UNIVERSITY OF CALIFORNIA, MERCED

Four studies of communicative, cognitive, and social factors in extremism and polarization

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Cognitive and Information Sciences

by

Matthew A. Turner

Committee in charge:

Paul E. Smaldino, Chair
Teenie Matlock
Christopher T. Kello
Jeffrey Yoshimi

2021

The dissertation of Matthew A. Turner is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

(Teenie Matlock)

(Christopher T. Kello)

(Jeffrey Yoshimi)

(Paul E. Smaldino, Chair)

University of California, Merced

2021

DEDICATION

To Phoebe Ruth.

As Daniel Tiger and his parents sing, "We gotta look a little closer to see just how things go." Taking care of you makes me happy, too—and caring for you has pushed me to try to understand what is going on between people in this crazy world, which will maybe contribute a little bit to making the world better for you and any future family members. Thank you for getting me out of my own head and back into thinking about what I can do to make the world a better place. Maybe some day this will inspire you in your own pursuits of new skills and a deeper understanding of the world.

EPIGRAPH

Toute vérité et toute action impliquent
un milieu et une subjectivité humaine.

All truths and actions emerge in the context
of human culture and human subjectivity.

—Jean Paul Sartre, "L'Existentialisme est un Humanisme", 1946

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I thank my parents for providing me with many rich experiences throughout childhood that set me on the path to being a scientist, and for their generous help and encouragement while completing this PhD program.

I sincerely thank Paul Smaldino for pushing me to improve with more patience than could be reasonably expected of anyone. I entered this program unaware of my poor sense of career direction and lack of discipline. I thought I could find success freewheeling across many different subdisciplines. Paul patiently peeled back the layers of insulation over my tender ego to help me see that I had a lot of work to do to improve to be the expert I wanted to be. At the same time Paul has made it clear that he believes I have valuable skills that can and should be developed further, and has helped me along immesnely with that.

My committe members have all endured my overly grand first draft ideas and rantings and patiently steered me towards more practical steps that helped me make real progress on tractable pieces of the grand ideas. Teenie Matlock (along with Paul Maglio) first took me on when I began the program. I appreciate Teenie's wisdom, encouragement, and tutelage on metaphor and language, and research in general. Her guidance has enabled me to hear and read political discourse in a vastly more enlightened way, even enabling me to study it scientifically, which I did not know people did before meeting her. I thank Chris Kello for his support, inspiration, and encouragement for my grand ideas about timescales in cognitive science even bringing me in to work with him on a computational project on the subject. I also vividly remember Chris giving me a thumbs up from down the hall after the Monday brownbag in support of my first modeling results in what became a journal article (M. A. Turner & Smaldino, 2018) and Study 4. I thank Jeff Yoshimi for two excellent courses (COGS 202 and COGS 269: *Phenomenology*) and for generously tutoring me in the philosophy of science, which has been a powerful addition to my theoretical toolkit.

Finally, I thank everyone in the CIS community. These past five years have been a stimulating, challenging, and fun experience that enabled me to learn about and join in the collective project of cognitive science.

VITA

2008	B. S. in Mathematics & Physics, Syracuse University, Syracuse, New York
2012	M. S. with Thesis in Applied Physics, Rice University, Houston, Texas
2012-2014	Data Engineer, Economic Modeling Specialists, Int'l, Moscow, Idaho
2014-2016	Research Software Developer II, Northwest Knowledge Network, University of Idaho, Moscow, Idaho
2016-2020	Graduate Teaching Assistant, University of California, Merced
2020-2021	Graduate Student Researcher, University of California, Merced
2021	Ph. D. in Cognitive and Information Sciences, University of California, Merced

PUBLICATIONS

Turner, M.A., and Smaldino, P.E. (2021). Most group polarization results may be simple conformity. In prep.

Turner, M. A., and Smaldino, P. E. (2021). Mechanistic Modeling for the Masses - commentary on Yarkoni, The generalizability crisis. *Behavioral and Brain Sciences*, Forthcoming. Retrieved from https://psyarxiv.com/8pj9n.

Smaldino, P.E., and Turner, M.A. (2021). Covert signaling is an adaptive communication strategy in diverse populations. Under review. Preprint: https://osf.io/preprints/socarxiv

Turner, M.A., and Smaldino, P.E. (2020). Stubborn extremism as a potential pathway to group polarization. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*. Online.

Smaldino, P.E., Turner, M.A., and Contreras Kallens, P. (2019). Open science and modified funding lotteries can impede the natural selection of bad science. *Royal Society Open Science*.

Turner, M.A., and Smaldino, P.E. (2018). Paths to polarization: how extreme views, miscommunication, and random chance drive opinion dynamics. *Complexity*.

Turner, M.A., Maglio, P.P., and Matlock, T. (2018). Metaphorical violence in political discourse. Under revision. Preprint at Mhttps://osf.io/preprints/socarxiv/t8yg9/.

ABSTRACT OF THE DISSERTATION

Four studies of communicative, cognitive, and social factors in extremism and polarization

by

Matthew A. Turner
Doctor of Philosophy in Cognitive and Information Sciences
University of California Merced, 2021
Paul E. Smaldino, Chair

Rising extremism and polarization threaten democratic institutions worldwide. As opposing factions become more extreme in their opinions, polarization increases—the chasm widens the chasm between fellow citizens, and common ground erodes, washed away down a river of vitriol, bitterness, and hate. What causes increased extremism and polarization? Due to the highly complex nature of human societies, this problem of explaining polarization must be broken down into many sub-problems, which themselves require complex systems thinking to address. Simplified models of social systems and rigorous analysis of empirical data, are necessary to build a thorough, coherent understanding of social behavior.

In this dissertation I present my findings from studying three sub-problems in explaining why and how extremism and polarization emerge. First, I focus narrowly on a communication strategy shown in behavioral studies to increase extremism, *metaphorical violence*, such as "Biden hit Trump over his tax returns in yesterday's debate." While we know the effects of violence metaphors, we do not understand their distribution in the wild, or what causes their usage to increase and decrease. I found that metaphorical violence use increased around the time of presidential debates and elections in the United States, and was correlated with presidentical candidates' tweets.

Second, I show that rising extremism in isolated social gropus may be simply

explained by the fact that extremists are more stubborn than centrists—however behavioral studies on the subject may contain ubiquitous false detections of rising extremism, which I demonstrate in Study 3.

Finally in Study 4, I developed and analyzed an empirically motivated, network theoretic, agent-based model of social influence at the societal level to understand how well we can predict polarization based on the effects of initial conditions, network structure, communication noise, and random chance on predictions of polarization.

Taken together these studies advance our understanding of communicative, cognitive, and social factors in the emergence of extremism and polarization.

Chapter 1

Introduction

High levels of political polarization seem to bring about or go along with hardening of partisan identites (Lee, 2015). As society becomes more polarized, political disagreement spills over and fouls up collaborative social behavior more generally (Iyengar, Lelkes, Levendusky, Malhotra, & Westwood, 2019), even making violent responses to verbal communication more likely (Kalmoe, 2014; Kalmoe, Gubler, & Wood, 2018; Mason, 2018). Why do extremism and polarization increase and decrease over time in different contexts? This simple question yields no simple answers. These questions have been studied scientifically for some decades now by researchers across perhaps a dozen diverse disciplines and sub-disciplines including political science (Mason, 2018; Boxell, Gentzkow, & Shapiro, 2020), sociology (Baldassarri & Bearman, 2007; Flache & Macy, 2011), economics (Schelling, 1971; Dixit & Weibull, 2007), cognitive science (Rollwage, Zmigrod, De-Wit, Dolan, & Fleming, 2019), and philosophy (O'Connor & Weatherall, 2018).

In this dissertation I focus on three more specific questions about increasing extremism and polarization. These questions help us understand rising extremism and polarization, and predict what situations will foster rising extremism and polarization. First, what is the prevalence of communication strategies on mass media known to increase extremism, specifically the use of *violence metaphors* to describe non-violent political events? Second, what causes observed increases in extremism among ideologically similar groups over time? Third, and finally, what are some fundamental cognitive capacities and social factors that are required for

polarization to occur, and what role do random chance and miscommunication play in the emergence of polarization? In the course of this work, I also identified a statistical problem that undermines many or possibly most published results on rising extremism among ideologically biased groups.

To answer these research questions, this dissertation studies the communicative, cognitive, and social factors that provide the human substrate for rising extremism and polarization (Jung et al., 2019; Rollwage et al., 2019).

The studies I present in this dissertation help delineate and explore points of contact between the various disciplinary approaches to studying polarization. As such, they required diverse theoretical, modeling, and computational methods, including structured corpus building and analysis for studying metaphor use on cable TV news, statistical modeling of the opinion generation and measurement process, and agent-based modeling of social influence processes.

In political communication, people are differentially influenced depending on what language is used, even down to choice of grammar (Matlock, 2012). For example, if the past participle is used to describe a politicians bad deeds (e.g., *he was imbezzling campaign funds*) this worsens people's opinions of the politician compared to communicating using the simple past tense (*he imbezzled campaign funds*). This dissertation focuses specifically on the choice of metaphor used in political discourse, which is a powerful method for framing political messages to rally political allies and identify, disparage, and target political enemies and other outgroup members (O'Brien, 2003; Charteris-Black, 2009; Landau, Meier, & Keefer, 2010).

There are also communication-independent cognitive and social factors at work in social influence that leads to extremism and polarization. Cognitively, we know that (1) similar individuals tend to find consensus with one another (French, 1956; DeGroot, 1974); (2) dissimilar individuals can push one another to be even more different (Cikara & Van Bavel, 2014; Bail et al., 2018); (3) social influence between similar others tends to be more attractive the more similar they are and more repulsive the more dissimilar they are (Lord, Ross, & Lepper, 1979; Ross, 2012); and (4) that extremists tend to be more stubborn than centrists (Reiss, Klackl,

Proulx, & Jonas, 2019; Zmigrod, Rentfrow, & Robbins, 2019);

Social relationships structure social interactions. Social relationships determine who interacts with whom, which is often analyzed using social networks (Watts, 1999). A major factor that determines an individual's social network is the tendency towards *homophily*, summarized in the heuristic that "birds of a feather flock together" (McPherson, Smith-Lovin, & Cook, 2001). This results in social network structures where similar individuals tend to interact more often compared to dissimilar individuals. This can have major impacts on societal opinion structure and dynamics when similar individuals interact more and more often, and dissimilar individuals sometimes come to not interact at all (Axelrod, 1997; Centola, González-Avella, Eguíluz, & San, 2007; DellaPosta, Shi, & Macy, 2015).

1.1 Introduction to the dissertation studies

This dissertation's modest contributions to the expansive literature on extremism and polarization are in breaking down the complex social influence system of society into three model sub-systems to understand and predict how communicative, cognitive, and social factors work together to contribute to extremism and polarization in different contexts. This has resulted in the four studies presented in this dissertation.

In Study 1 I develop a dynamic model of *violence metaphor* use (e.g., "Clinton hit Trump over his tax returns") on cable TV news to understand how this changes under the influence of the US presidential debates and elections in 2012 and 2016. Violence metaphors are important to understand because exposure to violence metaphors tends to increase anger towards and dislike of political opponents, including increased support of violence to achieve political goals (Kalmoe, 2014; Kalmoe et al., 2018). I found that violence metaphor usage was more reactive to the debates in 2016 and 2012, largely influenced by using violence metaphors to describe Twitter "attacks" (i.e. tweets) by one political candidate against the other.

In Study 2 I develop a novel, parsimonious explanation of a group-level process

that seems to increase extremism among like-minded group members, known as *group polarization* (Brown, 1986; Isenberg, 1986; Brown, 2000; Sunstein, 2002). Existing explanations of group polarization tend to rely on several auxiliary assumptions that may or may not be well supported, which make the explanations difficult to evaluate (Meehl, 1990). I use agent-based modeling to show that group polarization can be more parsimoniously explained by the empirically motivated assumption that people become more stubborn as their opinions become more extreme (Reiss et al., 2019; Zmigrod, Rentfrow, & Robbins, 2019; Kinder & Kalmoe, 2017).

Unfortunately many published detections of group polarization are plausibly false, which I demonstrate in Study 3 using a statistical model. False detections may occur in group polarization data because researchers failed to rigorously account for floor/ceiling effects introduced when ordinal behavioral data (e.g. Likert scale data) is analyzed using metric statistical models (e.g. a t-test to detect differences between normal distributions). Metric statistical models fail to account for the fact that very extreme opinions can become significantly more moderate when measured on an ordinal scale. Therefore, metric models can be tricked into thinking real psychological opinions have become more extreme among a group, when in reality the ordinal measurement scale failed to detect opinion shifts towards moderation among extremists, which masks a simpler process of consensus around the initial group mean opinion. I found that 92% of detections of group polarization across ten journal articles are plausibly false detections, which throws into question the reality of the group polarization.

The purpose of Study 3 is to understand how to measure changing extremism in experimental settings, though it may seem to undermine the findings of Study 2. If we are to develop a more complete theory of political opinion change, we need to be able to reliably measure that change. Study 2 still succeeds in predicting that increased extremism should emerge among small, socially isolated groups, whether that is called "group polarization" or not. Together, Studies 2 and 3 predict a social phenomenon, known as "group polarization", and explain how and why group polarization must be analyzed with ordinal statistical models that account

for the group polarization measurement procedure.

In Study 4 I adapt the same agent-based model from Study 2 to understand how well we can explain and pedict large-scale societal polarization under increased social network connectivity as might emerge from interacting with dissimilar others over the Internet (Bail et al., 2018), in addition to considering the effect of initial extremism, miscommunication, and path-dependence of social interaction order (i.e., who interacts with whom, and when). I found that while social network structure can bias society towards more or less polarization (Flache & Macy, 2011), this is highly path-dependent. Furthermore, there are critical values of initial extremism and communication noise that can override social network structure or path dependence to make either high levels of polarization or full consensus (i.e., no polarization) inevitable.

In the remainder of this introductory chapter I will introduce the twin problems of extremism and polarization, including definitions of both terms. I then introduce the importance of metaphor and framing more generally on rising extremism and polarization. After this, I introduce cognitive and social factors that, together, provide the social influence substrate in which extremism and polarization emerge and rise. Next, I explain my strategy for modeling the emergence of extremism and polarization since all four studies rely on mechanistic modeling of social influence processes. To close this chapter I give an overview of the four studies that comprise the rest of the dissertation.

1.2 Rising extremism and polarization

Among the popular press and politically-involved citizens, it seems obvious that polarization is increasing, and that this increase is a dangerous problem that needs to be solved. For example, when in 2014 the Pew Research Center found polarization in 2014 to be the highest in decades, journalist and Vox founder Ezra Klein (2014) found the statement obvious, writing "(E)veryone already knew that." Klein (2020) later explained further in a book for the popular press, Why we're polarized. However, whether or not polarization occurs depends on how

polarization is defined and the population being studied. In fact, there is a debate among political scientists whether polarization really occurs, but this is a matter of definition (Mason, 2015; Lelkes, 2016; Kinder & Kalmoe, 2017). Extremism and polarization must be well defined to study them scientifically.

At worst, extremism and political polarization leads to political violence and even civil war (Epstein, 2013; Freeman, 2018). However, different types of extremism and polarization may have different effects on society and governance (Lelkes, 2016). Lee (2015), for example, found that although partisan sorting had occurred in recent decades, there had been little degradation in legislative and other government outcomes. The rise of *affective polarization* between political groups has been found to increase as non-political preferences align among partisans as well—for example, preferences for leisure activities and entertainment are becoming increasingly correlated with ideology and party membership in the United States (Pew Research Center, 2014b; DellaPosta et al., 2015). Affective polarization includes the increasing dislike and distrust between opposing political parties, which spills over into non-political areas of life (Iyengar et al., 2019). Polarization is certainly on the rise, but this may be a correction to normal from several previous decades of unnaturally low levels of political sorting (Lee, 2015; Wood & Jordan, 2017). Similar trends and concerns can be observed worldwide (Borge-Holthoefer, Magdy, Darwish, & Weber, 2015; Morales, Borondo, Losada, & Benito, 2015; Romenskyy, Spaiser, Ihle, & Lobaskin, 2017; Zmigrod, Rentfrow, & Robbins, 2018). We in the United States may, however, have a particularly bad case of rising polarization: Boxell et al. (2020) found that among the United States and eight other OECD countries the United States had the largest increase in affective polarization over the past four decades. Further cross-cultural study is necessary to understand the fundamental human factors underlying extremism and polarization, especially societies that are not Western or democratic, with possibly lower standards of living (Henrich, Heine, & Norenzayan, 2010).

In this dissertation I aim to identify fundamental communicative, cognitive, and social factors that foster extremism and polarization, with limited contextual details. Of course context is important, but we cannot know how important with-

out first understanding baseline capacities and processes that lead to changes in extremism and polarization. Note that the above evidence for increasing extremism and polarization assumes the existence of political parties and ideologies—the Republican and Democratic parties, and conservative and liberal ideologies. Polarization in these studies is measured by the degree of sorting of ideologies and preferences into associated political parties (Mason, 2015) and increasing interpersonal dislike (Iyengar et al., 2019). To understand more fundamental factors than parties and ideologies, I take a more general approach that does not label individual opinions, and defines increased extremism and polarization formally in terms of opinion distributions. In this dissertation, *extremism* is defined by how extreme one's opinion is on some opinion scale, which represents how intensely or confidently someone believes in their own opinion on some topic. For example, giving a zero on a Likert scale often means that one neither agrees or disagrees with some statement. Strong disagreement or agreement is indicated by indicated the largest negative or positive values on the Likert scale. *Polarization* in this dissertation is conceptualized as the bimodality in opinions among a population, ignoring partisanship and political or ideological identities. Opinion bimodality can quantify ideological divergence among all members of society (Bramson et al., 2016; Lelkes, 2016).

Among all the ways in which individual preferences and opinions are sorted and polarized in the United States, a major one one that both reflects and drives rising polarization is the split in where partisans get their news (Pew Research Center, 2014a; Martin & Doumas, 2017). What is said on cable TV news and other mass media is extremely important, given the reach of mass media and the way mass media frames the terms of debate (Chong & Druckman, 2007). One important communication strategy is the use of different metaphors to frame different political messages, processes, and events. These framings influence the way politics is understood and discussed by news consumers. One's opinions about immigrants, for example, may depend on whether one has been exposed to metahpors that cast immigrants as "indigestible food, conquering hordes," or "waste materials" (O'Brien, 2003). Those who had been exposed to such metaphors may

later tend to favor stricter limits on immigration and harsher treatment for undocumented immigrants.

In Study 1, I quantify the change in frequency over time of a specific type of metaphor use, violence metaphors, across cable news channels MSNBC, CNN, and Fox News, around the time of the United States presidential debates and elections. Violence metaphors are important because they have been observed to push individuals to more extreme political opinions, even increasing support for real world political violence (Kalmoe, 2014; Kalmoe et al., 2018). Metaphorical violence is a prime strategy for inflaming partisan passions through statements such as "Trump has been getting *attacked* by the liberal democrats on Capital Hill," which one might hear by a commentator or anchor on Fox News.

Mass media frame the terms of political discourse, which spreads through interpersonal influence among ordinary citizens (Katz & Lazarsfeld, 1955). But how does interpersonal social influence work, and how do we know which social influence processes are essential for rising extremism and polarization to emerge? In this dissertation I break down interpersonal influence into formal computational models to analyze whether polarization emerges from that model, without assuming anything about polarization itself.

Interpersonal influence of opinions can be broken down into four important cognitive factors: (1) attractive and repulsive influence of opinions (French, 1956; Cikara & Van Bavel, 2014; Bail et al., 2018); (2) homophily, i.e., preferential assortment with like others (McPherson et al., 2001); (3) biased assimilation, i.e., heightened influence by similar others (Dandekar, Goel, & Lee, 2013); and (4) a correlation between stubbornness and extremity of opinions (Reiss et al., 2019; Zmigrod, Rentfrow, & Robbins, 2019). Other cognitive factors may include, e.g., personality traits that may be predictive of ideological or other opinion, attitude, belief, etc., preference (Zmigrod et al., 2018).

Social factors that modulate social influence processes are: (1) social networks, theoretical entities that represent a person's social relationships that structure who in a society interacts, when; and (2) the stochasticity of interpersonal influence, e.g., three people may frequent a certain bar and talk regularly on Fridays, but

who attends varies depending on essentially random factors like other obligations or obstacles to attending.

In Study 2 I incorporated these cognitive and social factors into a computational model, which led to the simulated emergence of "group polarization", the empirical observation that socially isolated, initially biased groups tend to become more extreme in their opinions over time. In Study 4 I examine critical tipping points of the model that might guarantee polarization or consensus and explore the limits of using this model (or any model) for predicting rising extremism and polarization.

I conceptualize polarization as an emergent property or phenomenon of society. Like the field of psychology generally, understanding and predicting extremism and polarization requires cross-disciplinary understanding that sometimes involves scientists working focusing on one system component and sometimes has scientists exploring the interfaces between system components (Brewer, 2013; Rollwage et al., 2019). By making general assumptions about how social influence works in society, it is possible to encapsulate many of the dimensions along which individuals are separated and how social influence regarding opinions (or beliefs, attitudes, etc.) on one dimension is correlated with social influence along other dimensions. We can add details to such a general model as necessary to understand, for example, how framing strategies change over time on cable TV news (Study 1) or how extremism rises in initially biased, socially isolated groups (Study 2). In Study 4, I use a general model of social influence to investigate the role of social network structure, initial extremism/polarization, communication noise, and random "path dependence" on the order of interpersonal interactions on the emergence of polarization.

1.3 Metaphor and framing in politics and polarization

In the first study of this dissertation, I analyze violence metaphor use on cable news around the times of the 2012 and 2016 United States presidential debates and elections. But what is metaphor? How is it used and what are the effects of

metaphor use in political communication, especially regarding political polarization? Metaphor has long been recognized as an important element in the study of political communication at least since Aristotle's time if not before. Aristotle saw metaphor as a special feature of especially talented orators' rhetoric (Aristotle, 1965; Kirby, 1997). In contrast, modern cognitive understanding of metaphor recognizes the ubiquity of metaphor as a critical cognitive tool evolved for conceptual scaffolding used for abstract thought (Lakoff & Johnson, 1980; Heyes, 2018b, 2018a).

The word metaphor comes from the ancient Greek word *metaphora* ($\mu\epsilon\tau\alpha\phi o\rho\acute{\alpha}$), meaning *transferrence*. Metaphor works by "transferring", or mapping in the mathematical sense, conceptual entailments from a more concrete concept, such as a fight, onto a more abstract concept, such as politics (Regier, 1996; Kövecses, 2010a; Lakoff, 2014). Politics is an abstract concept because it can describe many different situations, events, and processes. One never directly sees or feels politics. The outcomes of political decision are only felt indirectly in terms of increased or restricted liberty, or economic effects such as tax breaks or an improved economy. On the other hand, either being engaged in or observing physical conflict results in a cascade of immediate bodily effects, including body-to-body contact and possibly injury for fight participants. The conceptual entailments of casting politics as violence encodes the fact that politics generates similar feelings and physiological reactions to being in a fight, for example, including the physical sensations of elation or depression following a political win or a political defeat, and the adrenaline and other biophysical repsonses of the fight itself (Gallese & Lakoff, 2005; David, Lakoff, & Stickles, 2016).

Metaphor is one of several forms of linguistic *framing* that can powerfully influence our understanding of political events through strategic, pragmatic choice of words (Fillmore, 1982; Chong & Druckman, 2007; Lakoff, 2008; Charteris-Black, 2009; Fausey & Matlock, 2011; Matlock, 2012; Sagi, Diermeier, & Kaufmann, 2013; Cacciatore, Scheufele, & Iyengar, 2016). Embodied metaphors enable people to gain intuition about many different abstract concepts beyond politics. For example, we talk about "navigating" the internet, but this is really just typing

and clicking links or buttons (Matlock, Castro, Fleming, Gann, & Maglio, 2014). We often describe the passage of time in terms of physical motion, as in, "my dissertation defense date is fast approaching" (Matlock, Ramscar, & Boroditsky, 2005; Núñez, Cooperrider, Doan, & Wassmann, 2012; Flusberg, Matlock, & Thibodeau, 2017b). Embodied concepts such as rotations and extrusions permeate the abstract realm of mathematics (Lakoff & Núñez, 1997; Marghetis & Núñez, 2013).

Politicians and commentators have long used metaphor to motivate supporters and villify opponents (Charteris-Black, 2009). To take a current example, Fox News has recently been covering what they call "Classroom Warfare" over Critical Race Theory [1]. Anchors and commentators on Fox News variously cast anti-Critical Race Theory protestors as "an army of moms and parents" waging war "on the front lines of this fight." War metaphors for addressing the climate crisis seem to foster a greater sense of urgency for finding solutions to the crisis (Flusberg, Matlock, & Thibodeau, 2018). War metaphors have also been ubiquitous in attempts to mobilize public responses to mitigate the spread of COVID-19 (Castro Seixas, 2021). To understand the purpose and effectiveness of this metaphorical framing, we have to consider the *entailments* that go along with the WAR conceptual frame. Wars have at least two opposing belligerent groups— soldiers for each side are literally mortal enemies. In the context of American politics and to Fox News viewers, there are conservatives on the Fox News side and liberals on the other side. Patterns in news consumption reflect this, with Donald Trump voters watching Fox News far more than any other oultet in 2016, and Hillary Clinton voters watching MSNBC and CNN more than any other outlet (Prior, 2013; Pew Research Center, 2014a, 2017b).

Beyond the partisan "wars" that play out in the minds of American citizens based on what they see on American cable news, there are many other political issues and events in the USA and abroad that are described and understood using metaphorical language. For example, the metaphor casting Washington, D.C., as a swamp has been used over the years by both liberals and conservatives to cast

[1] Humorously summarized by The Daily Show here:
https://www.youtube.com/watch?v=7sGK33uTOpU

the other side as dirty and corrupt (Burgers, Jong Tjien Fa, & de Graaf, 2019). In another example, during the Gulf War of 1991, metaphors were used to cast Saddam Hussein as a madman and the United States as "givers" of freedom to Kuaiti citizens, which was metaphorically "taken away" with the Iraqi invasion (Lakoff, 1991). This is a metaphor since freedom is not a thing one can give, receive, or take away, as one would give another person water or food.

In the French and UK presses, metaphor use was observed to vary depending on the political context: when Obama won the 2008 US election, the UK and French presses framed Obama's victory as something predestined, casting Obama as a sort of savior ushering in a new era of US politics, saying things like "Obama walked on water." However, reporting on politics in Pakistan, when former Pakistani General and retiring President Pervez Musharaf's party lost in Pakistan's presidential elections, UK and French news outlets called it a "knockout" and generally used other violent and disparaging metaphors against the former president, despite him not even being a candidate (Burnes, 2011).

Empirical data from behavioral studies supports the inference that violence metaphors could contribute to readers and listeners' to resort to real world violence to attain their political goals. In a series of studies, Kalmoe (2014) and Kalmoe et al. (2018) showed that exposure to violent metaphors drove partisans further apart in terms of opinions, and exacerbated aggressive tendencies towards one's political out-group. These effects were most pronounced for the most aggressive members of society. Clearly violence metaphors need to be understood due to their possibly detrimental effects on political and social stability. This need is a major motivation for Study 1 that measures the dynamics of metaphorical violence usage on cable TV news and finds it to be correlated with candidate Twitter activity—this correlation is amplified in 2016 compared to 2012.

1.4 Cognitive and social factors in extremism and polarization

Extremism and polarization are emergent phenomena of social systems composed of individuals. Human beings are the most fundamental components in the models of social systems I develop in this dissertation. But humans are complex themselves, a composite of simpler cells properly organized to have the capacity for social influence of and by others, among many other capacities (Kello, Beltz, Holden, & Van Orden, 2007; Spivey, 2020). So, how can humans be treated as fundamental? The modeling strategy that solves this is to assume humans have only a small set of critical capacities essential for the social interaction and influence that leads to extremism and polarization (N. Cartwright, 1989; Smaldino, 2017a). Understanding the capacities for social interaction and influence requires multi-method, investigation spanning several disciplines in the form of computational, behavioral, and neurobiological studies. Also important for understanding the emergence of extremism and polarization are social factors, such as the effect of social relationship networks on emergent social phenomena. While communication is essential to increasing extremism and polarization, much can be understood about cognitive and social factors in polarization, indpendent of specific contextual details about interpersonal communication.

1.4.1 Cognitive factors in polarization

In this dissertation, I focus on cognitive factors in polarization I focus on a set of essential individual- and dyad-level cognitive capacities and processes that enable mutual social influence. The first essential capacity is the capacity for one's opinions to become more similar to others' opinions. The second essential capacity is the ability to become more different from those with whom we disagree. Whether two individuals are attracted to or repulsed from one anothers' opinions is often determined by their group membership—people tend to be attracted to in-group members' opinions and repulsed by out-group members' opinions. Therefore determining one's own and others' group membership is also an essential cognitive

capacity. Social influence can be modulated by one's degree of similarity or dissimilarity to in-group or out-group members—more similar views may be more attractive, e.g., or more different opinions more repulsive. Individuals may also vary in their susceptibility to social influence—for example in the model used in Studies 2 and 4 I assume those with more extreme opinions are less susceptible to social influence, i.e., they are more stubborn.

We know that humans tend to find agreement with one another and consensus often emerges within groups (Festinger, 1954; D. Cartwright & Harary, 1956; French, 1956). Consensus with (or conformity to) others' opinions has been shown to emerge even when direct evidence contradicts those opinions, as Asch (1955, 1956) found in his classic studies in which participants were fooled by confederates into going along with the crowd despite their own direct perception that the crowd was obviously wrong. Consensus can be problematic when consensus occurs around, e.g., false scientific beliefs and misinformation (K. J. S. Zollman, 2007; K. J. Zollman, 2013; O'Connor & Weatherall, 2018, 2019).

Often in intergroup social influence, members from different groups develop more different opinions over time when they interact, instead of becoming more similar (Tajfel & Turner, 1979; Sherif, 1988; Flache & Macy, 2011; Bail et al., 2018). Group membership may be determined by observable traits such as race, sex or gender identity, language, or style of dress, but it need not be. The "minimal group" experimental design has been used to design experiments that revealed that novel group membership specified by experimenters can almost immediately override observable indicators of group membership (Tajfel, Billig, Bundy, & Flament, 1971a; Billig & Tajfel, 1973; Tajfel, 1982). These quick changes in behavior are reflected by equally quick changes in brain activity, showing that neural responses to group membership are extremely plastic (Cikara & Van Bavel, 2014; Cikara, Van Bavel, Ingbretsen, & Lau, 2017). This is both a problem and an opportunity—it is a problem because people can be quickly hijacked to see their neighbors as "other", but an opportunity because people can be equally quickly converted to more prosocial behaviors, such as mitigating climate change, if they feel they are part of a group.

People often are more strongly attracted to others' opinions the more similar they are. That is, we tend to adopt our friends' opinions more readily than strangers' opinions because we know we agree with our friends on several other issues or topics. Conversely, individuals are often more repulsed by opposing views the more different other views are. For example, if someone we dislike buys a car, we will maybe be less likely to buy the car brand in the future. This is known as *biased assimilation* (Lord et al., 1979). In politics, individuals have been observed to be more influenced by presidential candidates in a debate who are perceived as similar to themselves. On large scales, it has been observed that food, hobby, and other preferences are becoming increasingly correlated with political ideologies such as conservativism and liberalism, leading to stereotypical "latte liberals" and "bird-hunting conservatives" (DellaPosta et al., 2015). Suhay and Erisen (2018) found that emotions may be critical in biased assimilation: they found that anger especially, along with other emotional states, "fuel(ed) biased reactions to issue arguments" in an online behavioral study.

The final cognitive factor in social influence I consider is that those with more extreme opinions tend to be more stubborn, i.e. less susceptible to social influence, than centrists. Evidence from several disciplines supports this assumption. On the one hand, longitudinal survey studies have found that a large portion of the population are centrists with low "opinion stability over time" (Converse, 2006; Zaller, 1992; Kinder & Kalmoe, 2017). Centrist opinions tend to be more susceptible to framing effects (Chong & Druckman, 2007) and to question ordering (Zaller, 1992). Extremists in the United States and United Kingdom were observed to be more cognitively inflexible than their centrist counterparts (Zmigrod, Rentfrow, & Robbins, 2019). Extremism has also been electrophysiologically linked to differences in responses to stimuli. In an EEG study, Reiss et al. (2019) found that ERP responses to anomolies in experimental stimuli were muted among participants with more extreme socio-political opinions compared to centrist participants.

Other approaches to studying the cognitive factors in extremism and polarization include analyzing the correlation between personality traits and ideological alignment (Rollwage et al., 2019), and considering cognitive factors of social influ-

ence for information exchange, instead opinion influence (K. Carley, 1990, 1991; Bala & Goyal, 1998). In the UK, for instance, Zmigrod et al. (2018) found that dependence on routines was positively correlated with subscribing to conservativism, nationalisim, and authoritarianism, which in turn were positively correlated with support for Brexit from the European Union. My work complements these approaches in that it considers simpler cognitive factors than personality traits, which I see as a composite of opinions, beliefs, knowledge, etc. Due to the complexity of the personality trait construct, it is difficult to tell whether personality traits and their relationship to national-scale ideologies and policies are biologically or culturally determined, and so possibly subject to change along with cultural context (Claidière & Whiten, 2012; Smaldino, Lukaszewski, von Rueden, & Gurven, 2019; Falandays & Smaldino, 2021).

1.4.2 Social factors contributing to polarization

If we want to understand how opinions change under social influence in the media and societal system outlined above, we must also consider social relationships. Social polarization and its opposite, consensus, strongly depend on who interacts with whom, and when (Flache & Mäs, 2008; M. A. Turner & Smaldino, 2018). But what determines these social relationships and how do these relationships change over time? How do we model these relationships scientifically?

Who we interact with is somewhat random and out of our control: it depends on our family membership, geographic location, participation in social activities (e.g. attending school, getting groceries, going to a restaurant), and more. In addition to these random factors, we also adjust our social relationships based on interpersonal affinity and similarity, i.e., we tend to prefer to interact with people we like and avoid people we dislike. Thinking of these evolving relationships as a social network modeled by a mathematical graph enables us to formally represent social relationships and harness graph theory to calculate and predict social facts and behavior. For example, graph theory can help us predict how quickly information (Milgram, 1967; Travers & Milgram, 1969), disease (Salathé et al., 2010; Block et al., 2020), violence (Epstein & Hammond, 2002), and innovations (Deroiain,

2002; Acemoglu, Ozdaglar, & Yildiz, 2011; Kreindler & Young, 2014) spread in groups and in society (Watts, 1999; Palla, Barabási, & Vicsek, 2007; Backstrom, Boldi, Rosa, Ugander, & Vigna, 2012; Wohlgemuth & Matache, 2014).

In social systems people tend to choose social interaction partners who are similar to themselves, a tendency known as *homophily*. As homophily increases among a population, this increases the chance that individuals interact with similar others, and decreases the chance that individuals will interact with dissimilar others (McPherson et al., 2001). Homophily, then, amplifies the cognitive factor of biased assimilation, since increased homophily tends to further insulate individuals from exposure to opposing viewpoints as biased assimilation causes individuals to ignore or reflexively dislike opposing viewpoints and uncritically incorporate information that supports their pre-existing opinions (N. P. Mark, 2003; Dandekar et al., 2013). Another social factor that affects social outcomes are power structures in which some people have a greater social influence than others. This may be represented as having a greater number of relationships with others, so that their opinions are more widely shared (French, 1956; Friedkin, 1986), or due to social status, or both (DeGroot, 1974).

Study 4 incorporates homophily and random chance in a network-theoretic model of social influence to understand how social networks contribute to the emergence of extremism and polarization—the model is formally introduced there. Individuals in a social network are represented by *nodes*, often drawn as dots or some other marker. Nodes can encapsulate an individual's identity in addition to traits such as group membership, accumulated resources (i.e. "payoffs"), etc. Sometimes these traits are visualized by changing the marker size, color, or shape of the node in network visualizations. Relationships between individuals are represented by *edges*, drawn as lines that connect individual nodes. Homophily and power differentials between individuals may also be represented in terms of edge *weights* on the graph. Any graph may have weighted edges which could stand for many different things; in navigation applications, for example, edge weight might represent the time it takes to reach one location from another.

1.5 Mechanistic models of emergent social phenomena

This dissertation relies on theory driven, empirically motivated mechanistic models to simplify the complex system of human social influence. But how do mechanistic models work, specifically for rigorous scientific investigation of emergent phenomena such as extremism and polarization? A hallmark of the scientific explanation of some phenomenon is that the explanation only posits the existence of theoretical entities, entity capacities, and relationships between entities (Kauffman, 1970; N. Cartwright, 1989; Craver, 2006; M. A. Turner & Smaldino, 2021). If the phenomenon of interest emerges from system dynamics specified by the entities and their capacities, then the model and its theoretical basis have some explanatory power. A phenomenon *emerges* when a statistical pattern is detected that is associated with that phenomenon, e.g., polarization is often measured as the bimodality of the distribution of individual opinions (i.e. attitudes, beliefs, etc.) in a society. The patterns of interest in this dissertation are static (polarization at a given point in time) and dynamic (Kelso, 1995), e.g. rising extremism and collective changes in violence metaphor use on cable news. It is not valid or explanatory to assume in advance the existence of the phenomenon.

In this dissertation I use a mechanistic model-based theoretical approach designed to explain observed patterns in mass media metaphorical violence use in Study 1; explain rising extremism in socially isolated groups in Study 2; demonstrate that many or perhaps most detections of rising extremism in socially isolated groups are false detections in Study 3; and to identify critical determinants social polarization in Study 4.

1.5.1 Emergent social phenomena

In this dissertation I focus on the emergence of rising extremism and polarization, which is theoretically influenced by the emergent dynamics of metaphorical violence use on cable news (one of many influential mass media communication strategies). Emergent social phenomena are identified by finding patterns in the

distributions of individual-level behaviors, opinions, traits, etc., among a population (Blau, 1974; Schelling, 2006). It is challenging to explain, with scientific rigor, how emergent social phenomena such as rising extremism and political polarization actually emerge from repeated instances of social influence (Watts, 2011). Social systems are complex systems of groups of various sizes, and individual humans themselves are complex emergent phenomena (Kello et al., 2007; Lazer et al., 2009).

In this work we have assumed that individuals "have" opinions. Polarization is calculated as the distributional variance (or similar measures of bimodality) of individual opinions (Bramson et al., 2016). A totally polarized society has exactly half of the population holding one of two extreme opinions, and the other half holding the opposing view. Other behaviors that lead to different emergent social phenomena include choosing where to live based on racial preferences (not racial animosity), which can result in emergent racial segregation (Schelling, 1971); publishing journal articles of differing validity which leads to systemic scientific problems (Smaldino, Turner, & Contreras Kallens, 2019); or writing statements and documents online that together form a system of cultural frames including harmful ethnic, gender, and racial biases and stereotypes (Caliskan, Bryson, & Narayanan, 2017; Garg, Schiebinger, Jurafsky, & Zou, 2018).

Emergent phenomena occur at all scales of social influence, including dyads and other small groups (Abney, Dale, et al., 2014). For example, dyads were found to synchronize with one another when working together on collaborative tasks, and "asynchronize" when in an adversarial relationship (Abney, Paxton, et al., 2014; Ramirez-Aristizabal, Médé, & Kello, 2018; Schloesser, Kello, & Marmelat, 2019; Schneider, Ramirez-Aristizabal, Gavilan, & Kello, 2020; Abney, Paxton, Dale, & Kello, 2021). In turn, individual humans are emergent properties of a complex electrochemical interaction of individual, differentiated cells (Schrödinger, 2012; Kello et al., 2007; Lazer et al., 2009). It is for this reason that I believe it is best to avoid thinking about "micro" and "macro" scales as seems to be popular among sociologists (Macy & Willer, 2002). I conceptualize the assumptions we must make about individual cognition and social interaction between dyads as "individual-

level" assumptions instead of "micro-motives" (Schelling, 2006). Similarly I prefer the concept of an *emergent phenomenon* to Schelling's concept of "macrobehavior".

Collective violence metaphor usage on cable TV news

The first emergent phenomenon I study is the frequency of violence metaphor usage across cable TV news outlets. I hypothesize and show in Study 1 that violence metaphor use varies depending on how soon there will be or how recently there has been a presidential debate or the presidential election. This approach complements similar approaches to studying time series of semantic content in mass media, social media, and other historical documents in order to understand how cognitive, cultural, and communicative frames covary with historical events (Nunn, 2012; Klingenstein, Hitchcock, & DeDeo, 2014; Hamilton, Leskovec, & Jurafsky, 2016a; Caliskan et al., 2017; Barron, Huang, Spang, & DeDeo, 2018; Garg et al., 2018). Partisan polarization can be identified by analyzing semantic differences in partisan communications (Gentzkow, Shapiro, & Taddy, 2019).

From the complex systems perspective, "pragmatic choice", i.e. what words to use when, is the result of many ongoing subprocesses that occur within different contexts (Gibbs & Van Orden, 2012). The collective attention of society becomes entrained on shared cultural events (Fusaroli et al., 2015). The utterances of news anchors, commentators, and pundits cannot be separated from their pragmatic purpose and societal context (Kövecses, 2010b). Collective violence metaphor use, then, can be considered an emergent property of social systems since it depends on complex interactions between individuals at varying time and population scales.

"Group polarization": rising extremism in small, socially isolated groups

Group polarization is the name given by social psychologists to the observation that novel, socailly isolated, collectively biased groups become more extreme in their opinions after deliberating on some topic (Brown, 1986; Isenberg, 1986; Brown, 2000; Sunstein, 2002). In group polarization, the emergent phenomenon is the rising extremism among the group. Furthermore, there is a higher-level emergent pattern of the magnitude of the group polarization effect—extremsim has

been observed to rise more when the group is already relatively extreme (Myers, 1982). Because of the complex interplay between individual-level cognition and social power dynamics (Friedkin, 1999), we can identify group polarization as an emergent social phenomenon as well.

Polarization

Polarization may be arrived at through a varity of cognitive and social mechanisms, though of course communication details can exacerbate polarization as already discussed. Polarization can increase through repulsive influence (Baldassarri & Bearman, 2007; Flache & Macy, 2011; Bail et al., 2018; M. A. Turner & Smaldino, 2018). When dissimilar individuals repulsively influence one another, their opinions become more extreme in opposite directions, marginally increasing opinion distribution bimodality (Mäs & Flache, 2013; M. A. Turner & Smaldino, 2020). Polarization can also increase through attractive influence only, e.g., through group polarization. In group polarization, isolated groups become more extreme as they find consensus. If one of two groups becomes more extreme, polarization also increases since bimodality will have increased.

1.5.2 Model-based theoretical approach

Scientific models are simplified versions of reality used to identify which components of complex systems are most important in the emergence of collective larger-scale phenomena (Kauffman, 1970; Wimsatt, 1972, 1997; Machamer, Darden, & Craver, 2000; Wimsatt, 2007; Smaldino, 2017a). The most explanatory models are *mechanistic models*, ones that explicitly identify the atomic theoretical entities in a system and how those entities influence one another (Machamer et al., 2000; Craver, 2006; M. A. Turner & Smaldino, 2021). In our case mechanistic models of societal and group systems explain that human individuals communicate with and influence those with whom they share social connections, with social connections represented as social network neighbors. Above I listed assumptions about how individuals process social influence and how social interactions are structured, which are further details incorporated in the model of social influence

used in Study 2 and Study 4. Mechanistic models are stronger still when they are formalized into mathematical notation and implemented computationally to make quantitative predictions of how different social phenomena emerge based on model assumptions.

Models in the dissertation studies

All four studies presented here use some form of mechanistic modeling to represent system dynamics that give rise to emergent phenomena. Mechanistic models may be expressed and implemented in a variety ways. In addition to developing detailed verbal models of how social influence and mass communication work, Studies 1 and 3 implement statistical models and fitting procedures to empirically determine inflection points in violence metaphor dynamics (Study 1) and to demonstrate that a high rate of experimental detections of rising extremism are plausibly false (Study 3). Studies 2 and 4 use agent-based models to understand which cognitive and social factors best explain and predict rising extremism and polarization, respectively.

Study 1 and Study 3 both use statistical models—in Study 1 the model is fit to observations, and in Study 3 the model generates simulated counterfactual data. In Study 1, to partially explain the dynamics of metaphorical violence use on cable TV news, I developed a dynamical model expressed as a statistical regression model where each news channel is in either a normal state or a transient excited or depressed state. In Study 3, I used a generative statistical model to simulate experimental group polarization data where pre- and post-deliberation opinions are drawn from distributions with the same mean. By simulating the measurement of these opinions, I show that floor/ceiling effects lead to a false detection of an opinion shift due to the process of consensus that reduces group opinion variance from pre- to post-deliberation.

Study 2 and Study 4 model different systems using the same underlying agent-based social influence model that incorporates the cognitive and social factors outlined above. Agent-based models of social systems start by defining a computational representation of a person, called an *agent*. To implement these models,

I wrote computer code that created a world in which computational agents were brought to life and made to interact with other agents according to rules and assumptions based on the cognitive and social factors outlined above. After thousands or millions of rounds of simulated social interaction I measured the distribution of opinions to calculate either a rise in extremism or increased polarization.

1.6 Overview of dissertation results

Now we have reviewed the overarching problems of extremism polarization that motivated this work, the theoretical foundations I draw on to address specific subproblems, and the analytical approach I take to studying different emergent social phenomena. I will now give an overview of the four studies presented in this dissertation.

First, Study 1 calculates the influence of political events on metaphorical violence use across three cable TV news channels in 2012 and 2016. I found that significant changes in metaphorical violence usage occur across cable news outlets MSNBC, CNN, and Fox News in both 2012 and 2016 around the time of the presidential debates. In 2012, changes were significant, but rather small and there was no discernable pattern across news outlets. In 2016, however, metaphor use increased dramatically across networks in the run-up to the 2016 elections. I hypothesized that Twitter usage was an important driver of this difference due to a close reading of cable news violence metaphors that seemed to cast candidate tweets as metaphorical violence. This was indeed the case, as revealed through a correlational analysis between Twitter activity by Republican and Democratic candidates and violence metaphor usage on cable news in both election cycles. Correlations in 2016 were more significant overall than in 2012. Candidate Twitter use accounted for over 30% of variance in overall violence metaphor use in 2016, compared to about 8% of variance in 2012. By understanding the dynamics of violence metaphors on cable news, we can better understand and predict downstream effects of violence metaphor use, including reduced capacity for rational thinking due to emotional overwhelm (Suhay & Erisen, 2018) and heightened risks

of political violence (Kalmoe, 2014; Kalmoe et al., 2018).

Study 2 focuses down from mass communications to group-level interpersonal social influence and shifts in extremism. In Study 2 I use agent-based modeling to show that group polarization may be driven by the cognitive factor of "stubborn extremism", where extremists are less susceptible to social influence (Reiss et al., 2019; Zmigrod, Rentfrow, & Robbins, 2019). To detect group polarization, researchers survey participant opinions, assemble participant groups that are biased in one opinion direction or another, and then re-survey participants. If the mean of the group opinions changes, then a group polarization opinion shift is said to occur. Existing explanations of group polarization include potentially problematic axuiliary assumptions that seem to lack robust empirical support (Meehl, 1990). Existing explanations also fail to explicitly account for the observation that the magnitude of opinion shifts are positively correlated with initial group extremism (Myers, 1982). In Study 2 I show that the stubborn extremism explanation also predicts this correlation between initial extremity and opinion shift. The stubborn extremism explanation seems more parsimonious in that it makes a single, simple, empirically valid assumption that is more explanatory than any alternatives.

In the course of Study 2 I found that many behavioral studies of group polarization used a problematic method for measuring group polarization. Specifically, many group polarization studies used metric statistical models on ordinal data, which is known to sometimes yield false inferences (Liddell & Kruschke, 2018). Study 3, then, develops a generative statistical model to show that, indeed, 92% of published detections of group polarization over ten journal articles are plausibly false detections. False detections occur due to the fact that extreme opinions are mapped to relatively moderate opinions, but this is not accounted for when using metric statistical models designed to detect whether two distributions of *continuous* variables are different (e.g., a t-test). Metric models can be tricked into detecting a difference between two distributions' means when the data are ordinal measurements of continuous *latent* psychological opinions from two distributions with the same mean but different variances. We know that opinion variance changes within

groups as groups discuss a topic: as group members find consensus, i.e., variance decreases.

Finally, in Study 4, I explore how well the model from Study 2 could be used to explain and predict society-level polarization, as opposed to increased extremism in small groups in Study 2. I found that (1) greater initial polarization often led to greater long-term polarization; (2) realistic small-world networks tend to result in higher levels of polarization than common alternative configurations; (3) more miscommunication leads to greater polarization; and (4) polarization outcomes are highly stochastic, i.e., the same initial configurations and parameter settings can result in a range of predictions from low to high levels of polarization solely due to the path dependence of interpersonal interactions over many time steps.

Chapter 2

Metaphorical violence in political discourse on US cable TV news

Metaphor is far more than a literary device. It is a fundamental cognitive ability that drives the human capacity for reasoning about states, situations, and actions in the world (Gibbs, 1994; Lakoff & Johnson, 1980). Metaphor—which involves understanding of abstract concepts in terms of relatively more basic ones—permeates political discourse (Lakoff, 2008; Matlock, 2012). Its ubiquity in everyday discourse is evident in the frequent use of statements such as "It's time to drain the swamp", "Obama sprinted toward victory on Election Day", and "Trump attacks Jeff Sessions over Russian probe methods". No one is releasing water. No one is running. No one is causing physical harm. How is metaphorical violence expressed, for instance, expressions with words such as "attack", "slaughter", and "hit", and how does such language influence political thought and communication? Here, we describe novel time-resolved observations and explanatory dynamical models of the use of metaphorical violence language in political discourse on U.S. cable television news in the period leading up to the two most recent presidential elections. Our results quantify the details and dynamics of the use of these metaphors, revealing how cable news shows act as reporters, promoters, expectation-setters, and ideological agents in different degrees in response to differing cultural situations. Our work has implications for shaping political discourse and influencing political attitudes.

2.1 Introduction

Conceptual metaphor theory holds that linguistic metaphors, such as "Costs are rising," reflect a process whereby one concept is structured in terms of another; in this case, costs are conceptualized in terms of physical verticality. In this way, metaphor is not just language; it is a way of thinking (Gibbs, 1994; Lakoff & Johnson, 1980; Thibodeau & Boroditsky, 2011) and it is intimately linked to emotions (Kövecses, 2010b) and grounded in bodily experience (Gallese & Lakoff, 2005). Because metaphor is so pervasive and because many people care about political matters, it is useful to consider how it is used and how it might shape public opinion on matters of national or international importance, such as climate change (Flusberg, Matlock, & Thibodeau, 2017a) and politics (Lakoff, 2008; Lakoff & Wehling, 2012).

We are especially interested in violence metaphors in the context of political discourse. We define *violence metaphors* as those that portray political concepts in terms of physical violence. Consider two statements from cable TV news in 2012 that both feature the word "attack":

(1) Because we want you to pay for your own birth control, that's an *attack* on your womb like we're flying a predator drone over your fallopian tubes and calling in a strike?[1]

(2) John McCain and his allies have been trying to turn the Benghazi attacks into a political scandal for the president since September.[2]

The first statement refers to political efforts to force employers to provide insurance that covers birth control for women. It is metaphorical because the womb is not physically assaulted. Here there is a mapping from a *source domain* of violence, associated with bodily harm, wars, battles, etc., to a *target domain* of argumentation, in this case, about who should pay for birth control. The second refers literally to a terrorist attack in the town of Benghazi; clearly, it is not

[1] Adam Carolla on *The O'Reilly Factor*, FOX News, September 10, 2012; https://goo.gl/jVBsqH

[2] Chris Matthews on *Hardball with Chris Matthews*, MSNBC, November 15, 2012; https://goo.gl/Pfs4Sc

metaphorical.

Metaphor can heighten emotions in political communication (Charteris-Black, 2009). Reporters seem to understand this well. They use metaphor to draw attention and create a reaction in readers and listeners (Lakoff, 2008). Americans have long been fascinated by the political theater afforded by television, and, over time, the media has come to frame debates as violent events (Schroeder, 2008). The trend toward increased spectacle and competitive framing continues; for instance, political campaigns are often portrayed as military campaigns (see Burnes, 2011, and Kalmoe, 2014). In the U.S. and elsewhere, political contests are now routinely conceptualized in terms of physical actions, often taken against another, such as footraces (see Matlock, 2013) or battles (see Flusberg, Matlock, and Thibodeau, 2018). Importantly, using metaphorical violence in political discourse has real consequences on reasoning; for instance, it can increase the tendency to polarize (see Kalmoe, Gubler, and Wood, 2018).

The influence and diverse range of ideological perspectives of U.S. cable television news make it an important system to understand. Interested in how metaphorically violent language would vary in reportage around debates leading up to a U.S. presidential election, we analyzed language used on the most-watched cable television news networks MSNBC, CNN, and Fox News (O'Connell, 2017). CNN and MSNBC are on the progressive end of the ideological spectrum, and Fox News, the conservative end (Pew Research Center, 2014a). Right before the 2016 presidential election, 40% of Trump voters said Fox News was their primary source of news, whereas 27% of Clinton voters said theirs was MSNBC or CNN (Pew Research Center, 2017b). For our analysis, we analyzed the use of metaphorical violence language on two different shows from each of these three networks during September 1 to November 30 in 2012 and 2016, periods in which four major political events occurred: three presidential debates and election day.

The main questions of interest concern how metaphorical violence was used leading up to election day: Which networks produce the most metaphorical violence language? Is this consistent across years? What is the contribution of each show to total use? What is the difference in how often metaphorical violence

language is used, and does this change across networks or years? Who is conceptualized more often as attacking and being attacked by metaphorical violence, and does this change across networks and time? In addition to revealing details of the use of metaphorical violence language on cable television news, informing the study of political communication and action, our results provide data for understanding a deep question in cognitive linguistics: to what extent and how does the cultural context influence which metaphors are used (Gibbs, 1997; Kövecses, 2010b)?

Our main results are a series of observations about the use of metaphorical violence language across different cultural situations and by different cultural actors. We expected metaphors to change over time in response to, or in anticipation of, the cultural events of the presidential debates and election day, and on the specific actions taken and language used by the candidates themselves. We also expected metaphor use to differ across the three networks given their differing ideologies (Lakoff, 2008), though we also expected some similarities across networks, because of shared cultural frames (Kövecses, 2010b).

To address these questions, we collected data from the Internet Archive's TV News Archive (TVNA), a curated library containing millions of short video clips from cable television news shows from the last decade. We collected data from the two most highly rated news shows on each network in each of the two study years. We relied on closed caption data provided by the TVNA to create textual transcripts of each show, and searched each transcript for words that signal, or instantiate, the source domain of violence, the *violence signal*. We considered only phrases that use one of three violence signals—*attack*, *beat*, or *hit*. If a violence signal was found in an episode of a show, a human reviewer then manually decided whether it represented metaphorical violence based on the context, annotating the text to identify subject, verb, and object of the phrase for all uses of metaphorical violence. Analyzing subject and object allowed us to determine who was portrayed as the aggressor and who was portrayed as the victim.

2.2 Methods

2.2.1 Data collection and annotation

Data were collected from the Internet Archive's TV News Archive (TVNA).[3] Using custom software to access, annotate, and analyze TVNA data, we could effectively download, review, and code hundreds of hours of news broadcasts.[4] We collected data from the two most highly rated news shows on each network in each of the two study years, relying on closed caption data to create textual transcripts of each show. We searched each transcript for words that signal violence, namely, *attack*, *beat*, or *hit*. If a violence signal was found, a human reviewer then manually decided whether it represented metaphorical violence based on context, annotating the text to identify subject, verb, and object of the phrase for all uses of metaphorical violence. Annotations were stored along the transcript, date and time, show, and network to enable later analyses.

We focused primarily on three violence signals which were far the most commonly used metaphorically among a list of twenty violence words that we initially considered. Our initial list was built based on a close reading of newspapers and other online news, and cable news transcripts. We assume there is one best interpretation of whether or not a statement is metaphorical. For this reason, we do not calculate inter-rater reliability. We have, however, made our full datasets available, including the original phrases found in cable news transcripts containing metaphorical and non-metaphorical violence, and all our annotations (https://osf.io/ypa8h/). Analyses can be re-run to see whether our results significantly change if the annotations change. Instructions for reviewing our analyses and performing your own are in the on GitHub (https://github.com/mt-digital/metvi-analysis).

We collected cable television news transcripts indexed by date, network, and show. We identified and counted daily metaphorical violence use based on the violence signals *attack*, *hit*, and *beat* (see Table 2.1). We counted the daily instances

[3] See https://archive.org/details/tv.

[4] Our custom software, Metacorps, is freely available at https://github.com/mt-digital/metacorps.

of Democratic presidential candidates (Barack Obama in 2012 and Hillary Clinton in 2016) and Republican presidential candidates (Mitt Romney in 2012 and Donald Trump in 2016) appearing as the aggressor or victim of metaphorical violence (see Table 2.3). Sometimes the aggressor and victim of metaphorical violence are clear, as a reporter on CNN's *Anderson Cooper 360* described Clinton criticizing Trump in the first debate as Clinton hitting Trump[5]:

(3) Clinton *hit* Trump for voicing support for invading Iraq and calling climate change a hoax.

The subject and object are not always explicitly specified in a single sentence, but often can be inferred. We include a reference to the video link on the Internet Archive so the reader can understand the context which leads us to our inferences. For instance, a guest on *The Rachel Maddow Show* described some of Donald Trump's comments as a metaphorical *attack* on Hillary Clinton, without saying their names explicitly in the sentence[6]

(4) One joke after another . . . was a political attack mildly veiled in humor.

[5] https://archive.org/details/CNNW_20160928_040000_Anderson_Cooper_360/start/2820/end/2880

[6] https://archive.org/details/MSNBCW_20161021_010000_The_Rachel_Maddow_Show/start/3000/end/3060

Violent Word	Network	$f^{(1)}$	$f^{(2)}$	Δ	total uses
hit	MSNBC	0.86	0.86	-0.00	67
	CNN	0.54	0.11	-0.81	34
	Fox News	0.57	0.33	-0.42	41
beat	MSNBC	1.03	1.64	0.59	89
	CNN	0.66	0.63	-0.04	51
	Fox News	0.83	0.53	-0.35	60
attack	MSNBC	1.30	3.14	1.42	127
	CNN	2.07	0.32	-0.85	128
	Fox News	2.08	2.00	-0.04	161

(a) 2012

Violent Word	Network	$f^{(1)}$	$f^{(2)}$	Δ	total uses
hit	MSNBC	0.16	0.06	-0.63	10
	CNN	0.27	0.45	0.64	25
	Fox News	0.46	1.36	1.97	56
beat	MSNBC	0.54	1.47	1.75	55
	CNN	0.50	0.79	0.59	45
	Fox News	0.48	0.88	0.84	45
attack	MSNBC	0.61	1.59	1.62	61
	CNN	1.16	2.59	1.23	126
	Fox News	1.08	4.32	2.99	160

(b) 2016

Table 2.1: Uses and Δ for violence signals on each network in 2012 and 2016.

2.2.2 Dynamical statistical model

We modeled change in frequency of metaphorical language use as an impulse function with two states:

$$f[t] = \begin{cases} f^{(1)} \text{ if } t \in T^{(1)} \\ f^{(2)} \text{ if } t \in T^{(2)} \end{cases} \tag{2.1}$$

Many more complicated models for change in frequency are possible. Here, we simply used the simplest model of change—that there is one state and then at some point later there is another. Of course, in general there many be fewer than two states or more than two states, and there is no reason to suppose there would be exactly two. Nevertheless, we have opted to use the simplest possible model, supposing there is change.

The dates for which we have data form a time series, T, of frequencies of use for each of the three networks in each of the two election years, six total. All shows do not air episodes every day. When neither of a network's two shows aired an episode on a given day, that day is not included in T. On a day that is included in T, there may be one or two episodes, so when considering dynamics, we plot and model the frequency of metaphorical violence use per episode. Frequency is simply the number of uses in a day divided by the number of episodes of the shows in that day. The time series are modeled as beginning at a mean frequency $f^{(1)}$ (State 1), then at some point later, the mean frequency changes to $f^{(2)}$ (State 2). Model fitting amounts to categorizing dates as either belonging to State 1 or State 2. These are subsets of T, with the State 1's dates denoted $T^{(1)}$ and State 2's dates denoted $T^{(2)}$:

$$T^{(2)} = \{t \in T : t^{(2)}_{first} \leq t \leq t^{(2)}_{last}\}$$

and

$$T^{(1)} = T \ \setminus \ T^{(2)}.$$

To fit parameters $t^{(2)}_{first}$ and $t^{(2)}_{last}$ to the model to minimize error, we used Bayesian

multi-model inference, which allowed us to quantify the likelihood that alternate parameterizations would better fit the observed data (Burnham, Anderson, & Huyvaert, 2011). Specifically, to determine the best-fitting model, we use Bayesian multimodel inference to infer which parameters are most likely to best represent the system dynamics (Burnham et al., 2011). Choosing a model with minimum AIC when all models have the same number of parameters is equivalent to selecting the model with minimum error, or maximum log-likelihood. Using the AIC allows us to calculate the relative likelihood of different parameterizations. Once the minimum AIC is found, call it AIC_{min}, the relative likelihood that model parameterization i outperforms the model with minimum AIC is $\exp(\frac{\text{AIC}_{min}-\text{AIC}_i}{2})$. The AIC on its own tells us nothing about how well the model matches the data, only how well the model performs relative to other models. An added feature of using this inference approach is that it reveals more about the system dynamics than if we were to simply select and use the model that minimized error. It also provides a foundation for future work that considers more complex metaphorical violence language dynamics.

Given a model we can calculate the fractional change in frequency, which we denote by Δ:

$$\Delta = \frac{f^{(2)} - f^{(1)}}{f^{(1)}}. \tag{2.2}$$

Δ, $t^{(2)}_{first}$, and $t^{(2)}_{last}$ enable us to compare changes in metaphorical violence language frequency across the networks and over time.

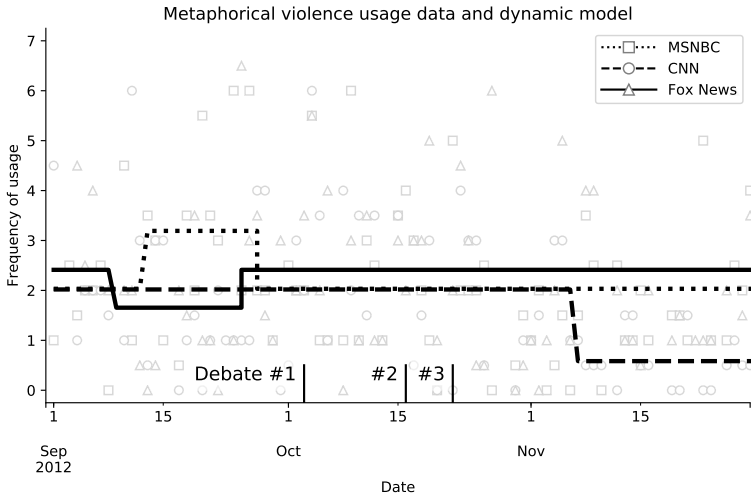
2.3 Analysis

Overall, we observed 758 uses of metaphorical violence language in 2012, and 583 in 2016. In 2012, the MSNBC show *Hardball* alone contained 208 metaphorical violence uses, whereas other MSNBC shows ranged from 60 to 120. Shows on CNN were more consistent, ranging from 99 to 118, as were shows on Fox News, ranging from 130 to 150. The distribution of specific violence signals across networks and shows was similar in both 2012 and 2016: *attack* was used most, *beat* next most, and *hit* least. Interestingly, in 2012 MSNBC led in total metaphorical violence

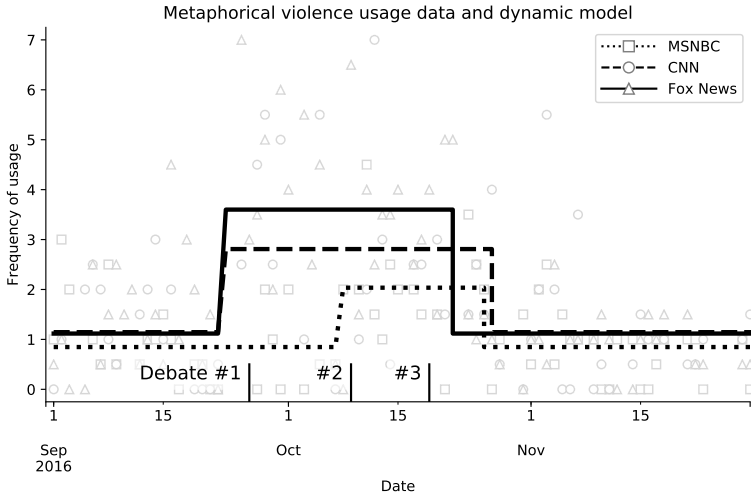
language use, and *hit* and *beat* were used more often than *attack* on that network. In both 2012 and 2016, the Republican candidate was both the aggressor and victim of metaphorical violence more often than the Democratic candidate. In 2016, Trump was characterized as doing metaphorical violence 102 times by Fox News, compared to 30 times for Clinton. This finding is consistent with other research that suggests conservatives more often conceptualize interpersonal relationships in terms of violence or are more likely to resort to violence in interpersonal relationships than progressives (Lakoff, 1996; Cohen, Nisbett, Schwarz, & Bowdle, 1996).

To compare the dynamics and time-course of metaphorical violence use across the different networks, we modeled change in frequency of use as an impulse function with two states (Equation 1). We fit our dynamical model for six time series, one for each network in each study year. Bayesian multi-model inference allowed us to identify the best-fit model and to quantify the relative likelihood of other parameterizations being better (all best-fits were significant). We next calculated change in relative frequency of metaphorical violence use, or Δ, across networks, violence signals, and clausal subject and object (Equation 2). We found both positive and negative values for Δ, meaning that metaphorical violence language did not increase uniformly within the study period across networks and years. Fox News and CNN had negative Δ in 2012. In the case of Fox News in 2012, metaphorical violence language decreased starting September 9 and ending September 25, the days leading up to the first presidential debate on October 3. CNN's use dipped after November 6, election day. In 2012, MSNBC was the only network with a positive change, starting on September 13 and ending September 27, just before the first debate. In 2016, Δ was positive and larger in magnitude for all three networks, with the start date of the elevated state overlapping to a much greater degree (see Figure 2.1). This reflects the differences in cable news viewership between 2012 and 2016: 67.2 million watched the first Obama-Romney debate in 2012 compared with 84 million for the first Clinton-Trump debate in 2016 (Perlberg, 2016). Part of this broad, synchronized excitement about the election may have been because of the big personalities of the two main contenders: Clinton was the first woman

candidate and a controversial first-lady. Candidate Trump was a rich, controversial television star.



(a) 2012



(b) 2016

Figure 2.1: Observed daily frequencies (markers) and best-fit models (lines). The dynamical impulse model is given in Equation 1. In four of the six network-year pairs, there is an increase in the frequency of metaphorical violence language in the three-month study period: MSNBC in 2012 and all three networks in 2016. However, two of the six network-year pairs showed decreases in frequency of metaphorical violence language use in one three-month period: CNN and Fox News in 2012.

MSNBC's positive Δ in 2012 resulted mainly from an increased use of the signal

Show (Network)	Total Uses
The Rachel Maddow Show (MSNBC)	93
Hardball With Chris Matthews (MSNBC)	208
Anderson Cooper 360 (CNN)	99
Piers Morgan Tonight (CNN)	118
The O'Reilly Factor (Fox News)	141
Hannity (Fox News)	133

(a) Total number of uses metaphorical violence language by news show in 2012

Show (Network)	Total Uses
The Rachel Maddow Show (MSNBC)	66
The Last Word with Lawrence O'Donnel (MSNBC)	80
Anderson Cooper 360 (CNN)	100
Erin Burnett OutFront (CNN)	118
The O'Reilly Factor (Fox News)	146
The Kelly File (Fox News)	148

(b) Total number of uses metaphorical violence language by news show in 2016

Table 2.2: Total uses by show in each of the two study years

attack. There was no change in use of the signal *hit* on MSNBC. CNN's use of *hit* and *attack* decreased by about 80%. On Fox News in 2012, most of the decrease in overall metaphorical violence use resulted from decreases in the use of *hit* and *beat*, with *attack* use remaining nearly constant. In 2016, Δ was positive for all networks. All Δ were positive for violence signals as well, with one exception: MSNBC's use of *hit* fell by 63%. MSNBC's use of *beat* and *hit* increased by a factor of almost 2. In 2016, CNN's use of *attack* accounted for most of its overall increase in metaphorical violence language use, and for Fox News, use of *attack* increased by nearly 300%.

In 2012, the candidates were involved in less of the metaphorical violence than

in 2016 (Table 2.3). Two of the three networks showed a decrease in overall metaphorical violence use at some point in the three-month study period in 2012. Even the increase in use on MSNBC was not as pronounced in 2012 as it was in 2016, with a Δ of 0.57 in 2012 and 1.40 in 2016. Changes in frequency of metaphorical violence language use were uniformly positive and larger in magnitude in 2016, beginning before the first presidential debate (September 26) and ending soon after the last debate (October 19). CNN and Fox News showed increased frequency on the same day, September 9, decreasing a few days apart, October 27 for CNN and October 22 for Fox News. MSNBC's frequency of the use of metaphorical violence language rose later, on October 8, but decreased around the same time as the other networks, on October 26.

Subject/Object	total uses	Network	$f^{(1)}$	$f^{(2)}$	reactivity	total uses
Subject=Barack Obama	88	MSNBC	0.49	0.38	-0.21	27
		CNN	0.58	0.12	-0.78	30
		Fox News	0.55	0.50	-0.08	31
Subject=Mitt Romney	100	MSNBC	0.51	0.77	0.51	33
		CNN	0.72	0.00	-1.00	36
		Fox News	0.52	0.57	0.09	31
Object=Barack Obama	113	MSNBC	0.65	0.69	0.06	41
		CNN	0.74	0.00	-1.00	39
		Fox News	0.54	0.50	-0.08	33
Object=Mitt Romney	136	MSNBC	0.63	0.77	0.22	41
		CNN	0.81	0.11	-0.86	44
		Fox News	0.83	0.79	-0.06	51

(a) 2012

Subject/Object	total uses	Network	$f^{(1)}$	$f^{(2)}$	reactivity	total uses
Subject=Hillary Clinton	74	MSNBC	0.25	0.18	-0.29	15
		CNN	0.39	0.52	0.33	29
		Fox News	0.25	0.80	2.20	30
Subject=Donald Trump	232	MSNBC	0.44	1.35	2.09	44
		CNN	0.81	1.97	1.44	86
		Fox News	0.75	2.88	2.84	102
Object=Hillary Clinton	107	MSNBC	0.21	0.41	0.98	17
		CNN	0.33	0.83	1.48	36
		Fox News	0.42	1.48	2.48	54
Object=Donald Trump	128	MSNBC	0.25	0.47	0.88	20
		CNN	0.58	0.79	0.36	44
		Fox News	0.70	1.44	1.06	64

(b) 2016

Table 2.3: Uses and Δ for Republican and Democratic candidates as subject and object of metaphorical violence.

2.4 Discussion

What might have caused the difference in timing and magnitude of changes in the level of metaphorical violence usage between 2012 and 2016? Fox News' decrease in metaphorical violence usage preceding the first debate seems to fit with the traditional role of lowering passions and expectations, and casting one's preferred candidate as the underdog (Schroeder, 2008). This explanation is supported by the content of the metaphors themselves. For example, on the September 17 episode of *Hannity*, Sean Hannity said,

(5) I want to see Romney *hit* harder. I want to see him ... take it right to (Obama).

A panelist followed up, telling Hannity, "If Romney had your passion, if Romney

had your intelligence, he would have a shot."[7] Then the next day, a contributor on *The O'Reilly Factor* said,

(6) Romney is not projecting strength. He put out a statement, he got *attacked*, and he crawled into a hole. He should have kept moving forward with what he was saying.[8]

The underdog strategy worked: Romney enjoyed a boost in the polls after the first debate. Before the debate, only 29% of survey respondents expected Romney would "do a better job" in the debate. After the debate, 72% of those who watched it thought Romney did a better job (Pew Research Center, 2012).

The increase in MSNBC's usage of metaphorical violence began September 13 and continued for two weeks in response to a statement made by Mitt Romney when he was criticizing the Obama administration's response to terrorist attacks on a U.S. compound in Benghazi, Libya. Romney called the administration's response "disgraceful" and claimed they "sympathize with those who waged the attacks." On September 13, Rachel Maddow described this statement as both "*attacking*" President Obama and

(7) *attacking* U.S. diplomatic personnel in the places that were being attacked.[9]

Maddow went on to quote a "senior republican foreign policy adviser" who said the Romney campaign was "just trying to score a cheap news cycle *hit* based on the embassy statement and now it's just completely blown up." On the same day, Chris Matthews wondered on *Hardball*,

(8) who is pushing that and saying, 'release the statement, *attack, attack, attack*'?[10]

Later on MSNBC, Rachel Maddow covered the controversial Massachusetts senate race between Republican Scott Brown and Democrat Elizabeth Warren. Maddow cast Warren, the Democrat, as the victim of metaphorical attacks. The controversy began in the first debate when Brown noted that Warren identified

[7] https://goo.gl/mc8aXk
[8] https://goo.gl/HJ6BWu
[9] https://goo.gl/QD2SFv
[10] https://goo.gl/DQULSG

herself as a Native American on school applications, but, Brown said, "You can see that she's not"[11]. Maddow first addressed this event in an interview with Rep. Barney Frank (D-Massachusetts), when she asked him his thoughts about

(9) personal *attacks* by Senator Brown against Elizabeth Warren.

Frank said it was a

(10) silly *attack* on the fact that she once said she was of Native American ancestry[12].

Over the next seven days, Maddow or a Maddow guest used metaphorical violence twelve times to describe Brown's comments on Warren's race at the debate.

As described, 2016 differed from 2012 in quantity and dynamics of use of metaphorical violence on cable television news. We now consider specific examples of metaphor use in 2016. Donald Trump made aggression an explicit character feature early on in the primary election campaign, claiming in January 2016, "I could stand in the middle of Fifth Avenue and shoot somebody and I wouldn't lose any voters, OK?"[13] Many of Trump's statements against others either originated on Twitter or were echoed on Twitter. These statements were reported in cable news as metaphorical violence. Below we first give some examples demonstrating the themes of metaphorical violence usage that made up the higher use for the three cable news channels we studied. As tweeting seemed to show up regularly in 2016, we end by calculating models relating metaphorical violence frequency to candidate tweeting.

In playing its role as cheerleaders and expectation-setters, CNN and Fox News anticipated a first debate where much metaphorical violence would be done by each candidate. This anticipation partly caused increased metaphorical violence usage. Fox News and CNN increased metaphorical violence usage on September 24 and September 25, respectively, just before the first presidential debate on September 26. It seems news outlets were expecting debates that resembled violence, not taking the underdog strategy for either candidate as in 2012. Both CNN and Fox

[11] https://www.c-span.org/video/?c4722477/attack-referred-msnbc

[12] https://archive.org/details/MSNBCW_20120921_040000_The_Rachel_Maddow_Show/start/2400/end/2460

[13] https://www.realclearpolitics.com/video/2016/01/23/trump_i_could_stand_in_the_middle_of_fifth_avenue_and_shoot_somebody_and_i_wouldnt_lose_any_voters.html

News noted the novelty of having a debate between candidates of different genders, and the new experience for Trump of debating a woman "of his same generation." On *Anderson Cooper 360* on CNN, various commenters said the following[14]

(11) Is he going to *hit* back if *attacked* tomorrow, or even if not *attacked*?

(12) There's a gender dynamic going on here. It'll be interesting to see whether he *attacks* her the way he *attacked* "Little" Marco (Rubio).

(13) (Trump) thrives on the *attack* . . . how that will work out when it's a woman of his same generation . . . that will be dramatic.

The same broadcast mentions a Trump tweet that referenced Gennifer Flowers' affair with Bill Clinton[15]. A commentator goes on to quote Jane Goodall, who said "Trump debates like a chimp in a dominance ritual." Cooper's guest explained that Trump "is not just arguing, but intimidating" his opponents[16]." On Fox news, September 25, host Megyn Kelly and guests on *The Kelly File* weighed in on debate strategy[17]:

(14) You have a column out saying she should get in his face and stay in his face . . . put him in the pain locker and shake it around. You think she should *attack, attack, and attack* some more. Doesn't she have to worry about people saying . . . sexist terms like she is a shrew, she's shrill?

(15) What she needs to do is *attack* him on many points calmly one after another.

(16) If I was Donald Trump I would really stay away from *attacking* Hillary Clinton.

During a man-on-the-street segment on *The O'Reilly Factor*, one passerby said Clinton "*beat* the [bleep] out of Donald Trump. It was like a boxing match, Hillary hit him 1, 2, bing bing."[18]

[14] https://archive.org/details/CNNW_20160926_000000_Anderson_Cooper_360
[15] https://archive.org/details/CNNW_20160926_000000_Anderson_Cooper_360/start/120/end/180
[16] https://archive.org/details/CNNW_20160926_000000_Anderson_Cooper_360/start/3180/end/3240
[17] https://archive.org/details/FOXNEWSW_20160926_010000_The_Kelly_File
[18] https://archive.org/details/FOXNEWSW_20160928_030000_The_OReilly_Factor/start/2940/end/3000

In the closing minutes of the first 2016 debate, Hillary Clinton introduced the story of former Miss Universe Alicia Machado. Machado won Miss Universe when Trump owned the competition in 1996. Clinton said Trump called Machado "Miss Piggy" because of Machado gained too much weight and "Miss Housekeeping" because Machado was born in Venezuela. This controversy reverberated throughout the rest of the presidential race. Trump spoke out on the Fox News morning show Fox and Friends the next morning, and on Twitter over the next few days, to defend his negative view of Machado. A reporter on *Anderson Cooper 360* cast Clinton's strategy as metaphorical violence on September 27

(17) The Clinton campaign had an ad ready to *hit* Trump[19].

On the morning of September 30, in the third of a series of three tweets about Machado, Trump called Machado "disgusting" and told readers to "check out sex tape."[20] The following quotes from the September 30 episode of *Erin Burnett OutFront*[21] demonstrate how metaphorical violence was used to describe the exchange of words on this issue, and the candidates' reactions and counter-reactions:

(18) Did the debate hurt Donald Trump and are his *attacks* on a former Miss Universe taking a toll?

(19) In a statement (Machado) says Trump's latest *attacks* are cheap lies with bad intentions.

(20) Trump is also *attacking* the media.

(21) Tonight Hillary Clinton hammering Donald Trump for his *attacks* on former Miss Universe Alicia Machado.

Regarding Trump's tweets, Clinton herself asked at a campaign rally, "Who gets up at three o'clock in the morning to engage in a Twitter *attack* against a former Miss Universe?"[22]

Two days before the second debate, October 7, the Washington Post published a video in which Trump brags that being famous enables him to sexually assault

[19] https://archive.org/details/CNNW_20160928_040000_Anderson_Cooper_360/start/1800/end/1860
[20] https://twitter.com/realdonaldtrump/status/781788223055994880
[21] https://archive.org/details/CNNW_20160930_230000_Erin_Burnett_OutFront
[22] https://archive.org/details/CNNW_20160930_230000_Erin_Burnett_OutFront/start/1020/end/1080

women[23]. Trump apologized for those words in a video posted to Twitter that night [24] Along with the apology in the same video, Trump accused, "Bill Clinton has actually abused women, and Hillary has bullied, attacked, shamed, and intimidated his victims," and foreshadowed "we will discuss this more in the coming days. See you at the debate on Sunday." Two quotes from a special edition of *Last Word* illustrate the coverage of this threat using metaphorical violence[25]. This is also further evidence that cable news uses metaphorical violence in their coverage of debate preparation and in expectation-setting.

(22) Clinton . . . has already been practicing for these *attacks* from Donald Trump . . . she already has her playbook.

(23) (Clinton's) team has been preparing for Donald Trump to throw every possible *attack* at her.

On October 9, the day of the second presidential debate, all three channels were in an elevated state of metaphorical violence usage. In pre-debate coverage on *The O'Reilly Factor*, Fox News anchor Tucker Carlson used metaphorical violence to describe his understanding of Trump's strategy

(24) (Donald Trump) has decided not simply to *attack* Hillary Clinton . . . but to *attack* basically the entire American establishment, the press . . . and basically the keepers of American standards.

Here Donald Trump is framed as a herculean aggressor, with the specific victims of his attacks being Clinton, the American establishment, the press, and those who value prototypical American norms of behavior.

On October 10, the day after the second debate, Donald Trump began criticizing Paul Ryan, the Speaker of the House, on Twitter. Trump wrote, "Paul Ryan should spend more time on balancing the budget, jobs and illegal immigration and not waste his time on fighting the Republican nominee." Ryan had said he was "sickened" by Trump's comments and decided to cancel a scheduled joint

[23] https://goo.gl/tk3ZNf

[24] https://twitter.com/realDonaldTrump/status/784609194234306560

[25] https://archive.org/details/MSNBCW_20161008_100000_The_Rachel_Maddow_Show; although the show ID says it's *The Rachel Maddow Show*, but it is in fact *The Last Word with Lawrence O'Donnell*

appearance with Trump (Fahrentold, 2016). Trump would criticize the Speaker five more times in the next six days on Twitter[26]. All three networks reported on this exchange using metaphorical violence. Juan Williams said this on October 15 on *The O'Reilly Factor*[27]:

(25) That is something that Donald Trump is spending time on, attacking Paul Ryan because Paul Ryan is distancing himself, but he's attacking a fellow Republican instead of broadening or shoring up his base with republicans.

In the following weeks, there were many more contentous issues which caused a series of "attacks" and counter-attacks. Among these was the Al Smith fundraising dinner, a tradition where each candidate is invited and expected to make light-hearted jokes at the other candidate's expense. Voices on MSNBC, and on the other two networks, felt Trump's jokes were mean-spirited and described the jokes as attacks. Here is one example from MSNBC's in which Senator Al Franken described some of Trump's jokes as attacks, with Clinton as the victim, at that dinner on the October 20 episode of *The Rachel Maddow Show*

(26) It takes skill to write a joke. And there were some where he just *attacked* her.

Many of these instances of metaphorical violence involve statements on Twitter. To understand the link between candidate tweeting and metaphorical violence usage, we fit a series of linear regressions with classes of metaphorical violence usage as the dependent variable and daily tweets issued by major candidates as the independent variable. This analysis also provides a further quantification of how broader cultural trends affect the timing and amount of metaphorical violence usage. This analysis demonstrates that metaphorical violence can be used as an indicator of communicative efficacy. In this case, metaphorical violence use provides a yardstick for the impact of candidate Twitter use in both election years. Our analysis confirms that, compared to 2012, 2016 was the year of the "Twit-

[26] https://www.thetrumparchive.com/?searchbox=%22Paul+Ryan%22&dates=%5B%222016-07-31%22%2C%222016-11-30%22%5D

[27] https://archive.org/details/FOXNEWSW_20161016_030000_The_OReilly_Factor/start/420/end/480

ter Election" (Heller, 2016), using metaphorical violence as a measure of Twitter impact.

In 2016, all linear model fits were significant across categories of metaphorical violence (Table 2.4). In 2012, there were still a number of statistically significant fits, but much less variance could be explained through Twitter use. About 1/3 of metaphorical violence use across all categories can be predicted from either Hillary Clinton's or Donald Trump's Twitter use in 2016. Across both years and all candidates, CNN was most reactive to Twitter use. Candidate Twitter use explained between 14% and 23% of the variance in metaphorical violence where the candidates were either the subject or object of metaphorical violence in both years.

2.5 Conclusion

Our efficient and effective approach to data collection and annotation enables new experiments aimed at understanding the dynamic relationship of language use in the media and voter attitudes. Consider that in one large scale study, online news agencies selected which news topics would be published when, and results showed that discussion of the chosen topics on social media correlated with publication of news stories (King, Schneer, & White, 2017). Whereas that study took years to implement, we believe many more natural experiments can be done using the approach we have outlined here (see also Fusaroli, et al., 2015). To understand the impact of metaphorical violence language—or any specific sort of language—we can record data from the Internet TV News Archive, concurrently polling test subjects to record, for instance, their recent TV viewing history, political opinions, use of metaphorical violence in prompts, and support for political violence to identify correlations (as in Kalmoe, 2014).

In summary, our data and analyses revealed similarities and differences in the use of metaphorical violence language on U.S. cable television news across networks and presidential election years. There were differences in how much metaphorical violence language was used and in the relative changes and timing of use across

	(Reg. coefficient, r^2) for tweets from			
Met. Vi. Category	@BarackObama	@MittRomney	@HillaryClinton	@realDonaldTrump
All	(0.01, 0.07)**	(0.11, 0.09)**	(0.04, 0.31)***	(0.06, 0.33)***
MSNBC	(0.01, 0.01)	(0.10, 0.04)	(0.02, 0.05)*	(0.05, 0.05)*
CNN	(0.02, 0.07)**	(0.15, 0.05)	(0.04, 0.20)***	(0.12, 0.20)***
Fox News	(0.01, 0.02)	(0.09, 0.02)	(0.04, 0.14)**	(0.05, 0.13)***
Self as subject	(0.01, 0.23)***	(0.06, 0.21)***	(0.01, 0.17)***	(0.03, 0.18)***
Other as object	(0.01, 0.16)***	(0.05, 0.14)***	(0.01, 0.20)***	(0.01, 0.14)***

*p<0.1; **p<0.05; ***p<0.01

Table 2.4: Regression coefficients, and significance indicators for linear models of metaphorical violence usage as a function of the number of tweets from individual candidates. The regression coefficient represents the additional metaphorical violence uses that occur with each message the candidate tweets. of variance that is represented through a linear relationship with candidate tweets. The 2016 candidates's Twitter use had a greater impact on metaphorical violence usage than the 2012 candidates's. In both years, Twitter use had a strong effect on metaphorical violence use where the tweeting candidate was cast as the subject of metaphorical violence, or where the other candidate was the object of metaphorical violence.

networks and years; for instance, in some cases, metaphorical violence language use increased around presidential debates, and in others, it decreased. There were similarities in the details of use of specific violence signals and in which party's candidate was most involved in metaphorical violence; for instance, *attack* was used most often, and Republican candidates were represented as either the aggressor or the victim of metaphorical violence more than Democratic candidates. Thus, our study has provided detail and perspective on the workings and dynamics of metaphorical violence in political discourse. Previously, metaphorical violence was known to be a feature of political communication, but its extent and dynamics were not known. We have shown that use of metaphorical violence language can change substantially over a short period, both in amount and in kind, in response to external actions and cultural events. We have shown that different political perspectives make different use of metaphorical violence language. Yet there is still a lot more we do not know. We know little about the relationship between metaphorical violence language used on television and actual violent actions. Some may infer cause and effect, as the suggestion that observing violence in video games leads to tolerance for and actions of violence in the real world (Calvert et al., 2017). Others may simply see the use of specific metaphorical language primarily for purposes of political persuasion (Charteris-Black, 2009; Mio, 1997). In a time of ever-more optimization and automation, we must consider carefully how to shape political discourse to create the desired outcomes. Our results are one step in the direction of understanding how use of specific language influences political attitudes.

Acknowledgements

We thank Oana David and Jamin Pelke for helpful and inspiring discussions. We also thank Isabella Methot, Gloria Quintana, and Amy Tang for assistance with annotating

Chapter 3

Stubborn extremism explains and predicts group polarization

Group polarization is the widely-observed phenomenon in which the opinions held by members of a small group become more extreme after the group discusses a topic. For example, conservative individuals become even more conservative, while liberal individuals become even more liberal. Social psychologists have offered competing explanations for this phenomenon. These typically require questionable assumptions about human psychology. Here, we posit a more parsimonious explanation: the stubbornness of extreme opinions. Using agent-based modeling, we demonstrate that such "stubborn extremism" gives rise to group polarization as observed across the literature on polarization. We conclude with an evaluation of stubborn extremism and existing explanations to identify opportunities for theoretical integration.

3.1 Introduction

Group polarization is a phenomenon in which the opinions held by members of a small group become more extreme after the group discusses a topic (Myers, 1982; Brown, 1986; Isenberg, 1986; Sunstein, 2002; Sieber & Ziegler, 2019). This phenomenon is socially important for many reasons. First, small groups of advisers often influence executive decisions in government and business. At the "grass

roots" level in politics, individuals discuss important issues first in small groups before they vote. Second, group polarization at the local level increases overall polarization at the societal level. Polarization, measured as the bimodality of the distribution of opinions in a group or society, increases whenever either of two opposed groups becomes more extreme (Bramson et al., 2016). Many studies of political polarization frame the issue in terms of intergroup conflict (Mason, 2018; Klein, 2020). However, we also must understand how group polarization can exacerbate political polarization through increased in-group extremism without an explicit out-group. Understanding the cognitive mechanisms supporting group polarization is therefore a matter of concern.

Social psychologists have offered several explanations for group polarization, but four are considered acceptable today (Sieber & Ziegler, 2019). First, *social comparison theory* posits that individuals' privately-held opinions tend be more extreme than those they express publicly, and exposure to consonant opinions gives them confidence to express their true opinions openly (Myers, 1982). It is not clear, however, when or if people really do hold more extreme views than they tend to express. Second, *persuasive arguments theory* posits that when individuals discuss a topic within an already-biased group, they accumulate more persuasive arguments supporting those biases, leading to a more extreme version (Bishop & Myers, 1974; Vinokur & Burstein, 1974). This is problematic because it lacks inclusion of arguments for moderation, explaining that moderation comes from knowing arguments for each polar opinion. Third, *self-categorizaztion theory* explains group polarization as emerging from the desire of individuals to consolidate group membership by expressing more extreme opinions, which further contrasts individuals from hypothetical out-groups (J. C. Turner & Wetherell, 1987; Abrams, Wetherell, Cochrane, Hogg, & Turner, 1990; McGarty, Turner, Hogg, David, & Wetherell, 1992). This is problematic because it is not clear how the calculation works that would determine how much one individual should shift their opinions to more clearly signal group membership. Finally, *social decision schemes theory* explains group polarization as the result of two main factors, namely the distribution of individual-level traits that determine individual opinions and the method

by which groups make collective decisions (Zuber, Crott, & Werner, 1992; Friedkin, 1999). If extremists are more powerful in groups on average, then group decisions on collective opinions will become more extreme after group deliberation. This is problematic because it may be difficult to determine a novel group's decision scheme *a priori*. These explanations may explain the empirical phenomenon of group polarization, though more formal modeling is required to bring precision to the underlying theories (Smaldino, 2017b, 2019).

We present an alternative explanation for group polarization that, while not mutually exclusive with the other theories discussed, manages to explain the phenomenon of group polarization without assuming anything about the intrinsic distribution of extreme opinions in human groups. This is important because the extant theories of group polarization outlined above make auxiliary assumptions that add needless complexity and are perhaps not well supported. We demonstrate that group polarization emerges when we assume the psychological property of *stubborn extremism*: as a person's opinion on some topic becomes more extreme, that opinion also becomes more stubborn, i.e. less susceptible to social influence. We support this explanation using a computational model of group polarization. Our model was originally developed for explaining how polarization emerges where two groups become more extremely opposed (Flache & Macy, 2011; M. A. Turner & Smaldino, 2018). The model incorporates both negative, repulsive social influence (Cikara & Van Bavel, 2014), assimilative influence, and the stubborn extremism assumption, though repulsive influence is not at work in group polarization because all opinions start out similarly valenced.

Theories can be thought of as one or a few central assumptions, or hypotheses to be supported, conjoined with auxiliary assumptions about how a phenomenon of interest emerges (group polarization being our phenomoenon of interest). Together all these assumptions form the antecedent in a logical implication, with the phenomenon of interest as the consequent. Parsimony is a useful measure of how many central or auxiliary assumptions a theory makes—more parsimonious theories explain the same empirical observations (group polarization here) with fewer assumptions. More parsimonious theories are generally preferred because they are

easier to interpret. One reason more parsimonious theories are often preferred are because they are easier to interpret.

We use computational modeling to simulate social behavior that leads to group polarization under a set of empirically motivated assumptions, including the stubborn extremism assumption. Such modeling enables us to build a simulated world where group polarization is not assumed, only a small set of basic psychological mechanisms. In this way, our computational model is a mechanistic model provides a concrete system that encapsulates a theory's assumptions that make up the antecedent of our scientific proposition. We use this approach to demonstrate that the relatively minimal assumptions of stubborn extremism, along with a few other context-independent cognitive factors, can lead to the emergence of group polarization in computer simulations.

Group polarization can emerge computationally by simply assuming agents hold binary opinions on a multitude of topics (Mueller & Tan, 2018; Banisch & Olbrich, 2019). However, most group polarization studies do not measure participants' binary opinions (e.g., for vs. opposed) on a multitude of topics, but rather measure opinions as falling on a range between strongly for and strongly opposed. Furthermore, the assumption of discrete opinions is problematic from a psychological perspective, since it is rare for quantum leaps in opinion to occur—more often we are influenced gradually over the course of many interactions (Baldassarri & Bearman, 2007, p.793). Our model is most similar to that of Baldassarri and Bearman (2007) in that stubbornness is a function of opinion extremity directly. Martins and Galam (2013) allow for agents to become more or less stubborn, but assume discrete opinions and a separate, continuous measure of open-mindedness/stubbornness. Most other opinion dynamics models that link stubbornness to extremism assume infinitely stubborn extreme agents (sometimes called "zealots") whose opinions are static and whose existence is specified *a priori* by the modeler (Galam & Jacobs, 2007; Mobilia, Petersen, & Redner, 2007; Arendt & Blaha, 2015; Mueller & Tan, 2018). Baldassarri and Bearman (2007) nearly make the connection between stubborn extremism and group polarization, but they mischaracterize group polarization and discuss it in terms of negative influ-

ence, writing "interaction with dissimilar others may increase distance, leading to group polarization" (p. 792). Group polarization experiments are designed so that this rarely, if ever, occurs. Instead, it is only interaction among relatively like-minded individuals that leads to the group polarization opinion shift.

Behavioral studies across disciplines support the stubborn extremism explanation. Zaller (1992) and Converse (2006) established that, at least at the time of their studies, most of the United States electorate, for example, were relatively ignorant of real political issues and easily swayed by momentary predilections and the framing of questions. Guazzini, Cini, Bagnoli, and Ramasco (2015) found that stubborn extremists drove the opinions in groups discussing the use of animals in laboratory experiments, and Lewandowsky, Pilditch, Madsen, Oreskes, and Risbey (2019) found that stubborn extremists have an outsized influence in the perpetuation of scientific misinformation regarding climate change. Group polarization opinion shifts have been observed to increase with the group's initial extremity (Teger & Pruitt, 1967; Myers & Arenson, 1972; Myers, 1982; Brown, 1986). This has only been tested in detail by Teger and Pruitt (1967) and Myers and Arenson (1972), apparently, and has not been established for political opinions. This could cause acceleration of political polarization. Some researchers have suggested that stubbornness is an attribute found generally among people, and is not limited to those with extreme opinions. However, support for this view often comes from studies in which opinions are operationalized as answers to general knowledge tests (such as found in a pub quiz), and not on opinions with political or ethical components in which subjective judgment plays a larger role (Moussaïd, Kämmer, Analytis, & Neth, 2013; Chacoma & Zanette, 2015). More direct empirical tests of the stubborn extremism explanation for group polarization are needed.

The rest of the paper is organized as follows. We first review evidence and explanations for group polarization in more detail. We will then introduce an agent-based model of opinion dynamics with stubborn extremists, which is adapted from previous work by Flache and Macy (2011), and we will demonstrate how the model supports the stubborn extremism hypothesis. We will then compare our model to the persuasive arguments model of Mäs and Flache (2013), and

show how our model can yield a fit to the empirical dataset they test that is at least as congruent. We conclude with limitations of our model's assumptions, and suggestions for future work.

3.2 Group polarization theory, methods, and results

Initially, the group polarization effect was thought only to apply to opinions about how much risk would be appropriate to take given some life decision (Wallach & Kogan, 1965; Teger & Pruitt, 1967; Stoner, 1968)—the so-called "risky shift". Moscovici and Zavalloni (1969) then showed that deliberation about political opinions also led to group polarization. Motivated by this and by Cartwright's calls for increased precision in group polarization theory (D. Cartwright, 1971, 1973), new explanatory mechanisms were proposed. The two explanations that survived to today are the social comparisons theory (Brown, 1974; Sanders & Baron, 1977; Myers, 1978) and persuasive arguments theory (Burnstein & Vinokur, 1973; Vinokur & Burstein, 1974; Burnstein & Vinokur, 1975). Around the time of Isenberg's (1986) review, a self-categorization explanation (J. C. Turner & Wetherell, 1987) of group polarization was developed and supported with new empirical studies (J. C. Turner, Wetherell, & Hogg, 1989; Abrams et al., 1990; Hogg, Turner, & Davidson, 1990; McGarty et al., 1992; Krizan & Baron, 2007). Following the development of social comparisons, persuasive arguments, and self-categorization theories, social decisions scheme theory that identifies social power structures as a dominant factor in the emergence of group polarization (Zuber et al., 1992; Friedkin, 1999). More recently, focus on the correlation between stubbornness and extremism has emerged as a simple, empirically-motivated explanation of group polarization (Mueller & Tan, 2018; Banisch & Olbrich, 2019). However, existing studies do not allow extremism to emerge naturally, but instead posit the problematic existence of infinitely stubborn extremists who are totally unsusceptible to social influence. Because it is unlikely that anyone is totally unsusceptible to social influence, we allow individuals to become more stubborn as they become

more extreme.

In this present study we create a model group polarization experiment with the goal of demonstrating the explanatory power of the stubborn extremism explanation. We believe that this explanation is more parsimonious in that it makes fewer, simpler assumptions about human behavior while accounting for both group polarization itself and the empirical observation that the group polarization opinion shift increases with initial group extremism. In order to understand the model and computational analysis it is necessary to first identify common experimental design elements in group polarization studies. Second, it is necessary to understand in more detail how stubborn extremism explains group polarization and how it relates to other theoretical explanations of group polarization. These topics are reviewed in the remainder of this section.

3.2.1 Common experimental design elements

Group polarization studies all follow the same general experimental paradigm, with slight variations to test particular theoretical explanations or real world situations. In this paradigm, participants first answer questionnaire items or somehow give their opinions or positions on some situation. Small groups typically of 2-6 participants are formed such that the mean opinion or position of group members is non-neutral, baised towards one or the other extreme of the measurement scale. Group formation is sometimes based on initial participant answers to the questionnaire, but sometimes uses some other method such as a different questionnaire (Myers & Bishop, 1970) or geographic location that is correlated with individual opinion (Schkade, Sunstein, & Hastie, 2010). Political questionnaires are common choices. For example, Moscovici and Zavalloni (1969) asked Parisian lycée students about their opinions of then-president Charles de Gaulle and of American foreign policy. More recently, Schkade et al. (2010) asked US residents of Colorado about affirmative action, same-sex civil unions, and global warming.

Many studies using questionnaires prompt participants to give their responses on an ordinal, Likert-type scale. Stoner's (1961; 1968) choice dilemma questionnaire was a 10-point ordinal scale, with 1 representing the most risk acceptance

and 10 representing the least risk acceptance. When French students answered "American economic aid is always used for political pressure", they marked a whole number on a seven-point scale from -3 (strongly disagree) to +3 (strongly agree), with zero representing neutral or no opinion. These scales do not always include 0 as the neutral point. Schkade et al. (2010) used a ten-point scale from 1 (disagree very strongly) to 10 (agree very strongly).

Non-questionnaire group polarization studies have used a variety of methods. In one approach, researchers simulate jury deliberations for an experimental design where participants give either opinions on whether a defendant is guilty or how much money for damages should be awarded (Kaplan, 1977; Kaplan & Miller, 1977; Schkade, Sunstein, & Kahneman, 2000; Schkade, Sunstein, & Hastie, 2007; Sunstein, 2000)[1]. Another approach studied group polarization in the context of gambling behavior in the game of blackjack (J. Blascovich & Ginsburg, 1974; J. I. M. Blascovich, Ginsburg, & Howe, 1975; J. Blascovich, Ginsburg, & Howe, 1976), which found that participants demonstrated opinion shifts to be more risky merely when exposed to other group members' bets.

In this study, we are only interested in the effect of explanatory assumptions and can ignore details of the measurement schemes. Therefore, we assume that we can directly observe people's opinions as we do in our computational models and analyses. Our model simulates the three stages of group polarization experiments identified above: (1) administer survey to participants to poll pre-deliberation opinions; (2) participants deliberate about their opinion in small groups; and (3) poll participants' post-deliberation opinions.

3.2.2 Theoretical explanations of group polarization

Below we review the four explanations or theories of group polarization assumed or evaluated by the case studies we investigated for false detections. We also review select empirical support for each explanation. The explanation of group polarization as due to the stubbornness of extremists comes from empiri-

[1]Schkade, et al., (2000), entitled "Deliberating about Dollars: The Severity Shift", was funded by Exxon Company, U.S.A., who have a clear interest in understanding what causes individuals to raise or lower the amount of damages they believe a responsible party should pay.

cally motivated modeling projects that have yet to be verified empirically in group polarization settings. Therefore, we do not review that here. Future work will use the results of the present paper to devise more appropriate measurement and statistical procedures that will help ensure the validity of future empirical studies.

Following our theoretical review, we identify and explain common experimental design elements and statistical methods used commonly across group polarization research independent of theoretical aims and assumptions. Then in this section we review findings from these decades of research, which overwhemingly support group polarization in general—each theory can boast supporting empirical evidence as well. This sets up the following section where we explain our model in mathematical detail that we will use to show that we should be highly skeptical of the broadly supportive evidence for group polarization.

Social comparisons

When researchers began searching for an explanation of group polarization in response to Cartwright's (1971; 1973) critiques of the "risky shift" literature, some adapted the extant

The social comparisons explanation of group polarization adapts the "theory of social comparison processes" (Festinger, 1954) of group-level social influence as an explanation. This theory assumes that when people interact in group settings, each individual infers what the prevailing social norms are, compares their own opinion to the social norm, and adjusts one's own opinions or behaviors so they are more socially accepted or celebrated. One testable corollary of this explanation is that no deliberation is required, *per se*. All that is required is "mere exposure" to others' opinions (Zajonc, 1968; Burgess & Sales, 1971; Bornstein, Kale, & Cornell, 1990; Montoya, Horton, Vevea, Citkowicz, & Lauber, 2017). Several studies have shown that non-verbal displays of individual opinions to the group is alone sufficient for inducing group polarization, without verbal deliberation (Teger & Pruitt, 1967; J. Blascovich, Veach, & Ginsburg, 1973; J. I. M. Blascovich et al., 1975; J. Blascovich et al., 1976; Sanders & Baron, 1977; Myers, 1978, 1982).

Just because mere exposure to others' opinions tends to lead to group polar-

ization does not necessarily support all auxiliary assumptions made by the social comparisons explanation (Meehl, 1990). It is not clear what the mechanism is by which individuals infer the group norm if it is not just the average. How is it, exactly, that individuals infer this more extreme than average group norm? Festinger (1954) assumes first that "there is a universal human drive to evaluate our opinions and abilities" (Brown, 2000, p. 78). But how ubiquitous is this drive to distinguish oneself through conformity? Clearly individuals vary in their drive to conform to social norms in general—how does this affect group polarization opinion shifts? Furthermore, achieving distinctiveness through conformity may have counterintuitive effects (Smaldino & Epstein, 2015a). Social comparisons theory fails to make contact with extensive literature on norms and norm change, which should be accounted for (Bicchieri, 2006; Bicchieri & Mercier, 2014; Bicchieri, 2017).

These are important questions to answer. Perhaps social comparisons offers a good starting point for a partial explanation of group polarization, but its epistemological status is shaky. It is therefore important that we understand how to properly measure opinion shifts to either support, refute, or revise and incorporate the social comparisons account into a broader explanatory model of group polarization.

Persuasive arguments

Persuasive arguments theory explains that opinion change is determined by the number and persuasiveness of arguments that support different poles of the opinion scale. Arguments, then, are central theoretical entities in this model alongside opinions. If there are more arguments favoring one polar opinion (disagree/agree) over another (Ebbesen & Bowers, 1974), or if arguments that exist for one polar opinion are more persuasive then the group will collectively move towards that polar opinion (Vinokur & Burstein, 1974; Burnstein & Vinokur, 1977). This theory assumes that for an argument to have an effect on a participant, that participants must not have heard the argument before (Bishop & Myers, 1974, see Equation on p. 96). Furthermore, the validity, or informativeness, is hypothesized to be the primary auxiliary factor in determining the magnitude of influence for a given

argument (Vinokur & Burnstein, 1978).

One problem with the persuasive arguments explanation is that only arguments are persuasive, not people. Perhaps, for example, there is a simple consistency in that more extreme individuals tend to be more persuasive than moderates, perhaps due to their confidence in their opinions. This assumption would actually explain observations made by Burnstein and Vinokur (1973) who found that insincere arguments are not influential. Another related problem is underspecified psycholinguistic mechanisms of social influence. Perhaps novelty and informativeness are two important factors in what makes an argument persuasive. Surely, though, there are other factors.

Self-categorization

The self-categorization explanation of group polarization posits that people conform to others' attitudes, opinions, or beliefs, by considering how best to "contrast" themselves with members of an out-group so as to consolidate their membership with an in-group (Tajfel et al., 1971a; Tajfel & Turner, 1979; J. C. Turner & Wetherell, 1987). Experiments testing the self-categorization hypothesis use the minimal group paradigm approach to understand differences in social influence (that leads to extremism) between in-group members versus out-group members. In one interesting counter-example to the persuasive arguments theory, the basic experimental design was used, but participants did not interact with a group—instead they were listened to tape recordings of arguments for or against some statement. Participants were told they would either be joining the group or that they were listening to members of an out-group. This changed whether opinion shifts were to a greater extreme they were already bised towards (in-group) or if participant opinions tended to shift away from their initial bias (out-group) (McGarty et al., 1992). Persuasive arguments theory does not account for group membership, so it could not have predicted this result. The minimal group approach continues to be applied today across cognitive sciences, especially in understanding the neuroscience of emotions towards novel in- and out-groups (Cikara & Van Bavel, 2014; Molenberghs & Morrison, 2014).

To explain group polarization, where there no explicit out-group, self-categorization theorists proposed that people engaging in social interaction mentally calculate the "metacontrast ratio", which is defined as a person's average distance in opinion space from all out-group members divided by that person's average opinion distance from all in-group members (McGarty et al., 1992, p. 3). This requires them to infer their average distance to the imagined outgroup. A person is then hypothesized to update their opinions to match the prototypical opinion, which is defined as "the pre-test mean where the mean is at the mid-point of the comparative context. . . ." This supposedly leads to group polarization, since "(a)s in-group responses shift. . . towards a more extreme position, then it becomes more likely that the prototype will tend to be more extreme than the mean in the same direction" (p. 4, *ibid*).

While neuroscientific studies implementing the minimal group paradigm support the assumption that differential social influence depends on whether an individual interacts with in-group or out-group members (Cikara & Van Bavel, 2014), it is not clear that it operates as hypothesized in self-categorization explanations of group polarization. Specifically, the assumption that people calculate metacontrast ratios and hypothetical in-group prototype opinions does not seem to be empirically supported. It is not clear to us how such a claim could be empirically supported. Another possible critique is that this reasoning seems to be circular: the in-group prototype begins as the pre-deliberation mean, but changes once opinions begin to change. This seems to sidestep the problem of how opinions change in the first place and why the average opinion tends to become more extreme. Finally, it seems that perhaps "prototype" in the self-categorization explanation is homologous in form and function to a "norm" in social comparisons theory. Future work should explore this connection in more detail to understand exactly how the two theories substantively differ.

Social decision schemes

Social decision schemes generally considers the social structure of groups to determine what opinions or behaviors group members will take in the course of group

interaction (Davis, 1973). In the social decision schemes framework, individual-level interaction strategies are hypothesized and specified. To understand social decision schemes, consider the following example adapted from Brown (2000, p. 195). Assume a group is trying to solve some problem. The group may be composed of three types of people: (1) people who are able to solve the problem, (2) people who can recognize a solution but not solve the problem themeselves, and (3) people who cannot solve the problem or recognize a correct solution. The group may adopt different decision rules, such as "Truth wins" (as long as one member has the solution, the group solves the problem), "Majority rule" (a majority of group members must know or recognize the solution), or "Unanimous" (all group members must know or recognize the solution). If we know the composition of the group in terms of these three types, then we can calculate the probability that a group solves the problem. According to the social decision schemes framework, if we observe how often a group solves a problem and we know the distribution of strategies, we can infer the decision rule used by the group.

In the context of group polarization, instead of recognizing solutions to problems, people are assumed to adopt a strategy of "risk wins", "conservatism wins", or "majority wins" in the context of the choice dilemma questionnaire (Laughlin & Earley, 1982; Zuber et al., 1992). Friedkin (1999) developed a network theoretic model that aligned with the social decision schemes approach, but focused on power structures that determine relative social influence. When extremists are more powerful, one would expect group polarization to emerge. Friedkin ran behavioral experiments to support his explanation, but unfortunately, several of Friedkin's results are *prima facie* null, since several of the confidence intervals around the opinion shift measurements include zero.

One issue with the social decision schemes approach seems to be that the emergence of distribution of strategies, and the strategies themselves, is not accounted for. How does such a norm as "risk wins" emerge? How is this not a "norm" or "prototype" as could be found in either the social comparisons or self-categorization explanations, respectively? Because norms may indeed be important for group polarization, future theorizing should consider how norms emerge and culturally

evolve (Bicchieri, 2006; Bicchieri & Mercier, 2014; Bicchieri, 2017).

3.3 The model

We developed an agent-based model to demonstrate the stubborn extremism model predicts group polarization patterns reviewed above. Our goal is to demonstrate that the relatively minimal assumption of stubborn extremism can predict observed patterns group polarization opinion shifts. This model allows for both positive and negative influence, wherein initially similar agents become more similar after interacting, while initially dissimilar agents become more polarized. The model is identical to that studied previously in Flache and Macy (2011) and M. A. Turner and Smaldino (2018), but is analyzed here with a different focus than was used in those studies.

We consider a population of N agents, who each have opinions on one topic. This model can account for social influence across multiple opinion topics, but one suffices for our purposes. Future work could consider the effect of deliberation on multiple opinions, which has been shown to foster cultural fragmentation (DellaPosta et al., 2015). Agent i's opinion at time t is written $o_{i,t} \in (-1, 1)$ and changes after i has interacted with its N_i network neighbors. The weight of social influence with each neighbor j is $w_{ij,t}$, with zero direct influence over non-neighbors. Weights depend on the Manhattan distance between agents i and j: $d_{ij,t} = |o_{i,t} - o_{j,t}|$. The specific operation of these social influence mechanisms is defined by the following dynamical equation

$$o_{i,t} = o_{i,t-1} + \Delta o_{i,t}(1 - |o_{i,t-1}|^{\alpha}) \tag{3.1}$$

where

$$\Delta o_{i,t} = \frac{1}{2N_i} \sum_j w_{ij,t}(o_{j,t} - o_{i,t}) \tag{3.2}$$

and

$$w_{ij,t} = 1 - d_{ij,t}. \tag{3.3}$$

Our model includes both positive and negative influence. Positive influence is when agents become increasingly similar to their dyad partner if the pair are sufficiently

similar to begin with ($d_{ij} < 1$). Negative influence is when interaction causes a dyad to become more different, to be repulsed away from one another toward more extreme regions of opinion space if the pair are sufficiently dissimilar to begin with ($d_{ij} > 1$). This is important for group polarzation because while a group overall may be biased towards one extreme, in general there may be group members who lean towards the opposite opinion pole—in these situations sometimes dyads become more different when they interact instead of more similar (Bail et al., 2018). The parameter α determines the degree to which extreme opinions are stubborn. In the analyses presented here, we use $\alpha = 1$. Stubborn extremism emerges in our model due to the smoothing factor $(1 - |o_{i,t-1}|)$, which is smaller when $|o_{i,t-1}|$ is larger. Therefore, more extreme opinions (larger $|o_{i,t-1}|$) are less susceptible to social influence than less extreme opinions (smaller $|o_{i,t-1}|$).

Our model generates a number of empirically-observed outcomes. First, we show that our model yields group polarization in an idealized generic case that resembles the studies of Moscovici and Zavalloni (1969), Myers and Bishop (1970), and Myers and Lamm (1975). For our computational experiments, we set the number of agents in the population to $N = 25^2$. The social network for this first experiment was fully connected, meaning all agents could potentially influence all other agents. Second, we represent the Mäs and Flache (2013) empirical experiment with our model and show our model predicts their empirical observations as accurately as their computational model of persuasive arguments theory.

3.3.1 Computational experiments

Our first experiment examined the correlation between initial mean opinion and shift magnitude. This also establishes that our model generates group polarization. Initial agent opinions were drawn from a normal distribution with $\sigma = 0.25$. In order to demonstrate that our model predicts the correlation between opinion shift and initial opinion extremity, we ran the model with seven different experimental conditions. Each of the seven conditions specified a different mean for the normal

[2]This is much larger than real group polarization experiments, but served to generate group polarization shifts in a shorter number of time steps for a proof of concept. This will need to be made realistic for the journal article.

distribution from which initial opinions were drawn, $\mu \in \{0.2, 0.3, \ldots, 0.8\}$. For each condition we ran 100 trials. Since opinions are bounded between ± 1 and group polarization experiments force group members to have opinions of the same valence, we re-mapped any drawn opinions greater than 1 to be $+1$ if the drawn opinion was greater than 1, and 0 if the drawn value was less than 0. Each model run consisted of 100 rounds of agent interactions. In one round of agent interaction, N agents are selected at random to update their opinions according to Equation 3.1. To model a typical group polarization experiment with open discussion, we assume a fully-connected network, so all agents influence one another.

Our second experiment was designed to generate the results of Mäs and Flache (2013). Here we utilized the multidimensionality of opinions to represent different "persuasive arguments" that participants held. To do this, we set $K = 12$, the total number of persuasive arguments available to each agent in Mäs and Flache's study, and initialized three of the twelve opinions to be non-zero. Recall that in their study, Mäs and Flache provided individuals with one of twelve pre-defined "arguments" they were to share with others to advocate for their opinion. Six of the twelve were chosen as pro-A arguments and six of the twelve were chosen as pro-B arguments. The pro-A arguments were given initial values of $-1/3$ and pro-B arguments given initial values of $1/3$. In our adaptation of this experimental setup, we are using each of K elements of agent i's opinion vector to represent the presence or absence of an argument. As in the Mäs and Flache study, group "A" members all received the same initial pro-B argument, and vice versa. To calculate each agent's scalar opinion based on its $K = 12$ "persuasive argument" components, we first normalize opinions so their absolute values sum to 1, and then averaged over all opinions. This is similar to the persuasive argument model that assumes an individual's opinion is an aggregate of the arguments they know for their position. This computational experiment mirrors Mäs and Flache's persuasive arguments model, but includes stubborn extremism. Furthermore, in our formulation, agents can partially agree or disagree with a given argument, unlike persuasive arguments which assumes an agent either knows an argument or not. For our computational experiment's outcome measure, we calculated the average over all agent opinions in

each group at each timestep, and then averaged those averages across 100 trials at each timestep, identical to Mäs and Flache's procedure for obtaining their results (Figures 5 and 6 of their paper).

3.3.2 Implementation

The model was implemented as an agent-based model written in plain Python with user-defined `Agent`, `Model`, and `Experiment` classes. We use NumPy and SciPy for numerical and scientific routines and functions. For full implementation details including instructions for installing and running model code and reproducing our results, please visit the GitHub repository, `Mhttps://github.com/mt-digital/group-polarization`. Our computational experiments easily run on a laptop.

3.4 Analysis

Our model predicts that more extreme initial group opinion results in larger shifts up to a certain extremity where the trend reverses (Figure 3.1). In terms of stubborn extremism, this general trend is expected because there will be more extremists when the initial mean is greater. These initial extremists exert a greater pull towards extremism when they are more numerous. However, when many agents are extreme and there are few neutral agents to be shifted to more extreme views, the shift begins to decrease in magnitude compared to the maximum shift over initial mean (occurs at initial mean of 0.8 in Figure 3.1).

Our model predicts group polarization as observed by Mäs and Flache (2013), but via the assumption of stubborn extremists instead of persuasive arguments. Our model predicts the same initial increase in the extremity of the average group opinion for both A- and B-Type agents as predicted and observed in Mäs and Flache (2013). Then when A-Types and B-Types interact with one another, our model predicts consensus emerges, as was observed by Mäs and Flache's experiments and predicted by their model (Figure 3.2 above; compare with Figure 6 Mäs and Flache (2013)). Note that, in our model, no explicit persuasive arguments

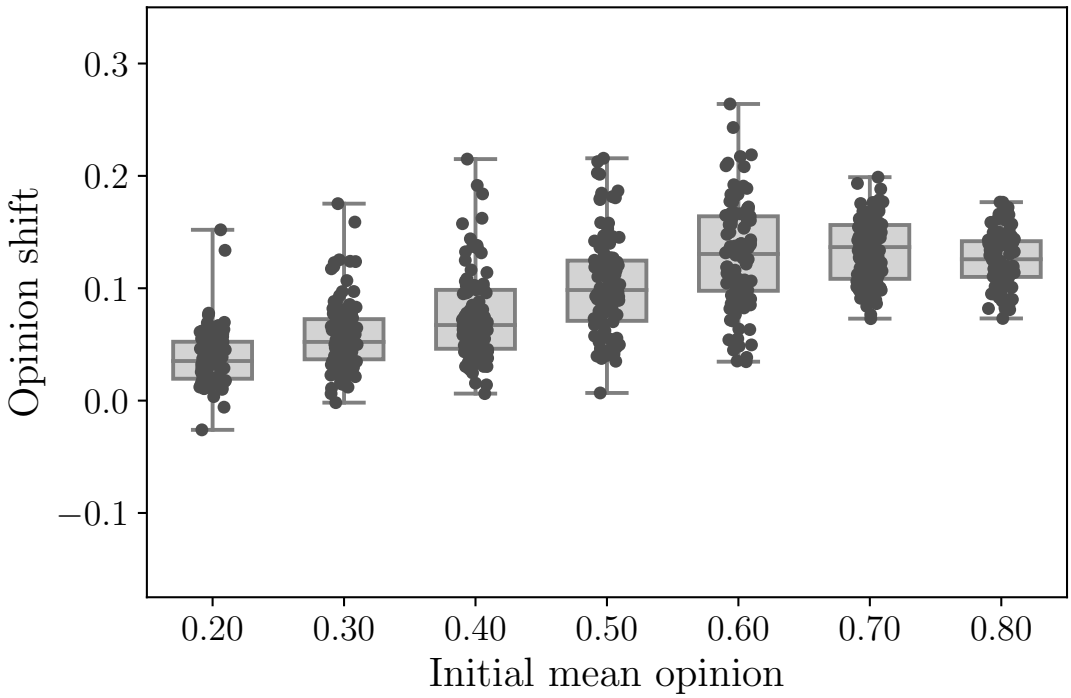


Figure 3.1: Group opinion shift when individuals' initial and final opinions are given on a continuous scale. Stubborn extremism leads to group polarization and predicts that opinion shift is positively correlated with mean initial group opinion.

are exchanged. Instead, each argument is represented as an opinion on a certain cultural topic. Influence occurs on all cultural topics, and similar group members draw one another closer in hypothetical 12-dimensional opinion space through attractive social influence and stubborn extremism, resulting in group polarization.

3.5 Discussion

We have shown that stubborn extremists are a feasible explanation for group polarization. Our model that incorporates this simple mechanism predicts behavior observed in a number of empirical studies. These empirical studies have often considered two alternative pathways to group polarization: *persuasive arguments* and *social comparisons*. The persuasive argument theory explains that group polarization occurs because individuals are exposed to more arguments supporting

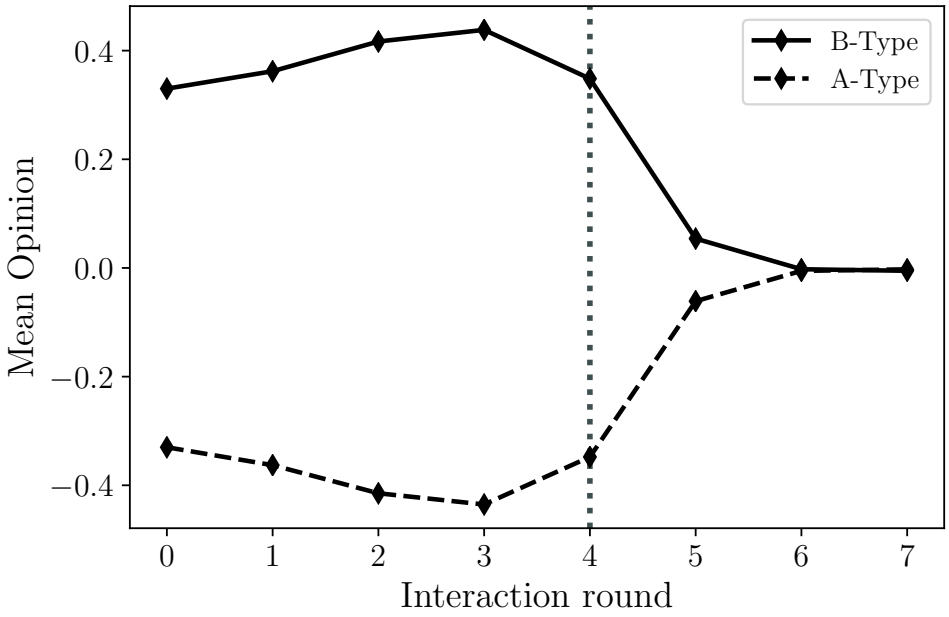


Figure 3.2: Our model's prediction of group opinions in the Mäs and Flache (2013) study. Within-group interactions are rounds 1-3, intergroup interactions are rounds 4-7.

their initial position in contrast with the opposing opinions, thereby strengthening that opinion. At the group level, this leads the average opinion to shift towards an extreme. Alternatively, social comparison theory posits that group polarization is due to group members calculating some optimal opinion to express publicly that takes into account both their private opinion and the perceived social consequences of expressing that opinion. The theory posits that, following group discussion, this optimal public opinion is usually judged to be more extreme than individuals' initially stated opinions.

First, to address persuasive arguments theory, it certainly matters what language and communication strategies are used. Linguistic frames modulate the perceived meanings of words and sentences (Fillmore, 1982; Chong & Druckman, 2007; Cacciatore et al., 2016). These frames often become norms that are shared, repeated, and modified by group members. In this process linguistic frames co-evolve with the meanings of words (Hamilton, Leskovec, & Jurafsky, 2016b; Garg et al., 2018; Hawkins, Goodman, Goldberg, & Griffiths, 2020). Metaphorical fram-

ing provides a particularly strong example of how language can lead to extremism. Kalmoe (2014); Kalmoe et al. (2018) found that using violence metaphors to describe political issues and events (e.g., "EPA regulation is *strangling* the economy") led participants to increase their support for real world violence to reach political goals—this effect was even more pronounced among the most trait aggressive participants.

Self-categorization theory is correct to assume that it is a fundamental human capacity to evaluate one's own and others' group membership status (Cikara & Van Bavel, 2014; Cikara et al., 2017). The desire to clearly belong to one's in-group may well motivate individuals to increase their extremism in such a way as to lead to more clear signals of group membership, whether that is from being drawn towards the direction others are tending, or to be more clearly different from a perceived out-group. Whether this is achieved through a calculation of the hypothesized "meta-contrast ratio" (J. C. Turner & Wetherell, 1987) is less clear. Using the meta-contrast ratio as a theoretical variable calculated in the brain lacks the sort of mechanical explanation of behavior as Bayesian cognitive models. To ensure the validity of the meta-contrast ratio, or any other theoretical psychological calculation, one must co-develop a mechanistic model of how the value is calculated (Jones & Love, 2011), which does not seem to be developed in self-categorization explanations of group polarization.

Social decision schemes models of group polarization posit that there exist individual-level decision making traits (e.g., the ability to find or identify a solution to some problem) and group-level decision making schemes (e.g., the group must unanimously vote to choose an opinion or behavior) (Brown, 2000). Power dynamics are an important component for determining the social decision scheme used by a group (Friedkin, 1999). If it is the case that that one can enumerate individual-level traits and group-level decision schemes and power structures, then the social decision scheme model can theoretically be used to predict group decisions, opinions, and resulting behavior (Zuber et al., 1992; Friedkin, 1999). If the social decision scheme model encodes or evolves extremists to be more powerful, then group polarization will emerge. If extremists dominate the conversation,

which seems like it may plausibly occur often, then group polarization will emerge. One issue here is the introduction of the social decision scheme construct, which itself would be subject to cultural evolutionary pressures depending on group constitution and estimated payoffs of different strategies (King-Casas et al., 2005). The idea of payoffs in a group polarization context is potentially problematic as well since there is no tangible benefit to finding consensus, becoming more extreme, etc. It can only be understood as emotionally beneficial.

We believe that the stubborn extremism explanation of group polarization is a more appropriate starting point since it seems more parsimonious and robustly supported than alternative explanations (Kinder & Kalmoe, 2017; Reiss et al., 2019; Zmigrod, Zmigrod, Rentfrow, & Robbins, 2019). The stubborn extremism explanation makes one simple assumption, which could be complemented by certain elements of existing explanations outlined above. Even if stubborn extremism explains group polarization in some contexts, it is not clear which contexts. Our work does not address this important outstanding question directly. Likely it will take multiple methods and approaches to understanding the subtleties of the effect of context on group polarization. Although there is evidence supporting the hypothesis that extreme opinions are more stubbornly held, we are aware of no research specifically investigating the relationship between stubbornly held opinions and group polarization. Future empirical work should evaluate the stubborn extremism hypothesis using a statistical model to detect correlation between opinion extremity and stubbornness.

Models of opinion dynamics should be able to explain a number of empirical phenomena, including but not limited to group polarization. Another program of future work, then, could be to perform similar computational experiments shown here using alternative, influential models of political polarization, such as Bayesian/information-theoretic models (e.g. Dixit and Weibull (2007)) or algorithmic models (e.g. Dandekar et al. (2013)).

Chapter 4

Most group polarization results may be simple conformity

"Group polarization" is said to occur when socially isolated groups become more extreme following deliberation on some topic. This has clear implications for politics and other social organizations since extremism tears at the fabric of society. The goal of the current paper is to raise an alarm that many published results may plausibly be false detections of group polarization. These false detections are caused by failing to account for how opinions are represented psychologically and measured in the physical world. Group polarization studies implicitly assume latent psychological opinions are continuous when they use t-tests to detect group polarization, as many or most do. We demonstrate that if we assume participant opinions are drawn from a continous distribution but reported on an ordinal scale, then common group polarization experiments could be reporting group polarization when groups really just converged to the pre-deliberation average. This may be masking interesting differences in social dynamics when the group is more moderate versus more extreme. Our analysis revealed other problems including a lack of specificity in process models of group polarization and a failure to account for important sources of variance (e.g., group membership and survey item) in statistical models. To ensure reliable group polarization results, appropriate statistical designs must be adopted.

> In our introductory social psychology course, we have for many years used the [group polarization experimental paradigm] as a laboratory exercise. The exercise works beautifully, but one must be careful to forewarn a class that [group polarization] does not occur with every group. . . and that the effect is not large.
>
> (Brown, 1986, p. 205)

> One of the most robust findings in social psychology is that of attitude polarization following discussion with like-minded others.
>
> (Cooper, Kelly, & Weaver, 2001, p. 267)

4.1 Introduction

Social and political extremism and polarization threaten democratic institutions worldwide. If we could explain how and predict when extremism emerges, we could brace for its ill effects and perhaps devise interventions to counter it. To explain how extremism emerges, we need to focus on specific instances where extremism does emerge, since social systems are complex systems. Social psychologists, sociologists, political scientists, and legal scholars for decades have tried to explain "group polarization", the name given to the specific phenomenon where small, socially isolated groups tend become even more extreme in their opinions if their initial opinions centered around some non-neutral mean opinion. Several theories explaining group polarization have emerged, supported by extensive empirical evidence demonstrating that these groups reliably shift their opinions to become more extreme after group deliberation (Brown, 1986, 2000; Schkade et al., 2010; Sieber & Ziegler, 2019). The scientific consensus seems to be that different theories are potentially valid since there exists supporting evidence for all (Brown, 2000). However, we show in this paper that the evidence for group polarization is weak at best, meaning that these theories may be explaining a non-effect.

We demonstrate in this paper that a potentially large fraction of these group polarization detections are plausibly false. This means that the decades of theorizing about group polarization may be for naught as there is no value explaining something that does not really exist.. This is highly concerning given that the regularity of detecting group polarization makes the phenomenon a celebrated social

psychological result (Brown, 1986, 2000) that has essentially never been seriously questioned as a real effect. One prominent author has even elevated it to a scientific "law" (Sunstein, 2002). Litte work has been done on group polarization in the past two decades, apparently because researchers thought it was real and explained well enough. Understanding group polarization, if it really exists, has broad impacts for society at large. We must have a solid empirical foundation to trust theoretical explanations of group polarization—our study suggests that foundation is cracked at best. Mechanistic modeling and appropriate Bayesian statistics can be used to elminate the problems we identify and explain here (Kruschke & Liddell, 2018a, 2018b; M. A. Turner & Smaldino, 2021).

Many group polarization studies' findings are plausibly false due to their use of $t-$tests to detect group opinion shifts measured with ordinal valued survey instruments. False detections plausibly occur due to the simpler process of consensus where shifts from more extreme to less extreme opinions are masked by ceiling effects, but shifts from less extreme to more extreme opinions are detected (Liddell & Kruschke, 2018). It is a common observation among group polarization researchers that consensus occurs just as would be expected, i.e., opinion variance decreases following deliberation (Asch, 1951, 1955; French, 1956; DeGroot, 1974). What makes group polarization special is that the consensus (mean) opinion has increased in extremity compared to the pre-deliberation opinion. However, using an ordinal scale introduces ceiling/floor effects so that those in, say, the 80th percentile and the 99.99th percentile opinions report the same opinion, or worse. When simple consensus occurs we expect the less extreme opinions tend to become more extreme and the more extreme opinions tend to be less extreme. If simple consensus occurs, but only the less extreme opinions' shifts are detected, then the average will apparently increase. If participants' internal, "latent" psychological opinions are measured either directly on a continuous scale (a minority of group polarization studies do) or indirectly somehow, then perhaps such problems could be avoided.

We are not the first to point out serious problems in the group polarization paradigm—D. Cartwright (1973) was concerned about poor theory and methods

in group polarization from the start when researchers believed the phenomenon only occurred in situations of risk determination, and claimed their research should be used for studying how critical decisions about, e.g., nuclear deterrence should be made. We hope our work here further pushes group polarization researchers, and social psychologists and others who measure opinion change, to develop sound theories supported by valid statistical inferences.

To understand the technical and theoretical importance of this work, it is necessary to first explain how group polarization experiments work. Below we review common methods of inducing and detecting group polarization among groups of participants and how individual opinions and group polarization opinion shifts are typically measured. After introducing group polarization methods, we will explain in detail how these methods plausibly lead to false group polarization detections. Then we develop our formal, generative statistical model that simulates false group polarization detections. We then present our results showing that over 90% of published detections of group polarization opinion shifts are plausibly false. We close with a discussion of whether group polarization is real and how to improve group polarization research going forward.

4.2 Group polarization theory, methods, and results

Initially, the group polarization effect was thought only to apply to opinions about how much risk would be appropriate to take given some life decision (Wallach & Kogan, 1965; Teger & Pruitt, 1967; Stoner, 1968). Moscovici and Zavalloni (1969) then showed that deliberation about political opinions also led to group polarization. At this point, motivated by D. Cartwright (1971) and D. Cartwright (1973), new explanatory mechanisms were proposed. The two explanations that have survived from that time are the social comparisons theory (Brown, 1974; Sanders & Baron, 1977; Myers, 1978) and persuasive arguments theory (Burnstein & Vinokur, 1973; Vinokur & Burstein, 1974; Burnstein & Vinokur, 1975; Vinokur & Burnstein, 1978). Around the time of Isenberg's (1986) review, a self-

categorization explanation (J. C. Turner & Wetherell, 1987) of group polarization was developed and supposedly supported empirically (J. C. Turner et al., 1989; Abrams et al., 1990; Hogg et al., 1990; McGarty et al., 1992; Krizan & Baron, 2007). There is also a "social decisions scheme" theory that identifies social power structures as a dominant factor in the emergence of group polarization (Zuber et al., 1992; Friedkin, 1999). More recently, focus on the correlation between stubbornness and extremism has emerged as a simple, empirically-motivated explanation of group polarization (Baldassarri & Bearman, 2007; Mueller & Tan, 2018; Banisch & Olbrich, 2019; M. A. Turner & Smaldino, 2020). The results we present here, while damning for many group polarization studies, will enable real progress to be made untangling this theoretical boondoggle.

4.2.1 Common experimental design elements

Group polarization studies tend to follow the same general experimental paradigm, with slight variations to test particular theoretical explanations or real world situations. Participants first answer questionnaire items or otherwise give their opinions or positions on some topic. Small groups typically of 2-6 participants are formed such that the mean opinion or position of group members is baised towards one or the other extreme of the measurement scale. Group formation is often based on initial participant answers to the questionnaire. Sometimes researchers use a different, but similar, questionnaire to make like-minded groups. For example, Myers and Bishop (1970) examined group polarization in the context of racial attitudes. To create groups with different levels of mean tolerance or racism, Myers and Bishop used a survey instrument to assess racial attitudes generally. Then they used a different questionnaire on racial policy opinions for deliberation topics, where pre- and post-deliberation survey responses were used not to pick groups, but to measure group polarization. In a different approach altogether, Schkade et al. (2010) relied on a correlation between geographic location and political opinions to create novel groups that were reliably biased towards liberal or conservative bias.

In the most common paradigm participants first answer one or several ques-

tionnaire items to determine their initial opinions on some deliberation topic. One widely used questionnaire is the choice dilemma questionnaire first used by Stoner (1961) to induce group polarization. The questionnaire prompts participants for their opinions on how much risk would be acceptable for certain life decisions, such as whether or not to pursue riskier research projects with higher payoffs compared to lower risk projects with lower payoffs. Political questionnaires are also common. For example, Moscovici and Zavalloni (1969) asked Parisian lycée students about their opinions of then-president Charles de Gaulle and of American foreign policy; Myers and Bishop (1970) asked about racial attitudes; Schkade et al. (2010) asked about affirmative action, same-sex civil unions, and global warming.

Most studies using questionnaires prompt participants to give their responses on an ordinal, Likert-type scale. Stoner's (1961; 1968) choice dilemma questionnaire was a 10-point ordinal scale, with 1 representing the most risk acceptance and 10 representing the least risk acceptance. More generally common Likert scales typically have participants rate (un)favorability of some entity in the world or degree of (dis)agreement with some statement of opinion or belief. For example, when French students answered "American economic aid is always used for political pressure", they marked a whole number on a seven-point scale from -3 (strongly disagree) to +3 (strongly agree), with zero representing neutral or no opinion. These scales do not always include 0 as the neutral point. Schkade et al. (2010) used a ten-point scale from 1 (disagree very strongly) to 10 (agree very strongly).

Non-questionnaire group polarization studies have used a variety of methods. In one approach, researchers simulate jury deliberations for an experimental design where participants give either opinions on whether a defendant is guilty or how much money for damages should be awarded (Kaplan, 1977; Kaplan & Miller, 1977; Schkade et al., 2000, 2007; Sunstein, 2000)[1]. Another approach studied group polarization in the context of gambling behavior in the game of blackjack (J. Blascovich & Ginsburg, 1974; J. I. M. Blascovich et al., 1975; J. Blascovich et al.,

[1]Schkade, et al., (2000), entitled "Deliberating about Dollars: The Severity Shift", was funded by Exxon Company, U.S.A., who have a clear interest in understanding what causes individuals to raise or lower the amount of damages they believe a responsible party should pay.

1976), which found that participants demonstrated opinion shifts to be more risky merely when exposed to other group members' bets. Another odd example of questionable *prima facie* validity is an experimental design that used an "autokinetic situation" where participants watched a flashlight move in a darkened room, then deliberated about how far the light moved after being told that longer measurements were more socially desirable (Baron & Roper, 1976). Our model does not apply to these studies, but there are many more studies that use ordinal scales. Furthermore, other problems such as not accounting for the multilevel structure of the data may subvert the validity of these studies.

4.2.2 Common statistical procedures and implicit assumptions

All group polarization studies we reviewed that used an ordinal opinion measurement scale also used a t-test to detect group polarization opinion shifts. t-tests are used to determine the probability that two datasets were drawn or generated from the same distribution. These tests assume that individual opinions are continuous and normally distributed. To determine whether two datasets came from the same distribution, a normal distribution is fit to each dataset. Then, the probability that the two datasets were drawn from the same distribution is proportional to the degree of overlap between the fitted distributions. In the studies we reviewed, pre- and post-deliberation distributions are always pooled over groups, and often by pooling over several items within one topic. For example, Moscovici and Zavalloni (1969) pool over 11 items in the topic about Charles de Gaulle and 12 items in the topic and deliberation about American policy. Schkade et al. (2010) provide a counterexample to this, where there is only one item per topic.

When t-tests are used to detect group polarization with ordinal observations, they are susceptible to false positives due to ignoring the effects of the measurement process (Liddell & Kruschke, 2018). The problem is that no matter how extreme a participant's latent opinion is, it will be reported as the maximal ordinal value. This means that an opinion in the 99th percentile of extremity may be mapped to the same value as an opinion in the 80th extremity percentile. This means

that if, for example, an extremist shifted their opinion towards moderation, the measurement scheme could not detect this—it would appear as if the opinion did not change at all.

4.3 Model

Our primary goal in this paper is to evaluate whether published positive detections of group polarization are reliably true, or, equivalently, plausibly false. We do this by first developing a generative model of group polarization experiments that simulates how opinions are reported and change, and how standard analytical techniques can generate the appearance of group polarization where none exists. Our model is based on the assumptions that (1) a participant's internal "latent" opinion on some topic can be represented as a real number varying continuously; (2) when a participant reports their opinion on an ordinal scale, the formulation of their latent opinion can be represented as a draw from a latent opinion distribution; and (3) participants faithfully convert their continuous latent opinion into whatever ordinal ratings scale (e.g. a Likert scale) the experimenters present them with. Note that these assumptions assume there are there are two forms of opinions. There are *latent* opinions that are somehow represented and formulated in a person's mind, but never directly observed. Then there are *observed* opinions that participants report on an ordinal scale. We also then have two distributions of opinions that do not in general have the same summary statistics (mean and variance).

We make these assumptions for the sake of consistency with the implicit assumptions made in psychological and social science studies of opinions. When someone gives their opinion on some topic it is the result of a complex psychological process that is sensitive to personal beliefs and experiences, and cultural and contextual factors. Because of this complexity, opinions may not in fact be readily mapped onto a unidirectional scale, continuous or ordinal. For our purposes we can ignore this possibility because our goal is to show that, under common assumptions of group polarization studies, many detections of group polarization

may plausibly be false detections.

Because simple conformity to the mean can be masked by ordinal measurements, a change in pre- and post-deliberation opinion variance can masquerade as group polarization, i.e., a change in *mean* from pre- to post-deliberation. Theoretically, we expect variance to decrease from pre- to post-deliberation as participants feel pressure to conform (Asch, 1951, 1955; French, 1956; DeGroot, 1974; Lorenz, 2009). Conformity has been observed across group polarizaiton studies, with many containing explicit instructions to find consensus with group members as part of the experimental design.

4.3.1 Formal model

Our formal model incorporates three main features we review now. First, we formalize our assumptions about what opinions are and how they are generated "internally" in model participants. Next, we formalize the measurement process where participants transform their internal, *latent* opinions to their reported opinions in one of several ordinal scale bins, e.g., a Likert scale. Finally, we develop a statistical model that can generate plausibly false detections of group polarization if that is possible, or fail if it is not possible, which instead would support a positive finding of group polarization. Formal models are important to develop because in doing so we specify and include those social influence components we hypothesize are important for studying phenomena of interest (Kauffman, 1970; N. Cartwright, 1999).

All model calculations are done in the large N limit. This enables us to perform exact calculations to directly find what pre- and post-deliberation variances could have generated false detections of group polarization. Theoretically, effect sizes calculated with finite N will be less reliable, if anything, so demonstrating that a false discovery occurs even in the large-N limit is a sort of formal proof that there exists a plausible combination of parameters that gives rise to a false group polarization discovery.

Each experimental condition that claims to detect group polarization is a "possible false detection" (Table 4.1). When we determine that a false detection is plau-

sible, that means we have no data to decisively say whether or not the published result is reliable, meaning we cannot count it as evidence of group polarization.

After we formally introduce the psychological representation of opinions, we will consider how a large collection of opinions becomes a distribution of observed ordinal scale opinion ratings, which in turn are used to calculate mean pre- and post-deliberation opinions and which, in experimental analyses, are tested against one another to detect a significant opinion shift due to group polarization. We will attempt to generate pre- and post-deliberation observed opinion distributions with different means, but that were generated from two latent distributions with the *same* mean. Different pre- and post-deliberation latent standard deviations are what cause different observed mean opinions to be generated, even though latent means are identical.

Opinions

We assume that participant i's internal, latent psychological opinion at time of reporting ($t \in \{pre, post\}$) is drawn from a normal distribution with mean μ_t and standard deviation σ_t,

$$o_{i,t} \sim \mathcal{N}(\mu_t, \sigma_t). \tag{4.1}$$

All group polarization studies we have reviewed, using both ordinal and continuous measures of opinion, make this same assumption, which implicitly pools participant data over groups, even though it is well-known that, e.g., the initial extremity of the group predicts the magnitude of the group polarization opinion shift (Myers, 1982). Studies such as Moscovici and Zavalloni (1969); Myers and Bishop (1970) also implicitly pool over opinion items, on which participants give several opinions, but these shifts are given only as an average over all items (and groups). Future work should examine the impact of this practice, which has been shown to lead to overgeneralizations and overestimations of other psychological effects (H. H. Clark, 1973; Yarkoni, 2021).

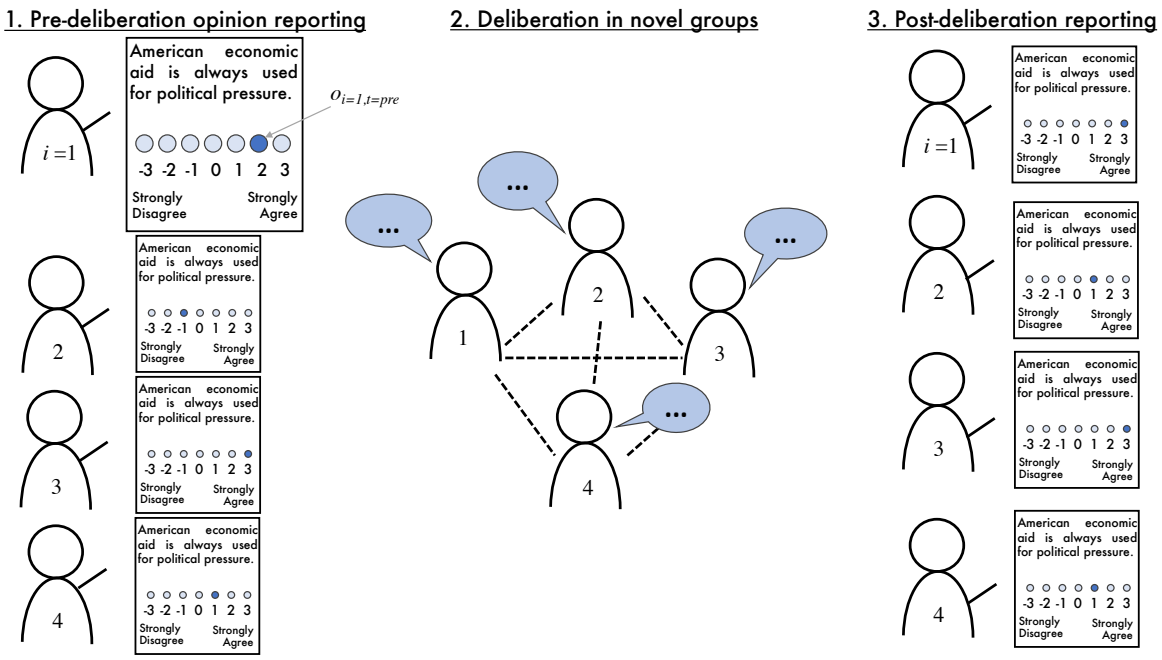


Figure 4.1: Schematic diagram of our model of a group polarization experiment. Many experiments add additional complexity, but this simple model suffices for studying the effect of measurement and statistical procedures on empirical results. For each of the ten case studies presented here In Step 1, participants have not yet met one another and so report their opinions independently of any experimental social influence. We denote $t = pre$ at this stage, referring to *pre*-deliberation. In Step 2, a discussion group is formed that has an overall bias in one direction or another. It is through discussion that opinions are hypothesized to change, i.e., group polarization occurs. At the third and final step, post-deliberation ($t = post$), participants again report their opinions, which, if group polarization has occurred, have increased in extremity overall.

Experiment model

There are many versions of the group polarization experiment, however they all share three main steps, which constitute our model here (J. C. Turner, 1987, p. 143) (Figure 4.1). First, typically before small deliberation groups are formed, participants are given a questionnaire on which the indicate their initial opinions on the item(s) on the experiment's topic(s) of discussion. We generate pre-deliberation data by first drawing a latent opinion from this distribution, then binning participant opinions into an ordinal opinion scale, which is described in more detail in the next subsection on the Measurement Model.

Next, participants are placed with a small discussion group with all or mostly others who share their bias, e.g., towards -3 or +3 on a seven-point Likert scale, and then the participants deliberate in these biased groups—though in some conditions participants may only display their opinion to others or some other sort of twist on communicating individual opinions. To form bias groups in our model we simply assume participant opinions are drawn from a non-neutral latent mean. Deliberation is simulated in the aggregate, with its effects modeled as a possible change in mean (if group polarization does indeed occur) and as a decrease in variance due to consensus/conformity processes. After deliberation, participants again report their opinions. To generate a false detection, we assume that the pre- and post-deliberation means are identical ($\mu_{pre} = \mu_{post}$), but their variances are not.

Measurement model

Our measurement model transforms a distribution of pre- or post-deliberation latent opinions into an ordinal-valued distribution of ordinal scale opinion measurements. This simulates the three step group polarization experimental design where participants do not directly report their continuous latent opinions, but instead report their opinions in terms of a finite set of ordinal bins. Formally, this is achieved by integrating over the probability density function (Equation 4.1) of opinions for each ordinal scale bin (Figure 4.2).

We assume that participants give their opinions in terms of one of K opinion

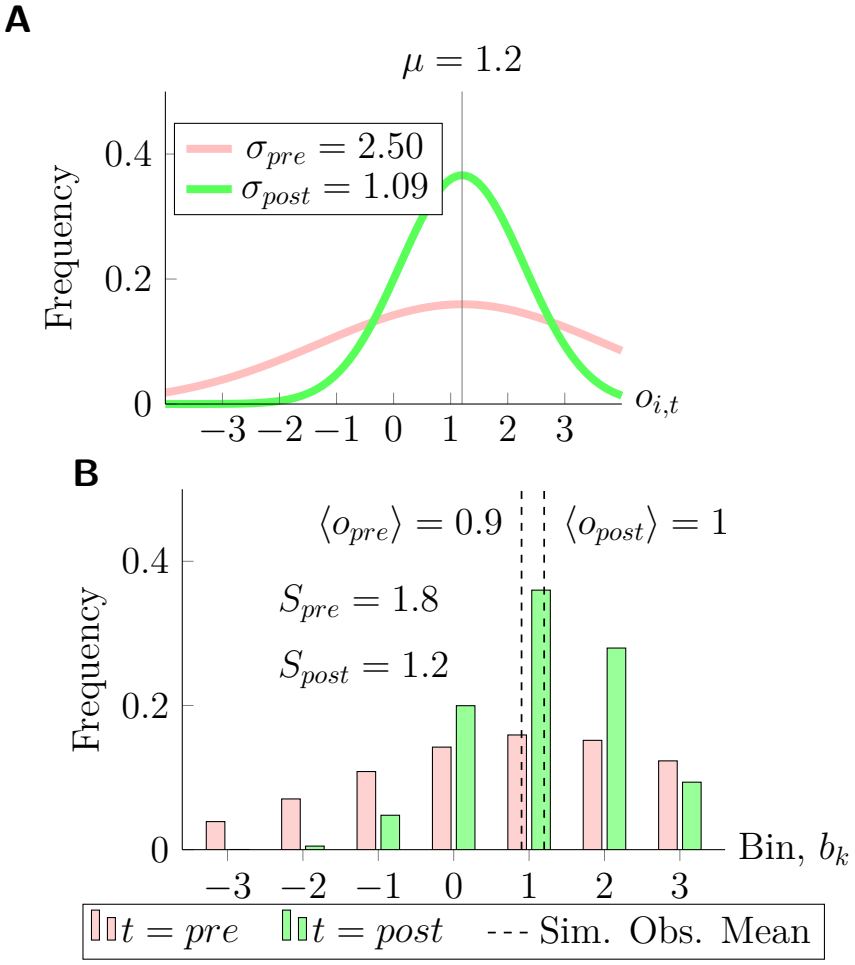


Figure 4.2: Example of a plausibly false detection of group polarization on the condition where the deliberation topic was whether or not participants approved of Charles DeGaulle's presidency, following one of two conditions in Moscovici and Zavalloni (1969). In (A) we show hypothetical pre- and post-deliberation latent distributions with identical means ($\mu_{pre} = \mu_{post} = \mu = 1.2$, but different latent standard deviations. Following the consensus process that invariably occurs in group polarization, the pre-deliberation standard deviation (σ_{pre}) is larger than the post-deliberation standard deviation (σ_{post}). In (B) we show how these latent distributions lead to observed pre- and post-deliberation opinion distributions with different means ($\langle o_{pre} \rangle < \langle o_{post} \rangle$).

bins, with each bin value denoted b_k and indexed by $k = 1, \ldots, K$. The array of all bin values a participant may choose is simply b. In the popular choice dilemma questionnaires the ten opinion bins are $b = \{1, 2, \ldots, 10\}$. In this case, we happen to have $b_k = k$. A seven-point Likert scale (e.g., -3 strongly disagree, 0 neutral, and +3 strongly agree) has $K = 7$ bins, $b = \{-3, -2, \ldots, 3\}$, i.e. $b_1 = -3$ and $b_{K=7} = 3$.

An individual reports an opinion in bin b_k if their latent opinion is within bin thresholds θ_{k-1} and θ_k. There are $K + 1$ thresholds, starting from $\theta_0 = -\infty$. Similarly, $\theta_K = \infty$. Other than $k = \{0, K\}$, $\theta_k = b_k + 0.5$. Taking the example of a seven-bin Likert scale, if $o_{i,t} = 1.4$, then participant i would report a binned opinion of $b_5 = 1$. In this case we assume for simplicity that except for thresholds at $\pm\infty$, thresholds are separated by 1 in "opinion space"—for more on the Cartesian representation of opinions see Blau (1974).

We model measurement of $N \to \infty$ participant opinions, which results in a histogram of frequency of opinions in each bin, i.e., $o_{i,t} = b_k$. The frequency of responses in each bin is the integral over the continuous normal probability density function from one bin threshold to another. This transforms the probability density function from $p(o_{i,t};\ \mu_t, \sigma_t)$ to the probability of observing a reported opinion of each bin value. The probability of observing an opinion in bin b_k is calculated by integrating the normal probability density function over the range of θ_{k-1} to θ_k. Formally, we write the probability of observing an opinion in bin k at time t as

$$\begin{aligned} p(o = b_k; \mu_t, \sigma_t, \theta) &= \int_{\theta_{k-1}}^{\theta_k} p(o; \mu_t, \sigma_t) do \\ &= \Phi\left(\frac{o - \mu_t}{\sigma_t}\right)\Big|_{o=\theta_{k-1}}^{\theta_k} \\ &= \Phi\left(\frac{\theta_k - \mu_t}{\sigma_t}\right) - \Phi\left(\frac{\theta_{k-1} - \mu_t}{\sigma_t}\right) \end{aligned} \tag{4.2}$$

where $\Phi(\frac{o-\mu}{\sigma})$ is the normalized normal cumulative distribution function over opinions o, shifted by an amount μ with standard deviation σ.

From this we can calculate the simulated expected value of observed opinions

at time $t \in \{pre, post\}$, written

$$\langle o_t \rangle = \sum_{k=1}^{K} b_k \cdot p(o = b_k; \mu_t, \sigma_t, \theta). \tag{4.3}$$

We differentiate this from the mean opinion observed in a particular experimental condition, which we write $\bar{o}_t$. Importantly for performing our investigation of whether published findings may be false detections, it is vanishingly rare that the latent mean and expected observed value are identical, i.e., it is rare to find $\mu_t = \langle o_t \rangle$. This occurs only for μ_t at the exact midpoint of the ordinal opinion measurement scale. This fact underlies our method for generating false detections explained in the following subsection.

False detection model

We use our model to re-evaluate the reported results to see if, in fact, the null hypothesis is plausible, i.e. there is plausibly no difference between pre- and post-deliberation means. To do this, we test data from published experiments assuming the null hypothesis, i.e. $\mu_{pre} = \mu_{post}$. We demonstrate that the null hypothesis is often plausible, i.e., that there was in fact no shift in group opinions. We do this by first finding a latent mean that generates the observed pre- and post-deliberation means ($\langle o_{i,pre} \rangle$ and $\langle o_{i,post} \rangle$) reported in published studies, for certain pre- and post-deliberation latent standard deviations (σ_{pre} and σ_{post}). The challenge is to identify which σ_t generate false detections.

To find σ_t we might first think to simply set the observed mean equal to the calculated mean, i.e., $\bar{o}_t$. However, it is not clear if this is tractable to solve directly. Therefore, we solve for σ_t numerically by finding the the σ_t that minimizes the squared error between the observed mean, $\bar{o}_t$, and the simulated observed mean, $\langle o_t \rangle$, i.e.

$$\begin{aligned} \sigma_t = \arg\min_{\sigma} \ (\bar{o}_t - \langle o_t \rangle)^2 = \arg\min_{\sigma} \ (\bar{o}_t - \sum_{k=1}^{K} b_k \cdot p(o = b_k; \mu_t, \sigma_t, \theta)^2 \\ \text{s.t. } |\bar{o}_t - \langle o_t \rangle| < \epsilon \end{aligned} \tag{4.4}$$

where ϵ is the error tolerance. We will find that different studies allow for finding σ_t with larger or smaller ϵ. Note that since all bins are unit distance from one

another, it is reasonable to set $\epsilon \sim 0.1$, especially since we are comparing large-N simulations with finite-N observations. However, we use the smallest possible ϵ we can as long as it is at most on the order of 0.1; the values of ϵ used for each condition is available in an Excel spreadsheet we have provided as supplemental information.

There are two ways for this search to fail. First, there could be total failure, i.e., no $\sigma_{pre,post}$ are found that generate pre- and post-deliberation distributions whose means match those reported in a given experimental condition. Second, a solution to Equation 4.4 may be found, but the solution σ is too large, which yields a highly "bi-polarized" distribution that is not feasible in most group polarization studies, which do not contain groups of opposing viewpoints.

4.3.2 Model implementation and analysis

We implemented the model in R using primarily built-in or open source packages (R Core Team, 2021). We programmed a simple hillclimbing algorithm to solve the optimization problem in Equation 4.4 to find which latent standard deviations σ_{pre} and σ_{post} generate the observed data for a given observed group polarization opinion shift.

To facilitate the use and re-use of this code, we also developed a Shiny web application[2] to specify observed pre- and post-deliberation mean opinions, a hypothesized latent mean, the measurement scale, and to vary hillclimbing parameters (step size and stopping condition). This app will display theoretical histograms of responses for the binned latent pre- and post-deliberation opinion distributions.

In each study, we tabulate the number of plausible false discoveries made out of the number of potential false discoveries, which is equal to the number of experimental conditions in each study. For example, in the study of Schkade et al. (2010) we analyze below, they calculate group polarization opinion shifts in six experimental conditions. The conditions arise from two geographical locations where groups were assembled (Boulder, CO and Colorado Springs, CO) and three deliberation topics (affirmative action, civil unions, and global warming). For this study,

[2]https://mt-digital.shinyapps.io/grouppolarizationstatmod/

we inspect each of the six conditions to determine if the observed shift reported for each condition is plausibly a false detection arising from simple a consensus process (reduction in opinion variance from pre- to post-deliberation) instead of a group polarization process. For each of the 62 experimental conditions we inspected across ten studies, we made this determination of plausibility of the null hypothesis and recorded the latent mean and latent pre- and post-deliberation standard deviations, and hillclimbing step size parameter, that gave rise to our counterexample data supporting our assertion of a plausibly false detection.

With all studies inspected this way, we obtained a table with columns Study, Experimental Condition, and whether the detection was Plausibly False. We then calculated the worst-case false detection rate for individual studies and a global worst case false discovery rate across all ten original published studies (files with original data and analyses will be available in a supplement).

We developed our model to analyze published results demonstrating group polarization and other opinion shifts to determine whether shift detections are actually plausibly false. We evaluated 62 experimental conditions across ten influential published studies. Our approach can and should be applied to more studies. This can be achieved by the following strategy that we used to perform our analysis presented in the next section. We can only provide the worst case false detection rate because we do not, and can not in any of the studies, know if plausibly false detections are false or not. This is not a comfort, since this means that plausibly false detections are unreliable, i.e., of no practical scientific value. If original source data had been provided then the data could have been re-analyzed with proper statistical methods, and perhaps discoveries of group polarization could be confirmed.

More specific information about how our model and hillclimbing algorithm for solving Equation 4.4 were implemented can be found in the Appendix.

4.4 Analysis

We now show our results of applying our model to analyze whether published detections of group polarization are false. We chose ten influential group polarization studies published between 1969 and 2010. Each study has one or more experimental conditions in which a group polarization opinion shift was hypothesized to occur. The studies reported experimental data and possibly associated statistical tests to support their hypothesis that group polarization occurred in groups subjected to these conditions.

Across the ten studies, we found that 92% of group polarization detections are plausibly false according to our model. No studies had a false detection rate below 50% (Table 4.1), This means that for each published group polarization paper, at least half of their group polarization detections are explained by simple conformity. Myers and Bishop (1970) and Myers and Lamm (1975) had the lowest plausibly false detection rates (67% and 50%, respectively), possibly due to their use of 18-point Likert scales and highly charged deliberation topics, including racism and gender roles and relations in the United States.

Table 4.1: Tabulation of worst-case scenario false discovery rates obtained by showing it is equally plausible to accept as reject the null hypothesis by generating pre- and post-deliberation reported data from two distributions with the same latent mean. Different observed means are generated due to the two latent distributions having different standard deviations, found through a hillclimbing optimization routine.

	# Plausible FDs	**# Exp. Cond.**	**Worst case FD rate**	**# Citations**
Abrams et al. (1990, Table 2)	10	10	1.00	1008
Burnstein and Vinokur (1973, Table 1)	5	5	1.00	240
Burnstein and Vinokur (1975, Table 2)	1	1	1.00	294
Friedkin (1999, Table 1)	8	8	1.00	220
Hogg et al. (1990, Table 2)	8	9	0.89	385
Krizan and Baron (2007, Table 2)	10	10	1.00	47
Moscovici and Zavalloni (1969, Table 4)	4	4	1.00	1515
Myers and Bishop (1970, Table 1)	2	3	0.67	303
Myers (1975, Tables 1 & 2)	4	8	0.50	118
Schkade et al. (2010, Table 1)	5	6	0.83	79
Total	**55**	**60**	**0.92**	**4211**

4.5 Discussion

In this paper we showed that many published studies presented plausibly false detections of group polarization that could be equivalently described as conformity to the initial group mean, not to a more extreme mean. This was enabled by the use of metric statistical models to make inferences about ordinal valued data, which induces ceiling effects that mask changes in opinions among the most extreme group members. Because the original data is not available, we cannot determine whether the observed effects are true or false detections, nor can anyone else. Unfortunately for the authors of these studies and for psychological science in general, this means that the results cannot be used to support the theoretical explanations of group polarization they were meant to. Furthermore, it causes us to doubt whether there

is a group polarization effect at all.

The immediate solution is clear: use appropriate statistical procedures for group polarization research that uses ordinal opinion measurement scales, but assumes opinions are countinuous. This means we must expand our statistical models to incorporate opinion binning. This can be achieved through the use of ordered probit models, or any other statistical model that treats observed data as ordinal, and generated from binning continuous opinions into categorical bins. In the course of our study we also found that statistical models of group polarization failed to account for the multilevel, and sometimes hierarchical, structure of group polarization data. This must also be accounted for in the design of valid, robust statistical models of group polarization.

4.5.1 Is group polarization real?

It may seem that we have little justification left for asserting the reality of group polarization. We have demonstrated many detections of group polarization are plausibly false. Furthermore, in the course of this study, we observed that significant sources of variance are regularly not accounted for in statistical models used in group polarization research. This results in a lack of multilevel structure in statistical models that tends to lead to overestimates of effect sizes and underestimates of confidence interval widths (H. H. Clark, 1973; Yarkoni, 2021). The theoretical weaknesses identified earlier and these facts may seem to kill off any potential reality of group polarization. No research data is available from previous studies, so the data cannot be re-analyzed. Although we may no longer count group polarization as an empirical reality (pending new work using appropriate statistical methods), there are several theoretical reasons to believe that properly designed studies will find, in certain cases, that group opinions become more extreme following deliberation.

Even if we expect to observe group polarization in some contexts with more rigorous methods, it is not clear which contexts. As Brown (1986) observed (quoted in the epigraph to this paper), group polarization often occurs, but inconsistently, and the effect is not always large. Explaining this context-dependence of group

polarization is, in our opinion, the next step for group polarization research. Valid statistical models are necessary to reliably move forward.

4.5.2 Statistical model features and implementation for valid group polarization measurement

Future research on group polarization needs a valid statistical measurement procedure for quantifying group polarization. As we have demonstrated, one requirement for a valid statistical procedure is that the data must be represented as ordinal measurements, not continuous and normally distributed. In the course of our work, we also observed that each group should have its own mean and variance in pre- and post-deliberation opinions, and similarly different items have been observed to vary in their response distributions. Failing to account for this multilevel structure is known to lead to overconfident overestimates of effect sizes (Gelman & Hill, 2007; Yarkoni, 2021). Therefore group polarization statistical models must include this multilevel structure to be valid. One model that can meet these needs is the ordered probit model that combines a normal model of latent opinions with an ordinal model of ordinal measurement data.

One statistical model that represents ordinal measurements of metric data is the ordered probit model (Kruschke, 2015, Ch. 23). The ordered probit model combines a normal model of latent psychological opinions (equivalently beliefs, attitudes, etc.) with an ordinal model of observed data. In addition to latent normal opinion distribution parameters mean and variance, there are additional parameters that represent the binning of opinions into ordinal survey responses. These are the thresholds, θ_k, which we were free to set constant in our generative model. To fit a multilevel ordered probit model, one must fit the model using Bayesian methods, which, unlike frequentist methods, can account for several, even hundreds, of groups across multiple levels (Liddell & Kruschke, 2018).

4.5.3 Open science to improve group polarization research

This current paper and project could have provided much stronger conclusions about the validity of published results if group polarization researchers had followed current open science best-practices. Open science practices, including open data sharing, data and metadata standards, and publishing analysis code, can improve scientific outcomes generally (E. M. Hart et al., 2016; Smaldino, Turner, & Contreras Kallens, 2019; Samuel & König-Ries, 2021). Our study would have been further streamlined if group polarization research data was stored in a central database, accessible through an API for automated gathering and analysis. If we had access to the original data formats, our paper would not have simply shown whether existing findings are plausibly false. Instead we could have re-analyzed the existing data with more appropriate ordinal statistical models (Liddell & Kruschke, 2018). However, if that data was haphazardly stored in disparate personal websites, or even just in separate Open Science Foundation data repositories, then the process would be extremely tedious, and analyses of additional datasets would be needlessly time consuming.

4.5.4 Conclusion

We developed a measurement and statistical model of group polarization that invalidated the results of several published studies when we analyzed those studies' supporting data. While not all observations of the group polarization effect are invalidated by our model, many of the ones we studied are widely referenced and high profile—even though some are decades old, they continue to motivate new work (Mäs & Flache, 2013; Keating, Van Boven, & Judd, 2016; Sieber & Ziegler, 2019; Pallavicini, Hallsson, & Kappel, 2021). Even the literature that does not apply continuous statistical models to ordinal data has separate problems, including possible theoretical inconsistencies, overgeneralizations from underrepresentative sampling and failing to account for important sources of variance, and a lack of publicly available data.

By examining the effects of measurement and statistics in detail, we demonstrate that future work on group polarization must use ordinal statistical models

to analyze ordinal data. This effort will be further supported by the adoption of open science practices for the further refinement of research methods and new analyses and theorizing.

Chapter 5

Paths to polarization: extreme views, miscommunication, and random chance

Understanding the social conditions that tend to increase or decrease polarization is important for many reasons. We study a network-structured agent-based model of opinion dynamics, extending a model previously introduced by Flache and Macy (2011), who found that polarization appeared to increase with the introduction of long-range ties but decrease with the number of salient opinions, which they called the population's "cultural complexity." We find the following. First, polarization is strongly path dependent and sensitive to stochastic variation. Second, polarization depends strongly on the initial distribution of opinions in the population. In the absence of extremists, polarization may be mitigated. Third, noisy communication can drive a population toward more extreme opinions and even cause acute polarization. Finally, the apparent reduction in polarization under increased "cultural complexity" arises via a particular property of the polarization measurement, under which a population containing a wider diversity of extreme views is deemed less polarized. This work has implications for understanding the population dynamics of beliefs, opinions, and polarization, as well as broader implications for the analysis of agent-based models of social phenomena.

5.1 Introduction

Diversity of opinions in a community is often difficult to maintain. Iterative exposure, norm enforcement, and psychological biases for conformity can drive consensus within a group (DeGroot, 1974; Deffuant, Neau, Amblard, & Weisbuch, 2000; Henrich & Boyd, 1998; Smaldino & Epstein, 2015b; Efferson, Lalive, Richerson, McElreath, & Lubell, 2008; Muthukrishna, Morgan, & Henrich, 2016). On the other hand, in-group bias, outgroup aversion, and the tendency to further differentiate ourselves from those deemed different may lead to the emergence of strong inter-group differences (Tajfel, Billig, Bundy, & Flament, 1971b; Lord et al., 1979; K. M. Carley, 1990; Axelrod, 1997; N. Mark, 1998; McElreath, Boyd, & Richerson, 2003; Dandekar et al., 2013; Gray et al., 2014; Smaldino et al., 2017). Such differences can lead to polarization in opinions under certain conditions. Understanding the social conditions that tend to increase or decrease polarization is important for many reasons. Primary among these is that a functioning democratic society depends on clear communication among the citizenry, which is impeded by the mismatch in norms, the differential interpretation of facts, and the dehumanization that polarization can engender (see Pew Research Center (2017a) for a current analysis of these dynamics in the United States). The maintenance of social differences in the form of cliques and clubs may be inevitable, but cooperation depends on transcending differences.

We take a network theoretic approach to studying the conditions for polarization in an agent-based model of opinion dynamics. Empirical research on the population dynamics of opinions is challenging and must be supplemented by formal modeling (Flache et al., 2017). Models reduce complex systems to ones that are tractable using mathematical or computational analysis, and allow for the exploration of replicate and counterfactual scenarios. Of course, the conclusions we draw from our models depend essentially on the assumptions of those models, and so caution must be taken when using model results to make inferences about empirical phenomenon. For example, Smaldino and Schank (2012) analyzed models of human mate choice and showed that very different individual decision rules could be fit to almost any empirical outcome by modulating assumptions about the pop-

ulation structure that had been ignored in prior analyses. When considering an important phenomena such as polarization, similar caution must be exercised, as we will demonstrate.

Our analysis extends the work of Flache and Macy (2011), who used a network-structured model of opinions and biased influence (hereafter the FM model) to study polarization. Network ties in this model exist between individuals as an indicator of social influence. Like several other models of opinions and beliefs, they operationalized the well-known phenomena of *biased assimilation* (Lord et al., 1979; Dandekar et al., 2013), the tendency for an individual to become more similar to those to whom they are similar, and to become more distinct from those with whom they already differ. Some empirical studies support the assumption of both positive and negative biased assimilation (Adams & Roscigno, 2005; P. S. Hart & Nisbet, 2012, e.g.). Other empirical studies failed to find evidence of negative biased assimilation at work where computational studies suggested it would be (Takács, Flache, & Mäs, 2016; Boxell, Gentzkow, & Shapiro, 2017, e.g.). Of course, if further empirical research turns out to invalidate that assumption, then our model conclusions must also be re-examined, as with any theoretical model (Smaldino, 2017a). Flache and Macy found that, when compared with a highly clustered population structure, the addition of long-range ties could dramatically increase polarization. When individuals were clustered into relatively isolated groups, they tended to converge to local consensus while maintaining diversity in the population at large. However, the addition of long-range ties increased exposure to substantially different opinions. Whether by attractive or repulsive forces, these long-range ties tended to drive opinions more toward their extreme values, resulting in increased polarization. Another important result was that the extent of "cultural complexity"—the number of orthogonal traits that are important to individuals in assessing their similarities and differences with others—mitigated polarization. When the number of traits was large, polarization was reduced. DellaPosta et al. (2015) used a variant of the FM model to explain data from the General Social Survey indicating that arbitrary traits tend to become associated with polarized identity groups, leading to often-puzzling stereotypes such as "latte-drinking liberals" and

"bird-hunting conservatives".

If we take the results of Flache and Macy (2011) at face value, two possible recommendations for the reduction of polarization readily emerge. First, we might try to reduce the number of long-range ties in our social network. This is made difficult due to the pervasive influence of internet social media (Center, 2016; Pew Research Center, 2018). Second, we might attempt to broaden the number of domains in the public discussion, so that points of agreement are easier to discover. This is also challenging, due to the increasingly fractured media landscape in which niche interests are increasing and common knowledge diminishing (Pew Research Center, 2014a). However, challenging is not the same thing as impossible. We must ask, then: How seriously should we take these recommendations? Might there be other solutions available?

To address these questions we perform new analyses of the FM model and reveal several additional factors influencing polarization. First, polarization is almost always a probabilistic occurrence. Even when parameter exploration appears to reveal regularities in polarization, specific outcomes are strongly path dependent. Indeed, there is often a wide range of possible outcomes even given identically repeatable starting conditions, due to stochasticity in the dynamics of interactions. This result highlights potential limits of our ability to make reliable predictions about polarization in any particular social system. Complex systems are often stochastic, and something that increases or decreases average polarization in a simulation is not guaranteed to do so in reality. Second, resultant polarization depends strongly on the initial distribution of opinions in the population. In the absence of extremists, polarization may be mitigated. This highlights the well-known danger of extremists and suggests new routes to avoiding polarization. More broadly, we show that too much diversity of extreme opinions makes polarization more likely. Third, noisy communication can drive a population toward more extreme opinions and even cause acute polarization. Cooperation and consensus-building depend on individuals finding common ground, which can be jeopardized even in the presence of unbiased error (H. Clark, 1996). Finally, we show that the apparent reduction in polarization under increased "cultural complexity" arises via

a particular property of the polarization measurement, under which a population containing a wider diversity of extreme views is deemed less polarized. Although this may often be a reasonable assumption, it highlights the need for caution in our measurement of complex social phenomena.

5.2 Model

5.2.1 Modeling individuals and their opinions

Our model is an extension of one presented by Flache and Macy (2011), and shares many general features with other models of opinion dynamics in structured populations (Nowak, Szamrej, & Latané, 1990; K. M. Carley, 1990; Axelrod, 1997; N. Mark, 1998, 2003; Dandekar et al., 2013; DellaPosta et al., 2015; Battiston, Nicosia, Latora, & San Miguel, 2017). The population is modeled as a network of individuals (or agents), each of whom is defined by a vector of opinions. The size of this vector, K, is called the "cultural complexity," and may be more descriptively explained as the number of opinions that are important to individuals in assessing their similarities and differences with others. Opinions can present political views, religious or moral values, artistic tastes, or myriad other beliefs. The opinion of agent i on issue k ($1 \leq k \leq K$), s_{ik}, is operationalized as a real number implicitly bounded in $[-1, 1]$ by smoothing (Equation 5.3). In Flache and Macy's original analysis, all opinions were initialized as random draws from the uniform distribution $U(-1, 1)$. In order to study the importance of initially extreme opinions, each initial opinion is here drawn instead from $U(-S, S)$, where $0 < S \leq 1$.

5.2.2 Modeling social influence

The aggregation of the K opinions held by an agent determines its coordinates in opinion space. We adopt the FM model's measure of distance between agents i and j,

$$d_{ij} = \frac{1}{K} \sum_{k=1}^{K} |s_{jk,t} - s_{ik,t}|. \tag{5.1}$$

Distance thus defined measures the average absolute difference across opinion coordinates. Agents are nodes in a network, with an edge between agents reflecting a relationship and an opportunity for the agents to influence one another. The magnitude and direction of that influence is characterized by the *weight* of each edge. Weights are determined by the relative opinions of the two agents, as measured by their distance, and so can change dynamically. Positive weights represent positive influence, in which agents become closer in their opinions, while negative weights represent the tendency toward differentiation. For descriptive convenience, if two agents are connected with a positive weight, they could be considered "friends" and if the weight is negative they could be considered "enemies." In reality, no assumptions about such clear social roles are necessary. The weight of an edge between agents i and j is given by

$$w_{ij,t+1} = 1 - d_{ij,t}. \tag{5.2}$$

So, if the opinions of agents i and j are separated by $d_{ij} < 1$, the agents are friends and will harmonize their opinions. If $d_{ij} > 1$, the agents are enemies, and will drive each other's opinions to more extreme levels. This weighting rule embodies the psychological phenomena of *biased assimilation*, in which similar individuals grow more similar and dissimilar individuals grow further apart after interacting (Lord et al., 1979). This is a common assumption in models of social influence (Hegselmann & Krause, 2002; Flache & Macy, 2011; Dandekar et al., 2013)). It should be noted that while the empirical evidence for biased assimilation is quite strong, and spans almost four decades, it is less clear how coherence on various opinions or beliefs affects influence on orthogonal opinions or beliefs. The assumption in this model is that it is only average distance in opinions that matters.

At time $t + 1$, agents update their opinions by adding the average influence from all neighbor agents. For each opinion k, agent i uses the following update rule:

$$s_{ik,t+1} = s_{ik,t} + \Delta s_{ik,t} \left(1 - \mathrm{sgn}(s_{ik,t}) s_{ik,t}\right), \tag{5.3}$$

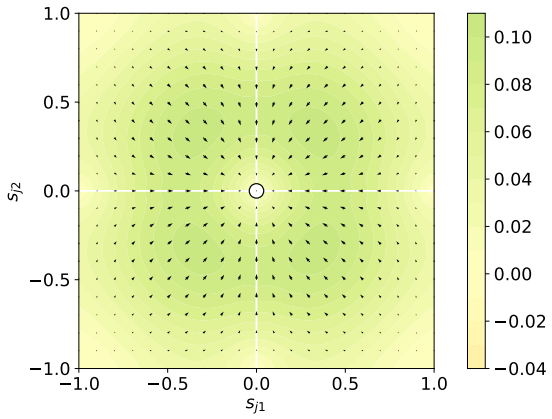
where

$$\Delta s_{ik,t} = \frac{1}{2N_i} \sum_{j \neq i} w_{ij,t}(s_{jk,t} - s_{ik,t}) + \epsilon. \tag{5.4}$$

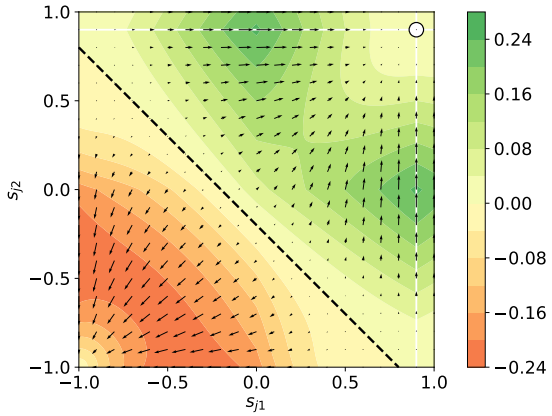
Here, N_i is the number of agents with which agent i shares an edge, and ϵ is a noise term that reflects errors in the communication of opinions. This term is in each instance drawn at random from a normal distribution with a mean of zero and a standard deviation of σ. We conceptualize updating to be the result of agents sensing the communicated opinions of neighbors. Furthermore, we conceptualize this σ as representing noise either in an agent sensing the opinions of other agents, noise in agents communicating their opinions, or both. In their original study Flache and Macy (2011) considered only scenarios without noise ($\sigma = 0$). Time in the model progressed in discrete time steps. At each time step, each agent's opinions were updated asynchronously in random order to avoid well-known artefacts that often accompany simultaneous agent updating.

It is worth noting a few immediate consequences of these update equations. First, agents with extreme opinions in dimension k will tend to make smaller changes to those opinions because of the smoothing factor $(1 - \text{sgn}(s_{ik,t})s_{ik,t})$. In other words, extreme opinions will be harder to change. Second, there are two opposing factors that modulate the magnitude of influence between two agents. On the one hand, edge weight is maximal when agents' opinions are very similar. On the other hand, $\Delta s_{ik,t}$ (which Flache and Macy refer to as the "raw" state change) increases the more agents' opinions differ, presumably because larger distances provide larger room for change, with a mathematical form drawn from psychological models of reinforcement learning (Rescorla & Wagner, 1972; Sutton & Barto, 1998). Influence will therefore be maximal for agents who are an intermediate distance apart in opinion space. To facilitate an intuitive understanding of dyadic interactions, we illustrate the strength of influence on agent opinions in $K = 2$ opinion space in Figure 5.1. We see that an agent with opinions at the origin of opinion space has only a moderate, attractive influence on other agent opinions in the opinion space. Agents at the corners of opinion space are barely influenced by a central opinion vector. When we consider the influence of an agent opinion nearer to the corner, at $\vec{s}_i = (0.9, 0.9)$, we see that there is a clear line where relationships

switch from friend to enemy ($s_{j2} = s_{j1} - 0.2$). Due to the co-mingling of effects described above, there is a varied and non-monotonic landscape of influence.



(a) Influence of agent at origin.



(b) Influence of agent at (0.9, 0.9).

Figure 5.1: Influence by one agent on another changes depending on the location of each agent. This illustrates the influence exerted by a central agent (white circle) on another agent at different locations in opinion space.

5.2.3 Measuring Polarization

There are a multitude of measures for polarization (Bramson et al., 2016) and no single measure is widely agreed upon. We follow Flache and Macy (2011) and define polarization at time t to be the variance of all distances between agents,

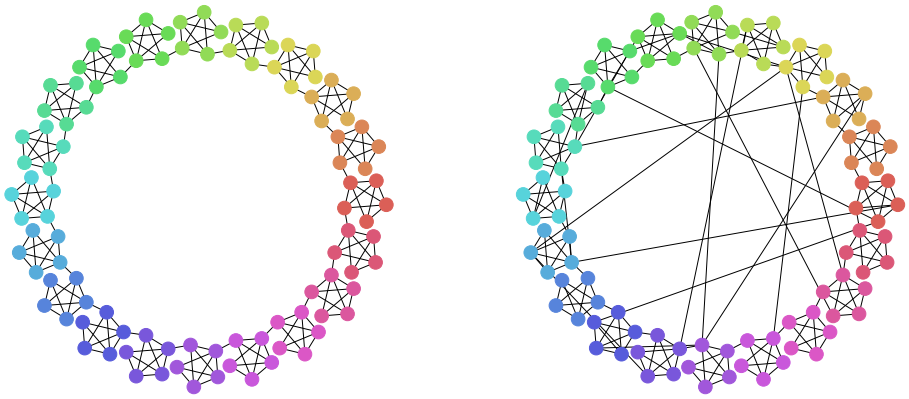
$$P_t = \text{var}(d_{ij,t}) \tag{5.5}$$

This metric has the advantage of simple interpretation. If half of all agents are in one corner of opinion space and the other half of agents are in the opposite corner, then the population is maximally polarized. As agent opinions spread to other corners and to other regions of opinion space, polarization will decrease. One disadvantage is that more general patterns of clustering, as would be detected using various machine learning clustering algorithms, will go undetected. In the final subsection of our Results, we illustrate another limitation of this metric. Nonetheless, we generally find that it is a useful and suitable operationalization for the concept of polarization.

5.2.4 Network structure

Our network structures are taken from Flache and Macy's (2011) Experiment 2. We begin with the connected caveman network structure introduced by Watts (1999). Specifically, we consider a network of $N = 100$ agents, grouped into 20 fully connected clusters (caves) of five agents each. These caves are arranged on a circle, and for each cave one edge is selected at random and rewired to connect to a random agent in the cave immediately to the right of the focal cave. This network has the appearance of tight-knit communities with weak ties to neighboring communities. The connected caveman network is highly clustered, meaning that if two agents are both neighbors of another single agent, there is a high probability that those two agents are also neighbors. However, relative path length is considerably greater in a connected caveman graph than for a totally random graph.

To assess the influence of adding long-range ties, we then consider a network for which 20 additional edges are added between randomly selected pairs of agents from across the entire network (Figure 5.2). Long-range ties are added at $t = 2000$ to give the local communities (caves) time to yield enclaves of conformity that differ slightly from their neighboring enclaves, following Flache and Macy (2011). The long range ties reduce the average path length of the network while retaining high clustering, yielding networks with "small-world" properties (Watts, 1999).



(a) Connected caveman graph before long-range ties added.

(b) After long-range ties added.

Figure 5.2: Connected caveman network with and without twenty long-range ties. Colors represent cave membership.

Finally, as a way to control for the effect of simply adding additional ties, we also consider the connected caveman network with *short-range* ties. In this case a randomly selected agent from each cave (who is not already connected to another cave) is connected to a random agent in the cave immediately to the right of the focal cave. Unless stated otherwise, all of our analyses were restricted to the connected caveman network with long-range ties, as this was the network structure found by Flache and Macy (2011) to maximize polarization.

5.2.5 Computational experiments

Below we present the results of our computational experiments. For all parameter combinations we ran 100 simulations of the model, with data collected after 10^4 time steps. This was always sufficient time for the system to settle down into a relatively stable pattern (true equilibria were not always reached due to the stochasticity inherent in the model). By calculating the difference in polarization on the final timestep for all simulations and finding all to be sufficiently small, we confirmed that 10^4 timesteps was sufficient to achieve stable behavior across

all simulations. We first replicate the major result of Flache and Macy (2011) that polarization increases with the addition of long-range ties but decreases with increasing cultural complexity, K. We then perform three sets of experiments:

1. *Quantifying variation.* We take a closer look at the variation among simulation runs, and explore path dependence on the road to polarization.

2. *Reducing extremism.* We investigate values of $S < 1$, in which the initial distribution of opinions is less extreme.

3. *Adding noise.* We investigate values of $\sigma > 0$, in which communication about opinions is noisy and influence is therefore more stochastic.

Unless stated otherwise, all simulations used a connected caveman network with random long-range ties, $S = 1$, and $\sigma = 0$. Model and analysis code is available on GitHub at `Mhttps://github.com/mt-digital/polarization`.

5.3 Results

In their original analysis of the FM model, Flache and Macy (2011) found two main causes of polarization. First, random long-range ties decreased the average path length of the network and increased the average polarization of the system across trials. Second, average polarization across trials decreased with increasing cultural complexity, K. We replicated these results, as illustrated in Figure 5.3. The remainder of this section is dedicated to novel results. The first three subsections show results of new analyses of the original FM model. The final subsection shows our analysis of the FM model modified to include communication noise.

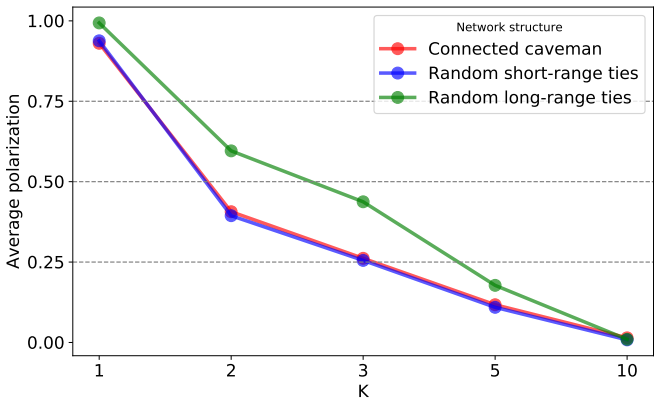
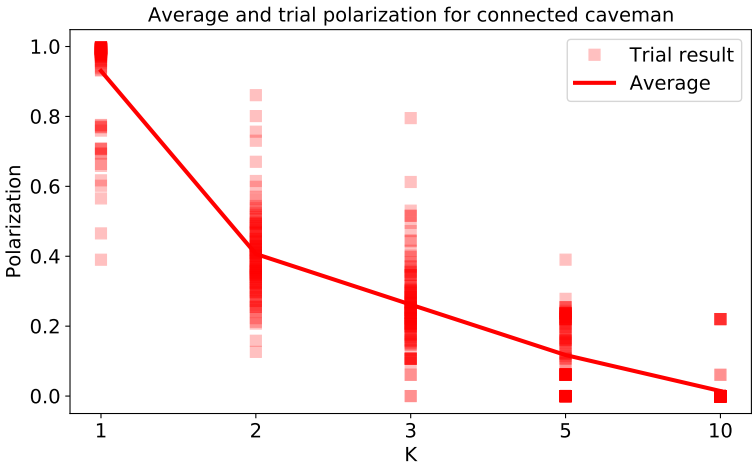


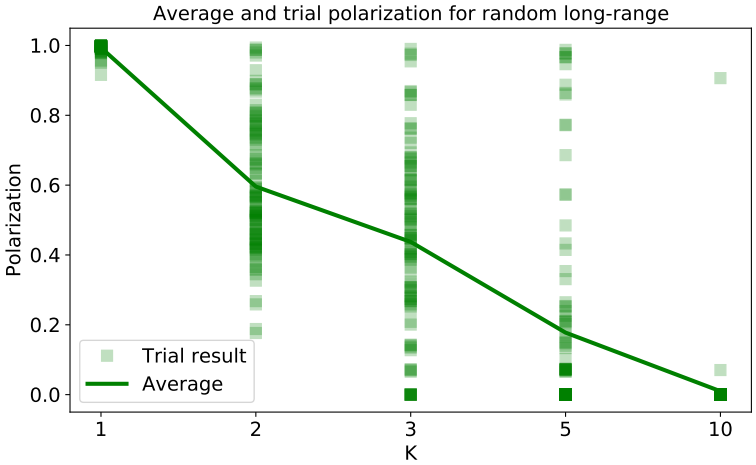
Figure 5.3: Reproduction of Figure 12b of Flache and Macy (2011). Average polarization decreases with K. However, as shown in subsequent figures, this does not mean trials with high polarization never obtain for large K. Average taken over 100 trials.

5.3.1 Polarization is probabilistic and path-dependent

Averages do not carry information about variation between trials. Here we explore that variation. Figure 5.4 shows the polarization for each of the individual trials averaged in Figure 5.3. We see a lot of variation around those averages, and that although polarization was low in all cases for large K, there are still individual trials for which polarization was high across all three network structures.



(a) Non-random connected caveman network.



(b) Randomized connected caveman network with long-range random ties added at iteration 2000.

Figure 5.4: Results of individual model runs under different network conditions. The averages of these were shown in Figure 5.3. Even in the non-random connected caveman structure, there is variation in the final polarization for different values of K. Highly polarized final states may obtain even for large K. 100 trials are shown for each network condition. Solid lines indicate the average across all trials.

In addition to the demonstrated influence of the overall network structure, three possible sources of variation in system polarization are (1) the initial distribution of agent opinions, (2) the initial distribution of how agent opinions are

clustered on the network, and (3) the update path—the order in which weights or agent opinions are updated. We performed additional analyses to investigate the contributions from each of these three factors, focusing on the initial distribution of agent opinions. We studied the non-random connected caveman network so as to keep network structure constant across trials, and for simplicity we restricted this analysis to $K = 2$. Due to the nature of our polarization measure, at initialization the system will have some non-zero degree of polarization, which will vary depending on the random draws of agents' initial opinions. Over 100 trials, we compare the initial polarization of the system to the final polarization. We found a significant, if relatively small, correlation between the initial and final polarization of agent opinions, $r^2 = .137$ (Figure 5.5). This means that the level of initial polarization accounts for only about 14% of the variation in final polarizations. It seems, then, that initial clustering of agent opinions and the stochasticity of the update path account for a large portion of the variability. In order to delineate the contributions of these two remaining factors to the overall variability in polarization, we considered the previously discussed simulations and ran 100 replicate trials with the initial conditions taken from the trials with the lowest and highest initial polarization. In other words, for each of two conditions, we ran replicate simulations with the exact same starting conditions between trials. Any variation in outcomes must therefore be due to stochasticity in the update paths. For example, if two opposing extremists influence a disjoint set of moderates disproportionately often, polarization will increase. The results are shown in Figure 5.6. Final polarization was clearly biased by the initial polarization (average final polarization across trials was 0.66 for the larger initial polarization, and 0.290 for the smaller initial polarization), but showed considerable variability. In other words, a large proportion of the variation between trials was due to stochasticity not in the initial configuration of the population, but to stochasticity in the transient dynamics of agent interactions.

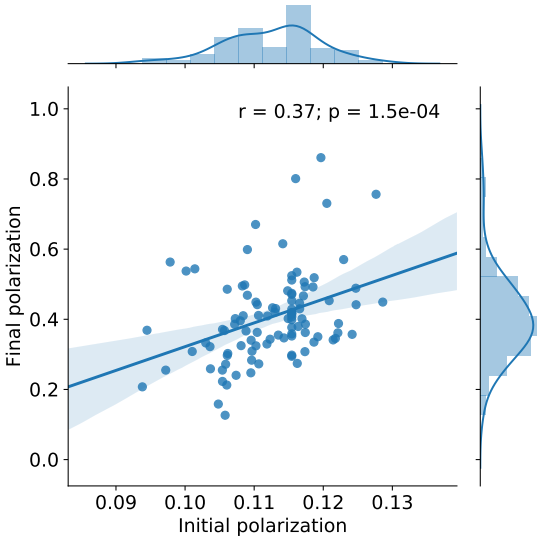


Figure 5.5: Regression of final polarization against initial polarization for $K = 2$ in the non-random connected caveman network configuration. Final polarizations are same as in the $K = 2$ column of Figure 5.4a. 100 trials are shown. The top histogram shows the distribution of initial polarization across trials. The right histogram shows the distribution of final polarization across trials.

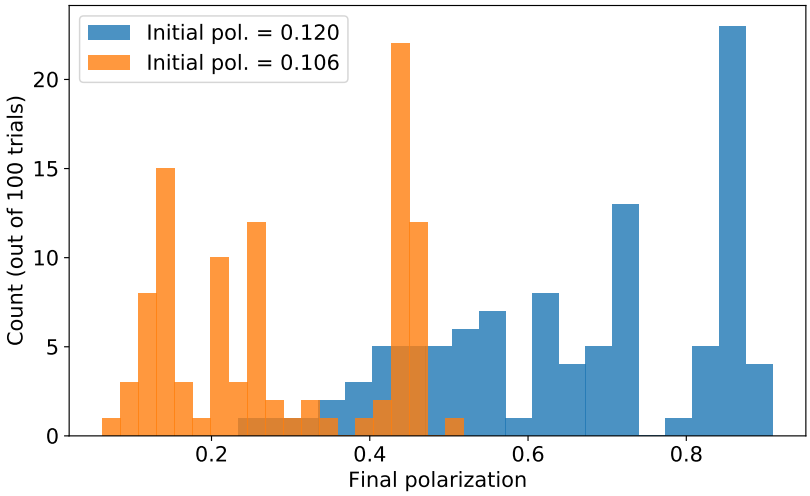


Figure 5.6: Distribution of final polarizations at $t = 10^4$ starting from initial conditions of either maximum or minimum polarization taken from the the connected caveman trials with $K = 2$.

5.3.2 The absence of initially extreme opinions reduces polarization

Next we extend our analysis of initial conditions further, by studying the breadth of opinions initially present in the population. Specifically, initial opinions were drawn from the uniform distribution $U(-S, S)$. Figures 5.7 and 5.8 show the mean and median polarization of the population as function of S, for $K = 2, \ldots, 6$. In general, the average final polarization decreased with smaller S for all values of K. The lines are not perfectly smooth due to the large variation in outcomes described in the previous section (see Figure 5.9).

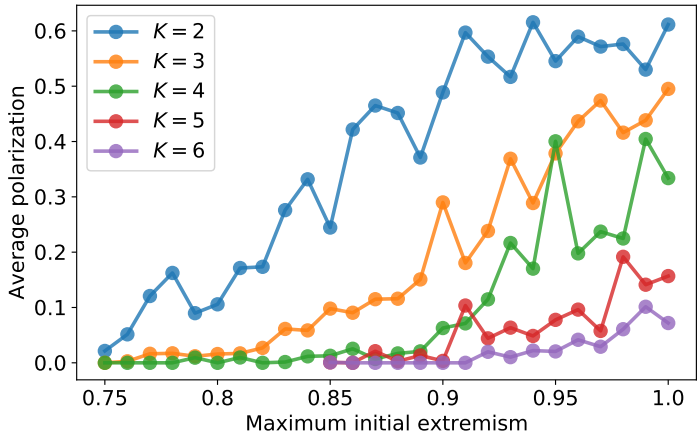


Figure 5.7: Average final polarization for different cultural complexities over maximum initial opinion magnitude, S. Averages are roughly zero for $S < 0.75$ for all cultural complexities.

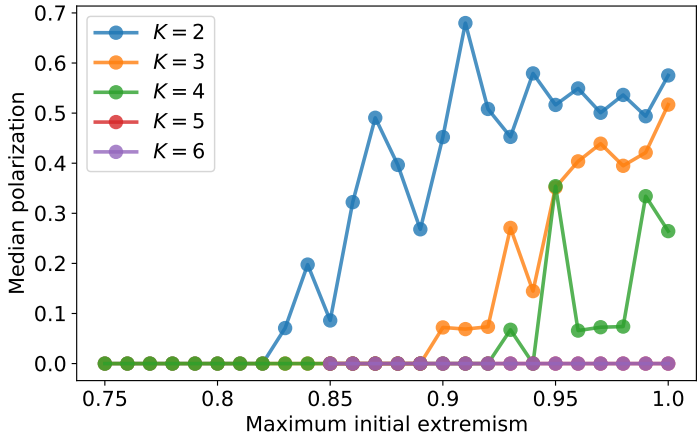


Figure 5.8: Median final polarization for different cultural complexities over maximum initial opinion magnitude, S. Median polarization for $K = 5$ and $K = 6$ are both flat at zero; $K = 5$ data is obscured by $K = 6$.

We again examined the within-condition variation in final polarization (Figure 5.9). Even when the average polarization was very small, we nevertheless saw instances of strongly polarized outcomes for $S < 1$ across all values of K. For small values of S, much more polarization occurred with small K. This further highlights the fact that initial conditions, in conjunction with the cultural complexity, bias the system towards larger or smaller levels of polarization, but do not eliminate the possibility of either conformity or extreme polarization.

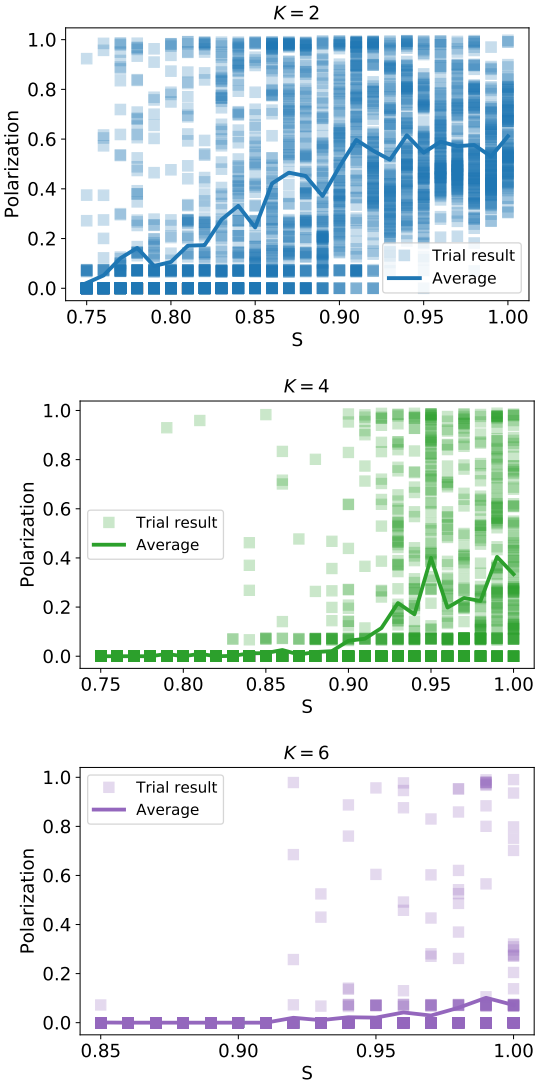


Figure 5.9: Final polarization of individual trial runs and averages from Figure 5.7 for a selection of K.

5.3.3 The meaning of polarization in high-dimensional opinion space

Clearly extreme positions are important in the FM model. Extremists are more stubborn (and therefore more influential) than centrists due to smoothing. Our analysis indicates that under a wide range of conditions, all opinions are likely to end up at extreme values. Indeed, the only stable states of the model are complete

consensus, which can be at any point in opinion space in the absence of noise, or for all opinions to be at extreme values. This brings us back to a key result of the FM model, which is that increased cultural complexity, K, decreases polarization. Recall that polarization is measured as the variance among distances between agent opinions. To what extent is this decrease in polarization with increased cultural complexity driven by the fact that, for larger K, there are simply more "corners" (extreme opinion values) for agent opinions to settle on?

We investigated this question by comparing polarization emerging from the dynamics of the FM model with polarization that occurs when agents are artificially placed on a random vertex of the K-dimensional opinion hypercube. We found the polarization for this combinatorial condition is $P_c \approx 1/K$ via Monte Carlo sampling with 100 agents and 1000 trials for each $K \in \{1, \ldots, 12\}$. In the Appendix we derive a formal proof that $P_c = 1/K$ exactly in the limit as $N \to \infty$.

When we compare the combinatorial result to the FM model results, we find that observed decrease in polarization with increased K follows the combinatorial results very closely (Figure 5.10). The connected caveman condition results in a lower polarization, on average, than P_c for all K that we tested. The random long-range condition results in an average polarization roughly equal to P_c for $K = 1$, higher average polarization than P_c from $K = 2$ to $K = 4$, and lower polarization for $K \geq 5$. The source of this jump from above-combinatorial to below-combinatorial is not clear, but is an interesting avenue for future work.

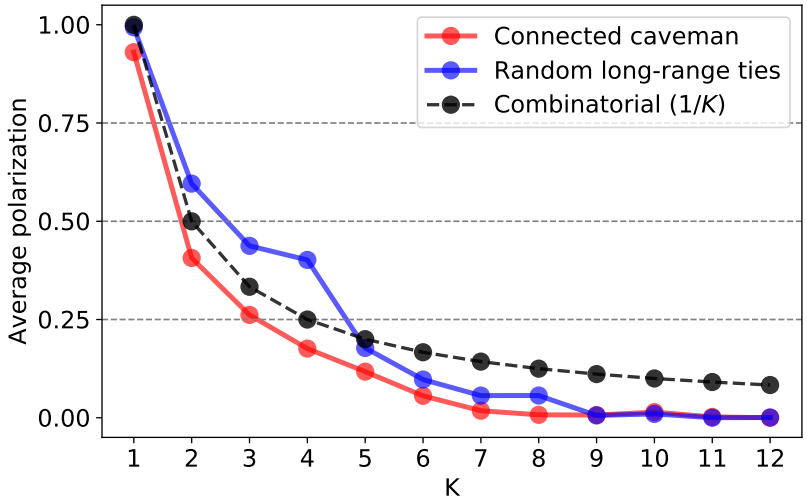


Figure 5.10: Polarization resulting from FM model simulations under connected caveman and random long-range tie conditions, compared with polarization resulting from agents arbitrarily choosing a corner of opinion space at random. Monte Carlo simulations revealed that polarization goes as $1/K$ if agents simply pick a corner at random. Random long-range and connected caveman data points are averaged from 100 trials with 10^4 iterations. Combinatorial condition data points are the average over 1000 trials and 10^4 iterations. Standard deviation around combinatorial trial averages was less than 10^{-2}.

5.3.4 Noisy communication increases polarization, particularly in the absence of initially extreme opinions

Up to this point, we have assumed that agents accurately express their own opinions and accurately receive information concerning the opinions of others. As this assumption is unlikely to fully hold in most cases of human interaction, it is important to assess the model's robustness to noisy communication. To do this, we introduced random error into the opinion update equation, so that every cultural feature communication channel, for every connected dyad, was modulated by a noise term, ϵ, drawn from a normal distribution with mean 0 and standard deviation σ. Let us call σ the "noise level." We varied the noise level from 0 to 0.2 in increments of 0.02. For each of these noise levels, we also varied S from 0.5 to

1.0 in steps of 0.05 for a total of 121 parameter pairs for each $K \in \{2, 3, 4, 5\}$. Note that we did not explicitly bound opinion components in the presence of noise. This led to us discarding 19 of the 60500 runs due to runaway opinions that diverged to infinity, and this was only for the highest noise levels used. These (discarded runs had noise levels of .18 or .2). Most parameter settings had only one discarded run if any, with one parameter setting having three discarded runs, lowering the number of samples to 97 from 100 for that parameter setting (K = 5, S = 0.95, and noise level= 0.2). This lack of smoothing had no effect on non-divergent model runs polarization outcomes, as polarization was less than or equal to 1.0 for all.

These experiments reveal an interesting pattern of results. A sufficiently large amount of noise produced high levels of polarization for low values of S, which never produced polarization in the absence of noise. Indeed, there appears to be a phase transition point for σ under low S, below which the system collapses to complete conformity and above which we see high levels of polarization (Figure 5.11). Across the values of K we tested, this threshold appealed to be around $\sigma = 0.8$, below which we never saw any polarization for low S (Figure 5.12). As S increases, however, the system behavior becomes less sensitive to noise, appearing to be completely insensitive to noise close to $S = 1$.

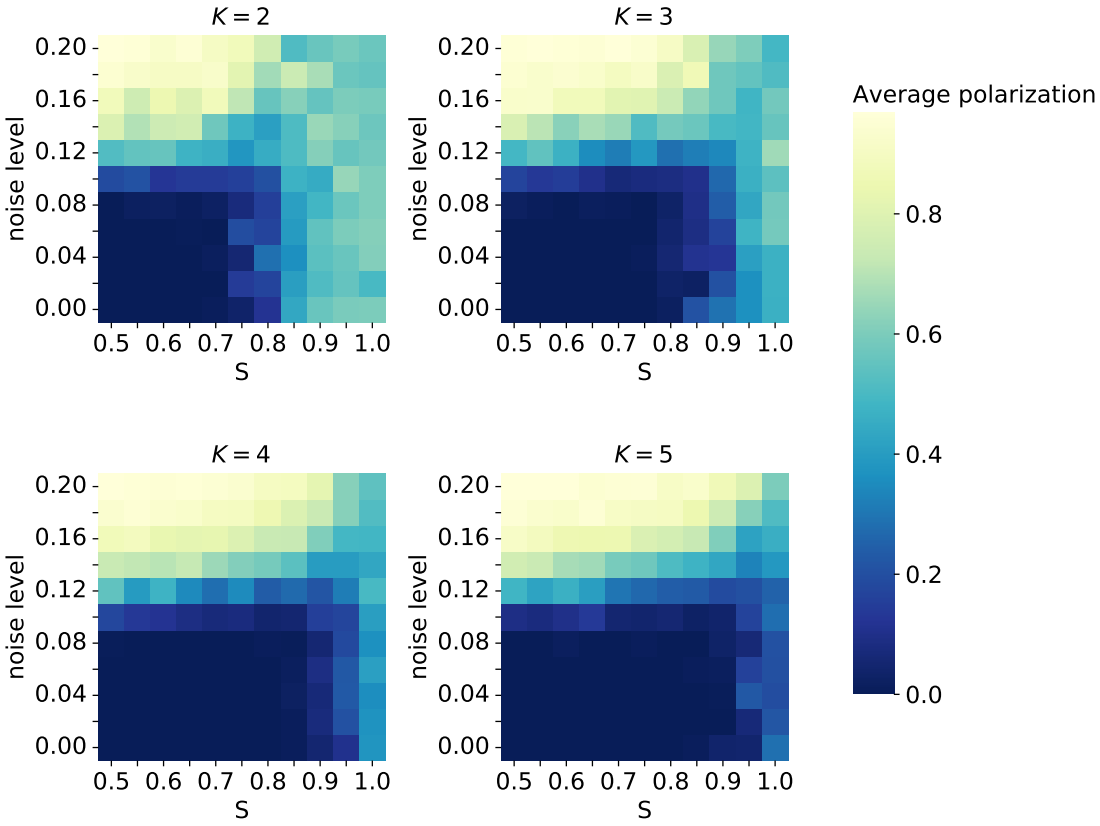


Figure 5.11: Final average polarization varies with both the width of the uniform distribution of initial opinion magnitudes and the noise level in the opinion updates. The value in each square of the heatmap is the average of 100 trials.

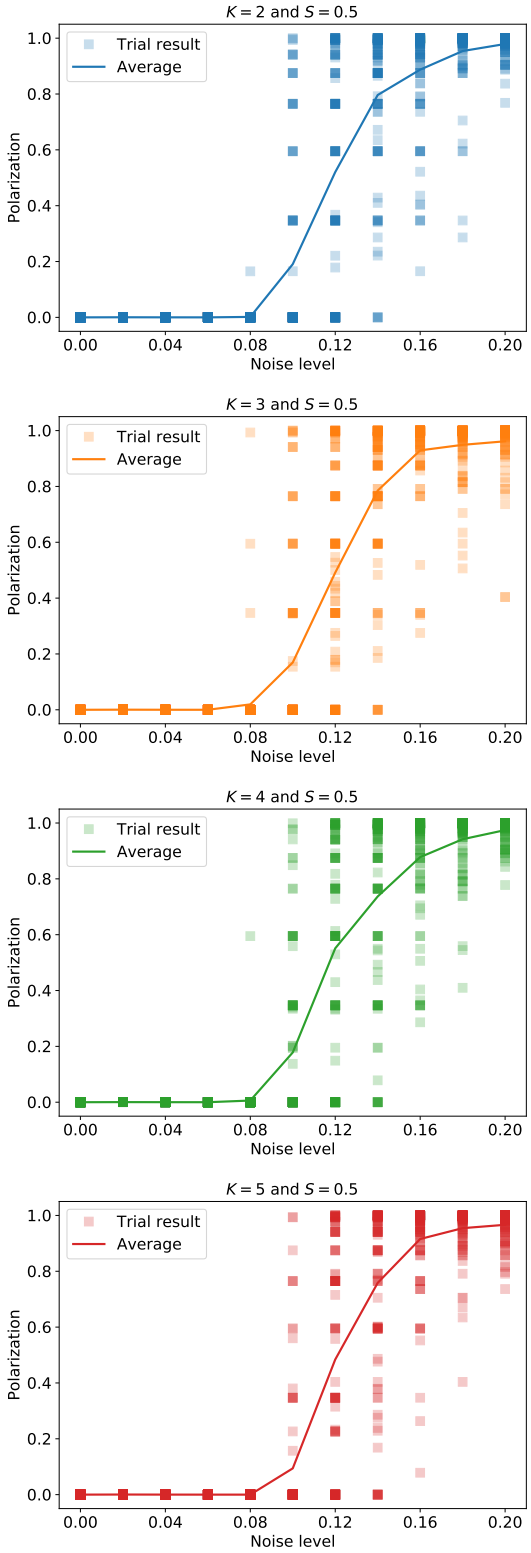


Figure 5.12: Final polarization of individual trial runs and averages from Figure 5.11 for $S = 0.5$ as a function of noise level, σ. As the noise level is increased, the system is increasingly biased towards larger final polarization outcomes.

Even though polarization is rare at moderate noise levels, extremism is not. A noise level of over 0.1 was required to reliably drive the system to polarization in our simulations, but lower noise levels led to consensus around an extreme location in opinion space rather than at a most centrist position. We infer this because the average agent distance from center increases to the maximum, 1.0, with noise levels of only 0.6 (Figure 5.13). Thus, we obtain the interesting result that even small amounts of communication noise can move the population to extremist positions.

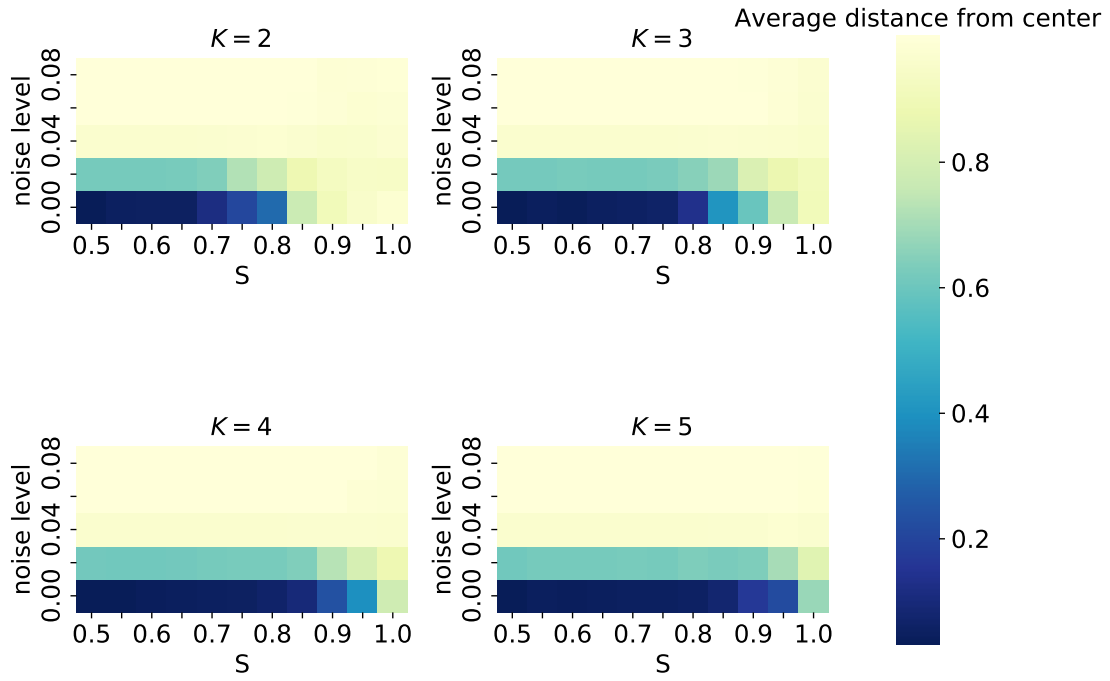


Figure 5.13: Noisy communication causes extremism without polarization before it causes extremism with polarization. For all K pictured, the average distance from center increases with moderate levels of noise, even though polarization has not increased, as shown in Figure 5.11. The value in each square of the heatmap is the average of 100 trials.

Figures 5.11 and 5.13 also illustrate a curious interaction between noise level, σ, and initial extremism, S. For smaller S, we observe clear phase transitions from centrist conformity to extremist conformity to polarization. For larger S, the populations responses are less clearly delineated. To help explain, we present illus-

trations of the spatiotemporal dynamics of the model for exemplar trials. Consider first a case of very low initial extremism, $S = 0.5$ (Figure 5.14). In the absence of noise, the system collapses around the center of opinion space at $t = 200$, and by $t = 3000$ has reached full consensus (Figure 5.14, top row). At the other extreme, under high levels of noise, $\sigma = 0.2$, agents reach a near-consensus by $t = 1000$ and remain there until $t = 2000$, when random long-range ties are added. At this point, agents are exposed to individuals with very slightly different sets of opinions, and those differences are amplified by the noise, leading to repulsion. This is sufficient to jolt the system away from conformity and into opposing camps moving towards opposing corners (Figure 5.14, bottom row).

For $\sigma = 0.08$ we found most simulations end in extreme consensus. That is, all opinions were at the extremes (± 1) rather than closer to zero, but these opinions were universally shared so that final polarization was zero. One such trial is shown in the middle row of Figure 5.14.This occurs because noise is sufficient to move the population toward the extremes (from which it is difficult to return to center), but agents remain sufficiently clustered so that all forces remain attractive rather than repulsive.

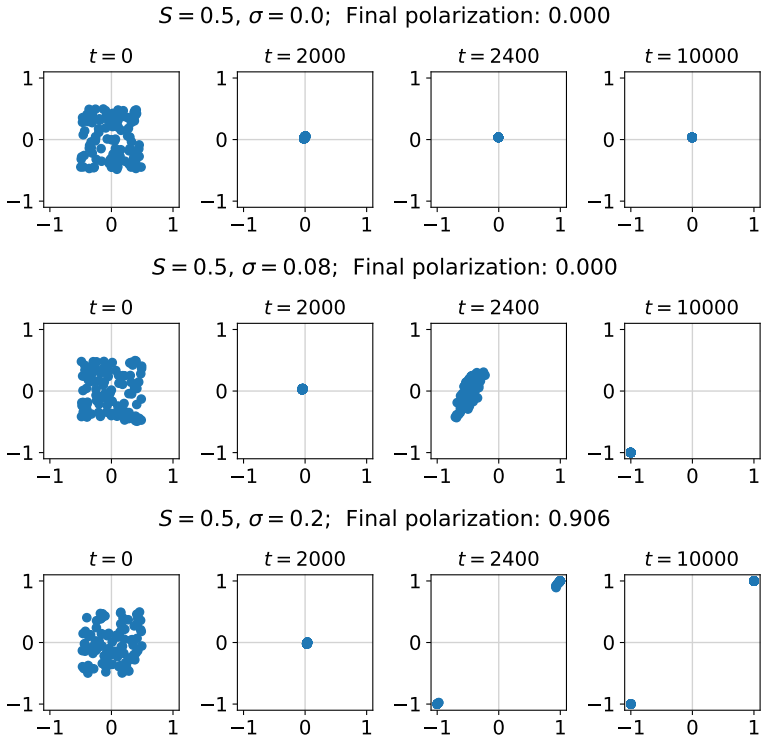


Figure 5.14: Exemplar spatiotemporal dynamics of agent opinion coordinates with $K = 2$ and $S = 0.5$ for $\sigma \in \{0.0, 0.08, 0.2\}$. There are three regimes. In the first, without noise, every simulation ends in centrist consensus (top row). In the presence of noise with $\sigma = 0.08$, agents find extremist consensus; in this trial agents found consensus around the point $(-1, -1)$. The third regime is the high polarization regime at the highest level of communication noise we tested, $\sigma = 0.2$. In this regime, agents split into opposing camps, led by first-mover extremists.

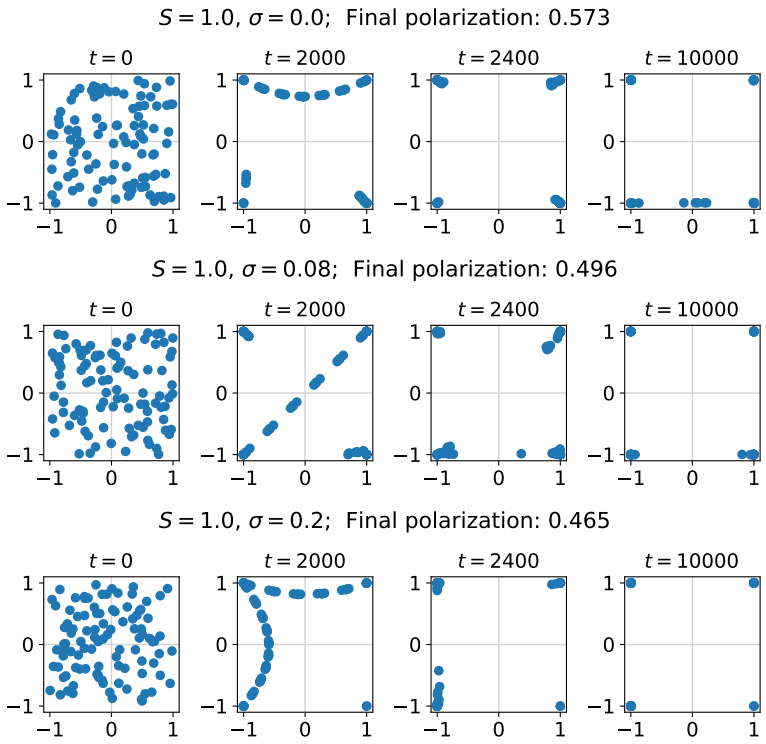
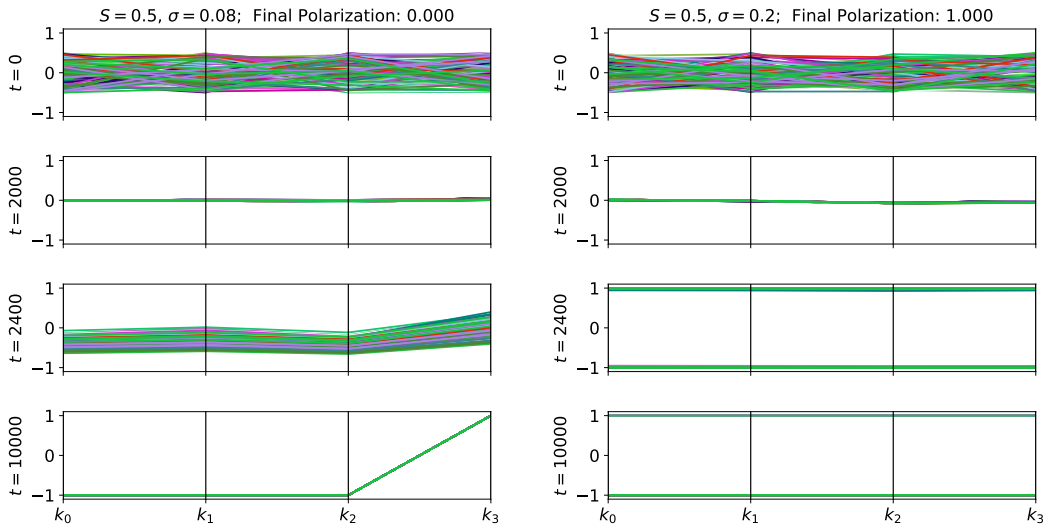


Figure 5.15: Exemplar spatiotemporal dynamics of agent opinion coordinates with $K = 2$ and $S = 1.0$ for $\sigma \in \{0.0, 0.08, 0.2\}$. Before the random long-range ties are added at $t = 2000$, extremists pull centrists to the extremes, but more centrist agent caves are balanced between more extreme caves. When long-range ties are added, the balance is broken and agents proceed to move to one of the extremes. Because at least some extremists held each of the corners, centrist agents do not move only to polar opposite corners, but in many cases to the nearest corner contained a neighboring (in the network sense) agent.



(a) Moderate noise, extreme consensus (b) More noise, polarization

Figure 5.16: Exemplar parallel coordinate timeseries for $K = 4$ and $S = 0.5$. Here the x-axis represents a single opinion coordinate, k_i, and the y-axis is the location of an agent for that coordinate. Each agent is represented by a line, colored by cave membership. With $\sigma = 0.1$, consensus emerges but at a corner of the opinion space.

When initial opinions are drawn from the full range of possibilities ($S = 1$), the system always achieves some degree of polarization. Because noise only serves to increase the likelihood of extreme opinions, this condition is unaffected by noise. Typical cases are shown in Figure 5.15. The behavior for $t \leq 2000$ is similar in all three cases: each cave reaches a local consensus, and the network of caves reaches a stable configuration. Some of the caves find consensus values at the corners. When random ties are added, the stable configuration is broken, and agents are pulled towards one of the four corners, where some caves have already been stably established. The caves in the corners do not move. Recall that a key assumption of the FM model is that extremist opinions influence centrist opinions more than centrists influence extremists. The noise is not strong enough to move extremists from extreme positions. In other words, in the presence of extreme opinions, network structure, not noise, dominates the dynamics. We extend the intuition to higher dimensions of opinions space using parallel coordinate plots,

visualizing time series of opinion dynamics for $K = 4$ (Figure 5.16).

5.4 Discussion

Humans are the quintessential cultural species. Our instinct to learn from others is a key reason for our domination of the planet (Henrich, 2015; Laland, 2017). An under-appreciated component of cultural learning concerns exacerbating differences and rejecting opinions when individuals are not likely to share one's current norms and beliefs. When those differences occur within a community, they can lead to discord. Many of us live in multicultural societies requiring cooperation and common ground, and so it natural to ask: when do we expect polarization, and is there anything we can do about it. Any suggestions based on our modeling efforts here should of course be compared with empirical studies. Hopefully these results stimulate further empirical work to understand when and why polarization emerges in real-world situations. One such opportunity for future work is to connect our findings to the political science literature on polarization (Sides & Hopkins, 2015), especially in relation to communication. If agents had different roles, such as elite agents (politicians and media) and common agents, we could model the effects of ideologically-biased news in political polarization (Prior, 2013; Pew Research Center, 2014a). Our results show that in the presence of sufficiently large communication noise and small-world networks, a situation we are arguably in today, a state of polarization is the only stable state (Figures 5.14, 5.15, and 5.16). It is interesting to consider this in light of one recent analysis suggesting that the United States Constitution was designed not just to accommodate polarization, but to foster it for the sake of stability (Wood & Jordan, 2017).

We have highlighted the stochastic nature of the system being modeled. A key conclusion is that empirical results of opinions on social networks may, when taken on a case-by-case basis, exhibit trends that bear little resemblance to those predicted by the model. This is not necessarily an invalidation of the model, but merely a consequence of the variability inherent in complex systems. That said, given enough data, key trends should emerge. We have confirmed Flache and

Macy's (2011) result that long-range ties increase polarization. As such, we might emphasize the importance of local communities being allowed to reach their own consensus. We have shown that decreasing initial extremism can reduce polarization, as one might expect. Achieving consensus in a community relies heavily on the absence of opinions at the extremes. However, this result is quite sensitive to noise in communication. A little bit of noise can shift consensus from centrist or ambivalent positions to more extreme views, while more noise can lead to polarization. Even if polarization is to be avoided, what about the intermediate case of "extreme consensus"? While it may be natural to view extreme opinions as undesirable, an alternative perspective is that they represent a more stable system of cultural coherence. Note that these findings contradict computational and mathematical studies of the bounded confidence model under the influence of noise, where sufficient noise breaks polarization and leads to disordered opinion spreading (Pineda, Toral, & Hernndez-García, 2009; Carro, Toral, & San Miguel, 2013; Kurahashi-Nakamura, Mäs, & Lorenz, 2016). This is because in the bounded confidence model, agents that are too far from one another do not interact. In the FM model, connected agents always interact, and the further apart they are in opinion space, the more strongly they repel one another in opinion space.

We confirmed Flache and Macy's (2011) result that increased "cultural complexity"—the number of opinions that are important to individuals in assessing their similarities and differences with others—decreased overall polarization. We also showed that this result stems directly from an increase in the number of permutations of extreme opinions individuals can hold when there are more items on which one can hold opinions. This might be viewed as a flaw in the metric of polarization used here. Alternatively, we believe it is reasonable to posit that a community with a wider diversity of views should be considered less polarized than a community with only a few suites of clustered opinions. In any case, this finding highlights the importance of a thorough understanding of ones distance measure when dealing with multidimensional opinions. Our analysis may in fact cast doubt on the interpretation by Flache and Macy (2011) that cultural complexity decreases opinion polarization, if one also rejects the interpretation that adding arbitrary traits on

which actors are indifferent should reduce their opinion distance.

As noted, the model we have studied is a simplified abstraction, and does not include many details that are important to the empirical reality of opinion dynamics. In general, theoretical modeling work should start simple, and gradually add heterogeneity as the simpler versions of the system in question become fully described. Future work should explore these sources of heterogeneity. First, we did not distinguish between private opinions and public productions representing those opinions (Nowak et al., 1990). Our operationalization of communication noise could be interpreted as a modulation of private opinion, but communication noise could also be interpreted as misunderstanding of perfectly-reproduced, publicly voiced opinions. People often communicate public opinions that differ from their private opinions when incentives for the parties involved are not aligned (Crawford & Sobel, 1982; Pinker, Nowak, & Lee, 2008; Smaldino, Flamson, & McElreath, 2018). Second, we ignored the structural influence of explicit identity groups. It could be argued that clustering of agent opinions implicitly defines an identity group. For example, DellaPosta et al. (2015) measured network autocorrelation to explain why people's preferences cluster together. This data-driven approach was offered as an attempt to explain arbitrary opinion clustering, as indicated by the paper's title, "Why do liberals drink lattes?". Nevertheless, explicit identity with groups and roles influences human behavior far beyond homophilic clustering (Barth, 1969; Berger & Heath, 2008; Smaldino, 2018). Third, we ignored individual differences in how individuals influence and are influenced. Some people may be stubborn while others are easily swayed. Some prestigious or charismatic individuals may have outsized influence while others are ineffective at communicating their opinions. Relatedly, individuals may also vary in their confidence in their opinions, which will influence the extent of their mutability and persuasion. The assumption that as agents become more extreme, their opinions become more stubborn, as formalized in Equation 5.3, may not always hold. Indeed, our work highlights the need for additional empirical work on how individuals alter their opinions as a function of how extreme those opinions are. Finally, the social networks used in our model are simplistic in both dynamics and structure. Ties in

many real world networks change with greater frequency than we modeled, providing new opportunities for social influence. Moreover, interactions and opinions are contextual. Individuals are embedded in multilayered social networks, in which the dynamics of opinions may be considerably more nuanced than indicated by our relatively static, single-layer network (Battiston et al., 2017; Smaldino, D'Souza, & Maoz, 2018).

In our study of the FM model we have found rich behaviors and theoretical lessons for understanding opinion dynamics. This work highlights the potential for complexity even in a very simple model of individual behavior, because network structure provides for path dependent effects and can be further influenced by initial conditions and noise. Our analytic approach highlights the value of systematic investigation of a model's explicit and tacit assumptions.

Appendix: Proof that polarization scales with $1/K$

We hypothesized that the decrease in polarization with increasing K observed in simulations of the FM model were driven by an increase in the number of permutations of binary vectors of length K, in which each element was -1 or 1. We supported this hypothesis in the main text with simulations in which agents were randomly initialized at such extreme positions in opinion space. Here we derive a formal proof that polarization in the FM model scales with $1/K$ if we assume that agents are randomly assigned a vector of "extreme" opinions, such that $\forall i, k,\ s_{ik} \in \{-1, 1\}$. To do this, we exactly calculate the polarization of a population where each agent occupies one of the 2^K corners of opinion space with K cultural features.

Recall that polarization is defined as the variance in pairwise distances between all agents. We define the *combinatorial polarization*, $P_c(K)$, as the polarization that arises from randomly placing each agent at one of the 2^K corners of opinion space with K cultural features, which is a K-hypercube, denoted Q_K. "Corners" of opinion space are simply vertices in the graph of Q_K. We computed this value numerically for $K \in \{1, \ldots, 12\}$ and found it tracks closely to $1/K$ (see Figure 5.10).

Here we demonstrate that $P_c = 1/K$ exactly as $N \to \infty$. To calculate $P_c(K)$, we need three elements. First, we need to calculate the distance between pairs of agents at different corners of Q_K. Second, we must count the number of agent pairs separated by the distance from one corner to another. We do this by first counting the number of subcubes of dimension L, or L-subcube. Then we count the number of maximally separated pairs in a subcube of L-subcube. Finally, we calculate the distance of maximally separated, or *antipodal* pairs, of agents in an L-subcube. We can then calculate the expected value of pairwise distances, $\langle d \rangle$, and the expected square of pairwise distance, $\langle d^2 \rangle$, from which we will have the combinatorial polarization

$$P_c = \langle d^2 \rangle - \langle d \rangle^2 \tag{5.6}$$

We will show that $P_c = \frac{1}{K}$ by showing that $\langle d \rangle = 1$ and $\langle d^2 \rangle = \frac{K+1}{K}$. Before we do that, we will derive functions to help us count the number of pairs separated by a particular distance, and to calculate distances between vertices on subcubes $Q_L \subseteq Q_K$. First, we denote the total number of pairwise distances as $n = \frac{N(N-1)}{2}$ where N is the number of agents. The number of L-subcubes $Q_L \subseteq Q_K$ is

$$n_s(L, K) = 2^{K-L} \binom{K}{L} \tag{5.7}$$

This results from the fact that at all 2^K vertices of Q_K, $\binom{K}{L}$ subcubes can be created by choosing L nodes adjacent to the vertex. This gives us $2^K \binom{K}{L}$ subcubes. This overcounts since each generated subcube was generated once for each of its 2^L vertices. So we must divide by a factor of 2^L, giving us the expression in Equation 5.7.

Within Q_L, the number of pairwise distances where agents occupy antipodal vertices is

$$n'_a(L, K) = 2^{L-1} \left(\frac{N}{2^K} \right)^2$$

There are 2^{L-1} pairs of antipodal vertices in Q_L. In the large N limit, agents are distributed in equal number to each vertex of Q_K. Then, the number of agents in a

single vertex is $\frac{N}{2^K}$, so the number of pairwise distances between any two antipodal pairs is $\left(\frac{N}{2^K}\right)^2$. The total number of antipodal pairs across all Q_L is then

$$n_a(L,K) = n'_a(L,K)n_s(L,K). \tag{5.8}$$

Finally, the distance between agent opinions $\vec{s}_1$ and $\vec{s}_2$ in antipodal vertices of Q_L is

$$d_a(L,K) = \frac{1}{K}\sum_{k=1}^{K}|s_{1k} - s_{2k}| = \frac{2L}{K} \tag{5.9}$$

since any antipodal vertices of Q_L share $K-L$ opinion coordinates, and the maximum magnitude of difference on a single opinion dimension is 2.

With these quantities we can write the expected value of pairwise distance,

$$\langle d\rangle = \frac{1}{n}\sum_{L=1}^{K} n_a(L,K)d_a(L,K). \tag{5.10}$$

Simplifying and taking $N \to \infty$, this becomes

$$\langle d\rangle = \frac{(K-1)!}{2^{K-1}}\sum_{L=1}^{K}\frac{1}{(K-L)!(L-1)!}$$

Using the identity

$$\sum_{L=1}^{K}\frac{1}{(K-L)!(L-1)!} = \frac{2^{K-1}}{(K-1)!},$$

we find $\langle d\rangle = 1$. Calculating $\langle d^2\rangle$ proceeds similarly, beginning with

$$\langle d^2\rangle = \frac{1}{n}\sum_{L=1}^{K} n_a(L,K)d_a(L,K)^2. \tag{5.11}$$

Simplifying and taking $N \to \infty$, this becomes

$$\langle d^2\rangle = \frac{(K-1)!}{2^{K-2}K}\sum_{L=1}^{K}\frac{L}{(K-L)!(L-1)!}.$$

With the identity

$$\sum_{L=1}^{K}\frac{L}{(K-L)!(L-1)!} = \frac{2^{K-2}(K+1)}{(K-1)!}$$

we find $\langle d^2\rangle = \frac{K+1}{K}$. So

$$P_c = \langle d^2\rangle - \langle d\rangle^2 = \frac{K+1}{K} - 1 = \frac{1}{K}. \tag{5.12}$$

Data Availability

Data used for our analyses is available for download (∼14GB) from `Mhttp://mt.digital/static/data/polarization_v0.1-data.tar`.

Conflicts of Interest

There are no conflicts of interest for either author.

Acknowledgements

Computational experiments were performed on the MERCED computing cluster, which is supported by the National Science Foundation [Grant No. ACI-1429783].

References

Abney, D. H., Dale, R., Yoshimi, J., Kello, C. T., Tylén, K., & Fusaroli, R. (2014). Joint perceptual decision-making: A case study in explanatory pluralism. *Frontiers in Psychology*, *5*(APR), 1–12. doi: M10.3389/fpsyg.2014.00330

Abney, D. H., Paxton, A., Dale, R., & Kello, C. T. (2021). Cooperation in sound and motion: Complexity matching in collaborative interaction. *Journal of Experimental Psychology: General*. doi: M10.1037/xge0001018

Abney, D. H., Paxton, A., Dale, R., Kello, C. T., Abney, D. H., Paxton, A., ... Kello, C. T. (2014). Complexity Matching in Dyadic Conversation. *Journal of Experimental Psychology: General*, *143*(6), 2304–2315.

Abrams, D., Wetherell, M., Cochrane, S., Hogg, M. A., & Turner, J. C. (1990). Knowing what to think by knowing who you are: Selfcategorization and the nature of norm formation, conformity and group polarization. *British Journal of Social Psychology*, *29*(2), 97–119. doi: M10.1111/j.2044-8309.1990.tb00892.x

Acemoglu, D., Ozdaglar, A., & Yildiz, E. (2011). Diffusion of innovations in social networks. *Proceedings of the IEEE Conference on Decision and Control*, 2329–2334. doi: M10.1109/CDC.2011.6160999

Adams, J., & Roscigno, V. J. (2005). White Supremacists, Oppositional Culture and the World Wide Web. *Social Forces*, *84*(2), 759–778. Retrieved from `Mhttps://academic.oup.com/sf/article-lookup/doi/10.1353/sof.2006.0001` doi: M10.1353/sof.2006.0001

Arendt, D. L., & Blaha, L. M. (2015). Opinions, influence, and zealotry: a computational study on stubbornness. *Computational and Mathematical Organization Theory*, *21*(2), 184–209. doi: M10.1007/s10588-015-9181-1

Aristotle. (1965). *Poetics*. Oxford: Blackwell.

Asch, S. E. (1951). Effects of group pressure upon the modification and distrotion of judgment. *Groups, leadership and men*.

Asch, S. E. (1955). Opinions and Social Pressure. *Scientific American*, *193*(5), 31–35. Retrieved from `Mhttp://www.freepatentsonline.com/WO2003027106.html` doi: M10.1038/scientificamerican1155-31

Asch, S. E. (1956). Studies of Independence and Conformity. *Psychological Monographs: General and Applied*, *70*(9).

Axelrod, R. (1997). The Dissemination of Culture: A Model with Local Convergence and Global Polarization. *Journal of Conflict Resolution*, *41*(2), 203–226. doi: M10.1177/0022002797041002001

Backstrom, L., Boldi, P., Rosa, M., Ugander, J., & Vigna, S. (2012). Four degrees of separation. In *Proceedings of the 4th annual acm web science conference* (Vol. 2, pp. 33–42). doi: M10.1038/nmat970

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Fallin Hunzaker, M. B., ... Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(37), 9216–9221. doi: M10.1073/pnas.1804840115

Bala, V., & Goyal, S. (1998). Learning from Neighbours. *Review of Economic Studies*, *65*(3), 595–621. doi: M10.1111/1467-937X.00059

Baldassarri, D., & Bearman, P. (2007). Dynamics of Political Polarization. *American Sociological Review*, *72*(5), 784–811. doi: M10.1177/000312240707200507

Banisch, S., & Olbrich, E. (2019, apr). Opinion polarization by learning from social feedback. *Journal of Mathematical Sociology*, *43*(2), 76–103. doi: M10.1080/0022250X.2018.1517761

Baron, R. S., & Roper, G. (1976). Reaffirmation of social comparison views of choice shifts: Averaging and extremity effects in an autokinetic situation. *Journal of Personality and Social Psychology*, *33*(5), 521–530. doi: M10.1037//0022-3514.33.5.521

Barron, A. T. J., Huang, J., Spang, R. L., & DeDeo, S. (2018). Individuals, Institutions, and Innovation in the Debates of the French Revolution. *Proc. Natl. Acad. Sci.*, *115*(18), 4607–4612. Retrieved from `Mhttp://arxiv.org/abs/1710.06867` doi: M10.1073/pnas.1717729115

Barth, F. (1969). *Ethnic groups and boundaries: The social organization of culture difference*. Little, Brown.

Battiston, F., Nicosia, V., Latora, V., & San Miguel, M. (2017). Layered social influence promotes multiculturality in the Axelrod model. *Scientific Reports*, *7*(1), 1809.

Berger, J., & Heath, C. (2008). Who drives divergence? identity signaling, outgroup dissimilarity, and the abandonment of cultural tastes. *Journal of Personality and Social Psychology*, *95*(3), 593.

Bicchieri, C. (2006). *The grammar of society: the nature and dynamics of social norms*. Cambridge: Cambridge University Press.

Bicchieri, C. (2017). *Norms in the Wild*. Oxford: Oxford University Press.

Bicchieri, C., & Mercier, H. (2014). Norms and Beliefs: How Change Occurs. *The Jerusalem Philosophical Quarterly*, *63*, 60–82.

Billig, M., & Tajfel, H. (1973). Social categorization and similarity in intergroup behaviour. *European Journal of Social Psychology*, *3*(1), 27–52. doi: M10.1002/ejsp.2420030103

Bishop, G. D., & Myers, D. G. (1974). Informational influence in group discussion. *Organizational Behavior and Human Performance*, *12*(1), 92–104. doi: M10.1016/0030-5073(74)90039-7

Blascovich, J., & Ginsburg, G. P. (1974). Emergent Norms and Choice Shifts Involving Risk. *Sociometry*, *37*(2), 205–218.

Blascovich, J., Ginsburg, G. P., & Howe, R. C. (1976). Blackjack, Choice Shifts in the Field. *Sociometry*, *39*(3), 274. doi: M10.2307/2786521

Blascovich, J., Veach, T. L., & Ginsburg, G. P. (1973). Blackjack and the Risky Shift. *Sociometry*, *36*(1), 42–55.

Blascovich, J. I. M., Ginsburg, G. P., & Howe, R. C. (1975). Blackjack and the Risky Shift , II: Monetary Stakes. *Journal of Experimental Social Psychology*,

11, 224–232.

Blau, P. M. (1974). Parameters of social structure. *American Sociological Review*, *39*(5), 615–635.

Block, P., Hoffman, M., Raabe, I. J., Dowd, J. B., Rahal, C., Kashyap, R., & Mills, M. C. (2020). Social network-based distancing strategies to flatten the COVID-19 curve in a post-lockdown world. *Nature Human Behaviour*, *4*(6), 588–596. Retrieved from Mhttp://dx.doi.org/10.1038/s41562-020-0898-6 doi: M10.1038/s41562-020-0898-6

Borge-Holthoefer, J., Magdy, W., Darwish, K., & Weber, I. (2015). Content and Network Dynamics Behind Egyptian Political Polarization on Twitter. In *Cscw '15: Computer supported cooperative work and social computing* (pp. 700–711). Retrieved from Mhttp://arxiv.org/abs/1410.3097 doi: M10.1145/2675133.2675163

Bornstein, R. F., Kale, A. R., & Cornell, K. R. (1990). Boredom as a Limiting Condition on the Mere Exposure Effect. *Journal of Personality and Social Psychology*, *58*(5), 791–800. doi: M10.1037/0022-3514.58.5.791

Boxell, L., Gentzkow, M., & Shapiro, J. M. (2017). Greater Internet use is not associated with faster growth in political polarization among US demographic groups. *Proceedings of the National Academy of Sciences*, *115*(3), 201706588. Retrieved from Mhttp://www.pnas.org/lookup/doi/10.1073/pnas.1706588114 doi: M10.1073/pnas.1706588114

Boxell, L., Gentzkow, M., & Shapiro, J. M. (2020). Cross-country trends in affective polarization (Working Paper 26669). *National Bureau of Economic Research*. Retrieved from Mhttps://www.nber.org/papers/w26669

Bramson, A., Grim, P., Singer, D. J., Fisher, S., Berger, W., Sack, G., & Flocken, C. (2016). Disambiguation of social polarization concepts and measures. *Journal of Mathematical Sociology*, *40*(2), 80–111. doi: M10.1080/0022250X.2016.1147443

Brewer, M. B. (2013). 25 Years Toward a Multilevel Science. *Perspectives on Psychological Science*, *8*(5), 554–555. doi: M10.1177/1745691613500996

Brown, R. (1974). Further comment on the risky shift. *American Psycholo-*

gist(June), 468–470.

Brown, R. (1986). Group polarization. In *Social psychology* (2nd ed., pp. 200–248). New York: Free Press.

Brown, R. (2000). *Group Processes*. Malden, MA: Blackwell.

Burgers, C., Jong Tjien Fa, M., & de Graaf, A. (2019). A tale of two swamps: Transformations of a metaphorical frame in online partisan media. *Journal of Pragmatics*, *141*, 57–66. Retrieved from `Mhttps://doi.org/10.1016/j.pragma.2018.12.018` doi: M10.1016/j.pragma.2018.12.018

Burgess, T. D. G., & Sales, S. M. (1971). Attitudinal effects of "mere exposure": a reevaluation. *Journal of Experimental Social Psychology*, *7*, 461–472.

Burnes, S. (2011). Metaphors in press reports of elections: Obama walked on water, but Musharraf was beaten by a knockout. *Journal of Pragmatics*, *43*(8), 2160–2175. doi: M10.1016/j.pragma.2011.01.010

Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, *65*(1), 23–35. doi: M10.1007/s00265-010-1029-6

Burnstein, E., & Vinokur, A. (1973). Testing two classes of theories about group induced shifts in individual choice. *Journal of Experimental Social Psychology*, *9*(2), 123–137. doi: M10.1016/0022-1031(73)90004-8

Burnstein, E., & Vinokur, A. (1975). What a person thinks upon learning he has chosen differently from others: Nice evidence for the persuasive-arguments explanation of choice shifts. *Journal of Experimental Social Psychology*, *11*(5), 412–426. doi: M10.1016/0022-1031(75)90045-1

Burnstein, E., & Vinokur, A. (1977). Persuasive argumentation and social comparison as determinants of attitude polarization. *Journal of Experimental Social Psychology*, *13*(4), 315–332. doi: M10.1016/0022-1031(77)90002-6

Cacciatore, M. A., Scheufele, D. A., & Iyengar, S. (2016). The End of Framing as we Know it ... and the Future of Media Effects. *Mass Communication and Society*, *19*(1), 7–23. Retrieved from `Mhttp://dx.doi.org/10.1080/15205436.2015.1068811` doi: M10.1080/15205436.2015.1068811

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. doi: M10.1126/science.aal4230

Calvert, S. L., Appelbaum, M., Dodge, K. A., Graham, S., Nagayama Hall, G. C., Hamby, S., ... Hedges, L. V. (2017). The american psychological association task force assessment of violent video games: Science in the service of public interest. *American Psychologist*, *72*(2), 126–143. doi: M10.1037/a0040413

Carley, K. (1990). *Group Stability: A Socio-Cognitive Approach* (Vol. 7). Retrieved from `Mhttp://casos.cs.cmu.edu/publications/papers/carley{_}1990{_}groupstability.pdf`

Carley, K. (1991). A Theory of Group Stability. *American Sociological Review*, *56*(3), 331–354.

Carley, K. M. (1990). Group stability: A socio-cognitive approach. *Advances in Group Processes*, *7*(1), 44.

Carro, A., Toral, R., & San Miguel, M. (2013). The Role of Noise and Initial Conditions in the Asymptotic Solution of a Bounded Confidence, Continuous-Opinion Model. *Journal of Statistical Physics*, *151*(1-2), 131–149. doi: M10.1007/s10955-012-0635-2

Cartwright, D. (1971). Risk taking by individuals and groups: An assessment of research employing choice dilemmas. *Journal of Personality and Social Psychology*, *20*(3), 361–378. doi: M10.1037/h0031912

Cartwright, D. (1973). Determinants of scientific progress: The case of research on the risky shift. *American Psychologist*, *28*(3), 222–231. doi: M10.1037/h0034445

Cartwright, D., & Harary, F. (1956). Structural balance: A generalization of Heider's Theory. *Psychological Review*, *63*(5), 277–292.

Cartwright, N. (1989). *Nature's Capacities and their Measurment*. Oxford: Oxford University Press.

Cartwright, N. (1999). *The Dappled World: A Study of the Boundaries of Science*. Cambridge, UK: Cambridge University Press.

Castro Seixas, E. (2021). War Metaphors in Political Communication on Covid-19.

Frontiers in Sociology, *5*(January), 1–11. doi: M10.3389/fsoc.2020.583680

Center, P. R. (2016, may). News Use A cross Social Media Platforms 2016. , 20.

Centola, D., González-Avella, J. C., Eguíluz, V. M., & San, M. (2007). Homophily, Cultural Drift, and the Co-Evolution of Cultural Groups. *Journal of Conflict Resolution*, *51*(6), 905–929.

Chacoma, A., & Zanette, D. H. (2015). Opinion formation by social influence: From experiments to modeling. *PLoS ONE*, *10*(10), 1–16. doi: M10.1371/journal.pone.0140406

Charteris-Black, J. (2009). Metaphor and Political Communication. In *Metaphor and discourse* (pp. 97–115). London: Palgrave Macmillan UK.

Chong, D., & Druckman, J. N. (2007). Framing Theory. *Annual Review of Political Science*, *10*(1), 103–126. Retrieved from `Mhttp://www.annualreviews.org/doi/10.1146/annurev.polisci.10.072805.103054` doi: M10.1002/sia.3423

Cikara, M., & Van Bavel, J. J. (2014). The Neuroscience of Intergroup Relations: An Integrative Review. *Perspectives on Psychological Science*, *9*(3), 245–274. doi: M10.1177/1745691614527464

Cikara, M., Van Bavel, J. J., Ingbretsen, Z. A., & Lau, T. (2017). Decoding "Us" and "Them": Neural representations of generalized group concepts. *Journal of Experimental Psychology: General*, *146*(5), 621–631. doi: M10.1037/xge0000287

Claidière, N., & Whiten, A. (2012). Integrating the study of conformity and culture in humans and nonhuman animals. *;w Psychological Bulletin*, *138*(1), 126–145. doi: M10.1037/a0025868

Clark, H. (1996). *Using language* (5th ed.). Cambridge, UK: Cambridge University Press.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359. doi: M10.1016/S0022-5371(73)80014-3

Cohen, D., Nisbett, R. E., Schwarz, N., & Bowdle, B. F. (1996). Insult, aggression, and the Southern culture of honor: An "experimental ethnogra-

phy". *Interpersonal Relations and Group Processes*, *70*(5), 945–960. doi: M10.1037/0022-3514.70.5.945

Converse, P. E. (2006). The Nature of Belief Systems in Mass Publics (1964). *Critical Review*, *18*(1-3), 1–74. doi: M10.4324/9780203505984-10

Cooper, J., Kelly, K. A., & Weaver, K. (2001). Attitudes, Norms, and Social Groups. In M. A. Hogg & R. S. Tindale (Eds.), *Blackwell handbook of social psychology: Group processes* (pp. 259–282). Malden, MA: Blackwell.

Craver, C. F. (2006). When mechanistic models explain. *Synthese*, *153*(3), 355–376. doi: M10.1007/s11229-006-9097-x

Crawford, V. P., & Sobel, J. (1982). Strategic information transmission. *Econometrica*, *50*, 1431–1451.

Dandekar, P., Goel, A., & Lee, D. T. (2013). Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(15), 5791–6. doi: M10.1073/pnas.1217220110

David, O., Lakoff, G., & Stickles, E. (2016). Cascades in metaphor and grammar: A case study of metaphors in the gun debate. *Construction and Frames*, *8*(2), 214–253. doi: M10.1075/cf.8.2.04dav

Davis, J. H. (1973). "Group decision and social interaction: A theory of social decision schemes": Errata. *Psychological Review*, *80*(4), 302. doi: M10.1037/h0020065

Deffuant, G., Neau, D., Amblard, F., & Weisbuch, G. (2000). Mixing beliefs among interacting agents. *Advances in Complex Systems*, *3*(01n04), 87–98.

DeGroot, M. H. (1974). Reaching a Consensus. *Journal of the American Statistical Association*, *69*(345), 118–121.

DellaPosta, D., Shi, Y., & Macy, M. (2015). Why Do Liberals Drink Lattes? *American Journal of Sociology*, *120*(5), 1473–1511. doi: M10.1086/681254

Deroiain, F. (2002). Formation of social networks and diffusion of innovations. *Research Policy*, *31*(5), 835–846.

Dixit, A. K., & Weibull, J. W. (2007). Political Polarization. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(2),

7351–7356. doi: M10.1073/pnas.0702071104

Ebbesen, E. B., & Bowers, R. J. (1974). Proportion of risky to conservative arguments in a group discussion and choice shift. *Journal of Personality and Social Psychology*, *29*(3), 316–327. doi: M10.1037/h0036005

Efferson, C., Lalive, R., Richerson, P. J., McElreath, R., & Lubell, M. (2008). Conformists and mavericks: the empirics of frequency-dependent cultural transmission. *Evolution and Human Behavior*, *29*(1), 56–64.

Epstein, J. M. (2013). *Agent_Zero: Toward Neurocognitive Foundations for Generative Social Science*. Princeton: Princeton University Press.

Epstein, J. M., & Hammond, R. a. (2002). Non-explanatory equilibria: An extremely simple game with (mostly) unattainable fixed points. *Complexity*, *7*(4), 18–22. Retrieved from `Mhttp://www3.interscience.wiley.com/cgi-bin/fulltext?ID=97519448{&}PLACEBO=IE.pdf{%}5CnEpstein{&}Hammond(2002).pdf` doi: M10.1002/cplx.10026

Fahrentold, D. A. (2016, oct). *Trump recorded having extremely lewd conversation about women in 2005.* Washington, D.C.. Retrieved from `Mhttps://www.washingtonpost.com/politics/trump-recorded-having-extremely-lewd-conversation-about-women-in-2005/2016/10/07/3b9ce776-8cb4-11e6-bf8a-3d26847eeed4{_}story.html?noredirect=on{&}utm{_}term=.55e4277f8734`

Falandays, J. B., & Smaldino, P. E. (2021). The Emergence of Cultural Attractors : An Agent-Based Model of Collective Cognitive Alignment. In *Proceedings of the cognitive science society 2021.*

Fausey, C. M., & Matlock, T. (2011). Can grammar win elections? *Political Psychology*, *32*(4), 563–574. doi: M10.1111/j.1467-9221.2010.00802.x

Festinger, L. (1954). A Theory of Social Comparison Processes. *Human Relations*, *7*(2), 117–140. doi: M10.1177/001872675400700202

Fillmore, C. J. (1982). Frame Semantics. In L. S. of Korea (Ed.), *Linguistics in the morning calm.* Soeul: Hanshin.

Flache, A., & Macy, M. W. (2011). Small Worlds and Cultural Polarization. *The Journal of Mathematical Sociology*, *35*(1-3), 146–176.

Flache, A., & Mäs, M. (2008). How to get the timing right. A computational model of the effects of the timing of contacts on team cohesion in demographically diverse teams. *Computational and Mathematical Organization Theory*, *14*(1), 23–51. doi: M10.1007/s10588-008-9019-1

Flache, A., Mäs, M., Feliciani, T., Chattoe-Brown, E., Deffuant, G., Huet, S., & Lorenz, J. (2017). Models of social influence: Towards the next frontiers. *Journal of Artificial Societies & Social Simulation*, *20*(4), 2. Retrieved from Mhttp://jasss.soc.surrey.ac.uk/20/4/2.html

Flusberg, S. J., Matlock, T., & Thibodeau, P. H. (2017a). Metaphors for the War (or Race) against Climate Change. *Environmental Communication*, *0*(0), 1–15. Retrieved from Mhttp://dx.doi.org/10.1080/17524032.2017.1289111 doi: M10.1080/17524032.2017.1289111

Flusberg, S. J., Matlock, T., & Thibodeau, P. H. (2017b). Thinking about the future : The role of spatial metaphors for time. In *Cognitive science society.*

Flusberg, S. J., Matlock, T., & Thibodeau, P. H. (2018). War metaphors in public discourse. *Metaphor and Symbol*, *33*(1), 1–18. Retrieved from Mhttps://doi.org/10.1080/10926488.2018.1407992 doi: M10.1080/10926488.2018.1407992

Freeman, J. B. (2018). *The Field of Blood: Violence in Congress and the Road to Civil War*. New York: Farrar, Straus, and Giroux.

French, J. R. P. (1956). A Formal Theory of Social Power. *Psychological Review*, *63*(3).

Friedkin, N. E. (1986). *A formal theory of social power* (Vol. 12) (No. 2). doi: M10.1080/0022250X.1986.9990008

Friedkin, N. E. (1999). Choice Shift and Group Polarization. *American Sociological Review*, *64*(6), 856–875.

Fusaroli, R., Perlman, M., Mislove, A., Paxton, A., Matlock, T., & Dale, R. (2015). Timescales of massive human entrainment. *PLoS ONE*, *10*(4), 1–19. doi: M10.1371/journal.pone.0122742

Galam, S., & Jacobs, F. (2007). The role of inflexible minorities in the breaking of democratic opinion dynamics. *Physica A: Statistical Mechanics and its*

Applications, *381*(1-2), 366–376. doi: M10.1016/j.physa.2007.03.034

Gallese, V., & Lakoff, G. (2005). The Brain's concepts: the role of the Sensory-motor system in conceptual knowledge. *Cognitive neuropsychology*, *22*(3), 455–79. Retrieved from `Mhttp://www.ncbi.nlm.nih.gov/pubmed/21038261` doi: M10.1080/02643290442000310

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644. doi: M10.1073/pnas.1720347115

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, UK: Cambridge University Press.

Gentzkow, M., Shapiro, J. M., & Taddy, M. (2019). Measuring Group Differences in HighDimensional Choices: Method and Application to Congressional Speech. *Econometrica*, *87*(4), 1307–1340. doi: M10.3982/ecta16566

Gibbs, R. W. (1994). *The Poetics of Mind*. New York: Cambridge University Press.

Gibbs, R. W. (1997). Taking metaphor out of our heads and putting it into the cultural world. In *Metaphor in cognitive linguistics: Selected papers from the 5th international cognitive linguistics conference* (pp. 145 – 166). Amsterdam: John Benjamins Publishing Company.

Gibbs, R. W., & Van Orden, G. (2012, jan). Pragmatic Choice in Conversation. *Topics in Cognitive Science*, *4*(1), 7–20. Retrieved from `Mhttp://doi.wiley.com/10.1111/j.1756-8765.2011.01172.x` doi: M10.1111/j.1756-8765.2011.01172.x

Gray, K., Rand, D. G., Ert, E., Lewis, K., Hershman, S., & Norton, M. I. (2014). The emergence of us and them in 80 lines of code: Modeling group genesis in homogeneous populations. *Psychological Science*, *25*(4), 982–990.

Guazzini, A., Cini, A., Bagnoli, F., & Ramasco, J. J. (2015). Opinion dynamics within a virtual small group: the stubbornness effect. *Frontiers in Physics*, *3*(September), 1–9. doi: M10.3389/fphy.2015.00065

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016a). Cultural Shift or Linguis-

tic Drift? Comparing Two Computational Measures of Semantic Change. , 2116–2121. doi: M10.18653/v1/d16-1229

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016b). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 1489–1501. doi: M10.18653/v1/P16-1141

Hart, E. M., Barmby, P., LeBauer, D., Michonneau, F., Mount, S., Mulrooney, P., ... Hollister, J. W. (2016). Ten Simple Rules for Digital Data Storage. *PLoS Computational Biology*, *12*(10), 1–12. doi: M10.1371/journal.pcbi.1005097

Hart, P. S., & Nisbet, E. C. (2012). Boomerang Effects in Science Communication: How Motivated Reasoning and Identity Cues Amplify Opinion Polarization About Climate Mitigation Policies. *Communication Research*, *39*(6), 701–723. doi: M10.1177/0093650211416646

Hawkins, R. D., Goodman, N. D., Goldberg, A. E., & Griffiths, T. L. (2020). Generalizing meanings from partners to populations : Hierarchical inference supports convention formation on networks. In *Proceedings of the 42nd annual conference of the cognitive science society.*

Hegselmann, R., & Krause, U. (2002). Opinion dynamics and bounded confidence: Models, analysis and simulation. *JASSS*, *5*(3). doi: Mciteulike-article-id: 613092

Heller, N. (2016, sep). The First Debate of the Twitter Election. *The New Yorker*. Retrieved from `Mhttps://www.newyorker.com/culture/cultural-comment/the-first-debate-of-the-twitter-election`

Henrich, J. (2015). *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.

Henrich, J., & Boyd, R. (1998). The evolution of conformist transmission and the emergence of between-group differences. *Evolution and Human Behavior*, *19*(4), 215–241.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-

3), 61–83. Retrieved from Mhttp://www.journals.cambridge.org/abstract{_}S0140525X0999152X doi: M10.1017/S0140525X0999152X

Heyes, C. (2018a). *Cognitive Gadgets: The cultural evolution of thinking*. Cambridge, MA: Belknap.

Heyes, C. (2018b). Enquire within: Cultural evolution and cognitive science. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1743), 1–9. doi: M10.1098/rstb.2017.0051

Hogg, M. A., Turner, J. C., & Davidson, B. (1990). Polarized Norms and Social Frames of Reference : A Test of the Self-Categorization Theory of Group Polarization. *Basic and Applied Social Psychology*, *11*(1), 77–100. doi: M10.1207/s15324834basp1101

Isenberg, D. J. (1986). Group polarization: A critical review and meta-analysis. *Journal of Personality and Social Psychology*, *50*(6), 1141–1151.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The Origins and Consequences of Affective Polarization in the United States. *Annual Review of Political Science*, *22*(1), 129–146. doi: M10.1146/annurev-polisci-051117-073034

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *The Behavioral and brain sciences*, *34*(4), 169–188. doi: M10.1017/S0140525X10003134

Jung, J., Grim, P., Singer, D. J., Bramson, A., Berger, W. J., Holman, B., & Kovaka, K. (2019). A multidisciplinary understanding of polarization. *American Psychologist*, *74*(3), 301–314. doi: M10.1037/amp0000450

Kalmoe, N. P. (2014). Fueling the fire: Violent metaphors, trait aggression, and support for political violence. *Political Communication*, *31*(4), 545–563. Retrieved from Mhttp://www.tandfonline.com/doi/abs/10.1080/10584609.2013.852642{#}.VdofNZdr9jc doi: M10.1080/10584609.2013.852642

Kalmoe, N. P., Gubler, J. R., & Wood, D. A. (2018). Toward Conflict or Compromise? How Violent Metaphors Polarize Partisan Issue Attitudes. *Political Communication*, *35*(3), 333–352. Retrieved from Mhttps://doi.org/

10.1080/10584609.2017.1341965 doi: M10.1080/10584609.2017.1341965

Kaplan, M. F. (1977). Discussion Polarization Effects in a Modified Jury Decision Paradigm: Informational Influences. *Sociometry*, *40*(3), 262. doi: M10.2307/3033533

Kaplan, M. F., & Miller, C. E. (1977). Judgments and Group Discussion: Effect of Presentation and Memory Factors on Polarization. *Sociometry*, *40*(4), 337. doi: M10.2307/3033482

Katz, E., & Lazarsfeld, P. F. (1955). *Personal Influence*. The Free Press.

Kauffman, S. A. (1970). Articulation of Parts Explanation in Biology and the Rational Search for Them. In *Psa: Proceedings of the biennial meeting of the philosophy of science association* (pp. 257–272).

Keating, J., Van Boven, L., & Judd, C. M. (2016). Partisan underestimation of the polarizing influence of group discussion. *Journal of Experimental Social Psychology*, *65*, 52–58. doi: M10.1016/j.jesp.2016.03.002

Kello, C. T., Beltz, B. C., Holden, J. G., & Van Orden, G. C. (2007). The emergent coordination of cognitive function. *Journal of experimental psychology. General*, *136*(4), 551–68. Retrieved from Mhttp://www.ncbi.nlm.nih.gov/pubmed/17999570 doi: M10.1037/0096-3445.136.4.551

Kelso, J. A. S. (1995). *Dynamic Patterns*. Cambridge, MA: MIT Press.

Kinder, D. R., & Kalmoe, N. P. (2017). *Neither Liberal nor Conservative*. Chicago: University of Chicago Press.

King, G., Schneer, B., & White, A. (2017). How the news media activate public expression and influence national agendas. *Science*, *358*(November), 776–780.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: Reputation and trust in a two-person economic exchange. *Science*, *308*(5718), 78–83. doi: M10.1126/science.1108062

Kirby, J. T. (1997). Aristotle on Metaphor. *The American Journal of Philology*, *118*(4), 517–554.

Klein, E. (2020). *Why We're Polarized*. New York: Simon and Schuster.

Klingenstein, S., Hitchcock, T., & DeDeo, S. (2014). The civilizing process in Lon-

don's Old Bailey. *Proceedings of the National Academy of Sciences*, *111*(26), 9419–9424. Retrieved from Mhttp://www.pnas.org/lookup/doi/10.1073/pnas.1405984111 doi: M10.1073/pnas.1405984111

Kövecses, Z. (2010a). *Metaphor: A Practical Introduction*. Oxford: Oxford University Press. doi: M10.1023/A:1023919116538

Kövecses, Z. (2010b). Metaphor, language and culture. *Delta*, *26*, 739–757. doi: M10.1075/babel.33.2.07fun

Kreindler, G. E., & Young, H. P. (2014). Rapid innovation diffusion in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(SUPPL.3), 10881–10888. doi: M10.1073/pnas.1400842111

Krizan, Z., & Baron, R. S. (2007). Group polarization and choice-dilemmas: how important is self-categorization? *European Journal of Social Psychology*, *37*, 191–201.

Kruschke, J. K. (2015). *Doing Bayesian Data Analysis: A tutorial with R, JAGS, and Stan*. Boston: Academic Press.

Kruschke, J. K., & Liddell, T. M. (2018a). Bayesian data analysis for newcomers. *Psychonomic Bulletin and Review*, *25*(1), 155–177. doi: M10.3758/s13423-017-1272-1

Kruschke, J. K., & Liddell, T. M. (2018b). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin and Review*, *25*(1), 178–206. doi: M10.3758/s13423-016-1221-4

Kurahashi-Nakamura, T., Mäs, M., & Lorenz, J. (2016). Robust clustering in generalized bounded confidence models. *Jasss*, *19*(4). doi: M10.18564/jasss.3220

Lakoff, G. (1991). Metaphor and war: the metaphor system used to justify war in the Gulf. *Peace Studies*, *23*(2), 25–32. Retrieved from Mhttp://www.jstor.org/stable/23609916

Lakoff, G. (1996). *Moral Politics*. Chicago: University of Chigago Press.

Lakoff, G. (2008). *The political mind: why you can't understand 21st-century politics with an 18th-century brain*.

Lakoff, G. (2014, dec). Mapping the brain's metaphor circuitry: metaphorical thought in everyday reason. *Frontiers in human neuroscience*, *8*(December), 958. Retrieved from Mhttp://journal.frontiersin.org/article/10.3389/fnhum.2014.00958/abstracthttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4267278{&}tool=pmcentrez{&}rendertype=abstract doi: M10.3389/fnhum.2014.00958

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. Chicago: University of Chicago Press.

Lakoff, G., & Núñez, R. E. (1997). The metaphorical structure of mathematics: Sketching out cognitive foundations for a mind-based. In *Mathematical reasoning: Analogies, metaphors, and images. studies in mathematical thinking and learning.* (pp. 21–89).

Lakoff, G., & Wehling, E. (2012). *The Little Blue Book: The Essential Guide to Thinking and Talking Democratic*. New York: Free Press.

Laland, K. N. (2017). *Darwin's unfinished symphony: How culture made the human mind*. Princeton University Press.

Landau, M. J., Meier, B. P., & Keefer, L. A. (2010). A Metaphor-Enriched Social Cognition. *Psychological Bulletin*, *136*(6), 1045–1067. doi: M10.1037/a0020970

Laughlin, P. R., & Earley, P. C. (1982). Social combination models, persuasive arguments theory, social comparison theory, and choice shift. *Journal of Personality and Social Psychology*, *42*(2), 273–280. doi: M10.1037/0022-3514.42.2.273

Lazer, D. M., Pentland, A. S., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., ... Barabasi, A. L. (2009). Computational social science. *Science*, *323*(5915), 721–723. doi: M10.1145/2556420.2556849

Lee, F. E. (2015). How Party Polarization Affects Governance. *Annual Review of Political Science*, *18*(1), 261–282. doi: M10.1146/annurev-polisci-072012-113747

Lelkes, Y. (2016). The polls-review: Mass polarization: Manifestations and mea-

surements. *Public Opinion Quarterly*, *80*, 392–410. doi: M10.1093/poq/nfw005

Lewandowsky, S., Pilditch, T. D., Madsen, J. K., Oreskes, N., & Risbey, J. S. (2019). Influence and seepage: An evidence-resistant minority can affect public opinion and scientific belief formation. *Cognition*, *188*(June 2018), 124–139. doi: M10.1016/j.cognition.2019.01.011

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348. doi: M10.1016/j.jesp.2018.08.009

Lord, C. G., Ross, L., & 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), 2098–2109. doi: M10.1037/0022-3514.37.11.2098

Lorenz, J. (2009). Heterogeneous Bounds of Confidence: Meet, Discuss and Find Consensus! *Complexity*, *15*(4). doi: M10.1002/cplx

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about Mechanisms. *Philosophy of Science*, *67*(1), 1–25.

Macy, M. W., & Willer, R. (2002). From Factors to Factors: Computational Sociology and Agent-Based Modeling. *Annual Review of Sociology*, *28*(1), 143–166. doi: M10.1146/annurev.soc.28.110601.141117

Marghetis, T., & Núñez, R. (2013). The motion behind the symbols: A vital role for dynamism in the conceptualization of limits and continuity in expert mathematics. *Topics in Cognitive Science*, *5*(2), 299–316. doi: M10.1111/tops.12013

Mark, N. (1998). Beyond individual differences: Social differentiation from first principles. *American Sociological Review*, 309–330.

Mark, N. (2003). Culture and competition: Homophily and distancing explanations for cultural niches. *American Sociological Review*, 319–345.

Mark, N. P. (2003). Culture and Competition: Homophily and Distancing Explanations for Cultural Niches. *American Sociological Review*, *68*(3), 319–345.

Martin, A. E., & Doumas, L. A. (2017). A mechanism for the cortical computation

of hierarchical linguistic structure. *PLoS Biology*, *15*(3), 1–23. doi: M10 .1371/journal.pbio.2000663

Martins, A. C., & Galam, S. (2013). Building up of individual inflexibility in opinion dynamics. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *87*(4), 1–8. doi: M10.1103/PhysRevE.87.042807

Mäs, M., & Flache, A. (2013). Differentiation without distancing. explaining bi-polarization of opinions without negative influence. *PLoS ONE*, *8*(11).

Mason, L. (2015). "I disrespectfully agree": The differential effects of partisan sorting on social and issue polarization. *American Journal of Political Science*, *59*(1), 128–145. doi: M10.1111/ajps.12089

Mason, L. (2018). *Uncivil Agreement*. Chicago: University of Chicago Press.

Matlock, T. (2012). Framing Political Messages with Grammar and Metaphor. *American Scientist*, *100*, 478–483.

Matlock, T., Castro, S., Fleming, M., Gann, T. M., & Maglio, P. P. (2014). Spatial Metaphors of Web Use. *Spatial Cognition and Computation*, *14*(4), 306–320. Retrieved from `Mhttp://www.tandfonline.com/action/journalInformation?journalCode=hscc20` doi: M10.1080/13875868.2014 .945587

Matlock, T., Ramscar, M., & Boroditsky, L. (2005). On the experiential link between spatial and temporal language. *Cognitive Science*, *29*(4), 655–664. doi: M10.1207/s15516709cog0000_17

McElreath, R., Boyd, R., & Richerson, P. J. (2003). Shared norms and the evolution of ethnic markers. *Current Anthropology*, *44*(1), 122–130.

McGarty, C., Turner, J. C., Hogg, M. A., David, B., & Wetherell, M. S. (1992). Group polarization as conformity to the prototypical group member. *British Journal of Social Psychology*, *31*(1), 1–19. doi: M10.1111/j.2044-8309.1992 .tb00952.x

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, *27*(1), 415–444. doi: M10.1146/annurev.soc.27.1.415

Meehl, P. E. (1990). Why summaries of research on psychological theories are

often uninterpretable. *Psychological Reports*, *66*(1), 195–244. doi: M10.2466/pr0.1990.66.1.195

Milgram, S. (1967). The Small-World Problem. *Psychology Today*, *1*(1), 61–67.

Mio, J. S. (1997). Metaphor and Politics. *Metaphor and Symbol*, *12*(2), 113–133.

Mobilia, M., Petersen, A., & Redner, S. (2007). On the role of zealotry in the voter model. *Journal of Statistical Mechanics: Theory and Experiment*, *2007*(08), P08029–P08029. doi: M10.1088/1742-5468/2007/08/p08029

Molenberghs, P., & Morrison, S. (2014). The role of the medial prefrontal cortex in social categorization. *Social Cognitive and Affective Neuroscience*, *9*(3), 292–296. doi: M10.1093/scan/nss135

Montoya, R. M., Horton, R. S., Vevea, J. L., Citkowicz, M., & Lauber, E. A. (2017). A re-examination of the mere exposure effect: The influence of repeated exposure on recognition, familiarity, and liking. *Psychological Bulletin*, *143*(5), 459–498. doi: M10.1037/bul0000085

Morales, A. J., Borondo, J., Losada, J. C., & Benito, R. M. (2015). Measuring political polarization: Twitter shows the two sides of Venezuela. *Chaos*, *25*(3). doi: M10.1063/1.4913758

Moscovici, S., & Zavalloni, M. (1969). The group as a polarizer of attitudes. *Journal of Personality and Social Psychology*, *12*(2), 125–135.

Moussaïd, M., Kämmer, J. E., Analytis, P. P., & Neth, H. (2013). Social influence and the collective dynamics of opinion formation. *PLoS ONE*, *8*(11). doi: M10.1371/journal.pone.0078433

Mueller, S. T., & Tan, Y.-Y. S. (2018). Cognitive perspectives on opinion dynamics: the role of knowledge in consensus formation, opinion divergence, and group polarization. *Journal of Computational Social Science*, *1*(1), 15–48. doi: M10.1007/s42001-017-0004-7

Muthukrishna, M., Morgan, T. J., & Henrich, J. (2016). The when and who of social learning and conformist transmission. *Evolution and Human Behavior*, *37*(1), 10–20.

Myers, D. G. (1975). Discussion-Induced Attitude Polarization. *Human Relations*, *28*(8), 699–714. doi: M10.1177/001872677502800802

Myers, D. G. (1978). Polarizing Effects of Social Comparison. *Journal of Experimental Social Psychology*, *14*, 554–563.

Myers, D. G. (1982). Polarizing Effects of Social Interaction. In H. Brandstätter, J. H. Davis, & G. Stocker-Kreichgauer (Eds.), *Group decision making.* London: Academic Press.

Myers, D. G., & Arenson, S. J. (1972). Enhancement of Dominant Risk Tendencies in Group Discussion. *Psychological Reports*, *30*(2), 615–623. doi: M10.2466/pr0.1972.30.2.615

Myers, D. G., & Bishop, G. D. (1970). Discussion Effects on Racial Attitudes. *Science*, *169*, 778–779.

Myers, D. G., & Lamm, H. (1975). The polarizing effect of group discussion. *American Scientist*, *63*(3), 297–303.

Nowak, A., Szamrej, J., & Latané, B. (1990). From private attitude to public opinion: A dynamic theory of social impact. *Psychological Review*, *97*(3), 362–376. Retrieved from Mhttp://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.97.3.362 doi: M10.1037/0033-295X.97.3.362

Núñez, R., Cooperrider, K., Doan, D., & Wassmann, J. (2012). Contours of time: Topographic construals of past, present, and future in the Yupno valley of Papua New Guinea. *Cognition*, *124*(1), 25–35. doi: M10.1016/j.cognition.2012.03.007

Nunn, N. (2012). Culture and the Historical Process. *Economic History of Developing Regions*, *27*(S1), S108–S126. Retrieved from Mhttp://www.tandfonline.com/doi/abs/10.1080/20780389.2012.664864 doi: M10.1080/20780389.2012.664864

O'Brien, G. (2003). Indigestible Food, Conquering Hordes, and Waste Materials: Metaphors of Immigrants and the Early Immigration Restriction Debate in the United States. *Metaphor and Symbol*, *18*(1), 33–47. Retrieved from Mhttp://www.tandfonline.com/action/journalInformation?journalCode=hmet20http://dx.doi.org/10.1207/S15327868MS1801{_}3 doi: M10.1207/s15327868ms1801_3

O'Connell, M. (2017, dec). *Fox News Holds No. 1, MSNBC Thrives*

During Wild Year for Cable News. Retrieved from Mhttps://www.hollywoodreporter.com/news/fox-news-holds-no-1-msnbc-thrives-wild-year-cable-news-1069621

O'Connor, C., & Weatherall, J. O. (2018). Scientific polarization. *European Journal for Philosophy of Science*, *8*(3), 855–875. doi: M10.1007/s13194-018-0213-9

O'Connor, C., & Weatherall, J. O. (2019). *The Misinformation Age*. New Haven: Yale University Press.

Palla, G., Barabási, A. L., & Vicsek, T. (2007). Quantifying social group evolution. *Nature*, *446*(7136), 664–667. doi: M10.1038/nature05670

Pallavicini, J., Hallsson, B., & Kappel, K. (2021). Polarization in groups of Bayesian agents. *Synthese*, *198*(1), 1–55. Retrieved from Mhttps://doi.org/10.1007/s11229-018-01978-w doi: M10.1007/s11229-018-01978-w

Perlberg, S. (2016, sep). *Presidential Debate Sets Viewership Record.* Retrieved from Mhttps://www.wsj.com/articles/debate-ratings-might-break-record-1474996186

Pew Research Center. (2012, oct). *Romney' s Strong Debate Performance Erases Obama' s Lead* (Tech. Rep.). Washington, D.C.: Pew Research Center. Retrieved from Mhttp://www.people-press.org/2012/10/08/romneys-strong-debate-performance-erases-obamas-lead

Pew Research Center. (2014a). Political Polarization and Media Habits. (October). Retrieved from Mhttp://www.journalism.org/2014/10/21/political-polarization-media-habits/{#}media-outlets-by-the-ideological-composition-of-their-audience doi: M202.419.4372

Pew Research Center. (2014b). *Political Polarization in the American Republic* (Vol. June; Tech. Rep.). Pew Research Center.

Pew Research Center. (2017a, oct). The Partisan Divide on Political Values Grows Even Wider. , 1–23. Retrieved from Mhttp://www.people-press.org/2017/10/05/the-partisan-divide-on-political-values-grows-even-wider/{#}overview{%}0Ahttp://www.people-press.org/2017/10/05/the-partisan-divide-on-political-values-grows-even-wider/

Pew Research Center. (2017b). *Trump, Clinton Voters Divided in Their Main Source for Election News | Pew Research Center* (Tech. Rep.). Retrieved from Mhttp://www.journalism.org/2017/01/18/trump-clinton-voters-divided-in-their-main-source-for-election-news/

Pew Research Center. (2018). *Anger beat love in Facebook reactions to lawmaker posts after 2016 election.* Retrieved 2018-07-21, from Mhttp://www.pewresearch.org/fact-tank/2018/07/18/anger-topped-love-facebook-after-2016-election/

Pineda, M., Toral, R., & Hernndez-García, E. (2009). Noisy continuous-opinion dynamics. *Journal of Statistical Mechanics: Theory and Experiment*, *2009*(8). doi: M10.1088/1742-5468/2009/08/P08001

Pinker, S., Nowak, M. A., & Lee, J. J. (2008). The logic of indirect speech. *Proceedings of the National Academy of Sciences*, *105*(3), 833–838.

Prior, M. (2013). Media and Political Polarization. *Annu. Rev. Polit. Sci*, *16*, 101–27. Retrieved from Mhttp://polisci.annualreviews.org doi: M10.1146/annurev-polisci-100711-135242

R Core Team. (2021). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from Mhttps://www.r-project.org/

Ramirez-Aristizabal, A. G., Médé, B., & Kello, C. T. (2018). Complexity matching in speech: Effects of speaking rate and naturalness. *Chaos, Solitons and Fractals*, *111*, 175–179. Retrieved from Mhttps://doi.org/10.1016/j.chaos.2018.04.021 doi: M10.1016/j.chaos.2018.04.021

Regier, T. (1996). *The Human Semantic Potential*. Cambridge, MA: MIT Press.

Reiss, S., Klackl, J., Proulx, T., & Jonas, E. (2019). Strength of socio-political attitudes moderates electrophysiological responses to perceptual anomalies. *PLoS ONE*, *14*(8), 1–17. doi: M10.1371/journal.pone.0220732

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations on the effectiveness of reinforcement and non-reinforcement. In *Classical conditioning II: Current research and theory* (p. 64-99). New York: Appleton-Century-Crofts.

Rollwage, M., Zmigrod, L., De-Wit, L., Dolan, R. J., & Fleming, S. M. (2019). What Underlies Political Polarization? A Manifesto for Computational Political Psychology. *Trends in Cognitive Sciences*, 1–3. doi: M10.1016/j.tics.2019.07.006

Romenskyy, M., Spaiser, V., Ihle, T., & Lobaskin, V. (2017). Polarized Ukraine 2014: Opinion and Territorial Split Demonstrated with the Bounded Confidence XY Model, Parameterized by Twitter Data. *arXiv*. Retrieved from `Mhttp://arxiv.org/abs/1706.00419`

Ross, L. (2012). Reflections on biased assimilation and belief polarization. *Critical Review*, *24*(2), 233–245. doi: M10.1080/08913811.2012.711025

Sagi, E., Diermeier, D., & Kaufmann, S. (2013). Identifying Issue Frames in Text. *PLoS ONE*, *8*(7), 1–9. doi: M10.1371/journal.pone.0069185

Salathé, M., Kazandjieva, M., Lee, J. W., Levis, P., Feldman, M. W., & Jones, J. H. (2010). A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(51), 22020–22025. doi: M10.1073/pnas.1009094108

Samuel, S., & König-Ries, B. (2021). Understanding experiments and research practices for reproducibility: An exploratory study. *PeerJ*, *9*, 1–26. doi: M10.7717/peerj.11140

Sanders, G. S., & Baron, R. S. (1977). Is social comparison irrelevant for producing choice shifts? *Journal of Experimental Social Psychology*, *13*(4), 303–314. doi: M10.1016/0022-1031(77)90001-4

Schelling, T. C. (1971). Dynamic Models of Segregation. *Journal of Mathematical Sociology*, *1*, 143–186. doi: M10.1080/0022250X.1971.9989794

Schelling, T. C. (2006). *Micromotives and macrobehavior* (2nd ed.). New York: W. W. Norton.

Schkade, D., Sunstein, C. R., & Hastie, R. (2007). What happened on deliberation day? *California Law Review*, *95*(3), 915–940. doi: M10.2139/ssrn.911646

Schkade, D., Sunstein, C. R., & Hastie, R. (2010). When deliberation produces extremism. *Critical Review*, *22*(2), 227–252. doi: M10.1080/08913811.2010.508634

Schkade, D., Sunstein, C. R., & Kahneman, D. (2000). Deliberating about Dollars: The Severity Shift. *Columbia Law Review*, *100*(4), 1139–1175.

Schloesser, D. S., Kello, C. T., & Marmelat, V. (2019). Complexity matching and coordination in individual and dyadic performance. *Human Movement Science*, *66*(May), 258–272. Retrieved from `Mhttps://doi.org/10.1016/j.humov.2019.04.010` doi: M10.1016/j.humov.2019.04.010

Schneider, S., Ramirez-Aristizabal, A. G., Gavilan, C., & Kello, C. T. (2020). Complexity matching and lexical matching in monolingual and bilingual conversations. *Bilingualism*, *23*(4), 845–857. doi: M10.1017/S1366728919000774

Schrödinger, E. (2012). *What is Life?* Cambridge, UK: Cambridge University Press.

Schroeder, A. (2008). *Presidential debates: fifty years of high-risk TV* (2nd ed.). New York: Columbia University Press.

Sherif, M. (1988). *The robbers cave experiment: Intergroup conflict and cooperation*. Middletown, CT: Wesleyan University Press.

Sides, J., & Hopkins, D. J. (Eds.). (2015). *Political Polarization in American Politics*. New York: Bloomsbury.

Sieber, J., & Ziegler, R. (2019). Group Polarization Revisited: A Processing Effort Account. *Personality and Social Psychology Bulletin*, *45*(10), 1482–1498. doi: M10.1177/0146167219833389

Smaldino, P. E. (2017a). Models are stupid, and we need more of them. *Computational Social Psychology*, 311–331. doi: M10.4324/9781315173726

Smaldino, P. E. (2017b). Models Are Stupid, and We Need More of Them. In R. R. Vallacher, A. Nowak, & S. J. Read (Eds.), *Computational models in social psychology*. Psychology Press.

Smaldino, P. E. (2018). Social identity and cooperation in cultural evolution. *Behavioural Processes*.

Smaldino, P. E. (2019). Better methods can't make up for mediocre theory. *Nature*, *575*(7781), 9. doi: M10.1038/d41586-019-03350-5

Smaldino, P. E., D'Souza, R. M., & Maoz, Z. (2018). Resilience by structural entrenchment: Dynamics of single-layer and multiplex networks following

sudden changes to tie costs. *Network Science*.

Smaldino, P. E., & Epstein, J. M. (2015a). Social conformity despite individual preferences for distinctiveness. *Royal Society Open Science*, *2*(3). doi: M10 .1098/rsos.140437

Smaldino, P. E., & Epstein, J. M. (2015b). Social conformity despite individual preferences for distinctiveness. *Royal Society Open Science*, *2*(3), 140437.

Smaldino, P. E., Flamson, T. J., & McElreath, R. (2018). The evolution of covert signaling. *Scientific Reports*, *8*(1), 4905.

Smaldino, P. E., Janssen, M. A., Hillis, V., Bednar, J., Smaldino, P. E., Janssen, M. A., ... Bednar, J. (2017). Adoption as a social marker : Innovation diffusion with outgroup aversion Adoption as a social marker : Innovation diffusion with outgroup aversion. *The Journal of Mathematical Sociology*, *41*(1), 26–45. Retrieved from `Mhttp://dx.doi.org/10.1080/0022250X.2016.1250083` doi: M10.1080/0022250X.2016.1250083

Smaldino, P. E., Lukaszewski, A., von Rueden, C., & Gurven, M. (2019). Niche diversity can explain cross-cultural differences in personality structure. *Nature Human Behaviour*, *3*(12), 1276–1283. Retrieved from `Mhttp://dx.doi.org/10.1038/s41562-019-0730-3` doi: M10.1038/s41562-019-0730-3

Smaldino, P. E., & Schank, J. C. (2012). Human mate choice is a complex system. *Complexity*, *17*(5), 11–22.

Smaldino, P. E., Turner, M. A., & Contreras Kallens, P. A. (2019). Open science and modified funding lotteries can impede the natural selection of bad science. *Royal Society Open Science*, *6*(8), 191249. doi: M10.1098/rsos.191249

Spivey, M. J. (2020). *Who You Are: The Science of Connectedness*. MIT Press. Retrieved from `Mhttps://direct.mit.edu/books/book/4659/Who-You-AreThe-Science-of-Connectedness` doi: Mhttps://doi.org/10.7551/mitpress/12818.001.0001

Stoner, J. A. (1961). *A comparison of individual and group decisions involving risk* (Master's). Massachusetts Institute of Technology.

Stoner, J. A. (1968). Risky and cautious shifts in group decisions: The influence of widely held values. *Journal of Experimental Social Psychology*, *4*(4), 442–

459. doi: M10.1016/0022-1031(68)90069-3

Suhay, E., & Erisen, C. (2018). The Role of Anger in the Biased Assimilation of Political Information. *Political Psychology*, *39*(4), 793–810. Retrieved from Mhttp://doi.wiley.com/10.1111/pops.12463 doi: M10.1111/pops.12463

Sunstein, C. R. (2000). Deliberative Trouble? Why Groups Go to Extremes. *Yale Law Journal*, *110*(1), 71–119. doi: M10.4324/9781315248592-4

Sunstein, C. R. (2002). The Law of Group Polarization. *Journal of Political Philosophy*, *10*(2), 175–195.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Tajfel, H. (1982). Social Psychology of Integroup Relations. *Annual Reviews of Psychology*, *33*, 1–39.

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971a). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, *1*(2), 149–178. doi: M10.1002/ejsp.2420010202

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971b). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, *1*(2), 149–178.

Tajfel, H., & Turner, J. (1979). An Integrative Theory of Intergroup Conflict. In W. G. Austin & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33–47). Monterey, CA: Brooks/Cole.

Takács, K., Flache, A., & Mäs, M. (2016). Discrepancy and disliking do not induce negative opinion shifts. *PLoS ONE*, *11*(6), 1–21. doi: M10.1371/journal.pone.0157948

Teger, A. I., & Pruitt, D. G. (1967). Components of group risk taking. *Journal of Experimental Social Psychology*, *3*(2), 189–205. doi: M10.1016/0022-1031(67)90022-4

Thibodeau, P. H., & Boroditsky, L. (2011). Metaphors We Think With: The Role of Metaphor in Reasoning. *PLoS ONE*, *6*(2). doi: M10.1371/journal.pone.0016782

Travers, J., & Milgram, S. (1969). An experimental study of the small world problem. *Sociometry*, *32*(4), 425–443.

Turner, J. C. (1987). *Rediscovering the Social Group: Self-categorization thoery.* Oxford, UK: Oxford University Press.

Turner, J. C., & Wetherell, M. S. (1987). Social Identity and Group Polarization. In *Rediscovering the social group: self-categorization theory* (pp. 142–170). Oxford, UK: Blackwell.

Turner, J. C., Wetherell, M. S., & Hogg, M. A. (1989). Referent informational influence and group polarization. *British Journal of Social Psychology*, *28*(2), 135–147. doi: M10.1111/j.2044-8309.1989.tb00855.x

Turner, M. A., & Smaldino, P. E. (2018). Paths to Polarization: How Extreme Views, Miscommunication, and Random Chance Drive Opinion Dynamics. *Complexity*.

Turner, M. A., & Smaldino, P. E. (2020). Stubborn extremism as a potential pathway to group polarization. In *Proceedings of the 42nd annual conference of the cognitive science society.* Toronto.

Turner, M. A., & Smaldino, P. E. (2021). Mechanistic Modeling for the Masses - commentary on Yarkoni, "The generalizability crisis". *Behavioral and Brain Sciences*. Retrieved from `Mhttps://psyarxiv.com/8pj9n`

Vinokur, A., & Burnstein, E. (1978). Novel argumentation and attitude change: The case of polarization following group discussion. *European Journal of Social Psychology*, *8*(3), 335–348. doi: M10.1002/ejsp.2420080306

Vinokur, A., & Burstein, E. (1974). Effects of partially shared persuasive arguments on group-induced shifts: A group-problem-solving approach. *Journal of Personality and Social Psychology*, *29*(3), 305–315. doi: M10.1037/h0036010

Wallach, M. A., & Kogan, N. (1965). The roles of information, discussion, and consensus in group risk taking. *Journal of Experimental Social Psychology*, *1*(1), 1–19. doi: M10.1016/0022-1031(65)90034-X

Watts, D. J. (1999). *Networks, Dynamics, and the Small World Phenomenon* (Vol. 105) (No. 2). Retrieved from `Mhttp://www.jstor.org/stable/10`

.1086/210318 doi: M10.1086/210318

Watts, D. J. (2011). *Everything is Obvious: once you know the answer*. New York: Crown Business.

Wimsatt, W. C. (1972). Complexity and Organization. In K. F. Schaffner & R. S. Cohen (Eds.), *Psa: Proceedings of the biennial meeting of the philosophy of science association* (Vol. 1972, pp. 67–86). Dordrecht, Holland: D. Reidel.

Wimsatt, W. C. (1997). Aggregativity: Reductive Heuristics for Finding Emergence. *Philosophy of Science*, *64*.

Wimsatt, W. C. (2007). *Re-Engineering Philosophy for Limited Beings*. Cambridge, MA: Harvard University Press.

Wohlgemuth, J., & Matache, M. T. (2014). Small-world properties of facebook group networks. *Complex Systems*, *23*(3), 197–225.

Wood, B. D., & Jordan, S. (2017). Polarization as the Norm of the American System. In *Party polarization in america* (pp. 300–314).

Yarkoni, T. (2021). The generalizability crisis. *Behavioral and Brain Sciences (forthcoming)*. Retrieved from Mhttps://doi.org/10.1017/S0140525X20001685

Zajonc, R. B. (1968). Attitudinal Effects of Mere Exposure. *Journal of Personality and Social Psychology*, *9*(2 PART 2), 1–27. doi: M10.1037/h0025848

Zaller, J. R. (1992). *The Nature and Origins of Mass Opinion*. New York: Cambridge University Press.

Zmigrod, L., Rentfrow, P. J., & Robbins, T. W. (2018). Cognitive underpinnings of nationalistic ideology in the context of Brexit. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(19), E4532–E4540. doi: M10.1073/pnas.1708960115

Zmigrod, L., Rentfrow, P. J., & Robbins, T. W. (2019). Cognitive inflexibility predicts extremist attitudes. *Frontiers in Psychology*, *10*(MAY), 1–13. doi: M10.3389/fpsyg.2019.00989

Zmigrod, L., Zmigrod, S., Rentfrow, P. J., & Robbins, T. W. (2019). The psychological roots of intellectual humility: The role of intelligence and cognitive flexibility. *Personality and Individual Differences*, *141*(January), 200–208.

doi: M10.1016/j.paid.2019.01.016

Zollman, K. J. (2013). Network epistemology: Communication in epistemic communities. *Philosophy Compass*, *8*(1), 15–27. doi: M10.1111/j.1747-9991.2012.00534.x

Zollman, K. J. S. (2007). The Communication Structure of Epistemic Communities. *Philosophy of Science*, *74*(5), 574–587. doi: M10.1086/525605

Zuber, J. A., Crott, H. W., & Werner, J. (1992). Choice Shift and Group Polarization: An Analysis of the Status of Arguments and Social Decision Schemes. *Journal of Personality and Social Psychology*, *62*(1), 50–61. doi: M10.1037/0022-3514.62.1.50