at'],'%Y-%m-%d %H:%M:%S'))
205     #print (results)
206
207
208     #print(len(tweets),len(quoteds_final),len(sites_final),len(dates_final
            ))
209
210     history = {"screen_name": user, "len_all_tweets":len(tweet_data['tweet
            ']),\
211     "len_all_urls":len(sites_final), "bot":user, #"quoteds":quoteds_final,
212     "sites":sites_final, "dates":dates_final,"RTs":RTs_final#,"tweetIDs":
            tweetIDs_final
213     }
214     with open("{}_processed/{}.txt".format(mypath,user), 'w') as outfile:
215       json.dump(history, outfile)
216     elapsed_time = time.time() - start_time
217     m, s = divmod(int(elapsed_time), 60)
218     h, m = divmod(m, 60)
219     print ('elapsed_time {0:d}:{1:d}:{2:d}'.format(h, m, s))
220     logging.info('elapsed_time {0:d}:{1:d}:{2:d}'.format(h, m, s))
```

Appendix B

Supplementary Information for Mitigating Backfire Effect Using Pacing and Leading

B.1 Keyword for Subject Acquisition

We show in supplementary Table B.1 the keywords used to find experiment subjects. We used the Twitter Search API to find tweets containing the keywords and the users posting the tweets become potential subjects.

1 RefugeesNotWelcome	12 StopIslam
2 Rapefugees	13 ISLAMIZATION
3 BanMuslims	14 UnderwearBomber
4 WhiteGenocide	15 NoRefugees
5 StopRefugees	16 StopIllegalMigration
6 CloseThePorts	17 AntiImmigration
7 ImmigrationInvasion	18 Reimmigration
8 MigrantCrime	19 NoRefugees
9 FreeTommy	20 NoIslam
10 QAnon	21 ProtectOurBorder
11 MAGA	

Table B.1: Hashtags used to identify target users

Phase	Argue Bot and Pace and Lead Bot
Phase 0	What an incredible experience #RyderCup18
Phase 0	Newcastle become the first team in #PL history to score twice against Man Utd at Old Trafford in the opening 10 minutes #MUNNEW
Phase 0	GOAAALLLL!! Shaqiri again playing a big part in the goal. Salah with a smashing finish to make it two!
Phase 0	Looking forward to Saturday already! #MondayMotivation

Table B.2: Tweets posted by the bots in phase zero of the experiment.

B.2 Example Bot Tweets

The experiment has four phases numbered zero to three. Phase zero is the incubation period where the bots post content which does not take a stance on immigration. The argue bot posts pro-immigration tweets in phases one, two, and three. The pace and lead bot posts anti-immigration tweets in phase one. In phase two its tweets express uncertainty about immigration or potential validity of pro-immigration arguments. In phase three the tweets are pro-immigration, similar to the argue bot. We constructed the tweets based on what we deemed a proper representation of the opinion for each phase.

Tables B.2,B.3, and B.4 shows randomly selected examples of the tweets posted by each bot in each phase of the experiment.

B.3 Bot Operation

The bots were active for two months before we started the experiment. We did this to make the bots seem real so that the targets would not be suspicious that they were being followed by a fake account. Their activity included having them first tweet some manually created messages. We also looked at trending topics and retweeted some of those posts, such as UEFA Europa League (we provide more example tweets in Table B.2). Each of the bots' locations were set to London, and they followed a number of common English Twitter accounts to give them the indication of living there.

Phase	Argue Bot	Pace and Lead Bot
Phase 1	Former Calais Jungle child refugee who was unlawfully refused safe passage to join his aunt in Britain still in France two years from the closure of the camp. Can we reunite him with his aunt?	Immigrants strike again. Muslim Uber driver Khaled Elsayedsa Ali charged in California with kidnapping four passengers. This needs to be stopped.
Phase 1	Unbelievable. A revised estimate of 56,800 migrants have died/-gone missing over the past four years.	Muslims attempt to derail high-speed train in Germany using steel wire. Threats in Arabic were found thereafter.
Phase 1	A win for refugees! Former refugee elected to US congress-woman.	Unacceptable. After mass Muslim migration into Germany, sex attacks are up 70% in Freiburg alone.
Phase 2	Chancellor Angela Merkel defends UN migration pact. A step in the right direction.	UK Government to sign UN migration pact. Interesting that Angela Merkel defends it, and rejects "nationalism in its purest form". I believe in her.
Phase 2	It's human rights day, and refugees across Europe face widespread human rights violations. Europe needs to do more to uphold natural human rights.	"Muslim imam performed call to worship during a Church of England cathedral's Armistice without permission." Crossed the line. However, it would probably be overlooked if it were the other way around, am i right?".
Phase 2	Now that's efficient and socially productive! Germany sets out new law to find skilled immigrants.	The UN migration pact, which would criminalize criticism of mass migration and redefine a refugee, will be signed by world leaders next week." Though not through public consent, the #ImmigrationMatters initiative did deliver guiding messages to the public.

Table B.3: Tweets posted by the bots in phases one and two of the experiment.

Phase	Argue Bot and Pace and Lead Bot
Phase 3	Pathetic. At the height of the Syrian refugee crisis in 2015, Syria's neighbors took in 10,000 refugees per DAY. Yet the UK Home Secretary just called the arrival of 75 asylum seekers by boat in 3 days a major incident.
Phase 3	Appalling? In 2018 at least 2,242 people have died in the Mediterranean Sea trying to reach Europe.
Phase 3	The sole survivor said he was left alone in the water for at least 1 day before a fishing boat found and rescued him.

Table B.4: Tweets posted by the bots in phase three of the experiment. Both bots tweeted pro-immigration tweets.

The bots started to follow the users we identified as anti-immigration people to gain followers. We made sure that no two bots were following the same user as this could arouse suspicion. To boost the follow back rate, the bots liked the users' tweets. To avoid bias before the experiment, all tweets the bots liked were not immigration related. The bots also unfollowed the users after some given time to prevent our following count from being inflated, and to keep a better ratio of followers to following which is desirable for appearing human and gaining followers. The "unfollow time" depends on user tweet frequency and was calculated in the following way. Let W_u denote how long the bot waits between following and unfollowing user u. The wait time should reflect how often a user checks Twitter and it should be shorter for more active users because we want to give the user time to log in and see that the bot had interacted with and followed them and then make the choice on whether or not to follow it. Also, we want the bot to wait at least one day before unfollowing and at most seven days to ensure that it would not wait too long or unfollow too soon. Let μ_u and σ_u be the mean and standard deviation of the inter-tweet time for user u. Small values for μ_u indicate that u is a active Twitter user and checks the app often. Then W_u is given by

$$W_u = \min(7 \text{ days}, \max(1 \text{ day}, \mu_u + 4\sigma_u)) \tag{B.1}$$

The bot would unfollow the user if the user was not following the bot when the wait time had elapsed. Users who followed the bot were not unfollowed and became

Bot	Treatment	Followed	Followed Back	Available
Alan Harper	White, Pacing/Leading	3045	636	578
Keegan Richardson	White, Arguing	3051	717	651
Carl Holtman	White, Control	817	125	107

Table B.5: Number of users who were followed by, followed back, and remained available for all phases of the experiment for each bot.

subjects for the experiment. For the remaining phases of the experiment all tweets from the bot would appear on their Twitter timeline.

Table B.5 shows the number of users each bot attempted to follow and the number of users who followed back and were available throughout the experiment. Users may not be available due to three reasons: (i) privacy settings, (ii) account deletion by user, (iii) account suspension by Twitter.

B.4 Covariate Balance Check

Table B.6 shows the followers and friend count of the study population. We performed a pair-wise t-test for all groups and we found that there is no statistically significant difference between any group means ($p < 0.05$).

	Followers Count		Friends Count	
	mean	std	mean	std
BOT				
A	6854	18510	6793	12962
AC	5184	9738	5622	9891
P	6057	11368	6155	9992
PC	4754	20319	4621	13950
Control	6367	7860	6474	7645

Table B.6: Descriptive characteristics of study population for each bot. The bot are labeled as follows: Control is the control bot, A is argue without contact, AC is argue with contact, P is pace and lead without contact, and PC is pace and lead with contact.

Phase	Treatment	Number of tweets	Number of tweets containing "illegals"
0	Control	24,156	40
0	Argue	97,277	149
0	Pace	86,236	159
0	Argue contact	50,474	102
0	Pace contact	47,467	77
1	Control	23,212	65
1	Argue	85295	194
1	Pace	83408	255
1	Argue contact	47880	177
1	Pace contact	49110	139
2	Control	19986	42
2	Argue	70877	179
2	Pace	68151	201
2	Argue contact	44114	171
2	Pace contact	47086	75
3	Control	25863	88
3	Argue	100079	425
3	Pace	90942	506
3	Argue contact	63382	375
3	Pace contact	56851	234
4	Control	23205	34
4	Argue	60262	159
4	Pace	47571	144
4	Argue contact	44462	146
4	Pace contact	42059	114

Table B.7: The number of tweets and tweets containing "illegals" in each phase and for each treatment group of the experiment.

B.5 Experiment Data

We show in Table B.7 the number of tweets and number of tweets with the word "illegals" in each phase and treatment group.

B.6 Spillover Effect

One source of contamination in our experiment could occur if a user retweeted the bot he followed, and then this retweet was seen by his follower who also followed a

different bot. This would cause the follower to receive treatments from two different bots, which is known as a spillover effect. Though retweets happen very rarely in our experiment, we still wanted to make sure the spillover effect does not affect our results.

In total, 18 users retweeted the bots during the experiment. This results in 213 users (including the 18 retweeters) in the experiment who may have experienced the spill over effect. We excluded these users to cross-validate our result. We run logistic regression on both the whole user set, as well as the refined user set, and compare the results in Figures B-1 and B-2. As seen in the coefficient plots, the results are quite similar and we do not see any appreciable spillover effect in the regression coefficients.

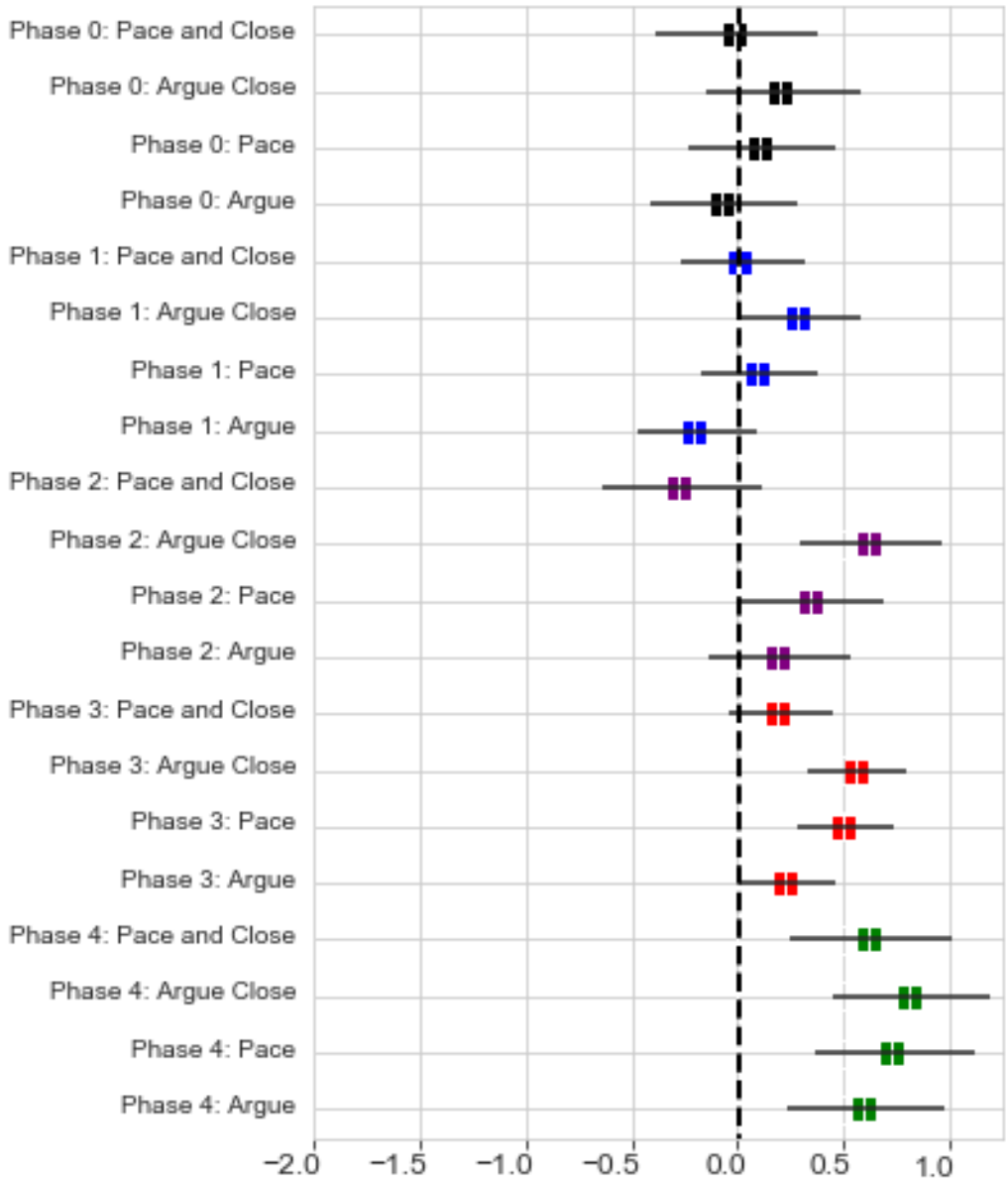


Figure B-1: Coefficient plots of regression with all users including those who may have experience the spillover effect. These are the results in the main paper.

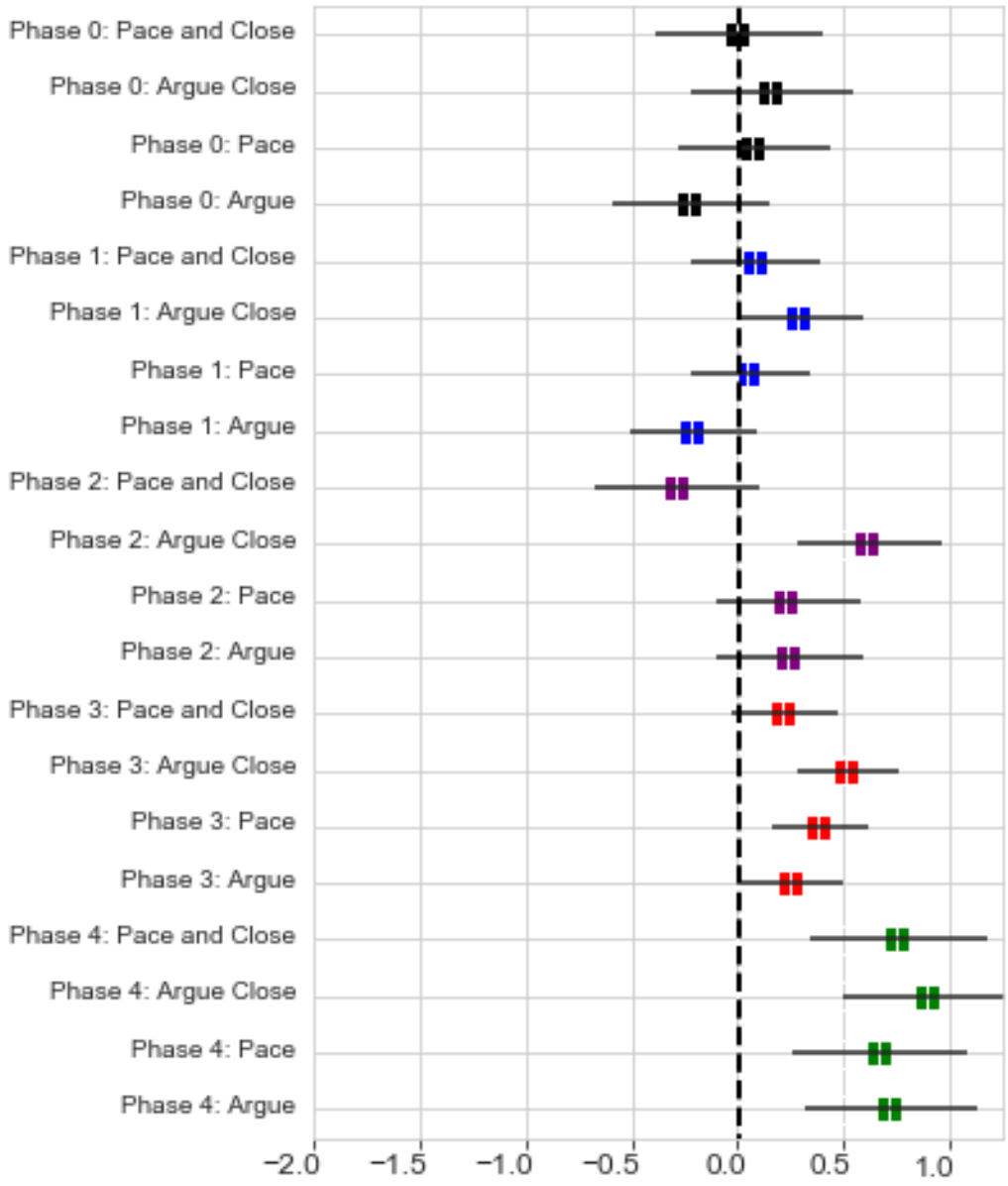


Figure B-2: Coefficient plots of regression with only refined users did not experience the spill over effect.

Bibliography

Akoglu, L.; Chandy, R.; and Faloutsos, C. 2013. Opinion fraud detection in online reviews by network effects. *ICWSM* 13(2-11):29.

Amichai-Hamburger, Y.; Kingsbury, M.; and Schneider, B. H. 2013. Friendship: An old concept with a new meaning? *Computers in Human Behavior* 29(1):33–39.

Backfried, G., and Shalunts, G. 2016. Sentiment analysis of media in german on the refugee crisis in europe. In *International Conference on Information Systems for Crisis Response and Management in Mediterranean Countries*, 234–241. Springer.

Bail, C. A.; Argyle, L. P.; Brown, T. W.; Bumpus, J. P.; Chen, H.; Hunzaker, M. F.; Lee, J.; Mann, M.; Merhout, F.; and Volfovsky, A. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* 115(37):9216–9221.

Bakshy, E.; Messing, S.; and Adamic, L. A. 2015. Exposure to ideologically diverse news and opinion on facebook. *Science* 348(6239):1130–1132.

Baldassarri, D., and Bearman, P. 2007. Dynamics of political polarization. *American sociological review* 72(5):784–811.

Baldassarri, D., and Gelman, A. 2008. Partisans without constraint: Political polarization and trends in american public opinion. *American Journal of Sociology* 114(2):408–446.

Barberá, P.; Jost, J. T.; Nagler, J.; Tucker, J. A.; and Bonneau, R. 2015. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science* 26(10):1531–1542.

Bayer, J. B.; Ellison, N. B.; Schoenebeck, S. Y.; and Falk, E. B. 2016. Sharing the small moments: ephemeral social interaction on snapchat. *Information, Communication & Society* 19(7):956–977.

Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57(1):289–300.

Bessi, A.; Petroni, F.; Del Vicario, M.; Zollo, F.; Anagnostopoulos, A.; Scala, A.; Caldarelli, G.; and Quattrociocchi, W. 2015. Viral misinformation: The role of homophily and polarization. In *WWW*, 355–356.

Bond, S. 2021. Donald trump sues facebook, youtube and twitter for alleged censorship. *URL: https://www.npr.org/2021/07/07/1013760153/donald-trump-says-he-is-suing-facebook-google-and-twitter-for-alleged-censorship*.

Boshmaf, Y.; Muslukhov, I.; Beznosov, K.; and Ripeanu, M. 2013. Design and analysis of a social botnet. *Computer Networks: The International Journal of Computer and Telecommunications Networking* 57:556–578.

Boulianne, S. 2015. Social media use and participation: A meta-analysis of current research. *Information, communication & society* 18(5):524–538.

Boyd, D. M., and Ellison, N. B. 2007. Social network sites: Definition, history, and scholarship. *Journal of computer-mediated Communication* 13(1):210–230.

Brennen, J. S.; Simon, F.; Howard, P. N.; and Nielsen, R. K. 2020. Types, sources, and claims of covid-19 misinformation. *Reuters Institute* 7:3–1.

Burger, J. M.; Soroka, S.; Gonzago, K.; Murphy, E.; and Somervell, E. 2001. The effect of fleeting attraction on compliance to requests. *Personality and Social Psychology Bulletin* 27(12):1578–1586.

Bursztyn, L.; Rao, A.; Roth, C.; and Yanagizawa-Drott, D. 2020. Misinformation during a pandemic. *University of Chicago, Becker Friedman Institute for Economics Working Paper* (2020-44).

Carpini, M. X. D. 2004. Mediating democratic engagement: The impact of communications on citizens' involvement in political and civic life.

Catanzaro, M.; Caldarelli, G.; and Pietronero, L. 2004. Assortative model for social networks. *Physical review e* 70(3):037101.

Catherine Herridge, Graham Kates, L. G. G. After years of trying to curb qanon messaging, twitter has now suspended more than 150,000 accounts. *CBS News*.

Cheng, J.; Romero, D. M.; Meeder, B.; and Kleinberg, J. 2011. Predicting reciprocity in social networks. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, 49–56. IEEE.

Cheng, J.; Bernstein, M.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, 1217–1230.

Cialdini, R. B., and Trost, M. R. 1998. Social influence: Social norms, conformity and compliance.

Coletto, M.; Esuli, A.; Lucchese, C.; Muntean, C. I.; Nardini, F. M.; Perego, R.; and Renso, C. 2016. Sentiment-enhanced multidimensional analysis of online social networks: Perception of the mediterranean refugees crisis. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 1270–1277. IEEE Press.

Colleoni, E.; Rozza, A.; and Arvidsson, A. 2014. Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of communication* 64(2):317–332.

Conway, B. A.; Kenski, K.; and Wang, D. 2015. The rise of twitter in the political campaign: Searching for intermedia agenda-setting effects in the presidential primary. *Journal of Computer-Mediated Communication* 20(4):363–380.

Coombs, W. T. 2014. *Ongoing crisis communication: Planning, managing, and responding*. Sage Publications.

Currarini, S.; Jackson, M. O.; and Pin, P. 2009. An economic model of friendship: Homophily, minorities, and segregation. *Econometrica* 77(4):1003–1045.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Davis, C. A.; Varol, O.; Ferrara, E.; Flammini, A.; and Menczer, F. 2016. Botornot: A system to evaluate social bots. In *Proceedings of the 25th international conference companion on world wide web*, 273–274. MISSING PUBLISHER.

Depoux, A.; Martin, S.; Karafillakis, E.; Preet, R.; Wilder-Smith, A.; and Larson, H. 2020. The pandemic of social media panic travels faster than the covid-19 outbreak.

DiMaggio, P.; Evans, J.; and Bryson, B. 1996. Have american's social attitudes become more polarized? *American journal of Sociology* 102(3):690–755.

Dimitrova, D. V.; Shehata, A.; Strömbäck, J.; and Nord, L. W. 2014. The effects of digital media on political knowledge and participation in election campaigns: Evidence from panel data. *Communication research* 41(1):95–118.

Ding, H., and Zhang, J. 2010. Social media and participatory risk communication during the h1n1 flu epidemic: A comparative study of the united states and china. *China Media Research* 6(4):80–91.

D.M. Romero, J. K. 2010. The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter. In *4th Int'l AAAI Conference on Weblogs and Social Media*.

Dwivedi, Y. K.; Kapoor, K. K.; and Chen, H. 2015. Social media marketing and advertising. *The Marketing Review* 15(3):289–309.

Eady, G.; Nagler, J.; Bonneau, R.; and Tucker, J. 2019a. Political information sharing and ideological polarization. *Midwest Political Science Association, Chicago*.

Eady, G.; Nagler, J.; Guess, A.; Zilinsky, J.; and Tucker, J. A. 2019b. How many people live in political bubbles on social media? evidence from linked survey and twitter data. *Sage Open* 9(1):2158244019832705.

F. Nagle, L. S. 2009. Can friends be trusted? exploring privacy in online social networks. In *IEEE International Conference on Advanaces in Social Networking Analysis and Mining*, 312–315.

Flaxman, S.; Goel, S.; and Rao, J. M. 2016. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly* 80(S1):298–320.

Gentzkow, M., and Shapiro, J. M. 2011. Ideological segregation online and offline. *The Quarterly Journal of Economics* 126(4):1799–1839.

Gillani, N.; Yuan, A.; Saveski, M.; Vosoughi, S.; and Roy, D. 2018. Me, my echo chamber, and i: introspection on social media polarization. In *Proceedings of the 2018 World Wide Web Conference*, 823–831.

Goolsby, R. 2010. Social media as crisis platform: The future of community maps/crisis maps. *ACM Transactions on Intelligent Systems and Technology (TIST)* 1(1):1–11.

Granovetter, M. S. 1977. The strength of weak ties. In *Social networks*. Elsevier. 347–367.

Grinberg, N.; Joseph, K.; Friedland, L.; Swire-Thompson, B.; and Lazer, D. 2019. Fake news on twitter during the 2016 us presidential election. *Science* 363(6425):374–378.

Guacho, G. B.; Abdali, S.; Shah, N.; and Papalexakis, E. E. 2018. Semi-supervised content-based detection of misinformation via tensor embeddings. In *ASONAM*, 322–325. IEEE.

Guess, A.; Nagler, J.; and Tucker, J. 2019. Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science advances* 5(1):eaau4586.

Habib, H.; Shah, N.; and Vaish, R. 2019. Impact of contextual factors on snapchat public sharing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13.

Higgins, M. J.; Sävje, F.; and Sekhon, J. S. 2016. Improving massive experiments with threshold blocking. *Proceedings of the National Academy of Sciences* 113(27):7369–7376.

Hiltz, S. R.; Diaz, P.; and Mark, G. 2011. Introduction: Social media and collaborative systems for crisis management. *ACM Trans. Comput.-Hum. Interact.* 18(4).

Huynh, T. L., et al. 2020. The covid-19 risk perception: A survey on socioeconomics and media attention. *Econ. Bull* 40(1):758–764.

Imran, M.; Castillo, C.; Diaz, F.; and Vieweg, S. 2015. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)* 47(4):1–38.

Juhász, L., and Hochmair, H. H. 2018. Analyzing the spatial and temporal dynamics of snapchat. In *AnaLysis, Integration, Vision, Engagement (VGI-ALIVE) Workshop*.

Jungherr, A. 2016. Twitter use in election campaigns: A systematic literature review. *Journal of information technology & politics* 13(1):72–91.

Kaghazgaran, P.; Bos, M.; Neves, L.; and Shah, N. 2020. Social factors in closed-network content consumption. *CIKM*.

Kalsnes, B. 2016. The social media paradox explained: Comparing political parties' facebook strategy versus practice. *Social Media+ Society* 2(2):2056305116644616.

Katz, J. E., and Crocker, E. T. 2015. Selfies| selfies and photo messaging as visual conversation: Reports from the united states, united kingdom and china. *International Journal of Communication* 9:12.

Killick, R.; Fearnhead, P.; and Eckley, I. A. 2012. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association* 107(500):1590–1598.

Kim, B. 2020. Effects of social grooming on incivility in covid-19. *Cyberpsychology, Behavior, and Social Networking*.

Koeze, E., and Popper, N. 2020. The virus changed the way we internet. `https://www.nytimes.com/interactive/2020/04/07/technology/coronavirus-internet-use.html`. Accessed: 2020-06-10.

Koopman, C. 2021. Cgo tech poll. *URL: https://www.thecgo.org/research/tech-poll/*.

Krivitsky, P. N.; Handcock, M. S.; Raftery, A. E.; and Hoff, P. D. 2009. Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social networks* 31(3):204–213.

Kruikemeier, S.; Van Noort, G.; Vliegenthart, R.; and De Vreese, C. H. 2013. Getting closer: The effects of personalized and interactive online political communication. *European journal of communication* 28(1):53–66.

Kumar, S., and Shah, N. 2018. False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559*.

Lamba, H., and Shah, N. 2019. Modeling dwell time engagement on visual multimedia. In *KDD*, 1104–1113.

Larson, H. J. 2018. The biggest pandemic risk? viral misinformation. *Nature* 562(7726):309–310.

Lau, D. C., and Murnighan, J. K. 1998. Demographic diversity and faultlines: The compositional dynamics of organizational groups. *Academy of management review* 23(2):325–340.

Lechner, M., et al. 2011. *The estimation of causal effects by difference-in-difference methods*. Now.

Lin, Y.-H.; Liu, C.-H.; and Chiu, Y.-C. 2020. Google searches for the keywords of "wash hands" predict the speed of national spread of covid-19 outbreak among 21 countries. *Brain, Behavior, and Immunity*.

Lord, C. G.; Ross, L.; and Lepper, M. R. 1979. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology* 37(11):2098.

Mason, L. 2018. *Uncivil agreement: How politics became our identity*. University of Chicago Press.

McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 27(1):415–444.

Metaxa-Kakavouli, D.; Maas, P.; and Aldrich, D. P. 2018. How social ties influence hurricane evacuation behavior. *Proceedings of the ACM on Human-Computer Interaction* 2(CSCW):1–16.

Miller K, C. K. 2020. The covid tracking project. `https://covidtracking.com/`. Accessed: 2020-05-16.

Mosleh, M.; Martel, C.; Eckles, D.; and Rand, D. 2021a. Perverse downstream consequences of debunking: Being corrected by another user for posting false political news increases subsequent sharing of low quality, partisan, and toxic content in a twitter field experiment. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13.

Mosleh, M.; Martel, C.; Eckles, D.; and Rand, D. G. 2021b. Shared partisanship dramatically increases social tie formation in a twitter field experiment. *Proceedings of the National Academy of Sciences* 118(7).

Munger, K. 2017. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior* 39(3):629–649.

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

Öztürk, N., and Ayvaz, S. 2018. Sentiment analysis on twitter: A text mining approach to the syrian refugee crisis. *Telematics and Informatics* 35(1):136–147.

Palen, L., and Anderson, K. M. 2016a. Crisis informatics—new data for extraordinary times. *Science* 353(6296):224–225.

Palen, L., and Anderson, K. M. 2016b. Crisis informatics—new data for extraordinary times. *Science* 353(6296):224–225.

Pandit, S.; Chau, D. H.; Wang, S.; and Faloutsos, C. 2007. Netprobe: a fast and scalable system for fraud detection in online auction networks. In *WWW*, 201–210.

Pennycook, G., and Rand, D. G. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* 116(7):2521–2526.

Pennycook, G.; McPhetres, J.; Zhang, Y.; Lu, J. G.; and Rand, D. G. 2020. Fighting covid-19 misinformation on social media: experimental evidence for a scalable accuracy-nudge intervention. *Psychological science* 31(7):770–780.

Pennycook, G.; Epstein, Z.; Mosleh, M.; Arechar, A. A.; Eckles, D.; and Rand, D. G. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature* 592(7855):590–595.

Perspective api. *URL: https://developers.perspectiveapi.com/s/about-the-api.*

Perrin, A. 2015. Social media usage. *Pew research center* 52–68.

Rind, B., and Strohmetz, D. 2001. Effect on restaurant tipping of presenting customers with an interesting task and of reciprocity. *Journal of Applied Social Psychology* 31(7):1379–1384.

Rind, B. 1997. Effects of interest arousal on compliance with a request for help. *Basic and Applied Social Psychology* 19(1):49–59.

Rios, J. S.; Ketterer, D. J.; and Wohn, D. Y. 2018. How users choose a face lens on snapchat. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 321–324.

Saha, K.; Liu, Y.; Vincent, N.; Chowdhury, F. A.; Neves, L.; Shah, N.; and Bos, M. W. 2021. Advertiming matters: Examining user ad consumption for effective ad allocations on social media. *CHI*.

Shah, N. 2020. Scale-free, attributed and class-assortative graph generation to facilitate introspection of graph neural networks. *KDD Mining and Learning with Graphs*.

Singh, L.; Bansal, S.; Bode, L.; Budak, C.; Chi, G.; Kawintiranon, K.; Padden, C.; Vanarsdall, R.; Vraga, E.; and Wang, Y. 2020. A first look at covid-19 information and misinformation sharing on twitter. *arXiv preprint arXiv:2003.13907*.

Smith, M. S., and Giraud-Carrier, C. 2010. Bonding vs. bridging social capital: A case study in twitter. In *2010 IEEE Second International Conference on Social Computing*, 385–392.

Snap Inc. 2020. Snap inc. q2 2020 earnings. `https://investor.snap.com/events-and-presentations/events`. Accessed: 2020-06-01.

Strekalova, Y. A. 2017. Health risk information engagement and amplification on social media: News about an emerging pandemic on facebook. *Health Education & Behavior* 44(2):332–339. PMID: 27413028.

Taber, C. S., and Lodge, M. 2006. Motivated skepticism in the evaluation of political beliefs. *American journal of political science* 50(3):755–769.

Tajfel, H.; Turner, J. C.; Austin, W. G.; and Worchel, S. 1979. An integrative theory of intergroup conflict. *Organizational identity: A reader* 56:65.

Tang, X.; Liu, Y.; Shah, N.; Shi, X.; Mitra, P.; and Wang, S. 2020. Knowing your fate: Friendship, action and temporal explanations for user engagement prediction on social apps. *KDD*.

Ugander, J.; Backstrom, L.; Marlow, C.; and Kleinberg, J. 2012. Structural diversity in social contagion. *Proceedings of the National Academy of Sciences* 109(16):5962–5966.

Ulvi, O.; Lippincott, N.; Khan, M. H.; Mehal, P.; Bass, M.; Lambert, K.; Lentz, E.; and Haque, U. 2019. The role of social and mainstream media during storms. *Journal of Public Health and Emergency* 3(0).

Uski, S., and Lampinen, A. 2016. Social norms and self-presentation on social network sites: Profile work in action. *New media & society* 18(3):447–464.

van Green, T. 2020. Few americans are confident in tech companies to prevent misuse of their platforms in the 2020 election. *URL: https://www.pewresearch.org/fact-tank/2020/09/09/few-americans-are-confident-in-tech-companies-to-prevent-misuse-of-their-platforms-in-the-2020-election/*.

Verstraete, G. 2016. It's about time. disappearing images and stories in snapchat. *Image & Narrative* 17(4).

Vogels, E. A.; Perrin, A.; and Anderson, M. 2020. Most americans think social media sites censor political viewpoints. *URL: https://www.pewresearch.org/internet/2020/08/19/most-americans-think-social-media-sites-censor-political-viewpoints/*.

Wang, H.; Can, D.; Kazemzadeh, A.; Bar, F.; and Narayanan, S. 2012. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 system demonstrations*, 115–120.

Welhausen, C. A. 2015. Visualizing a non-pandemic: Considerations for communicating public health risks in intercultural contexts. *Technical Communication* 62(4):244–257.

Wiederhold, B. K. 2020. Social media use during social distancing. *Cyberpsychology, Behavior, and Social Networking* 23(5):275–276.