TEXT, CULTURE TIME: BRINGING COMPLEX SYSTEMS TO BEAR ON CULTURAL SOCIOLOGY THROUGH COMPUTATIONAL MIXED METHODS

Zackary Okun Dunivin

Submitted to the faculty of the Graduate School in partial
fulfillment of the requirements
for the degrees of
Doctor of Philosophy
In the department of Informatics
&
Doctor of Philosophy
In the department of Sociology
Indiana University
December 2024

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

Staša Milojević, PhD

Fabio Rojas, PhD

Fillipo Raddicchi, PhD

Koji Chavez, PhD

Zackary Okun Dunivin

TEXT, CULTURE TIME: BRINGING COMPLEX SYSTEMS TO BEAR ON CULTURAL SOCIOLOGY THROUGH COMPUTATIONAL MIXED METHODS

This doctoral dissertation is a treatise on the use of computational data and methods to study culture and cultural change using my own empirical and methodological work as a substrate. I begin by showing that many of the challenges of studying culture originate in the complex nature of culture itself. Drawing on theories from the literature on complex systems, I first show that the complex adaptive systems framework has considerable overlap with contemporary sociological theories of culture including cognitive schemas, symbolic boundaries, and cultural environments. I then explore several aspects of complex systems as they are manifest in culture, and discuss how computationally derived data, such as trace data and digital archives, and computational methods, such as natural language processing and simulation, can be leveraged to capture this complexity. Five subsequent chapters demonstrate the use of computation to understand culture with studies I have conducted independently or leading a team of collaborators. Each piece is accompanied by a discussion of the methodological and data choices made for these studies noting how computation allows for the pursuit of questions that are otherwise difficult to analyze. In particular, I show how combining computational, qualitative, and traditional quantitative methods leverages the strengths of each to understand culture and culture change by capturing, rather than abstracting away, complexity.

TABLE OF CONTENTS

Acceptance Page ii

Abstract iii

Chapter 1: Introduction 1

1.1 Overview 1

1.2 Integrating complex systems with cultural sociology 2

1.3 What do I mean by computational mixed methods? 9

1.4 Theoretical and methodological contributions to cultural sociology 11

Chapter 2: Social media data with qualitative analysis: White nationalists on Twitter 16

2.1 Introduction 16

2.2 The method described 22

2.3 The method applied: White nationalism on Twitter 30

2.4 Discussion 41

Chapter 3: Trace data for temporal and semantic resolution: BLM discursive shifts 47

3.1 Introduction 47

3.2 Results 53

3.3 Discussion 65

Chapter 4: Complexity theory without complex models: The 27 Club myth 73

4.1 Introduction . . . 73

4.2 Results . . . 76

4.3 Discussion . . . 84

Chapter 5: Simulating cultural evolution: Dynamics of covert signaling . . . 87

5.1 Introduction . . . 87

5.2 The analytical model . . . 94

5.3 The agent-based model . . . 96

5.4 Results . . . 104

5.5 Discussion . . . 112

Chapter 6: Complex textual interpretation at scale: Qualitative coding using LLMs . . . 118

6.1 Introduction . . . 118

6.3 Results . . . 124

6.4 Discussion . . . 130

Chapter 7: Conclusion . . . 134

7.1 Where computation fits in cultural research . . . 134

7.2 Where complexity fits in cultural research . . . 136

7.3 Some Limitations . . . 137

References . . . 140

Appendix A: Chapter 3 . . . 155

Vita

CHAPTER 1
INTRODUCTION

1.1 Overview

There is a tendency for sociologists to say "all is culture". Truly our social world is abound with culture. It structures our individual and collective realities as schemas for interpretation and action. But as a scientist (and a social being) seeing culture is not always so easy, particularly in ways that are tractable to quantitative methods. One of the great strengths of quantitative methods is that they can provide a lens on systemic patterns that are hard to view from our human perspectives. Here a dilemma emerges: culture structures our social systems, systems can be brought into view with quantification, but culture is often intractable to quantitative methods.

This dissertation presents a variety of perspectives on how to study culture. Culture is not a novel object of study. Anthropologists have been at it for 150 years. It has been a prominent strain of sociology for over 50 (Grindstaff et al., 2010). More recently psychologists (Shweder, 1999), economists (Guiso et al., 2006), and all manner of other social scientists (Ferrell, 1999; Schudson, 1989; Kreuter and McClure, 2004) have recognized how deliberate attention to culture can enhance their inquiry. Standing next to such a massive and enduring object, what can this one researcher's paltry manuscript offer?

My contribution to the study of culture is twofold. The first is a theoretical contribution connecting the problems of culture as framed by sociologists to the theory of complex systems. The second lies in drawing attention to the use of computational tools to circumvent some traditional problems presented by culture. These problems are not necessarily unique to culture and astute readers may draw parallels to other perspectives on the social world.

- Culture is high-dimensional and relational. Quantitative methods are most adept at analyzing systems with low dimensionality, independent (rather than interdependent) variables, and discrete categories.

- Culture is best understood in situ. Experiments and surveys often cannot capture the complexity of culture.
- Culture emerges from individual minds and actions. It is often only observable in the actions of individuals, but we can only understand it in the context of the broader social system. Thus meso- and macro-scale phenomena are essential to understanding culture.

Through the empirical and methodological work I have conducted over the past 7 years, I will show computational methods can gain traction on some of these problems. However, I am not a computational absolutist or imperialist. I have seen firsthand how different approaches have their own epistemic, pragmatic, and rhetorical strengths and weaknesses. Thus, I pay particular attention to how computational methods may be married with traditional qualitative and quantitative methods to produce deeper and more robust accounts of cultural phenomena.

1.2 Integrating complex systems with cultural sociology

Contemporary sociological approaches to the study of culture come in many flavors, but generally they tend to fall into three camps. Most broadly, cultural schemas (e.g., DiMaggio, 1997), frames (e.g., Wood et al., 2018), and toolkits (e.g., Swidler, 1986) relate to how individuals interpret themselves and their environments, and subsequently, what action to take in response to stimuli. Symbolic boundaries (e.g., Lamont and Molnár, 2002) are narrower and relate specifically to how and where individuals draw the line between the collectives they belong to and outsiders. Both these approaches tend to draw attention to individual processes, though the extent to which cognition is explicitly recognized varies greatly. More recently, Bail (2014) has proposed "cultural environments" as a framework for understanding the cultural milieus in which we find ourselves, which are larger, broader, and more pluralistic than can be held in any individual. Bail and, similarly, Murthy (2012a) privilege the analysis of large text data sets crowdsourced through digital platforms, arguing they as holding traces of this cultural environment such that it is visible and tractable to analysis, particularly quantitative analysis.

I do not attempt to introduce a new framework for studying culture here. These theoretical tools above are more than sufficient for the study of culture from a sociological perspective. In fact I draw explicitly on symbolic boundaries in Chapters 2, cognitive schemas in Chapter 4, and cultural frames in Chapters 3 and

6. Symbolic boundaries are implicit in Chapter 5, which was written for a cognitive science and psychology audience.

Given the abundance of good sociological theory of culture, my theoretical contribution is instead to introduce complex systems theory as framework for understanding and augmenting existing theory in the cultural sociology. Further, in understanding the language of complex systems, we are able to explicate why culture is resistant to study. Happily, shining a light on the problems of complexity in the study of culture also provides some solutions in the form of conceptual and computational tools from complex systems and its sibling disciplines that similarly descended from cybernetics, the study of control, learning, and adaptation (Heims, 1991; Tabilo Alvarez and Ramírez-Correa, 2023).

Though it has struggled to establish itself as a discipline in the traditional sense, complex systems has a long and moderately successful history in and out of academia. Rather than survey the field, I will instead introduce some of its concepts as solutions to understanding the challenges to studying culture I pose earlier in the introduction. First, however, I will introduce the concept of the complex adaptive system and explain how it overlaps with traditional cultural sociological approaches.

1.2.1 Cultural sociological theories are articulations of the complex adaptive systems framework

The Complex Adaptive Systems (CAS) paradigm has its roots in cybernetics, but was largely developed by researchers with the Santa Fe Institute in the 1980s, notably John Holland and Murray Gell-Mann (Holland, 1992; Gell-Mann, 1994). The framework has numerous articulations, but at its core is an entity with sensing, interpreting, and behaving capacities, which are determined by a schema (Gell-Mann, 1994; Miller and Page, 2009). Thought it may offend our humanist sensibilities, the CAS can be thought of as a variously (at the upper ends, incomprehensibly) complex computer system. It takes in input, which it then interprets through its program to arrive at some output. Beyond this, the complex adaptive system does just that: it adapts, or learns, when its schema proves ineffective at producing appropriate behaviors with desired outcomes. When repeatedly confronted by regular information that does not fit its schema, or leads to negative outcomes, the system reconfigures its schema through association of this new information with interpretations and subsequent action that lead to more desirable results.

The CAS framework has been adopted as a small, but healthy minority perspective across the theoretical and applied social sciences, particularly in political science, economics, and management/operations research. One of the strengths of the framework is that it works equally well for competitive and cooperative

social dynamics, and can be applied across scales of social organization from individuals to collectives to collectives of collectives (Miller and Page, 2009). Sociology has been reluctant to admit the perspective explicitly (the social systems theory of Luhmann is essentially a parallel development of complex adaptive systems, which grew out of a distinct branch of cybernetics, though is undeniably fringe in sociology (Varela et al., 1974; Luhmann, 1990; Mingers, 2002).)

At this point a sociologist may object to the complex adaptive systems framework because it appears thoroughly rationalist. Rationalist accounts of behavior have historically been problematic for sociologists, especially those who study culture, even beyond the problems presented by more limited framing presented by rational choice theory (Kiser and Hechter, 1998; Vaughan, 1998). By rationalism, I refer to a more expansive definition that is linked to functionalism, rather than rational choice theory, which entails individual human actors consciously deliberating action. Instead, by rationalism I simply mean that schemas that determine interpretation and action are consciously or unconsciously adopted in anticipation of greater reward in a multidimensional value-space. Thus it is easy to skirt hardline rationalist or functionalist commitments and still maintain consistency with the CAS framework. On the one hand, adoption of cultural schemas (or frames or toolkits) is obviously rational because it helps actors interpret the cultural environment. Culture surrounds us, structuring our external realities, and without appropriate interpretive schemas it is difficult to participate in society. Acculturation is a natural and critical part of child development and socialization more broadly (e.g., professionalization), to the point that culture is typically not explicitly invoked in the literature on socialization unless it is in a "cross-cultural" context. The rationality of cultural learning is, of course, easiest to observe in cross-cultural situations, where a cultural schema enters a cultural environment for which it is not adapted, such as immigration (e.g., Gibson, 2001; Salant and Lauderdale, 2003) or corporate mergers (e.g., Olie, 1994). However, rationality of cultural adoption may also be less obvious. For example, individuals may learn cultural schemas through incentives from other group members, e.g., increased cooperation with those who display the cultural schema, as in Chapter 5 of this dissertation, which models the evolution of identity signals. Cultural schemas may also be adopted through bounded rationality (Simon, 1990). In situations where it is hard to know which action to take, it is rational to adopt schemas displayed by those around you who appear adept and knowledgeable (DiMaggio and Powell, 1983).

My goal in introducing complex adaptive systems is not to advocate for a rationalist or functionalist approach to cultural sociology. Neither is it really to contend that the CAS framework is necessary or even particularly useful for practitioners who study culture from a scientific or quantitative perspective. Rather, it

is to show that that complex systems theorists have for a quite a long while been operating through paradigms that are consistent with the frameworks that have been developed in cultural sociology. Having identified general isomorphism between CAS and the different flavors of cultural sociological theory, I use the broader discourse of complex systems to bridge cultural sociology to a set of methodological approaches as well as, to a lesser degree, concepts. The following subsections of this chapter accordingly approach problems in cultural systems and their representations in data through the language of complexity. Moreover, and more to the point, I explain how methodologies originating with complex systems and allied fields such as neuroscience, systems biology, operations research, and computer science, can offer complementary and sometimes fuller views of culture and cultural change to social scientists across all fields.

1.2.2 Culture is high-dimensional and relational

Culture is difficult to study even when we can see it. This is because cultural elements and especially the cultural systems they comprise challenge human capacities to identify attributes and categories. Firstly, the number of attributes is large, what I call here "high-dimensional". Culture comprises practices, objects, language, values, norms, ideas, etc, and each of these can be described by dozens, if not hundreds or thousands, of characteristics. Traditional statistical models struggle with high-dimensionality. Those unfamiliar with machine learning (ML) tend to think of it as designed for the problems posed by big data, that is, large N. But mathematically, the approaches across machine learning are most distinct from other analytical strategies in that they facilitate analysis of data with large K, that is, the number of variables. From early machine learning extensions of linear regression, such as Support Vector Machines, to classification algorithms, such as random forest and K-means, ML methods differ from traditional statistical methods in that they effectively and efficiently analyze data sets with hundreds or thousands of variables. This is to say nothing of deep learning approaches, which currently can compute parameter spaces in the billions. Chapter 2 of this dissertation shows how computational methods can be exploited to locate meaningful linguistic signals, words and phrases, from a pool of hundreds of thousands of potential types of signals.

Secondly, these high-dimensional systems are relational. The meaning of each variable or its contribution influencing some particular structure or outcome in the world, changes in relation to the values of the other variables. This is implicitly what is meant by qualitative scholars when they say that a phenomenon is "contextual" and must be studied qualitatively. It is not that the phenomenon is inherently resistant to quantification, but rather that the quantitative methods most social scientists are familiar with, i.e., frequentist

statistics, are unable to compute or represent the complexity of relationships that underlie the data. Rosen (1987) makes an analogous critique of dynamical systems modeling from physics as an approach to understanding biological systems, which from genetics to ecology, are, like social systems, high dimensional and relational. Relationality is also the principle behind intersectionality, which states that social identities are reliably shaped by the other identities that are embodied within a person (Crenshaw, 1989). In the language of mathematics, interactions between identities are non-additive, non-linear, or non-parametric (in order of increasing complexity) and as such determining relationships empirically is intrinsically difficult. Rosen goes so far as to say they are indeterminable, presenting an inherently pessimistic view of complexity.

Rosen's pessimism is overstated, however. In the subsequent decades an abundance of new tools have been developed to compute complex systems, enabled by the exponential progress in computational power. Equally important are the novel and readily available data sources enabled by computer systems. Typically, social scientists associate these new data sources with social media (e.g., Chapters 2 and 3 of this dissertation), but computational data collection extends to other crowdsourced data, such as Wikipedia and Google search usage (e.g., Chapters 2,3, and 5), digitized archives (e.g., Chapters 3 and 6), and others not represented by this dissertation including communication and travel networks (e.g., Balcan et al., 2009), organizational records (e.g., Chakraborti et al., 2024), and electronic health records (e.g., Perry et al., 2019).

1.2.3 Culture is best understood in situ

Due to its high-dimensional and relational nature, most data collection methods struggle to capture culture. Therefore, ethnography has long been the preferred tool of cultural researchers (Geertz, 1973). Observation of natural human behavior, as well as reflections on behavior and mental states through interview, allow for the fullness of human action, interaction, and subjectivity to emerge to the researcher. Therefore is the entire field of cultural anthropology organized around mastering and performing ethnography. Cultural sociology is a comparatively small field, but ethnography is still its principal tool. Though there are many flavors of ethnographic theory in sociology, I will briefly treat three here: symbolic interactionism, grounded theory, and institutional ethnography. These frameworks advance the position that complex systems methodologies can expand our view of culture through principles shared with ethnographic method and theory.

Symbolic interactionism is the theoretical position that the social is best understood through observing humans communicating and interpreting through vocabularies of shared symbols (Blumer, 1986). As such symbolic interactionism is an explicitly relational approach and requires extensive observation of natural

behavior, as opposed to the controlled, artificial behavior of psychological experiments. Social media offer a wealth of interaction data and though the symbols are largely limited to text, they are especially tractable to analysis. Two chapters of this dissertation probe how interaction gives rise to meaningful symbols. Chapter 2 employs a massive observation of interactions and utterances on Twitter, whereas Chapter 5 simulates interaction to abstract away content and focus purely on symbol emergence.

Grounded theory is an ethnographic process whereby research questions and the conceptual frameworks (categories of behaviors, dispositions, interpretations, etc.) to address them emerge through an iterative process observation and conceptual development/refinement (Strauss and Corbin, 1997). This involves qualitative coding of behaviors, interactions, or interpretations, and thus lends itself well to the study of textual records, e.g., digital archives, interview transcripts, social media. Nelson (2017) extends grounded theory beyond close reading of text, to computational methods of deriving qualitative codes. Chapter 2 contains a commentary on Nelson's "Computational Grounded Theory", supporting the iterative approach of grounded theory as applied to computational text analysis, but arguing that natural language processing has broader connections to grounded theory than she claims.

Institutional ethnography is a sociological framework advanced by Dorothy E. Smith, emphasizing the use of textual records to capture human behavior in situ, with an emphasis on gender, interaction and subjectivity (Smith, 2005). In many ways this is the precursor to what has become known as "digital ethnography", which exploits the massive amounts of human-generated data, especially text, aided by computers and the Internet (Murthy, 2008). Though the data originate with computers, digital ethnography is itself agnostic about the use of qualitative or computational tools to perform ethnography. In Chapter 2 I use computational tools to locate lexical symbols, but instead use traditional close-reading techniques to derive qualitative codes. Conversely, in Chapter 6 I produce qualitative codes through a completely conventional workflow, with no upstream computation beyond data collection, and then use a machine learning tool, a large language model (LLM), to apply the codes to the data set.

Ethnography is the gold standard in cultural inquiry because historically it was the only way to capture complexity in cultural systems. But ethnography is incredibly costly to perform, often requiring years of on-the-ground observation and interviews. Even with all this effort, ethnography often struggles to characterize large systems, generalize across cases, or track changes across years or decades. While computational methods will likely never equal ethnography in the richness of the models they yield, they help to resolve both the problems of time, scale, and scope. As such, the computational methods I mention here should

not be understood as a replacement for ethnography, but an expansion of the methodological toolkit for understanding culture that shares important characteristics with ethnography. Thus computation is a powerful component of the greater social scientific project to understand culture and cultural processes.

1.2.4 Culture is emergent and occurs at all organizational scales

Social science is deeply concerned with scale, as it examines human behavior and societal structures at various levels of analysis. Often the problem of scale is implicit, as the level of analysis is determined by well-defined subfields. Within sociology, social psychology primarily focuses on the micro-scale, examining individual processes and small group interactions. At the other end of the spectrum, demography, stratification, and institutional scholarship engage with macro-scale structures, exploring large-scale patterns and systems. The meso-scale, which lies between these two poles, is more challenging, as it requires an understanding of how micro-scale phenomena aggregate to produce macro-scale structures. Symbolic interactionism, for instance, attempts to bridge this gap by studying how individuals use symbols and interactions to create shared meanings and cultural norms. However, it often stops short of fully explaining how these interactions coalesce into the larger, macro-scale phenomena that structure societies.

The discipline of complex systems is expressly concerned with scale. Simon (1962), following Weaver (1948), locates complexity at the meso-scale, as many different interacting micro-scale components give rise to macro-scale phenomena. The philosophical concept of emergence refers to the case where properties of the whole system are not also properties of the system's constituent parts (Pepper, 1926; Bedau, 1997). Averages are not emergent properties. A market, for instance, calculates price by averaging constituent individual's perception of value. By contrast, a social hierarchy is an emergent property arising from the relationships and interactions between individuals. While each individual in the hierarchy holds a specific position, the structure itself cannot be fully understood by examining individual roles alone, as it reflects a complex network of social interactions. Similarly, a symbolic boundary is a property of a collective, but is not itself a property of the individuals who constitute or interact with that collective. It is important to note that individuals often internalize models of these macro-structures, which is precisely the intuition behind the frameworks of cognitive schemas or cultural toolkits. But often these models are incomplete or inaccessible to the individuals minds which hold them, meaning that only through observing behaviors, particularly interaction, can we understand the system as a whole.

One tool for exploring emergence in social systems that originated in the complexity community is

agential modeling (Holland and Miller, 1991; Smaldino, 2023). More commonly known as agent-based modeling (ABM), the technique involves simulating a population of agents that interact with each other according to a set of rules. The researcher then defines a set of population-level metrics and observes the population dynamics across different configurations of parameters. This allows for hypothesis testing of which dynamics emerge under which conditions. It is important, however, that the conditions and rules determining agent behavior accurately reflect the real systems in question in order for the results of the simulation to be meaningful. In this sense, agent-based modeling may be thought of as a tool for hypothesis generation rather than hypothesis testing.

Such simulations, which were once costly and required highly specialized knowledge, have since been democratized. An undergraduate student can run millions of simulations with millions of agents from their personal computer, far exceeding what the best funded modelers could accomplish at the advent of agential modeling in the 1980s. While agent-based modeling has been relegated to fairly insular communities within each social science, and I am myself skeptical of many of its uses, there are cases where it has contributed to our understanding of the social world (e.g., Helbing et al., 2000; Watts and Strogatz, 1998; Hong and Page, 2004). Chapter 5 presents an agent-based model of the evolution of covert signals, i.e., symbolic boundaries in the presence of a hostile outgroup.

1.3 What do I mean by computational mixed methods?

While this dissertation focuses on how computational methods can offer powerful, new perspectives on cultural sociology, throughout the empirical studies presented I employ methods from a variety of other broad paradigms. It is undeniable that computation has tremendously altered both science and the social, which is partly evidenced by the rise of computational fields and subfields. Figure 1.1 gives an overview of these fields and their interactions. The role of computation in research extends beyond analytical methods to data through new methods of collection, generation, sharing, and processing, all of which are upstream of data analysis. While I privilege computational data and methods in this dissertation, my approach to research relies heavily on combining computation with traditional methods and data. Table 1.1 shows a set of data collection, processing, and analysis, and where each was employed in this dissertation. With regard to understanding cultural complexity, I gain purchase on complexity at different points in the analytical pipeline. Some times this involves high-dimensional crowdsourced data, other times close reading, and still others it may be achieved through complex models. The analytical and rhetorical strengths of the

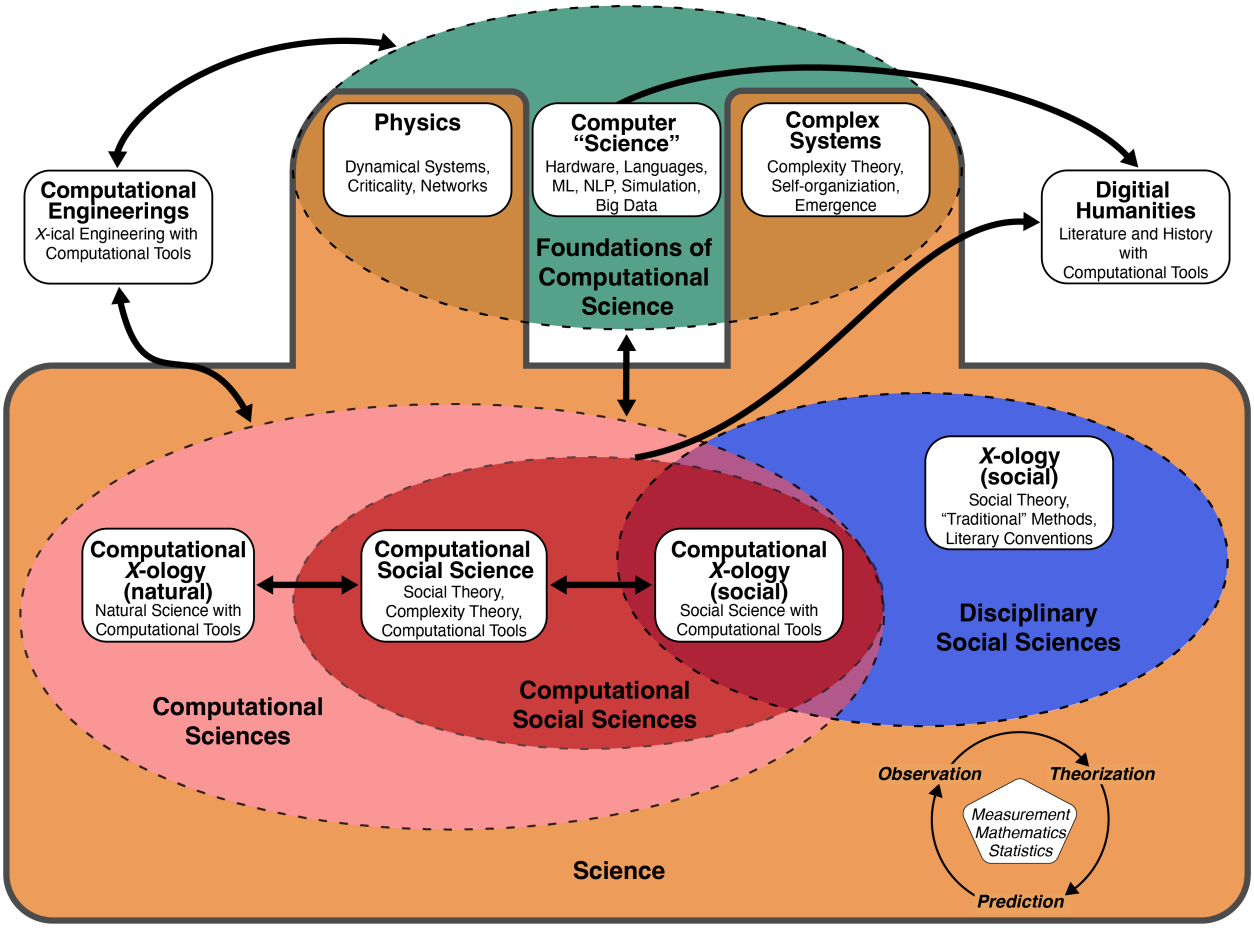

Figure 1.1: Computational social science and adjacent fields.

various methodological paradigms combine to produce research that is fuller and, one hopes, more widely accessible, than operating withinin one paradigm in isolation from the others. The following paragraphs address the diversity of these approaches and the advantages they confer on making complexity tractable to analysis.

Many "computational" data sources preceded the information revolution, but the volume, resolution, and accessibility have greatly increased since the 1990s. Similarly, much of this data does not require novel analytical tools and can be leveraged with a combination of computational processing tools and standard quantitative *and* qualitative methods, even when we privilege complex relationships in the data. Chapter 2 shows how simple computational tools can facilitate qualitative analysis of a dialogic community, identifying about 1,000 informative and widely used verbal identity signals across several dozen manually determined thematic categories. Similarly, I employ statistical methods to probe high-dimensional and relationally-derived behavior. Chapter 3 shows how sociopolitical concepts are engaged in discourse through a combi-

nation of close-reading, crowdsourced data, and frequentist statistics. Chapter 4 similarly infers the outcome of a contagion, the spread of a contemporary myth, using crowdsourced data and Bayesian statistics.

In some cases statistic models have become blurred with computational methods. While frequentist statistics are undoubtedly aided by computational implementation of maximum likelihood estimation, they are and historically were computed by hand. However, I employ several statistical models which are impossible to estimate without the use of computers. Chapter 4 estimates complex models using Bayesian inference, which originated with Bayes theorem in the 17th century, was adapted for statistical estimation through algorithmic developments in the mid-20th century, including cybernetician John von Neumann, and made feasible by the advancement of computing power in the 1990s and 2000s (Carlin and Chib, 1995; Hugh, 2017). Similarly, I fit frequentist generalized additive regression models in Chapter 3 to estimate non-parametric terms to capture cyclical trends in temporal data.

Most importantly, I employ close reading in each of the studies that deal with text and discourse. In Chapters 2 and 6 this comes in the form of traditional qualitative code development. In Chapter 3 close reading informed the dictionary-based method for locating thematic categories in the discourse on Black Lives Matter. A study I conducted with medical informaticians and engineers, which was not included in this dissertation, is perhaps the most complete and straightforward unity of crowdsourced data, computational data processing, frequentist statistics, and close-reading, though the medical informatics framing is considerably less relatable to a sociological or broadly social scientific audience (Dunivin et al., 2020).

Deep substantive knowledge is critical to nearly all meaningful social science, and computational social scientists without traditional disciplinary training more often than not lack the necessary knowledge of their subjects. This can lead to shallow hypotheses and meaningless operationalizations of social entities. Not only does a mixed methodological approach enhance computational work, a mixed disciplinary and theoretical foundation is also critical to responsible and meaningful science. Dual training is undeniably the best way to achieve a well-rounded approach to research, but mixed disciplinary teams are often sufficient, given time to develop bridges between concepts and approaches across fields.

1.4 Theoretical and methodological contributions to cultural sociology

This dissertation comprises five chapters detailing studies of cultural processes, two of which are principally focused on methodological advances toward the study of culture. Each of these studies presents

Table 1.1: Overview of methods used in this dissertation.

Method	**Chapter 2**	**Chapter 3**	**Chapter 4**	**Chapter 5**	**Chapter 6**
Close Reading	✓	✓			✓
Qualitative Coding	✓				✓
Crowdsourced Data	✓	✓	✓		
Web Scraping		✓			✓
API Requests	✓	✓	✓		
Natural Language Processing	✓				✓
Frequentist Statistics	✓	✓			
Bayesian Statistics			✓		
Non-parametric Statistics		✓	✓		
Machine Learning				✓	
LLM Prompt Engineering					✓
Agent-based Modeling				✓	
Dynamical Systems Modeling				✓	
Signal Processing		✓			
Time Series Modeling		✓		✓	

different contributions toward sociological and broader social-scientific knowledge. The remainder of the introduction briefly introduces each study and highlights either or both of theoretical and methodological developments offered by their findings.

Chapter 2 presents both a methodological and theoretical advancement to Nelson's Computational Grounded Theory (2017). I develop a novel flavor of a simple computational text processing technique to locate symbolic boundaries around collective identities in social media data. This demonstrates how cultural environments (Bail, 2014) or dialogic communities (Murthy, 2012a) as captured by trace data are a *tractable* and *accessible* site of symbolic boundaries. The literature on symbolic boundaries presupposes that dialogic communities will contain an abundance of social identities delineated by symbolic boundaries. However, the use of computational text processing methods shows how these data are rich and ripe for sociological analysis. Further, by performing qualitative analysis on the symbolic markers of the White nationalist discourse, I extend Computational Grounded Theory with an eye toward mixed methods, and in

particular qualitative practitioners. Whereas the original articulation suggests that CGT is suitable only for quantitative analysis, I make a meta-methodological case that CGT can be a component of methodologically diverse analytical pipelines, illustrated by the qualitative analysis as part of my application of CGT.

Chapter 3 principally shows how protest can be employed by a social movement to propagate a cultural frame. By analyzing Google Trends, Twitter, Wikipedia Pageviews, and news media data, I show how the Black Lives Matter movement injected terms representing different facets of its political agenda, including policy goals, historical framing, and emotionally resonant rallying cries, into the public discourse. This is particularly important because societal change is often precipitated by conceptual shifts among broad sectors of the public or elites. Along these lines, I advocate to scholars of BLM and racial justice to expand the lens of their inquiry beyond policy to discourse and culture.

Chapter 4 illustrates how several concepts from complex systems can shape cultural change. I examine the myth that famous persons are more likely to die at age 27 than other similar ages, commonly known as the 27 Club myth. Using Wikipedia Pageviews data I show that while this is not the case, those who died at 27 are more famous than we would expect looking at other similar ages. Paradoxically, this real 27 Club effect actually makes people who die at 27 more visible, which creates the appearance that the 27 myth is in fact true. Thus, belief in the 27 Club myth brought the 27 Club into being in a very real way. This illustrates at the collective scale the Thomas Theorem, which states "If men define situations as real, they are real in their consequences" (1938), and similarly Merton's concept of the self-fulfilling prophecy (1948). While this phenomenon, which I describe as "memetic reification", is an extension of classical sociological theory, the mechanisms that produced it originate with, or are at least most thoroughly theorized by, complex systems science. Specifically, while cultural change almost universally occurs through diffusion via social learning (Goldberg and Stein, 2018), the concept of stigmergy (Mittal, 2013) captures how environmental signals, here links to 27 Club member's Wikipedia pages, inflate the cultural presence of those who died at age 27. Further, the initial event which precipitated the 27 Club myth was an extremely unlikely event wherein 4 famous musicians died at age 27 within the span of 2 years. This illustrates path dependence in cultural evolution, whereby an initial random or meaningless occurrence sparks a causal chain of events which are themselves predictable. Path dependence originates in the historical literature and has spread to social science subdisciplines that deal with historical events, but is best formalized in the study of chaos and complexity, wherein events that are impossible to predict can have great consequences on system evolution (Page, 2006).

Chapter 5 presents a computational model of the evolution of covert identity signals, which extends formal models of static covert signals (Smaldino et al., 2018; Smaldino and Turner, 2022). I developed an agent-based diffusion model comprising two groups of actors, a marginalized ingroup and a potentially hostile outgroup, which are both incentivized to learn to identify the ingroup by arbitrary signals given by the ingroup. I show how parameters which replicate real features of group signaling dynamics such as rewards for identification of ingroup (positive reinforcement), punishment for detection by the outgroup (negative reinforcement), and the rate of ingroup-outgroup interaction relative to ingroup-ingroup interaction, produce different dynamic patterns of signal use, abandonment, and novel signal emergence. This theoretical model articulates the conditions under which we will see patterns such as the cycling of identity markers, as signals which are learned by the outgroup are abandoned and replaced by the ingroup. However, as this is the first model of its kind, I put forward a very simple system which lacks many interesting features of real systems such as multiple outgroups, networked interaction, symbolic context, and heterogeneous traits affecting signal adoption. This model doubles as a formalization of Simmel's (1904) model of fashion, where the ingroup can be thought of as social elites who wish to remain distinct from an outgroup of aspiring social climbers seeking to mimic the elites. Both articulations of the model are specific cases of the emergence of symbolic boundaries.

Chapter 6 differs from the preceding chapters in that its contributions are purely methodological. I show that a large language model, with carefully crafted prompting, is capable of identifying sophisticated categories in paragraph-long passages of text. By replicating a wholly manual workflow, I demonstrate how to robustly adapt the process of qualitative coding to allow for machine-application of human-derived codes. This development in automated text classification has eluded workers in natural language processing for decades, and owes its success to advances in machine learning and computational power rather than my own ingenuity. Unlike Chapter 2, which pairs a method with a case study for illustration, Chapter 6 contains no analysis. However the substrate on which I tested the method is expressly cultural. I developed codes to capture cultural frames which are invoked to discuss W.E.B. Dubois in New York Times articles. Further, as text corpora are an important site of rich cultural data, this methodological development shows promise for those scholars of culture who employ qualitative coding of text.

In addition to these substantive contributions, each study is framed by a commentary on the methodological choices I made throughout the study. Specifically, I emphasize how I leverage computational tools to perform scientific inquiry into the cultural forces that shape and respond to our social world. Particular

attention is given to the pairing of computational and traditional methods to produce more comprehensive analyses than any single analytical paradigm could in isolation.

CHAPTER 2

SOCIAL MEDIA DATA WITH QUALITATIVE ANALYSIS: WHITE NATIONALISTS ON TWITTER[1,2]

2.1 Introduction

Social groups and group affiliations are central to much social research. However, systematically identifying the symbols marking the boundaries around a particular group is often difficult, even for expert observers or group members themselves. In this chapter I present an extension of a set of simple computational techniques to automatically extract large and highly specific lexicons (lists of words) representing a symbolic boundary around a social category. The primary contribution of this approach is that it does not require high-quality labeled data and considers challenges and affordances presented by social media trace data. To illustrate the method and its application, I present a qualitative analysis of the rhetoric of White nationalists on the social media platform Twitter, which reveals both historical motifs and novel trends. Finally, I explore how computational methods can augment and expedite, rather than replace, traditional qualitative methods applied to text such as close reading and manual coding.

I first gained exposure to computational text analysis a decade ago. It was 2014 and word2vec was an exciting new technology at the cusp of what can now be described as a groundswell of deep learning. I pivoted from molecular and evolutionary biology to cultural evolution and sociocultural systems, inspired by my first taste of computational tools. By the end of my first year of doctoral studies I had become largely disillusioned with computational text analysis for general social science applications. Much of the work I saw was, I felt, irresponsibly executed, topic models in particular.

[1] This study was subsequently split into two pieces which were under review at the time this dissertation was published: "A lexical approach to locating symbolic boundaries around cultural identities on social media" of which I am the sole author and "White genocide, ethnocrisis, and far-right identity politics: Constructing white nationalist identity through Twitter discourse" with Amanda Lanigan.

[2] Special thanks to Vincent Wong who spent two years working through different formulations of these data, largely following me on what turned out to be a snipe hunt as we tried to capture the process of becoming a White nationalist on Twitter. Thanks also to many professors and students at IU who helped me complete this study, which emerged from the failure of the former. Fabio Rojas, Bernice Pescosolido, Brian Powell, and Stephen Bernard all reviewed drafts and provided helpful comments.

I wrote the essay in this chapter for my Master's Thesis in Sociology, and the latter third is a response to narrowness of scope and inadequacy of natural language processing in social science. The core of this critique is that natural language processing is much more broadly applicable than the core community describes it, and that computational tools would be more widely and more responsibly used if promoted as means of augmenting traditional workflows instead of replacing them. While most computational scholars do not explicitly denigrate qualitative approaches to text analysis, their failure to engage with pre-existing communities of scholars implies a disposition toward their obsolescence.

I have endeavored to center my methodological choices in this dissertation. This entails linking each empirical study through a broader theme of integrating computational and traditional methods to produce work that leverages the strengths of each, and the pragmatics of reaching larger, more diverse audiences. In the studies in the following chapters, the methodological considerations are almost entirely implicit and left to to the reader's imagination. In Chapter 2 methods are centered, and thus making it an especially apt entree to this dissertation.

2.1.1 Background

Boundaries are often invoked in the context of cultural sociology. This is not necessarily because boundaries are more likely to play a role in cultural processes, but rather that determining boundaries is often uncomplicated outside obviously cultural contexts. There are, however, both within and without of cultural sociology a large set of situations where boundary delineation is a barrier to social research. In some cases, unspecified boundaries can be attributed to the nature of the data. More often than not, though, such ambiguity arises from properties of the system itself. The diversity, intersectionality, subjectivity, and fluidity of these boundaries imbue our social lives with richness and texture, and may also motivate sociologists as we endeavor to understand the processes governing the social. The same properties of social boundaries that ignite the sociological imagination can impose enormous time investments if we are to derive scientifically adequate and meaningful accounts of those boundaries.

Historically, boundary specification in sociology has entailed qualitative methodologies such as participant observation, interviews, and content analysis (e.g. McCurdy and Uldam, 2014; Jorgensen, 1989; Watson and Weinberg, 1982; Kleinman et al., 1994; Gilbert, 2002; Lee and Roth, 2004).[3] Over the past 30

[3]Symbolic boundaries are often implicitly located. The terms "role" and "identity" tend to be good indicators that research is centering boundaries rather than situations, behaviors, or dispositions.

years, workers in various disciplines, including Sociology, have developed computational tools and methodologies to automate boundary specification. These approaches tend to fall into either of two camps: network analysis, which delineates the boundaries around groups of actors based on the relations that tie them to one another, and content or text analysis, which leverages trace data to elucidate the symbolic markers which impose or constitute those boundaries. As quantitative solutions to the boundary "specification" or "demarcation" problem were originally articulated in the context of social networks (Laumann et al., 1989), I briefly treat its history in networks before abstracting network methods as a specific case of classificatory solutions to boundary specification. I demonstrate why these methods are inadequate for a large set of cases of boundary specification which instead warrant a (computational) content analytical approach.

In this paper I recontextualize and extend an existing method of content analysis, "frequency-based" lexicon extraction (Monroe et al., 2008), to handle especially difficult cases of boundary specification in textual records. Generally this can be understood as the problem of symbolic boundary specification in unlabeled data, i.e., data for which the social category or categories are unknown to the researcher. This is often the case for textual records generated by a large group of authors. Most often these data are sourced from social media platforms, but administrative records, parliamentary transcriptions, and journalistic or correspondence archives may also be "unlabeled" for the purposes of a particular research question. Methods for extracting group-specific lexicons yield the most reliable and high-resolution approaches to symbolic boundary detection in text. However, existing methods of lexicon detection demand labeled data, and are specified for cases where boundaries are easily separable, mutually exclusive, and limited to just several social contexts. Social media trace data by contrast are often very loosely structured with few formal boundaries. Social media data are further complicated by small (and thus low information) documents, which can make patterns difficult to detect, and viral text, which can erroneously amplify lexical patterns. The method I present here is designed specifically to meet the challenges posed by social media and other corpora comprising many murky boundaries. As a pedagogical exercise I apply this method to locate symbolic boundary markers delineating the White nationalist discourse on the social media platform Twitter. This case is subject to multiple, equally valid theoretical frames, among them social movements, dialogic/textual communities, and ideological/intellectual traditions.

Finally, I situate this computational method of boundary specification in the sociological project more broadly. I follow Nelson's (2020) effort to rigorously articulate a sociological metamethodology of computational text analysis, which she calls "computational grounded theory" (CGT). CGT formalizes and empha-

sizes the role of human interpretation in otherwise "quantitative" analyses, unifying the reproducibility and precision of computational text analysis with the nuance and theoretic grounding of quantitatively derived concepts. In CGT, neither computation nor interpretation take primacy. Rather, they recursively inform one another throughout the process of developing the computational analysis. Here, I extend the jointly quantitative/qualitative ethos of computational grounded theory beyond the process of quantitative inference. Computational content analysis need not the beginning and end of a sociological inference. Rather, content analysis should be regarded as independent of the quantitative/qualitative cleavage. Patterns extracted through computation can inform downstream analyses that follow either analytical paradigm. I am especially concerned with how qualitative practitioners understand the potential for computational tools to play a secondary role in sociological inquiry. While Nelson characterizes deep reading as augmenting computational analysis, I suggest computation can augment qualitative analysis. I deliberately make this case as a response to valid concerns that computational social scientists (and funding agencies) have proposed computation as a replacement for qualitative analysis. Qualitative researchers can exploit the efficiency and breadth of computational processing to facilitate their own qualitative analyses. By embracing computational methods, and perhaps even driving their development, practitioners of qualitative methods best position themselves to weather attacks on their paradigm and preempt charges of Luddism.

2.1.2 Symbolic boundaries and the boundary specification problem

Boundaries are an abstract concept referring to the properties of social and cultural categories or the existence of the categories themselves. Lamont and Molnár (2002) propound two types of boundaries, symbolic and social, which are useful in understanding different methods of boundary specification. They characterize social boundaries as "conceptual distinctions made by social actors to categorize objects, people, practices, and even time and space." Social boundaries are defined as "objectified forms of social differences manifested in unequal access to and unequal distribution of resources (material and nonmaterial) and social opportunities" or less critically "groupings of individuals" in the sense of social network clustering or modularity. Symbolic boundaries, through homophily and other mechanisms, instantiate social boundaries.

Knowing where these boundaries lie is a precondition to much social research. In some cases it is easy to identify. Punks (Fox, 1987) and Hasidim (Tavory, 2010) are easily identifiable by their manner of dress. University department or political party affiliation may be gleaned from organizational documentation or a simple survey. However, when social situations are complicated by multiple salient identities, poorly de-

fined or unknown labels, and large symbolic vocabularies, locating social boundaries or symbolic boundary markers is nontrivial. Accordingly, multiple and diverse literatures have developed both explicit and implicit solutions to various boundary specification problems. Ethnography, interview, and close reading represent the bulk of the qualitative approaches to boundary specification. The quantitative side is represented by computational methodologies: network analysis, natural language processing, and machine learning. This brief review will treat only quantitative methods, though the literature on qualitative approaches is as great, if not greater.

Boundary specification is commonly treated as a classification problem. In social sciences quantitative classification almost always relates to social boundaries: given a set of actors, which actors should be tagged with what labels? Classification may be supervised (we know the labels for a subset of the total population) or unsupervised (we don't know the labels, or aren't leveraging them to train a classifier.) We may seek a simple binary classification or look for multiple categories. If there are multiple categories are these categories overlapping or mutually exclusive? Lastly, the data used to impute classifications may be of qualitatively different types. Specifically, much actor classification in sociology is based on network data, a set of social relations linking actors to one another. Other approaches rely on non-network variables, most often demographic data, but also an actor's behavioral, interactional, or dispositional attributes.

The networks literature(s) are a primary site of social boundary specification. Laumann et al. (1989) articulate a "boundary specification problem" in defining system boundaries in the context of social networks. In a network the size of the human population, where should we make the cut so as to get a subgraph that can offer a vantage on the particular facet of social life? I describe here a second order boundary specification problem. After defining a system, or having one defined for us by constraints in our data, can quantitative methods identify the symbolic or social boundaries intersecting the population? In the case of social boundaries, this is simply a matter of separating the group into distinct, typically nonoverlapping partitions. Locating symbolic boundaries by contrast entails finding the traits that define particular social categories.

Social boundary delineation occupies a large portion of the literature known as "network science." The network scientific approach is generally referred to as "community detection," a diverse set of methods for partitioning network nodes into groupings based on network structure (relations between the nodes) (Danon et al., 2005; Fortunato, 2010). Community detection algorithms come in many flavors, each with strengths and weaknesses and particular structures which they are better suited to identify (Peel et al., 2017).

There exist also non-network methods for social boundary detection which are generally referred to as

"classifiers." Clustering algorithms are unsupervised classifiers that specify a partition across the data without being given examples of particular classes. Common clustering algorithms are k-means or k-medians, random forest, DBSCAN, and Gaussian mixture models. Community detection can be thought of as an approach to clustering based on network structure. Supervised classifiers are typically some form of binary regression. In social boundary detection, the practitioner privileges the model's predictive capacity (the "computer scientific approach,") whereas in symbolic boundary detection, it is the strength of the association of each variable with the group which concerns the modeler (the "social scientific approach.") In other words the classification itself delineates a social boundary and the variables which are predictive of or correlated with a particular category can be thought to represent a symbolic boundary.

2.1.3 The approach advanced here

This chapter describes a method of identifying symbolic boundaries, or, more specifically, identifying a set of symbolic boundary markers from unlabeled data. This method produces a group-specific lexicon, a set of words that mark the symbolic boundary around a particular community. Monroe et al. (2008) present a thorough account of techniques and decision points for identifying differential word usage between two groups. The approach I present builds on a set of non-model-based approaches reviewed in Monroe et al. in two ways. First, their treatment assumes texts have been assigned high-quality labels denoting which group they belong to. The method I present here side-steps this problem by cheaply assigning low-quality labels to the text while still producing excellent results. Second, I address several challenges and affordances of social media and other "big" and situationally unconstrained data sets. By "situationally unconstrained" I mean that the data express a broad range of social situations and contexts. Political debate, organizational documents, and academic literatures are likely to be far more focused than the activity of a set of Twitter or Facebook users, because a single person's social media activity may cover all these in addition to remarks on their media consumption, relationships, mental health, etc. Further, social media is syntactically unconstrained, comprising formal and informal verbiage, including conversation, which leads to short documents and challenges in human and computer interpretation. I have no computational solution for the interpretive challenges, but do address some problems related to short documents.

Monroe et al. argue for model-based approaches for lexicon extraction to the exclusion of non-model-based approaches. They assert that their Bayesian models provide point estimates for odds ratios. However, their treatment is specified for cases that look fairly different from the "big" or trace data considered here.

In part this is because U.S. Congressional debate corpus they analyze contains less group-signalling jargon, which their results repeatedly demonstrate. But the more salient difference between Monroe et al.s' formulation and the Twitter data I analyze here is the enormity of trace data (10^8 records, 10^6 authors.) The combination of these two factors means there are many more robust signals in these data than in Monroe et al.'s. Therefore we can use much more stringent criteria to admit a symbol to our lexicon, and still end up with large and thorough (though not "complete") set of boundary markers.

The Bayesian approach is ill suited to this data for several reasons. The most serious incongruity is that without high-quality labeled data, point estimates for differential term usage are not meaningful in a strictly statistical sense. An alternative approach could be to manually apply in- and out-group labels. But this raises not only the efficiency bottleneck, but also difficulty in human interpretation of short documents and the fact that even a committed in-group member may dedicate only a small fraction of their online behavior to group-related discourse. This is further complicated because commitment changes over time, which means that while ruling "in" might be easy, type II error will be high if based only on a small sample of an author's total output in the corpus.

2.2 The method described

This study expands upon a set of computational methods for identifying a large set of group-identity symbols in textual records. As reviewed in Monroe et al. (2008), the method advanced here is a "frequency-based" approach to lexicon extraction. The frequency-based approach employs odds ratios to evaluate differences in term use between two groups, but does not rely on regression modeling to estimate statistical significance.

I advance here two developments to frequency-based methods to account for some challenges presented by social media trace data. First, I explain the importance of removing multiple copies of the same document, which account for a very large fraction of documents on platforms such as Twitter and Tumblr. More importantly, I describe a process for producing high-quality group-specific lexicons in cases where the group affiliation of utterances or authors is unknown.

2.2.1 Preprocessing the text: Removing copied utterances and identifying phrases in the corpus

Preprocessing text is an important step in any quantitative text analysis. Exactly which transformations are warranted is determined largely by the analytical algorithm(s) we will run over the text and to a lesser degree by particularities of the text itself. Common preprocessing operations include converting to lower case,

Table 2.1: Overview of the general methodological approach

Step	Description
1	Select data that contains high density of a given in-group and relevant out-groups
2	Select words used only by the in-group
	• Small set of *very specific* terms.
3	Identify all authors who know these words (or use them frequently)
4	(Optional) Repeat Steps 2 and 3 for a particular out-group
	• Otherwise, out-group is the set of all authors outside the in-group.
5	Find the words that are used by these authors and not the out-group
	• Rank by odds ratio
6	Manually verify the list of words extracted in Step 5.
	• If low-quality reiterate from prior step.
	• If high-quality, perform sociological inquiry incorporating the lexicon.

removing punctuation and stop words, partitioning text into sub-units (such as paragraphs or sentences), and stemming words. I recommend setting characters to lower case and removing punctuation. Stemming could be useful for amplifying particular signals and organizing the lexicon.

Several additional preprocessing steps are critical before applying the method. The first, n-gram extraction, is common to many tasks and data. N-gram extraction refers to the identification of word sequences of length n, typically bigrams and trigrams, that occur at much higher frequency than would be predicted from the frequency of the individual components. Since lexical identity symbols can be quite specific, it is imperative that we identify common phrases in our data. Alternatively, we may include all bigrams and trigrams in our set of words for a greater computational cost. The second additional preprocessing step is peculiar to social media, Twitter (among others) in particular. This is the removal of viral text. Platforms with sharing features (e.g., Twitter, Facebook, and Tumblr) are full of copied text. A document which has been copied many times in a data set is likely to unduly amplify particular rare words that happen to occur in that post. Since it is likely that members of the same group are closely linked in a network and are (multiply) motivated to share similar content, this has the effect of boosting the frequency of the words in a viral post within the in-group and likely not in the out-group. We must therefore screen out copied text prior to analysis. In fact viral amplification can similarly bias the results of n-gram extraction, thus copied posts

should be excluded from the set of posts used to identify phrases.

Importantly, much of this preprocessing should occur prior to locating the in- and out-group, since labeling is based on the use of particular words. Preprocessing is primarily intended to merge and amplify signals that would otherwise be distinct due only to differences in morphology, i.e., words that differ only in number, tense, part of speech, etc.

2.2.2 Locating the in-group in the data

This method was developed specifically to address unlabeled data. There are several potential barriers to labeling data. The primary reason is that building a training set is expensive. The time it takes to label data is a huge burden, which increases with the difficulty of classification problem, i.e., the strength and frequency of the predictive features. In the case of labeling for lexicon extraction, labeling individual documents would require an enormous data set, because of the number of features (words) and each word feature's relative infrequency in the corpus. Labeling training data must be performed by "experts" who can accurately identify the classes of interest. For many problems, such as image recognition, average adults are suitable experts. This is the intuition behind most crowdsourcing, especially through Amazon's Mechanical Turk. In the case of social identity labeling, many groups require specialized knowledge that precludes crowdsourcing. Furthermore, multiple, dynamic identities complicate the process of tagging individual authors, increasing the likelihood of false negatives. Perhaps the actor only embodies a particular identity for a short period of their total period of activity, or draws on that identity consistently throughout their history, but only in a small fraction of their total activity, which may include a large absolute number of posts, indicating a robust, if less salient identity.

An additional problem in classification is that an identity which is active in a particular post may be difficult to detect. The test case described here, the White nationalist discourse on Twitter, for which this method was developed, presents two sources of detection difficulty. The first is that tweets, i.e., the documents, are very short. This makes deducing the exact meaning of the post difficult because context is limited. The second is that the Internet, and White nationalists in particular, is rife with irony. Even with a fair amount of context, pinning down the exact meaning(s) of 140 (or even 280) characters can be extremely tricky. In practice, these challenges mean throwing out a lot of data that may be informative, which means that not all of the "labeling hours" actually translate to "training data hours." A further source of wasted effort is the large proportion of posts that may be completely irrelevant to demarcating the in- and out-groups we wish

to distinguish. Posts about movies, sports, or a day at work may not signal anything relevant about White nationalism or any other identity we might want to distinguish from White nationalism.

One solution to the challenges of labeling such diverse data as social media records is to pursue an unsupervised approach, which does not rely on labeled data. Unsupervised learning detects correlations between features, and partitions entities based on highly correlated values across subsets of features. In this case the features we train on and the entities we wish to distinguish are identical; we care about the sets of correlated feature-values, here (crudely) the presence or absence of a word, rather than the partition over the set of documents or users. Perhaps the most obvious way to partition users or authors is to identify the partition that relates to the group from whom we intend to extract a lexicon, and then examine which features (words) the algorithm used to draw the boundary. A more sophisticated approach might bypass the categorization of documents and authors entirely to find associations between sets of words. In text analysis, this is commonly accomplished by topic modeling. Topic modeling achieves mixed results, often struggling with fine-grained distinctions, and is especially difficult on small documents, such as tweets. An alternative unsupervised approach might be to apply a clustering on word vector embeddings, which are themselves unsupervised features attached to words based on co-occurrence. While such an approach may in fact identify one or multiple partitions of words representing the group of interest, results are far from guaranteed.

The preceding paragraphs outline the challenge posed by labeling training data and uncertain results achieved by unsupervised methods. The approach presented here circumvents both problems by automating the process of applying low-quality labels to the data set. In reality, it circumvents only the problem of the expense of human coding, sacrificing coding accuracy. However, as I will explain later, the following steps of the method are such that we can be largely unconcerned by this rather "lazy" solution.

The remainder of the Methods describes an automated process for extracting low-quality labels in textual data. This process is analogous to a single-step snowball sample to identify a community from network data. Rather than identifying authors as seeds, we select a set of "seed" terms, which are highly correlated with the group from whom we seek to extract a lexicon. Having selected these terms, we can identify all the authors who have used these words, effectively taking one step on a bipartite network of authors and terms. More stringent requirements may be set, such as knowing all terms, or setting a threshold for "knowing" that is greater than 1, e.g., must have used the term 10 times. In subsequent steps in extracting the group-specific lexicon, we might opt to exclude authors who meet the simple criterion (knowing one word,) but not more

stringent ones, as neither reliably in the in-group or out-group.

The set of seed terms need not be large, but the chosen terms should occur fairly frequently in data. There is no hard-and-fast rule about how frequent is frequent enough. The practitioner must exercise discretion in determining a sufficient set of words to capture most of the in-group in the sample while screening out relevant out-groups. Having a good sense of what proportion of the sample are members of the in-group can facilitate this judgment. Similarly, having selected seeds terms, determining the parameters for inclusion in the in-group or out-group based on the strength of association between author and seed terms requires informed intuitions about the relation between identity and term use. Ultimately the quality of the resultant lexicon will indicate whether the seeds and parameters were appropriate.

The approach to locating the in-group within our data is equally suited to locating an out-group. If we opt not to select an out-group, the out-group defaults to every author in our sample who has not been identified as a member of the in-group or removed from the sample due to ambiguous membership in the in-group. It is important to consider who constitutes the out-group, as they should be as similar to the in-group as possible apart from those terms that truly demarcate the in-group from the out-group. Identities are multiple and interdependent. In a simple sense of interdependence, covariance, I mean that the presence or degree of one identity in an actor informs the probability of each other identity.[4] Considering the case of White nationalists on Twitter, it is difficult to define an out-group that has the same degree of overlap with the rhetoric of the social media platform 4chan, which was the epicenter of the development of post-ironic trolling, anti-PC culture, and casual racism and misogyny on the Internet. Covariance between multiple identities cannot be completely controlled for. However, we can do our best to mitigate its effects by choosing an out-group that differs from as the in-group as little as possible, while still maintaining a hard distinction with respect to the identity we are trying to characterize. Qualitative investigation of the lexicon can help to identify covarying identities and subsequently manually remove words representing those identities from our lexicon.

2.2.3 Extracting a lexicon by ranked odds ratio

Having defined in- and out-group of authors and preprocessed documents as specified above, we can proceed with making the relative frequency calculations. A similar account of these calculations is given by Monroe et al. (2008). As with most steps in this process, the practitioner must apply their own judgment at

[4] In a complex sense of interdependence, identities are non-additive, meaning the effect of one identity is moderated by the presence of other identities. This is what is meant by intersectionality, which is often misinterpretted as simply multiple identities embodied within the same individual. Here, I do not treat intersectionality, as it is almost definitionally intractable to quantification and must be ceded to other epistemic paradigms.

multiple decision points, based both on intuition and an iterative process of lexicon extraction and parameter refinement, as described by Nelson (2020.) However, the results are frequently robust to a fairly broad set of specifications. In other words, multiple parameter settings can produce very similar lexicons.

This method identifies a set of words that are associated with a particular group identity. Thus it compares some sort of strength of association (here a frequency) of each term with some group of interest against some relevant out-group. Term frequencies are calculated by counting of the instances of N terms for each of the two groups, τ^{in} & τ^{out}. In its simplest formulation, this would be raw count of each term, however we might want to count the number of authors who know (i.e., have used) each term, or the number of authors who have used a term at least some number of times. I have found that number of authors who know a term yields the best results, as it prevents a small set of authors from inflating the counts; taking a logarithm of the each author's term counts would likely have a similar effect. Once we have counts for each term, we need to scale the counts to account for population size and productivity. The simplest way to do this is to divide the counts by total number of words in each group, which simultaneously controls for group size, number of documents, and document length. After scaling the counts we divide the in-group frequency by the out-group frequency to calculate the relative frequency. Equation 2.1 describes the calculation of the relative frequency vector, $\varphi \in \mathbb{Q}^N$ (A complementary algorithmic account is given by Algorithm 1.) This equation applies to any of the previously described methods of calculating the term count vectors, $\tau \in \mathbb{Z}^N$, and sum of author productivity vectors, $\omega = \sum_{i=1}^{k} \pi_i \,|\, \pi \in \mathbb{Z}^k$.

$$\varphi = \left(\frac{\tau_i^{in} \omega^{out}}{\tau_i^{out} \omega^{in}} \right)_{i=1}^{N} \tag{2.1}$$

Following the process outlined above is likely to identify very high relative frequencies for rare words that are not salient boundary markers due to their infrequency. In order to screen out infrequent words from a lexicon, we can impose a threshold such as total number of authors who must know the word in order for it to be included in the lexicon.

2.2.4 Qualitatively verifying a candidate lexicon

The description of this method has thus far detailed the acquisition and preprocessing of data and the calculation of relative frequencies for every word that occurs in the data set. These relative frequencies are intended to indicate how much more often a particular group uses a word compared to some out-group; a

Algorithm 1: ODDS RATIO For a set of terms, calculates the odds of the in-group's use of each term relative to some out-group's

Input: A set of terms, $T = \{t_1, t_2, \ldots, t_N\}$, a pair of sets of authors, A^{in} and A^{out} | $A = \{a_1, a_2, \ldots, a_k\}$, and pair of 2D matrices, C^{in} and C^{out} | $C = \begin{bmatrix} c_{11} & \ldots & c_{1N} \\ \vdots & \ddots & \\ c_{k1} & & c_{kN} \end{bmatrix}$, where c_{ij} is the number times the author, A_i, has used the term, T_j.

Output: A vector, φ, of the odds ratio of use of each term for all authors in A^{in} relative to all authors in A^{out}

scaled for the productivity, ω , of all authors in the respective groups

$\varphi \leftarrow T * 0$ (initialize the relative frequency vector)

$\omega^{in} \leftarrow \sum_{j=1}^{k^{in}} C_{,j}^{in}$ (total number of words uttered by the in-group)

$\omega^{out} \leftarrow \sum_{j=1}^{k^{out}} C_{,j}^{out}$ (total number of words uttered by the out-group)

for $j \leftarrow 1$ ***to*** N **do**

- $\tau^{in} \leftarrow \sum_{i=1}^{k^{in}} C_{ij}^{in}$ (count of in-group term use for all terms)
- $\tau^{out} \leftarrow \sum_{i=1}^{k^{out}} C_{ij}^{out}$ (count of out-group term use for all terms)
- $\varphi_j \leftarrow \frac{\tau^{in}\omega^{out}}{\tau^{out}\omega^{in}}$ (scale by author productivity and calculate odds ratio for each term)

return φ

subset of the words are to be assumed to represent a boundary separating one particular group from either a particular or general other. The efficacy of the configuration of a particular data processing "pipeline," i.e., the choices the researcher makes at each of the many decision points detailed above, must not be taken for granted. Instead the researcher must evaluate extracted lexicons (and their parameterizations) through their qualitative understanding of the social context.

The set of words and corresponding relative frequencies does not itself constitute a (candidate) lexical boundary. The words must be ranked by relative frequency and those with the greatest differential expression taken as a boundary around a social category in question. The first and most important of these steps is to rank the words by relative frequency and read through them in descending order. The words at the top of this list should represent the strongest group identity signals (by definition the seed terms, as they completely separate the two groups) and followed by progressively weaker ones.[5] The most import thing here is to ensure that the list seems to represent the boundary well. The top of the list should contain only terms that crisply demarcate the in-group, but as we move further down, good signals may become intermingled with ambiguous ones or indicators of other identities due to multiple identity covariance that

[5]The opposite pole, by contrast, represents terms that separate the out-group from the in-group, though, depending on how the out-group is defined and sampled this may not be a substantively informative boundary.

was not controlled for in the out-group definition. It is likely too that the list will be somewhat stratified by topic, with overlapping bands that represent various facets of the identity. If this pattern does not emerge, and instead the terms seem too general, represent a different boundary, or the relative frequency differences seem too small, this is an indication that at least one step was poorly specified. It may be that the in- or out-group were not defined well, the terms in the relative frequency calculation need to be tweaked, or even that the data do not actually capture the groups they were intended to. It is easiest to try tweaking the relative frequency calculation, since it is the furthest downstream in the process, and the hardest to intuit.

The importance of qualitative interpretations of the social phenomena in question cannot be understated. The complete failure of the frequency calculations to extract a group-specific lexicon is often so easy to identify that invocation of qualitative observation or theory may go completely unnoticed by the researcher even as they make it. In cases where the candidate lexicon is generally more coherent, the qualitative interpretation becomes more explicit, as the evaluation of the lexicon is more difficult, and thus demands deliberate and nuanced hermeneutics. Aberrant words may be erroneous, indicating that the extraction process was poorly configured. Yet surprising phenomena are often precisely what we hope to flush out through inquiry. Distinguishing fallacious associations from unexpected discoveries is unavoidably fraught, and is often insoluble through quantitative inference. Close reading is the best tool for resolving such a dilemma. We can, however, let the computer guide us through this step through a targeted close reading. In a targeted reading, the researcher mines instances of the term or terms in question in order to quickly access many utterances containing the term. This targeted reading may lead to further investigation of particular authors or other terms that can provide more context for the role of the original term or some discursive category it may be taken to represent. Such investigations serve a dual purpose. In addition to the immediate goal of identifying a lexical boundary, the researcher also deepens their understanding of the social phenomena in question, as in a non-computer assisted close reading.

Qualitative investigation may solidify the researcher's confidence in a particular set of terms, but more often than not it raises questions or concerns that inspire another round of parameter selection and frequency calculation. Progressively the researcher hones in a suitable lexicon, while also building a qualitative analysis of the subject. Nelson's (2020) "computational grounded theory" elegantly describes this iterative process for "efficient, rigorous, and fully reproduceable" text analysis. The preceding paragraphs correspond to the "pattern refinement" step of her method, in which deep reading is employed to verify the success of the lexicon extraction.

2.3 The method applied: White nationalism on Twitter

I advance my methodological case by demonstrating the power of lexicon extraction for the White nationalist discourse. Here, I treat White supremacy as a genre of reactionary political ideologies that explicitly justifies Western wealth and hegemony and the maintenance of racially and culturally "pure" communities.[6] The academic study of White supremacy and White nationalism has a long history in Sociology, Anthropology, and Political Science, and, more recently, in Ethnic and Gender studies. I opt for the term "White nationalism" or "White identity politics" rather than "White supremacy" both because they are the preferred (if euphemistic) self-identifiers for participants in the discourse, and because they reduce confusion with political-economic structures of White supremacy. Further, these terms include White identity politics which claim no racial or cultural hierarchy beyond ethnonationalism (i.e., Whiteness should ne privileged only in "White societies") or place Whites at a lower position in their racial hierarchies (such as prominent race scientist and evolutionary psychologist Kevin MacDonald (MacDonald, 1994)). This short study of lexical boundaries and boundary work among White nationalists illustrates the use of the method by revealing both established and novel motifs in right discourses of Whiteness.

Sociological accounts of White supremacy have tended to emphasize social movements organizing (Blee, 2017), including recruitment (e.g., Blee, 1996; Futrell et al., 2006), "demonstration" or public-facing political action (e.g., Beck, 2008), organizational networks (e.g., Burris et al., 2000; Simi and Futrell, 2006; Berlet and Vysotsky, 2006), individual psychologies and social dynamics (e.g., Berbrier, 2000). Throughout this literature there is treatment of identity and ideology. In contrast to the social movements literature, Ferber (1999) offers a post-structuralist account of the White supremacy movement as a "discursive construct," which aims to define the boundary around Whiteness and instantiate an idealist positionality of Whiteness and a politics of "White advocacy." The approach I take here is more closely aligned with Ferber's than any other aforementioned, as post-structuralism accounts for the role of discourse in instantiating identity and ideology.

The discourse around defining Whiteness and justifying a White identity politics follows many motifs with long histories in the White supremacist movement. Race mixing, demographic composition, and immigration are central to White nationalist discourse because they threaten the maintenance of biological, social, and political-economic boundaries around Whiteness. Similarly, White identity politics are concerned with

[6]This overlaps, but is distinct from political-economic structures that prioritize White persons and Whiteness.

who challenges these boundaries, be they immigrants, perceived discursive rivals (Jews/leftists/humanists), or the colonized subjects against whom White supremacy was original constructed (African diasporants / indigenous persons.) Like all political discourses, the response to these opponents is to demarcate them and locate them within relational political ontologies. Nevertheless, White supremacy is particularly violent in its attempts to delegitimize, employing dehumanization as its primary discursive tool.

The case study I present here reinforces the dominance of these historical motifs and uncovers new ones. I extract a White nationalist lexicon through the approach described in the methods and analyze it through a combination of deep reading and summary statistics. Novel motifs include an appropriation of the rhetoric of left identity politics, evolutionary psychological argument, "lexical Blackface," entanglements with the allied or adjacent discourses of 4chan and organized misogyny, and the prevalence of a (post-)ironic rhetorical mode. Further, by contrasting White/ethno/ultra nationalism with "ordinary" nationalism, the absence or deemphasis of particular rhetorical motifs helps to delineate the defining features of White identity politics from features representative of far-right nationalism more generally.

Considerable attention has been paid to the use of the Internet as a tool for White supremacist recruiting and producing solidarity (Burris et al., 2000; Adams and Roscigno, 2005; De Koster and Houtman, 2008; Caiani and Kröll, 2015). Less attention has been paid to its role in discursive boundary making. Murthy (2012b) in characterizing online discourses as Bakhtinian "dialogic communities" has drawn attention to social media as both a site and model system of discourse. Online communities are well-suited to studying both the global, e.g., social and symbolic boundaries, and local, e.g., interaction and rhetorical tactics, of discourse, as well as mesoscopic features such as network community structure and diffusion dynamics.[7] The Internet has become a ubiquitous, and, for some, even primary site of social action, particularly for political discourse and "fringe" communities, of which White nationalism is both. The study-within-a-study I present here draws attention to how White identity politics have evolved over the past decade, while also retaining their discursive core. Furthermore, it demonstrates the efficiency and breadth of a particular computational approach to characterize a rhetorical space. Finally, it advances the methodological and theoretical position of social media in the sociological imagination.

[7]Recommendation algorithms occur at the intersection of all of these, as they simultaneous are constructed by patterned action and structure the environment, and thus constrain action.

2.3.1 Data

The target population for this study is a set of Twitter users who are engaged in the White nationalist discourse through their Twitter accounts. In order to make inferences about these users, I defined a control population of users who might have become White nationalists but did not, or did not yet. In order to get both core and peripheral members, I employed a coarse sampling method. I selected 29 seed accounts of prominent White nationalists and collected a list of their followers. All but one are public figures, publications, or publication editors who are identified as White nationalists in at least one journalistic or hate-group monitoring source. The two most prominent White nationalists on Twitter, Richard Spencer and Klan leader David Duke, were excluded over concerns they were too much a part of the general far-right discourse and would contribute to the sample many non-White nationalist accounts and few genuine White nationalists, as most committed White nationalists would follow at least one of the less prominent seeds.

This sampling approach produced a population of 172,944 unique Twitter accounts in September 2017. I then used Twitter's REST API to gather the user's past 3,200 tweets, the maximum number of historical tweets available for user timelines. Of the original 172,944 users in the aggregated ego network, I collected 146,210 users' timelines, approximately 211 million tweets. I was unable to pull data from users whose accounts had been deleted since the time of sampling or were private and thus inaccessible.

2.3.2 Lexicon extraction

Every author in the corpus followed at least one White nationalist Twitter account. Therefore, these data almost certainly capture both genuine White nationalists as well as those peripheral to White nationalism, but not necessarily engaged in the discourse. I leveraged this assumption to automatically apply in- and out-group labels in preparation for the relative frequency calculation. The in-group is carefully defined using a set of 4 terms and associated hashtags summarized in Table 2.2. These terms were carefully selected after dozens of hours of close reading White nationalist Twitter accounts and other contemporary media. All users who have used at least one of these terms in an original utterance (*tweet*, rather than *retweet*) are taken as the in-group. This is a very loose boundary. There is strong reason to believe it captures both non-White nationalists and excludes some genuine White nationalists. However, the objective is to produce in-group and out-group labels that are overwhelmingly correct, not perfectly so.

In the process of extracting lexicons and measuring descriptive statistics (much of which is unreported here), I formulated two distinct out-groups (Table 2.3). One is simply the negation of the in-group: all

Table 2.2: Counts of seed terms in the corpus. "Raw" count is the total number appearances in the corpus. "Users" is the number of authors who have used a term at least once. The first set of terms are used to identify white nationalists for the purposes of extracting a group-specific lexicon. The second set of terms are used to identify "mainstream" nationalists.

	All Tweets		**No Retweets**	
Term	**Raw**	**Users**	**Raw**	**Users**
14 words	5,000	2,456	3,597	1,387
#14words	2,438	976	1,453	345
1488	5,531	3,490	3,852	2,324
#1488	816	444	685	332
antiwhite	87,514	25,072	37,422	9,402
#antiwhite	19,109	4,675	10,280	1,045
white genocide	41,220	15,145	20,494	7,401
#whitegenocide	117,042	15,280	45,386	4,443
Term	**Raw**	**Users**	**Raw**	**Users**
MAGA	97,031	29,958	45,282	26,971
#MAGA	1,574,769	57,407	476,421	14,175

Table 2.3: Group size for various group formulations based on usage of seed terms. Users are treated as part of a group if they have used one of the terms in a given set of seed terms (WN: white nationalist; MAGA: "mainstream" nationalist).

Category	**Users**	**Users (no retweets)**
all	146,210	–
WN	36,624	17,213
WN only	8,792	9,362
MAGA	60,993	33,565
MAGA only	33,161	25,714
WN & MAGA	27,832	7,851
neither	76,425	103,283

users who have never tweeted one of the seed terms (but may have retweeted them.) A second formulation attempts to capture only "mainstream" nationalists, or those nationalists who are not engaged in a rhetoric of White identity politics. Here the out-group is defined as all authors who have tweeted in support of candidate/President Trump with "(#)MAGA" (Make America Great Again), but have not tweeted one of the White nationalist seed terms. Both out-group formulations produced equally strong lexicons. The lexicon analyzed below was generated by taking as the out-group all authors who had not tweeted a seed term.

I produced many lexicons before settling on the one reported and analyzed here. The documents for this lexicon included only original utterances, i.e., no retweets. Both the term counts and scaling parameter were the simplest configuration detailed in the methods. The counts for each term were "raw" counts in that a term will be counted each time it appears in the corpus, rather than once per author or document, and the scaling counts were for each group was the sum of these raw counts (the total number of term tokens in each group's corpus.) Two thresholds were selected after a cursory reading of the lexicon, as these thresholds may be arbitrarily chosen at any point after the relative frequency calculation. Terms that appear in the final lexicon must have been used by at least 250 authors irrespective of group and have an odds ratio greater than 5 (5 times as much use among "White nationalists" as "non-White nationalists.") The White nationalist lexicon extracted through these parameters is extensive and broad. Strong boundary markers can be found beyond each of these thresholds, however, I judged the list produced by these parameters sufficiently comprehensive, general, and unmarred by ambiguous or erroneous symbols for the purpose of the following analysis.

2.3.3 Deep reading the lexicon: what can we learn about White nationalists?

A large and diverse group-specific lexicon has many possible uses. One of the simplest is an object of study in itself. In the following paragraphs, I analyze a lexicon generated through the method described above. Importantly, this analysis does not differ notably from a traditional, i.e., non-computer assisted, close reading of a group-specific lexicon. As noted in the Methods, in cases where I encountered an unfamiliar term, I occasionally used targeted close reading to rapidly bring up a set of examples of the term usage in the corpus. Apart from this, the analysis is a traditional close reading of a set of terms in the text. However, this set of terms could only have been generated using a computational method, and is far superior, as well as more efficient, than assembling a lexicon through qualitative methods themselves. In other words, the *description* of the symbolic boundary, i.e., the lexicon, is extracted through computational/quantitative methods, but the

inference around the symbolic boundary is executed through traditional, qualitative analysis.

I hand-coded a lexicon of 841 terms, which were selected with an arbitrary odds-ratio cutoff of 5. A majority of the terms up to an odds-ratio of 3 are still strong boundary markers, e.g, implicit 3.46, orcs 3.37, White lives 3.29, melanin 3.27, death squads 3.26, oppression olympics 3.16, separatism 3.09, Black supremacy 3.05, sickle cell 3.05, kosher 3.00. The final coding scheme has 27 categories (summarized in Table 2.4.) Several other codes are not reported due to infrequency and are aggregated under "Other." Each term was assigned at least one code; many were assigned multiple. Table 2.5 lists examples of the categories which will be discussed in the deep reading below. The complete lexicon and encodings are omitted from this manuscript due to space constraints. Nevertheless, there is considerable overlap between particular codes. Similarly, many terms are captured multiple times in the lexicon, due largely to pluralization, tense, parts of speech, and spelling variants. The summary statistics do not compress these variants, but my deep reading and Table 2.5 largely ignore morphological variation.

Far-Right Identity Politics

I follow Gray's (2018) characterization of the White nationalist discourse as a "White, male identity politics." This is a new formulation of the White supremacist movement that has adopted the rhetoric of left emancipatory politics. This is a response to both the institutionalization of identity politics and multiculturalism in the 1990's, and the decline of the post-war industrial socioeconomic order since the 1980s.

"Anti-White," "racist anti-White," "[anti-]White hate," "[anti-]White violence," "White rights," and "White oppression" appropriate the rhetoric of anti-racism to develop a "White identity politics." White nationalists developed their own term for identity politics, "identitarianism." They use identitarianism to refer not only to White identity politics, but all identity-based movements. In this way Zionism is an "identitarian" movement, much as Israel is an "ethnostate."

The rhetoric of the left is also employed to combat left politics. "#Whitelivesmatter" is a response to #Blacklivesmatter, much like the mainstream far-right adopted #alllivesmatter and #bluelivesmatter. Similarly, White nationalists refer to "Jewish" and "Black" privilege to undermine the notion of White privilege. Similarly, the so-called men's rights or meninist movement has also adopted the term female privilege, but it does not appear in this lexicon. "Diversity means" is a rhetorical trope employed to construct an association between multiculturalism and "White oppression," e.g., "Diversity means White genocide."

A set of meta-discursive terms appears near the top of the lexicon. "White propaganda" and "White

rhetoric" both engage a sense of *doing* politics. "White agenda," "White interests," and "White advocacy" motivate a White identity politics. The phrase "White pill" references the "red pill," which originates with the discourse of misogyny. The red pill refers to a scene in the *The Matrix* film franchise where the protagonist consumes a red pill and is liberated from the oppressive false consciousness of a computer-simulated world. While initially associated with men developing an emancipatory gender awareness, the term has taken on a life of its own with "red pilling" or "pilling" referring to the instillation of an emancipatory ideology in oneself or others, with various pill colors corresponding to particular ideologies.

A set of "dispossession" terms describes the oppression of Whites. "Deracinated," "rootless," and "demoralization" all describe alienation. "Alienation" itself is not used due to its association with the left. "Degeneracy" and "dispossession" refer to "White genocide," representing the destruction of "White culture" and subversion of White power. White supremacists have used White genocide to refer to race-mixing and being "outbred" by non-Whites, but the cultural meaning has also been used in the discourse for some time (Ferber, 1999). Degeneracy (*Entartung*, *degenarazione*) dates back to the 19th-century and generally refers to the decay of social, moral, or biological order. It has been widely used in race scientific, fascist imperialist, and anti-modernist argument, notably in Nazi Germany. Similarly "cultural Bolshevism/Marxism" links contemporary left identity politics to the rhetoric of the Third Reich, while simultaneously referencing the pseudo-academic term "post-modern neo-Marxism" popularized by evolutionary psychologist and masculinist self-help guru, Jordan Peterson.

Slurs

There is a large and growing interdisciplinary literature on "hate speech," which draws from political science, psychology, communications/media studies, law/criminology, and, recently, computer science (Tsesis, 2002; Douglas, 2007; Hine et al., 2017; Chetty and Alathur, 2018; Fortuna and Nunes, 2018). The hate speech paradigm emphasizes violent speech that targets minorities. Lexically, hate speech is epitomized by slurs. However, much hate speech can only be understood through larger linguistic constructions e.g., "I hate [minority group]" or "[Locale] would be better if [minority group] weren't here." I focus on the lexical in part as a response to computer science, which tends to employ lexical approaches, as they do not require sophisticated natural language understanding (NLU) still beyond our technical capacity.[8]

Many slurs appear in the lexicon, but they represent only a small fraction of it, and do not occupy the

[8]This was written several years prior to the development of generative transformer models generally know as large language models.

Table 2.4: Summary statistics for hand-coded categories. Categories are displayed in descending order by mean odds. Full Lexicon (**bold**) shows summary statistics for the entire white nationalist lexicon (excluding the seed terms.)

Category	# Terms	Mean Occurrences	Median Occurrences	Mean Rank	Mean Odds
Whiteness	129	4855	956	359	14.49
WN Media	103	1758	1058	341	13.72
Code	29	1498	1040	279	12.59
Left Discourse	29	2049	1000	320	12.29
South Africa	11	1075	901	261	11.83
Race Mixing	18	1334	1165	383	11.54
Governance	28	1907	1060	301	11.29
Jews	79	6760	1176	342	10.85
Irony	37	2535	1117	331	10.84
Full Lexicon	**841**	**3076**	**1024**	**421**	**10.46**
Nazi	31	1539	871	401	9.99
History	15	1116	1033	397	9.95
Boundary Work	30	3070	1047	390	9.80
Demographics	70	1586	821	416	9.71
Evolutionary Psychology	10	714	668	403	9.71
Slur	41	1752	859	409	9.48
Leftism	33	3108	1051	478	8.82
Anti-black	18	2036	923	435	8.80
Islam	18	2811	1416	561	8.69
Holy War	8	3545	1396	564	8.68
Dispossession	6	3097	1227	416	8.61
Blackface	32	2232	1046	477	8.52
Other	55	2258	949	466	8.41
Biological	28	1247	669	469	8.25
Ethnicity	51	2918	848	501	8.17
4chan	67	4127	1273	508	7.74
Misogyny	42	4388	1244	515	7.70
Anti-gay	9	1688	1803	595	7.10

highest echelons, i.e, the most distinctive terms. Among the slurs in the lexicon, the highest odds ratios are less common and novel slurs. More common slurs are used by racists who are not actively engaged in racist discourse or White identity politics. Slurs for Jews are the exception, since antisemitism is more prevalent among White nationalists than the far-right who are not engaged with White nationalism.

Demographics

The lexicon I extracted reveals a sophisticated and extensive preoccupation with demographics. These range from relatively neutral, e.g., "demographic trends," "urbanite," "naturalization act," "birth rates," to pointed, e.g., "White minority," "self-segregate," "homogeneous society," "diversity quotas," to inflammatory, e.g., "demographic replacement," "population removal," "migrant invasion," "import millions," "flood Europe." Notably, mainstream anti-immigrant rhetoric is absent from this list. This is less because White nationalists avoid "ordinary" (quotations my own) far-right nationalist terms than it is the out-group is predominantly made up of this second group. Therefore the high frequency in both in- and out-group neutralize one another.

Race Rationalized

Race is rationalized through biological and psycho-biological terminology. I.Q. is often central to White supremacist argument, as well as the target of personal insult (as is typical of the broader rationalist online discourse, which is overwhelmingly male.) Genetic determinism and evolutionary argument, e.g., "admixture," "genetic differences," and "conserved [trait]," are employed to justify White racial superiority and isolationist immigration stances. Race science terminology, e.g., "caucasoid," "negroid," "subspecies," and "physiognomy," is common. Demographic rhetoric is also biologized, e.g., "outbreed," simultaneously rationalizing and dehumanizing.

Rationalized racial rhetoric has been employed by White supremacists as long as race science has existed. Its use among White nationalists is thoroughly documented (Ferber, 1999). What is new, and is apparent from analyzing this lexicon, is that racism is itself rationalized. Evolutionary psychological discourse has provided a scientific line of argument justifying racial and ethnic hierarchies and segregation. The evolutionary psychological community, while suspect for other reasons, is largely not directly responsible for segregationist or racial hierarchical hypotheses, though it is, along with behavioral genetics, undoubtedly the greatest source of race science in the academy. "Pattern recognition" is an evolutionary explanation for recognizing difference in out-groups. Evolutionary psychology is used both to denigrate particular ethnic groups (biology/culture) as inherently deficient, e.g., "time preference," but also to argue for ethnic homo-

Table 2.5: Themes identified through hand coding a White nationalist lexicon. Terms are listed in rank order by odds ratio, i.e., the most distinctive words among white nationalists are listed first. Alternative forms are left out of the sample unless they are informative, whereby variants are indicated with parentheses, e.g., (low) birth rates, as both "birth rates" and "low birth rates" appear in the lexicon.

Theme	**Sample Terms**
Jews (irony)	cohencidence, schlomo, goy, gorillion, kvetch, shiksa, shekel, oy vey, shabbos, talmudic
Slur	muds, negroid, sheboon, shitskin, heeb, jewess, chimpout, negress, nogs, lampshades, noods, jewed, kike, nigs, skypes, chimping, rapefugee, mudslime, swarthy, beaner, spic, niglet, negros, pajeet, bantz, darkies, nigger, gook, sand nigger
Blackface	gibsmedat, dindu, gibs, (wuz) kangz, nuffins, rayciss, ebil, anudda, obongo, dindu nuffin, sheeeit, wypipo, nibba, tbh fam, (b)ooga, famalam, (kill) whitey
Boundary Work	jewish supremacist, alt lite, cuckservative, race traitor, white leftist, white jews, wigger, hapa, new right, #frogtwitter, black nationalism, #altrightmeans, albinos, black supremacist, black racist
Consciousness Raising	white propoganda, white rhetoric, white agenda, white interests, white advocacy, white pill, white movement, white identity
Left Discourse	(racist) anti-white, diversity means, jewish privilege, #whitelivesmatter, white identity, white hate, white violence, identitarianism, white rights, white victims, white racism, white identity politics, european identity, #whiteguilt, racial identity, #blackprivilege, ethnic identity, white oppression
Dispossession	white genocide, dispossession, deracinated, degeneracy, rootless, demoralization
Political Theory	reaching levels, vibrancy, ethno, greatest strength, serious country, slave morality, fasces, horseshoe theory, institutional power, balkanize, cultural enrichment, overton window, healthy society, accelerationist
Demographics	demographic replacement, ethnostates, white minority, third world immigration, homelands, naturalization act, population removal, physical replacement, homogeneous society, (low) birth rates, white flight, urbanite, self segregate, immigration act, migrant invasion, import millions, diversity quotas, homogeneous, outbreed, flood europe, native population, demographic trends, overrespresented, great replacement, #deportthemall
Biological	negroids, race realist, racial differences, dysgenics, iq differences, human biodiversity, subspecies, caucasoid, conserved, biological reality, outbreed, average/low iq, admixture, genetic differences, physiognomy
Evolutionary Psychology	pathological altruism, group preference, (high) time preference, high trust, parasitism, social cohesion, pattern recognition, egalitarian
Islam	white sharia, #pegida, (mass) muslim immigration, islamification, #notallmuslims, muslim rape gangs, mohammedans, londonistan, muslim rapists
4chan	\pol, pozzed, larp, cuck, (red/black/white) pill, soyboy, groyper, poasting, pepo, shit tier, sperg, wife's son, normies, neets, austically, press f, 8chan, gersh, henlo, cummies, incel, wew (lad), autists, manlets, edgelords, shitlord
Masculinism	cuck, (red/black) pill, soyboy, #mgtow, wife's son, shit test, manlet, incel, beta cuck

geneity as biologically necessary for "social cohesion" or a "high trust" society. The biological drive for "group preference" is a racialized extension of kin selection, a leading evolutionary account of altruism (Eberhard, 1975; Foster et al., 2006). "Pathological altruism" is the misapplication of the evolutionary drive that enables "parasitic" out-groups to subvert society. This is a biologized account of the White savior, except the White savior develops a racial awareness and recognizes that saviordom will destroy him, i.e., "White genocide." The relatively recent emergence of evolutionary psychology as a respectable race science is documented in Saini's (2019) book on the resurgence of race science under the banners of "human biodiversity" and "race realism," both of which appear in the lexicon.

Some (Gray, 2018; Saini, 2019, e.g.,) describe a biologized view of race as requisite to White supremacist ideology. I acknowledge that biological essentialism is typical and dominant, but not necessary either to espouse White nationalist rhetoric or develop an intellectually rigorous White nationalist worldview. White nationalists pose a society centering the "ethnos" as the anchor of social cohesion and cultural coherence. In this sense the biological and sociological arguments for the ethnostate merge. The former argues that *Homo sapiens'* evolutionary history demands ethnic segregation, the latter that homo socius requires a shared culture and "homeland."

Irony and Post-irony

Irony is a recurrent theme of the White nationalist lexicon. In part this is because irony has become a dominant feature of much Internet discourse by digital natives. White nationalists on Twitter heavily skew young and male, so irony is expected. But ironic discourse also stems from contemporary White nationalism's connection with the anonymous forum 4chan, on which (post-)irony is the dominant discursive mode (Nagle, 2017; Merrin, 2019; DeCook, 2020). By post-irony I refer to the use of irony that deliberately sends differing, even contradictory, messages to different audiences, simultaneously intends ironic and sincere interpretations of a message for certain audiences, or attempts to confuse and annoy readers who cannot interpret the irony.[9] Decook (2020) calls this "coded irony," which she characterizes as the ethos underlying the troll culture of 4chan and the alt-right.

The roots of a strain of White nationalism in 4chan are attested by a set of terms specific to 4chan, though not necessarily the discourse of White identity politics. Others may have originated there, but they are unknown to me. Many of these are inherently or potentially post-ironic.

[9]Others have applied the term "post-irony" to various literary responses to irony, including the embrace of sincerity (Collins, 1993; Konstantinou, 2009; Hoffmann, 2016).

Terms regarding Jews abound the lexicon. The set displayed in Table 2.5 specifically regard the use of irony in association with Jews. Perhaps surprisingly, many Yiddish words are employed to mock and subtly reference Jews. Others include the portmanteau "cohencidence," referencing Jewish conspiracy, "schlomo," a novel slur for a Jew, and "gorillion," which mockingly gestures at inflationary Holocaust death counts. The novel slur "lampshade" has its origin in a myth about how Nazi officers used the skins of executed prisoners.

Lexical Blackface

A majority of anti-Black terms engage a form of ironic digital Blackface that is explicitly racist. Most of these terms employ a fictitious African American patois, which dehumanizes in both form and content. "Gibsmedat" (give me that) and "dindu nuffin" (didn't do nothing) are novel slurs that originated on 4chan. "Tbh fam" (to be honest, fam) appropriates AAVE slang, though may simply reflect the relative youth of the White nationalist cohort as compared to the out-group. "Famalam" is genuine (though somewhat uncommon) AAVE slang, but is consistent with the ironic patois. "Wypipo" also plays on the patois, but originated in radical Black online discourse to refer to Whites (Smalls, 2018), before migrating to broader left anti-racist circles with a similar usage.

2.4 Discussion

This paper extends the so-called "frequency-based" toolkit for group-specific lexicon extraction. Prior formulations of the approach require that authors or documents be correctly labeled as belonging to the groups of interest. In the case of very large data sets, labeling by hand is usually infeasible and even when manageable may offer only a small subset of the data, eschewing the benefits of big data. My primary methodological contribution lies in demonstrating that low-quality automatic group labels can be substituted for high-quality manually coded ones. The results presented here as well as the full lexicon given in the appendix show that large and comprehensive lexicons may be generated through the process described in the Methods. Bolstering support for computational lexicon extraction, this list is much more comprehensive than any that could be generated through close reading or crowd-sourcing via survey. Compared with earlier applications of the frequency method on relatively small corpora, the big data approach obviates the need for a statistical test for significance as advocated by Monroe et al. (2008), as very large samples with a simple frequency threshold prevents erroneous associations.

This technique is described as an "automated" method of lexicon extraction. However, the automatic component is actually just a small part of an iterative process wherein the results of the lexicon extraction

must be validated within the researcher's intuitive understanding of the social phenomena in question. Nelson's (2020) computational grounded theory articulates this process in detail (summarized in Figure 2.1.) In computational grounded theory, the parameters used to define the lexicon are refined until substantial portions of the lexicon are brought in line with intuition and erroneous markers that can be explained away are no longer present. Nelson stresses that computational grounded theory is more rigorous and reproducible than traditional content analysis. Part of the "rigor" or "objectivity" of the approach is that the computational tools do not enable piecemeal excision of aberrations. Since the parameters are specified at a very abstract and theoretically removed level, targeting very particular challenges to theory for excision is difficult. If for instance an iteration of the lexicon extraction produces a subset of words that are theoretically problematic, say White nationalists appear to be talking a lot about Cajun food, we must either justify the presence Cajun cooking in our account of the boundary of White nationalism, or make some adjustment to the extraction process. This can be done either by tweaking the parameters for lexicon extraction or redefining the set of actors we take as the in- and out-group. The parameters of extraction used here (and those of any other computational text processing method) are very distant from concepts such as Cajun cooking, so any solution that removes Cajun food must be very general, resulting in widespread changes across the lexicon.

This chapter begins by framing group-specific lexicons as a symbolic boundary around a social category. The framework of boundary-specification problems, while originally articulated in the context of networks and social boundaries (Laumann et al., 1989), provides a scaffolding for understanding the problem of specifying symbolic boundaries as the flow of decisions around system representation (data acquisition), system description (lexicon extraction), and system analysis (inference using the lexicon) highlight how boundary specification is in fact a central problem in social research, and one that makes the same demand of the researcher across problems. "Where should I draw the boundary around this system so as to capture and make tractable the social phenomena I am trying to understand?" This presents several interrelated problems:

- Location: Was the boundary drawn around the relevant set of actors?
- Scope: Were the relevant details recorded given a particular set of methods?
- Scale: Is the system representation (the data) large enough to ensure conclusions are representative of genuine patterns?

Boundary specification, both here and in its original context, is generally a problem of the first type. However, it should be plain that these each of these questions must be resolved in order to execute any social inquiry. Happily, the answers are often obvious. But in many cases (more often than most believe) we encounter non-trivial, even prohibitively recalcitrant problems. Social research necessitates observational and theoretical intuitions around social situations. Science, whether conceived through a positivist or constructivist frame, entails recursive processes of observation and theorizing. This recursion is typically understood as a collective process of knowledge production, but it can be found throughout the nested hierarchy comprising rigorous inquiry. Boundary specification represents a fairly low-level recursion, wherein the boundary is validated by intuition while at the same time challenging it.

Nelson succeeds in articulating this recursive relationship, demonstrating that contextual understandings based on theory and qualitative analysis are essential to responsibly undertaking a (nominally) computational analysis. However, she characterizes computational text processing as ultimately building toward quantitative analysis. This portrayal of computational tools as quantitative in nature greatly undersells the utility of computation to facilitate textual analyses. In particular, lexicons such as the one derived here can guide the development of a broad variety of quantitative and qualitative studies other than verifying patterns extracted through computation. These lexicons can be used to inform interview questions or strategies, facilitate participant observation, construct a survey, design an experiment, or contextualize conversation analysis or other deep readings. Figure 2.1 situates the process of computational grounded theory in a much broader methodological field.

On its face, the characterization of computational text processing as a quantitative method is sensible. Computers are fundamentally quantitative entities, adding and subtract bits of information. We teach computation as a mathematical or engineering discipline, and, as social scientists, conveniently calculate regression analyses with them. Nevertheless, most of what we do with computers is not "quantitative:" we write emails and blogs, we read the news and books, we do graphic design and video editing, we record and generate music. As social scientists, we use computers to conduct experiments, give surveys, code text and interviews, and scrape trace data, none of which we characterize as quantitative.

If so much we do with computers is not quantitative, why do we treat NLP as a quantitative paradigm? A category error emerges when we recognize that there is an implicit understanding that quantitative methods are inherently inferential or necessarily feed into a quantitative inference. Inference is, however, only one of a number of steps in scientific inquiry. Data must be acquired and processed, and then described or

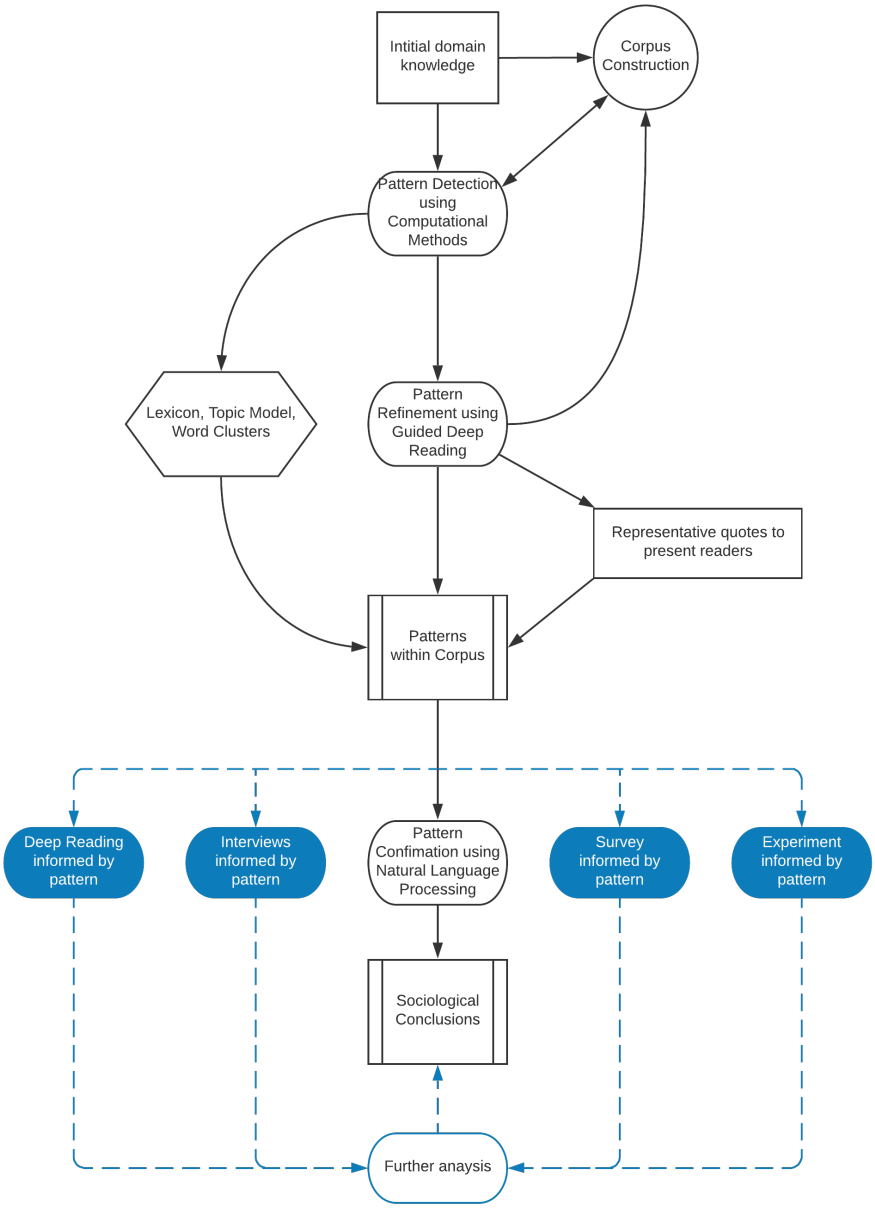

Figure 2.1: Extension of Nelson's (2020) computational grounded theory framework. Shaded and dashed structures (blue in digital) represent proposed alternative processes beyond the CGT framework. All other structures (Black in digital) represent a slightly modified version of the original figure.

analyzed.[10] While many computational tools deliver analysis, most are merely processors. Accordingly, we refer to natural language *processing* rather than *analysis*. Techniques such as topic modeling, vector embeddings, knowledge graphs, and the approach advanced here are not in themselves inferential tools. Rather they are (quantitative) approaches to categorizing or organizing words. Once processed, we may feed this organization into a quantitative (à la Nelson) or qualitative (this study) analysis. Alternatively, the organization may not lead immediately to analysis at all, and rather be employed to collect or process new data. Quantitative analysis need not be the ultimate destination of the inquiry.

2.5 Conclusion

Sociology has remained fairly suspicious of these and other computational methods.[11] As we look to other disciplines to identify triumphs and pitfalls, a critical metamethodological literature is slowly emerging to steer the discipline toward responsible application of computational tools. Personal computing, improvements in computational processing power and memory capacity, the Internet, archival and bureaucratic digitization, and "social media" have coalesced to provide both an unprecedented volume of social behavior records and the means to process those data. These technological and social-structural transitions far outpace the impact computational methods have had on content analysis, Sociology, or academic knowledge production more generally.[12] Computational text processing and analytic methods such as the one advanced here have the potential to enable inquiry which is heretofore intractable to existing methods and facilitate traditional content analysis by improving breadth and efficiency.

As with any tool, careful consideration must be made in order to avoid misapplication of a method or its specifications. Perhaps unlike most other methods, NLP presents a tremendous opportunity to integrate human interpretative and synthetic abilities with mechanical speed, reliability, and precision. However, reconsideration of existing quantitative paradigms will likely reveal that theorizing and qualitative observation are in fact operating throughout so-called quantitative processes. Recognizing and explicating the mutually informative relationship of quantitative and qualitative approaches can only improve the rigor of our science, and, one can hope, continue to break down barriers between paradigms, leading to more innovative

[10] "Theory" (though not "theorizing") is invoked at all of these stages, as theory provides the interpretive frame through which we decide how to execute these steps.

[11] Kozlowski et al. (2019) may indicate that Sociology is becoming more accepting of computational methods.

[12] We have also, Sociology in particular, greatly underestimated how computers and Internet communication have restructured aspects of society.

and impactful research.

CHAPTER 3

TRACE DATA FOR TEMPORAL AND SEMANTIC RESOLUTION: BLM DISCURSIVE SHIFTS[1,2]

3.1 Introduction

In the chapter I show that Black Lives Matter (BLM) protests shift public discourse towards the movement's agenda as captured by social media and news reports. I find that BLM protests dramatically amplify the use of terms associated with the BLM agenda throughout the movement's history. Longitudinal data show that terms denoting the movement's theoretically distinctive ideas, such as "systemic racism," receive more attention during waves of protest. I show that these shocks have notable impact beyond intense, or "viral," periods of nationwide protest. Together, these findings indicate that BLM has successfully leveraged protest events to engender lasting changes in the ways that Americans discuss racial inequality.

This study makes use of diverse computer-mediated data and two computational analysis methods to approximate patterns of the spread of ideas and conceptual linkages that emerged in conjunction with Black Lives Matter organizing. As with the previous chapter, the analysis largely uses traditional methods, here frequentist statistics, while the data source and processing are computational. The primary strength of these data (Google Trends, Wikipedia Page Views, Twitter, news media) is that they provided large N data with high temporal resolution (up to the day), which allowed me to correlate discursive change with protest events. Doing the same with survey data would require constant monitoring of a panel, which is infeasible. Beyond this, the particular variables I tracked reflected concepts related to a BLM agenda. The major

[1]This study was previously published as "Black Lives Matter protests shift public discourse" in PNAS on March 3rd 2022. It was developed in collaboration with Harry Yan, Jelani Ince, and Fabio Rojas.

[2]I owe great gratitude to Fabio, who recruited me to a second iteration the BLM team, which is now in its fourth year. Thanks, Jelani, for fighting alongside me to present a more radical framing of BLM despite its "unscientific" nature. Thank you, Harry, for your diligent counsel on every analysis and figure. I am massively indebted to Michael Schultz for helping devise the regression model that took this paper from an intuition to an analysis. And similarly to Maria Pope for pointing me to co-activation. Thanks also to the Racial Justice Research fund at Indiana University for sponsoring this work. The following scholars contributed support and critical commentary: Victor Ray, Pamela Oliver, Eric Grodsky, James Shanahan, Kenzie Givens, Bradi Heaberlin, Neal Caren, Rob Potter, Marit Rehavi, David S. Meyer, Clem Brooks, and Brian Powell.

finding of this paper is that by 2020 our collective response to a highly publicized murder of a Black man by police included concepts so far removed from racialized policing as White supremacy, systemic racism, and redlining. It is difficult to imagine choosing to track such variables in conjunction with BLM events in 2014.

Here, I employed two methods of computational analysis for this study. One is the generalized additive regression model (GAM), one of two typical methods of fitting non-parametric functions in a regression model, the other being Bayesian models. The other is coactivation, a method of multi-channel signal processing developed for use in EEG and fMRI studies. While the coactivation better aligns with the inter- and cross-disciplinary approach of computational social science, I found that regression modeling not only produced more intuitive results, but was also better equipped for fine-grained hypothesis testing. This is not a critique of such non-traditional methods in social science, but rather another case where regression proved to be not only adequate, but well-suited to analysis of "computational" data. In fact, this should not come as a surprise, as much machine learning is in fact linear regression optimized for large N and large K data sets.

While the analysis is a hybrid of traditional and computaional approaches, the data are squarely in the tradition of computational social science. The Google Trends data required not only the use of an API, but also some creative problem solving to circumvent restrictions on the data imposed by Google. Namely, Google does give access to high resolution (daily) query volume in periods larger than 8 months and all volumes are relative to a local maximum. This means that measurements from one period are not commensurate with another, and it is difficult to get relative volumes across separate queries. The solution to this was to collect overlapping 8-month chunks of data employ period-over-period change as the variable of interest for the analyses. Change is always accurate within the original window, and ratios retain their meaning outside the window. A second challenge using this data was the irregular size of time windows for protest periods. Most protests occur over the span of about two weeks, and the main analysis required that all time be binned into similar chunks. Though the results were robust to this binning, I bootstrapped these regression models to alleviate this concern.

A second lesson to be learned regards the meaning of the data sources themselves. I analyzed Google Trends, Wikipedia Page Views, Twitter, and news media to track the zeitgeist around BLM. Each of these data sources has strengths and weaknesses and captures different aspects of the discourse. The relatively small volume of news media data make analysis fairly tractable and they are informationally rich. At the

time, historical Twitter data were fairly accessible through the API, but are no longer. Twitter became the de facto public data source for massive observational studies from everything from mental health to political polarization. Its strength is that it was very large, fairly accessible, and moderately information rich (limited to 140, later 280, characters). But Twitter data are weak in that they capture just a subset of the population, and are fundamentally undemocratic: a small number of users produce an outsized proportion of the content. Google Trends by contrast are highly democratic. Everyone uses Google, and typically one does not Google the same query many times, as is the case for political messaging on Twitter. For the purposes of this study, tracking discursive change, Google has the added strength of capturing newcomers to an idea: typically we search for a concept we are unfamiliar with. Wikipedia Page Views have a similar character and represent a step beyond search, and are often the first place we go to learn more about something in the world we are unfamiliar with.

3.1.1 Background

Before social change, there is discussion of social change. For this reason, social scientists who study protest outcomes focus on agenda setting. Once protests and direct action draw attention to a movement's goals, allies in the media, government, and the private sector may introduce policies that institutionalize the movement's objectives. It is for this reason that social movement scholars in sociology, political science, and mass communication research have investigated the link between contentious political behavior and public discourse.

A rich literature documents the different ways that protests generate attention for political issues. McAdam and Su (2002) show that increased anti-Vietnam War protest resulted in more Congressional hearings and Guillion's multiple studies of the Civil Rights movement shows protest for Black rights was associated with more discussion of voting and housing rights by the White House, which was followed by a wide range of administrative policies and legislative efforts (Gillion, 2013, 2016; Noel, 2014). The political scientist Hans Noel has argued that anti-alcohol activists in the late 19th century were able to project their message into newspapers, which set the stage for the passage of the 18th amendment in 1919 (Noel, 2014).

Even though scholars have long understood that protest can lead to changes in public discourse and political agendas, the rise of Black Lives Matter (BLM), and the appearance of antiracist culture in the 2010s raises new questions about the link between activism and discourse change (Gillion, 2013; Lee, 2002; Weaver, 2008; Giugni, 2007; Mann, 1993). How does protest translate into behavioral changes in online

platforms such as Google, Twitter, and Wikipedia?

This study addresses a series of questions about how street protest is followed by large shifts in public attention as captured by digital platforms. By doing so, I contribute to an ongoing scholarly analysis of the BLM movement. Early studies examined the emergence of the movement and summarized its history (Boyles, 2019; Kudesia, 2021). In subsequent years, research on BLM has documented patterns of protest, identified the social contexts that trigger protests such as police shootings, and whether protests are associated with Black political institutions such as NAACP offices or Black mayors (Williamson et al., 2018). Recently, BLM research has measured the movement's policy and electoral impacts. Prior studies have shown an association between BLM protest and Democratic vote shares in the 2020 election (Klein Teeselink and Melios, 2021), as well as the reduction of police shooting fatalities (Campbell, 2024).

I expand this assessment with an analysis of how BLM protests lead to increased use of antiracist vocabulary on multiple digital platforms. Prior large-scale quantitative research on the movement's cultural impacts are scant. A handful of earlier studies examined the use of hashtags related to BLM on Twitter, while others claimed that BLM reduced bias in society with data from volunteers who took an implicit bias test online (Sawyer and Gampa, 2018; Ince et al., 2017; Brown et al., 2017; Wilkins et al., 2019).

By establishing a link between BLM's political rallies and increased use of antiracist terminology, I show how political movements change society beyond the political sphere. Scholars have shown that protest has non-political impacts such as changing school curricula (Rojas, 2010), encouraging "ethical consumerism" (Bartley et al., 2015; Summers, 2016), and suppressing the stock prices of firms that employ unethical labor practices (King and Soule, 2007). Similarly, BLM aims to change American society by encouraging people to use terms such as "systemic racism," "White supremacy," and "mass incarceration," which are drawn from antiracist theory. This theory argues that social institutions, such as the criminal justice system, reproduce inequality by penalizing social behaviors associated with minority groups and creating differences that encourage society at large to see racial disparities as natural and inevitable (Bonilla-Silva, 2006; Feagin, 2013). Antiracism discourse is distinctive in that it does not view racism as an individual pathology or dysfunction.

This analysis contributes to a larger discussion within the social sciences about the relationship between political action and cultural change. This literature depicts a complex and multi-directional process. Political actions, such as protest, draw attention to issues and shift agendas, while changes in the way that people frame political issues also enable political action. Prior research, for example, has examined how particular

forms of media, such as film, are associated with increased protest. Vasi et al.'s 2015 study showed how the screening of environmental films is associated with short term increases in local anti-fracking protest (Vasi et al., 2015). Other research shows how social movement frames can be adopted by political groups in successful campaigns for policy change. Using state level data on political campaigns, McCammon et al's study of women's jury activists showed that the way that legal activists framed women's participation in juries is associated with new legislation (McCammon et al., 2007). Similarly, Black Lives Matter relies on frames that already exist. Activists describe their movement as a continuation of earlier Black freedom struggles (Clark et al., 2018). The present analysis focuses on what happens after movements emerge and they attempt to bring their ideas to the public through widely accessible digital platforms. I add to this longstanding scholarly discussion with a large scale computational study of the association between street-level protest and discursive change.

3.1.2 Research questions and hypotheses

The primary purpose of this paper is to show that political actions, such as BLM protests, can trigger sustained attention to antiracist ideas. To document this effect, i tracked the use of antiracist terms in four different publicly available datasets. I then correlated vocabulary use data with public data on recent BLM protests that were organized in response to police violence. I use these data to answer the following research questions: 1) Is there an increase in the use of antiracist vocabulary on digital platforms in the period after the start of the BLM movement that is substantially higher than before? 2) How is BLM protest associated with increased attention paid to the movement as measured by the volume of Google searches for the phrase "Black Lives Matter?" 3) When protest increases attention to the phrase "Black Lives Matter," does it also increase attention to other terms associated with antiracist politics? 4) How does attention to the phrase "Black Lives Matter" and related terms change over multiple waves of protest? 5) Which antiracist terms experienced sustained attention in the period after the George Floyd protests? 6) Did increased and sustained attention given to Black Lives Matter "spill over" to other related terms? 7) How is Black Lives Matter protest associated with the use of terms representing the movement's opposition?

3.1.3 Data sources used in this study

This study aims to empirically measure and assess how BLM protest may be associated with increased attention given to antiracist terms on digital platforms. However, there is no single way to measure how issues are discussed at the national level. Therefore, I have opted for a research strategy where I use multiple

publicly available data sources, each of which has strengths and weaknesses. Throughout this study I largely rely on Google search volume as a measure of public attention. Google search represents popular, rather than elite, attention that is a response to encountering ideas that emerge from a wide range of contexts such as news, face-to-face conversation, or social media. Furthermore, people often use Google when they want to learn about an issue for the first time. I also use Twitter mentions, Wikipedia page visits, and national news media mentions as indicators for attention to Black Lives Matter. Twitter is a social media platform that is accessible to researchers and that has become a focal point for online political discussions. As a widely used reference, people use Wikipedia to understand current events and issues of ongoing concern, which allows me to assess typical, or baseline, levels of attention given to certain issues. News items, as measured with data from the Media Cloud API (Roberts et al., 2021), reflect both daily "news cycles" as well as articles prepared over a longer period. News media data measures elite discourse, as the ideas appearing in news sources reflect what editors and journalists find newsworthy. One analysis briefly uses data from Google books to assess long term trends in the use of antiracist terms by book authors.

3.1.4 Terms representing antiracist discourse

To answer questions about the heterogeneous effects of protest, I analyze Google search and Wikipedia page visit data for a wide range of antiracist terms. Using statements made by activists, sociological theory, and the direct observation of BLM events, I created a list of terms (given by Table A.1) that capture the main components of antiracist discourse such as its mottos, key figures, policy issues, and comparison with prior movements. While I recognize that no list of terms can completely capture every element of a new vocabulary, these terms capture the major themes associated with BLM and antiracism more generally.

Contemporary social movement theory suggests that the selection of terms should reflect how a movement frames issues because framing is one of the core tasks performed by activists. Social movement researchers have argued that framing consists of three elements: 1) diagnosing social problems; 2) suggestions of which movement tactics are appropriate, and 3) a prognosis describing the possible outcomes of collective action and which policies are to be effected (Benford and Snow, 2000).

BLM's diagnosis of a major problem in American society was formulated in response to the killing of African-Americans at the hands of police. The initial diagnosis of this problem that BLM proffered was that *victims* of police homicide, such as Michael Brown, Tamir Rice, and Sandra Bland, symbolized a larger pattern of systemic racism. For this reason, BLM activists often evoke their names and I used them in my

analysis. The tactics were denoted by the popular *slogans*, including the name of the movement. The initial remedy that the movement seeks is directly related to *policy* changes to address the injustice of policing and carceral systems such as "defund the police."

Early interviews with BLM organizers showed the tactic of reframing the movement as the "new-wave" of the Civil Rights movement and the intention to broaden the movement's scope by referencing systemic racism and *White supremacy* (Clark et al., 2018). Therefore, I also paid attention to *historical contexts and figures*, and the contemporary discussion of *identity politics*. The comparison between terms used by BLM and keywords associated with earlier Civil Rights discourse illuminates how today's activism has expanded and redefined popular discourse on race.

3.2 Results

3.2.1 The Black Lives Matter movement coincides with increased attention given to Black emancipatory rhetoric

To set the stage for answering questions about BLM protest, I show a dramatic increase in use of terms associated with BLM and antiracist theory in the 2010s. Figure 3.1 describes the shift in use of terms associated with Black Lives Matter, from multiple starting points until 2020. Figure 3.1 compares the rise of antiracist terminology with other forms of discourse about racial inequality such as the earlier Civil Rights Movement. To assess the robustness of the trends, I track these terms in three databases: Google's digital book archive, Google searches, and news media.

To facilitate comparison, Figure 3.1 starts with data concerning the use of selected terms in the Google books database. This comparison is helpful to two reasons. First, book publishing is less susceptible to short term fluctuations that affect social media and is therefore an indicator of long term trends. Second, Google books data underscores the point that people were already changing the way they discussed race before the rise of BLM. The data indicate that the policy issues motivating BLM, such as mass incarceration, began to attract much higher levels of attention as early as 2005. Thus, protest is a process that abruptly elevates certain terms in an environment that is becoming relatively favorable to the movement.

Terms associated with previous forms of racial discourse show either stability, such as "multiculturalism" within national news media, or modest decreases or increases in usage such as "desegregation" as used in multiple platforms. In contrast, antiracist vocabulary terms show a consistent super-linear trend, where post-2013 usages dramatically increase. Antiracist terms demonstrate a relatively low level of use and then

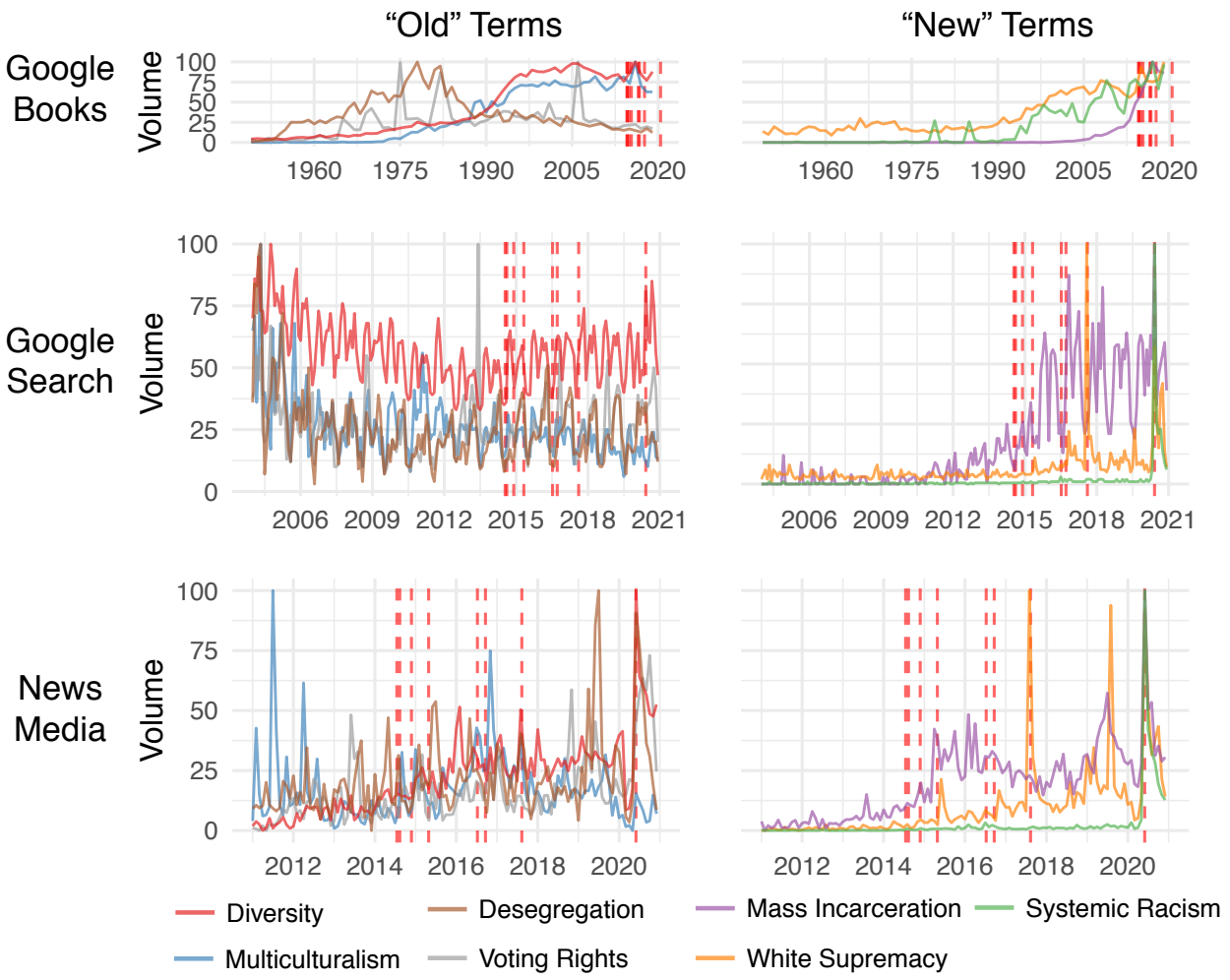

Figure 3.1: Trends in the use of selected antiracist terms. Normalized monthly volume for pre-BLM anti-racist terms (left) and post-BLM anti-racist terms. Top row: Google N-grams instances of each term (annual). Middle row: Google searches for each term. Bottom row: National news articles mentioning each term. Red dashed lines correspond to major BLM protest event.

a large upturn in the BLM era. These data suggest an affirmative answer to Question 1. BLM protest is followed by a distinct increase in usage of antiracist terms.

3.2.2 Black Lives Matter protest tracks with Google searches for "Black Lives Matter"

Here I answer Question 2 directly by examining the link between protest and antiracist terms. Figure 3.2 shows the association between BLM protests and Google searches for the term "Black Lives Matter." The top panel overlays three time series: a count of the estimated number of participants in BLM protests in the United States summed into 8–12 day bins, the volume of Google searches for "Black Lives Matter" with the same binning, and a series representing sums across 8–12 day bins of Google Search terms. The search

volume data is scaled 0–100, as Google does not provide absolute search volume data. Data are binned in order to smooth the noise and capture "protest periods" over which protest builds from events in one or several cities to a contagious event across the country. The choice of bin size is further explained in the next subsection of the results.

The middle panel shows that abrupt increases, or "spikes," in searches for "Black Lives Matter" often co-occur with protest events, which are denoted with dashed vertical lines. This second panel takes the data presented in the first panel and simplifies it into a single time series denoted the post-over-prior-period ratio ($\ln(V_t/V_{t-1})$) in search volume, V. This second panel shows that protests are often followed by attention that is often an order of magnitude or greater than search volume in the preceding period.

Question 3 asks if the link between protest and lexical change is limited to "Black Lives Matter" or whether protest also boosts other components of the antiracist lexicon. The bottom row of Figure 3.2 expands the analysis and answers this question. I look at the impact of BLM protest on the use of 41 related antiracist terms. Table A.1 lists these terms, which includes names of police shooting victims, slogans, and policy issues. The bottom two panels of Figure 3.2 show the distribution of relative change for all 41 Google Search terms combined has a higher mean and a larger right hand tail during time periods with protest. The fourth panel, which focuses on the right hand side of the distribution, shows that days with very "spiky" shifts in attention are much more likely to occur on days with major protests, such as protests associated with the murder of George Floyd and Philando Castile. This suggests that the answer to Question 3 is yes — protest is linked with attention increases to many terms associated with the antiracism movement.

3.2.3 Google search terms related to BLM co-activate

Figure 3.3 provides further evidence that BLM protest is not only generating attention for the term "Black Lives Matter," but a wider spectrum of antiracist discourse. To demonstrate the concurrent movement of terms, I plotted the co-activation of these words, which I define as the degree to which a bundle of terms experiences a simultaneous and abrupt shift in attention represented in Google searches. Co-activation captures the idea that multiple measures simultaneously, and abruptly, increase. In contrast, it may be possible that the increase in vocabulary use in Figure 3.2 reflects only handful of terms.

Following Liu and Duyn (2013), I define co-activation (C) of a group of terms (G) to be the square root of the sum of the pairwise products of daily search term volumes ($v_i v_j$): $C = \sqrt{\sum_{i,j \in G} v_i v_j}$. Co-activation defined in this way is a measure of synchrony designed for "spiky" time series, such as electrical impulse

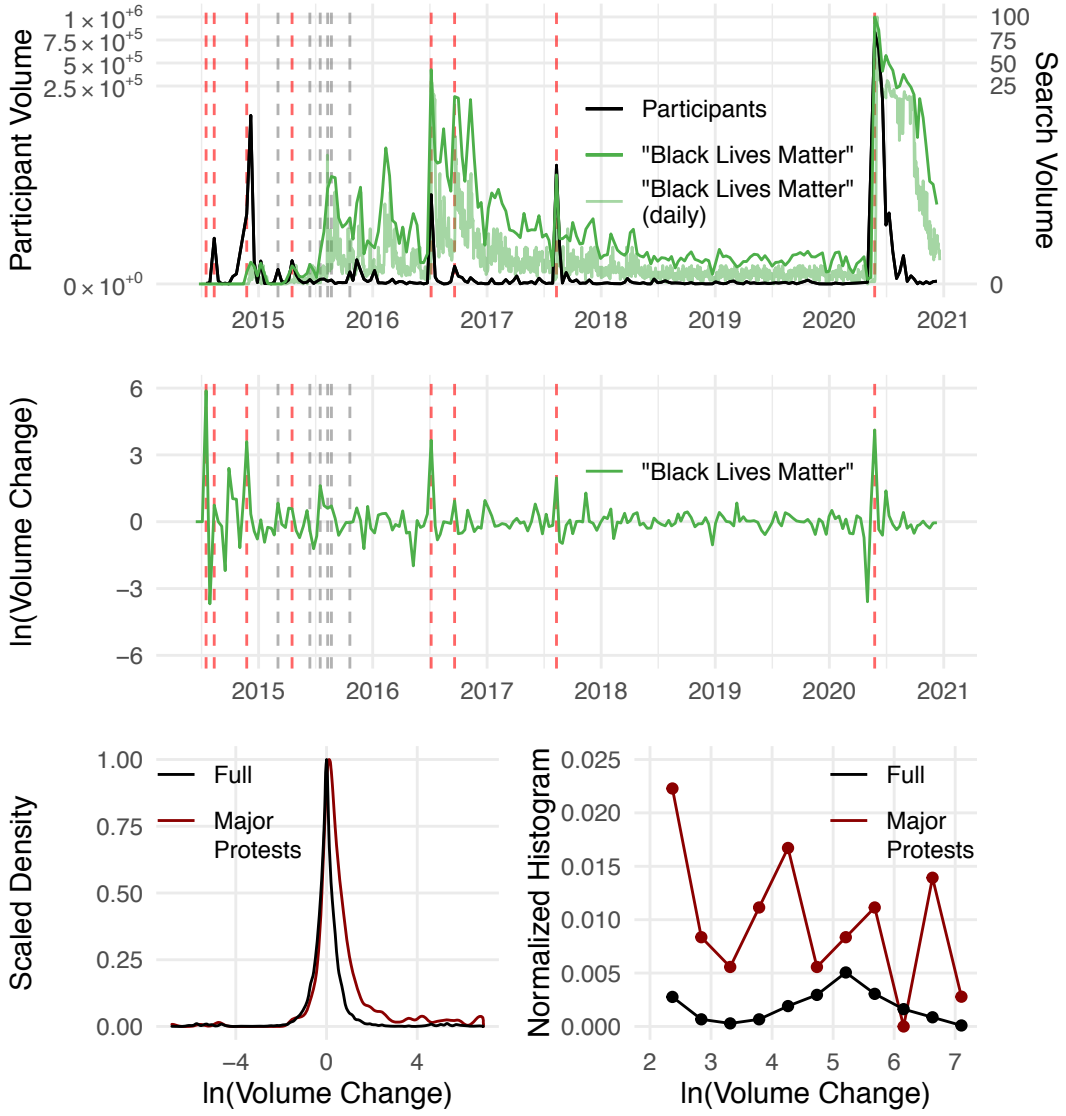

Figure 3.2: Protests and Google searches are bursty and correlated. Top: Number of participants at BLM-related protests and volume of Google searches for "Black Lives Matter" (scaled 0–100, as Google Trends does not provide absolute search volumes.) Middle: $\ln(V_t/V_{t-1})$ change in volume of Google searches, V. Unless otherwise stated, time has been binned such that each point t spans a period of 8 to 12 days designed to fully encapsulate, rather than straddle protest periods, which occur irregularly. Bottom Left: Scaled distribution of relative volume of Google searches. Relative volume for all 41 BLM-related search terms during all periods are shown in black, whereas the dark red line shows only protest periods. Bottom Right: The right tail of a histogram of the data in the bottom left. In the top and middle plots, red and grey dashed lines correspond to major and minor protest events respectively (minor protest events for only 2015 are pictured.) In the top plot the upper 3/4$^{\text{ths}}$ of the vertical axis have been contracted to make visible the variation in the lower 4$^{\text{th}}$.

data as found in electroencephalography (EEG) research.

I compute the co-activation of the 41 terms listed in Table A.1. To answer Question 3 and establish that these terms are unusually synchronized on days with BLM protest, one needs to show that co-activation is higher than would be anticipated given the normal fluctuations observed in online search behavior. To estimate whether co-activation on a particular day is unexpected, I fit a normal distribution for a moving 60-day window and compare the daily values to that distribution using a ratio of observed to expected, significance-tested by a z-score. For legibility, Figure 3.3 shows the log-ratio of the co-activation of terms within a 60-day window of time.

Since weekdays tend to have greater activity than weekends, I mask the 60-day window such that weekdays are compared to only weekdays, and weekends to only weekends. Thus, the number of observations in the 60-day window is much smaller for weekends, but the observations still span the same period. Figure 3.3 shows the ratio of daily co-activation to mean co-activation of the prior 60 days for all pairs of terms from 2014 to 2020. As search engines are often used in educational contexts, there will be natural fluctuations associated with school calendars such as weekends and summer/winter breaks.

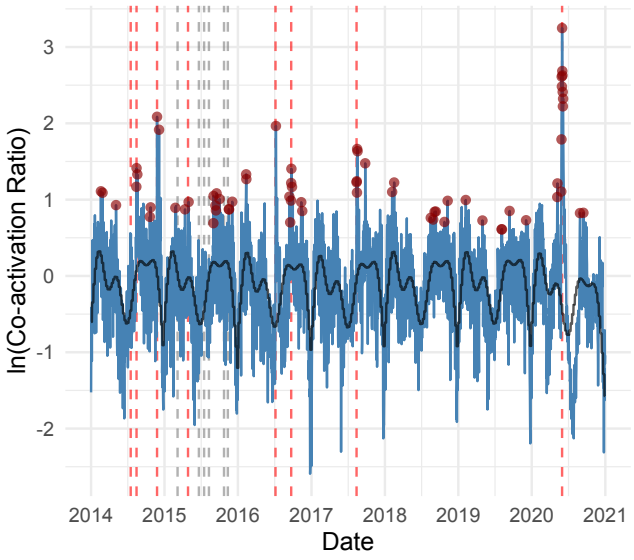

Figure 3.3: Co-activation of Google search terms. Blue lines show the log-ratio of daily co-activation to mean co-activation (60-day window) for all 41 terms. Black line shows GAM prediction from calendar week and absolute time. Red points show significant spikes in co-activation (P<0.001). Major and minor protest events are indicated by red and grey dashed lines respectively.

To account for these temporal variations, I also present the results of a generalized additive model (GAM) that estimates the effect of time on the co-activation of the selected antiracist terms. The model has

two components. The first component comprises non-parametric functions of time, which accounts for daily variations in attention due to academic calendars and other factors. The other component includes multiple linear effects for periods of time with high protest levels, such as June 2020.

The black line in Figure 3.3 visualizes predicted co-activation and shows that the abrupt increases in attention can't be attributed solely to the regular variation in daily use of search engines. The full results of this model are given in Table A.3 in the Supporting Information.

The four largest spikes in collective attention given to antiracism reflect major protest events. The first spike occurs in late 2014, following the hearings for the killers of Michael Brown and Eric Garner, which did not result in indictments, and the killing of Tamir Rice. The second spike occurs in July 2016, which are protests following the killings of Alton Sterling and Philando Castile. The third spike appears in August 2017, corresponding to counter-protests following the Unite the Right rally in Charlottesville. The largest spike is in 2020, which is related to the murder of George Floyd.

The model also allows us to understand the nuanced temporal structure of attention given to the antiracist lexicon. The data show that co-activation increases with the start of the academic term in August, declines sharply during school breaks and holidays, such as winter recess, peaks once again during February, which is Black History Month, and declines sharply during the summer recess. This suggests that some of the variation in attention given to antiracism reflects both routine and modest seasonal variation due to school assignments and the drastic increases seen during demonstrations. This pattern is consistent with prior research on the use of the Google search engine and Wikipedia which finds that school assignments are one of the most common uses for these digital reference tools (Singer et al., 2017). Future research can attempt to quantify how of the typical day-to-day attention given to these terms is related to scholastic activities.

3.2.4 Spikes in attention to terms related to the BLM agenda get larger and more diverse over time

I have established that terms associated with Black Lives Matter spike during protests. Movements, however, are dynamic and the impact of protests may vary over time. To answer Question 4 about the time varying effects of protest, I subdivide the search terms into thematic categories. By examining these categories separately, I can understand the discursive evolution of the antiracism movement while taking into account the unique role that particular protests have in creating a platform for selected subsets of ideas.

Search terms have been grouped into 8 categories representing different components of the BLM agenda (see Table A.1.) Each category comprises 2-8 terms. The next analysis disaggregates the co-activation

analysis by protest event and type of term. Figure 4 shows the way that protest built up and cemented the visibility of the movement itself. Using a generalized additive model, I estimated attention shifts during four protest waves: the initial protest wave subsequent to the deaths of Eric Garner, Michael Brown, Tamir Rice and Freddie Gray; the protests following the deaths of Alton Sterling and Philando Castile and the first wave of National Anthem protests; the protests in response the Unite the Right rally in Charlottesville, Virginia; and the demonstrations in response to George Floyd's murder. The baseline category consists of all other time periods outside of these protest waves, including less prominent protests. Then, I looked at the increase of attention to selected categories in these four time periods.

Here I address a number of modeling issues. First, the spread of protests over a multi-day period raises questions about selecting a unit of time during which attention increases may occur because protest waves occur over multiple days and, in some cases, weeks. Typically these begin with single city protest events organized in response to a catalyzing event such as the shooting of a Black civilian by police. As these initial protests grow in attention, they spread to other cities, and attendance grows in an exponential manner before enthusiasm peaks. As shown in Figures 3.2 and 3.3, the subsequent shifts in attention are "bursty" in that they are abrupt and clustered in time. I did not model the data with averages of moving time windows because that would artificially decrease the variance of theoretically important events: abrupt attention increases would be averaged out with many days of very little attention (days before the occurrence of the catalytic event.) Similarly, there are problems with using standardized units of time such as weeks or months because protest waves often cross over multiple days and weeks and they do so in non-uniform ways. Furthermore, protests began at unpredictable times, which suggests that uniform time bins cannot capture complete protest cycles.

As such, I developed an approach to binning that captures full protest cycles. For each major protest, beginning with the date of the initial event, I measure how many days it takes for 80% of terms to reach their maximum value in the 2 weeks following the event. If any term reaches its maximum value in an 8-month period (defined by Google's maximum window for daily search data) outside the initial 80% marker, that day becomes the end point for the protest. The duration and catalyzing events for the protests are given by Table A.2 in the Supporting Information. To produce commensurate periods for dates without protest, I bin the interceding days into periods of duration sampled from the protest periods. In order to limit the effect of random binning on the interceding dates, I bootstrap the binning 500 times. GAM results are reported for coefficients averaged over these 500 data sets in Figure 3.4. The GAM models predict search volume

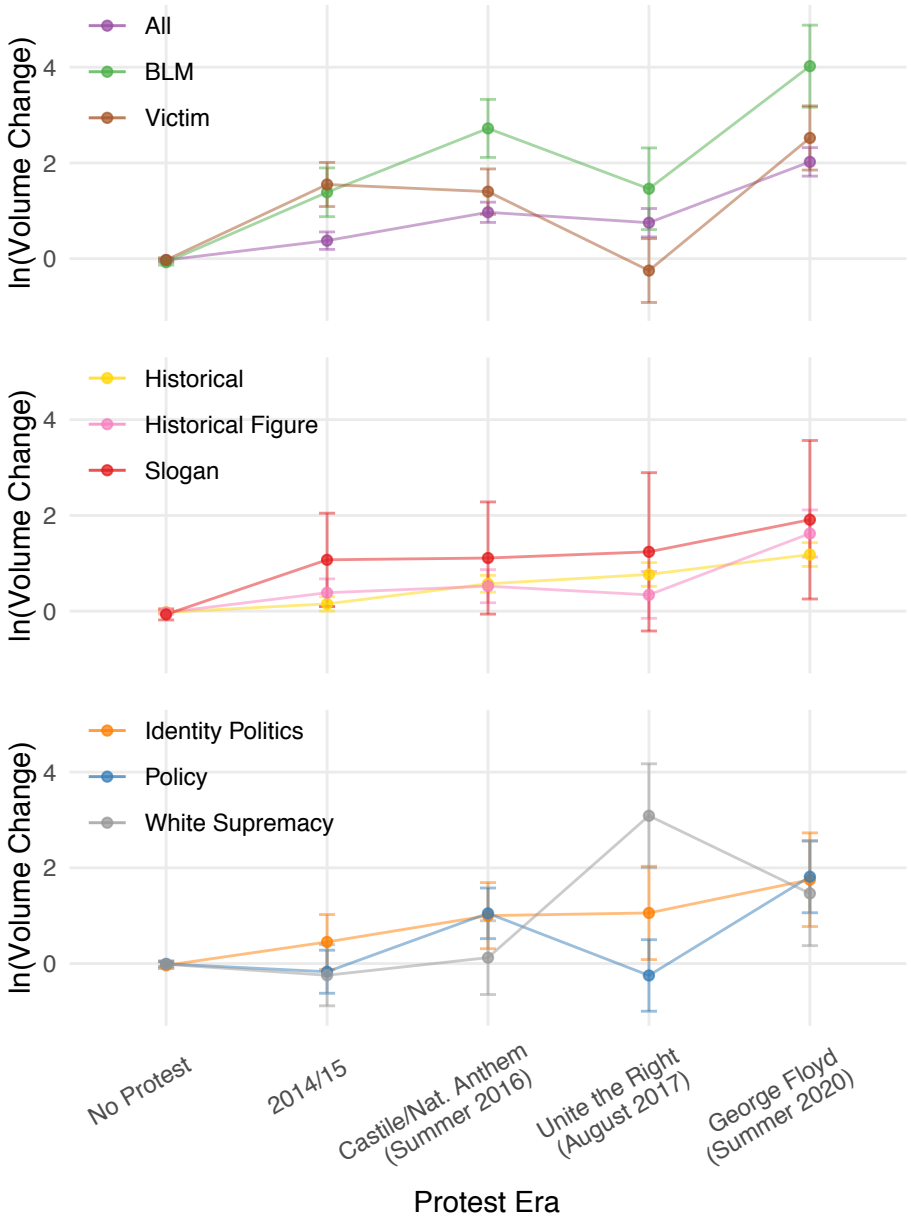

Figure 3.4: Expected change in search volume during protest. Generalized additive models (GAM) of expected change in search volume during a major protest during 4 protest "eras." Search terms are grouped into 8 categories, with an additional group, "All," representing every search term. The first point in every plot represents model intercept: the expected change in search volume when a major protest is not occurring at any time between July 2014 and December 2020. All other points are the bootstrapped coefficients for expected change in search volume of protests of a particular era. Whiskers show 95% confidence intervals.

increases during Black Lives Matter protest periods. I fit a model for each of the 8 thematic categories, as well as a model for all terms. Each of these models has three components related to time (absolute time, calendar year, and one time step lag of the dependent variable.) The averaged output of these bootstrapped models are given by Table A.4 in the Supporting Information.

Generally, all of the models represented by Figure 3.4 demonstrate a super-linear increase in expected volume during protests. My models indicate that more terms are spiking and that the spikes are growing larger. Furthermore, we can gain a better understanding of *which* protests are responsible for elevating particular types of ideas. During the 2014/2015 era, protest boosts all categories except "Policy" and "White Supremacy." During the 2016 protests, some "Policy" terms are expected to spike, but "White Supremacy" terms show wide error bars and there is no statistically significant effect of protest for related terms.

In 2017, however, BLM protests following the Charlottesville Unite the Right event show a somewhat different pattern. Most notably, searches for terms associated with "White Supremacy" spike dramatically. Searches for the "Black Lives Matter" category decrease because the discourse is not focused around "Police Shootings" or police homicide more generally. Similarly, searches for past lynching "Victims" do not spike. Searches for terms relating to the "History" of Black struggle, such as "Jim Crow" laws, do spike, but searches for "Historical Figures" do not, nor do terms for "Policy." "Identity Politics" terms do spike, however the 95% confidence interval is wide and comes close to 0, suggesting that some terms within the category spike and others don't. Finally the George Floyd protests show the greatest predicted change in volume for all categories except "White Supremacy."

3.2.5 Interest in BLM is sustained after the George Floyd protests

We have established that BLM-related searches spike dramatically during protests. Is attention to the BLM agenda sustained beyond protest periods? This section addresses Questions 5 and 6 — does protest trigger sustained attention shifts for antiracism terms and does this "spill over" into other terms? Figure 3.5 answers this question by demonstrating protracted interest in BLM following the George Floyd protests. Rather than focus on differences between protest waves, I examine one time period, August to December 2020, and ask which terms experienced simultaneous increases in total attention and increases in attention relative to the baseline for the same period in the previous year. This analysis uses Wikipedia search data because that platform provides raw counts of search terms, which allows me to standardize the attention measurements and directly compare terms. The X-axis indicates the logarithm of the ratio of Wikipedia searches in the

post- and pre-George Floyd protest eras. The Y-axis shows the raw frequency of searches for each term. This allows for simultaneously visualizing popularity and the growth of a term after protest. Figure 3.5 shows a wide range of terms associated with BLM and antiracism more generally, all of which have analogues in the Google search terms used earlier in the study.

I test for sustained interest in Black Lives Matter following the Spring 2020 George Floyd protests. Using Wikimedia's REST API, I collect daily page visits for 35 pages corresponding to the list of antiracist terms and terms associated with Civil Rights. I compare search behavior in two time periods: August to December 2020 and 2019. I chose August to December because there will be a natural decrease in attention in the immediate aftermath of protest. We also want to focus on the period of time when students are likely to be searching for terms related to antiracism.

In order to prevent attention surges from biasing estimates away from their baseline, I apply a locally weighted scatterplot smoothing (LOWESS) filter with a 60-day window to the daily volume time series. Further, I scale by average daily English Wikipedia visits for each period to account for increased Wikipedia use over time. Smoothing and scaling have slight effects on the results, whereas accounting for seasonality is critical to the accuracy of estimates.

I report post- vs. pre-George Floyd ratios in daily traffic and use a t-test to determine whether average daily visits increased following the protests in a statistically significant way. Additionally, in Table A.6 (see Supporting Information) I report the change in daily volume for news media mentions, Google searches, and Twitter mentions of a set analogous terms using the same method. I find that many pages related to BLM have greater daily traffic post-George Floyd protests of 2020.

The three terms that experienced the most growth, relative to their pre-George Floyd baseline, are "Black Lives Matter," "Prison abolition," and "Systemic racism." The terms that received the most page visits are "Black Lives Matter," "Jim Crow Laws," and "Martin Luther King, Jr." It is notable that "Martin Luther King, Jr." remained popular but its post-protest increase is very small. This is also true for the page "Slavery." In contrast, "Black Lives Matter" and "Jim Crow laws" saw very large increases compared to the baseline. Similarly "Prison abolition movement" and "Prison abolition" show increased attention, whereas the page "Prison reform" had fewer visits in 2020 than in 2019.

It is worth examining terms where attention did not increase post-George Floyd. These include a number of policy issues that predate the Black Lives Matter movement, such as decriminalization, prison reform, and housing segregation. An exception is "Redlining," which refers to governmental and non-governmental

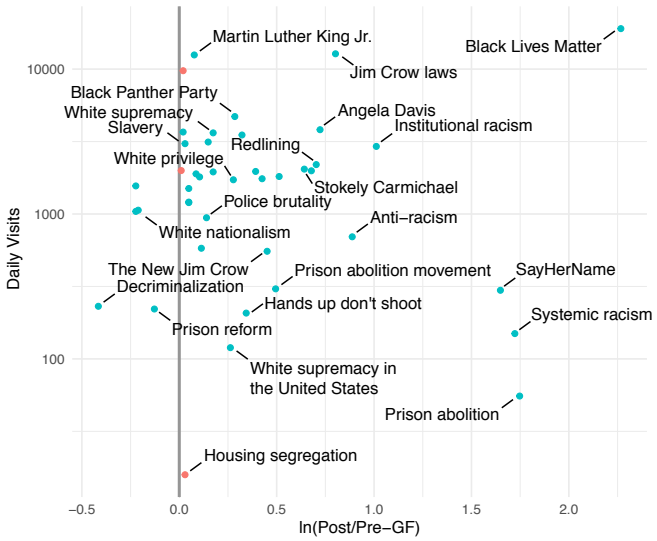

Figure 3.5: Shift in frequency of Wikipedia page visits after the George Floyd protests. Horizontal axis shows the ratio of expected daily visits during August to December of 2020 compared to the same calendar period of 2019. Vertical axis shows the expected number of page visits in August to December of 2020 (log scale). Blue dots are significant at P<0.001 as estimated by t-test.

practices relating to housing segregation. This term shows a moderate increase in page visits. The data also show evidence for the hypothesis that BLM protest may have moderately stigmatized White nationalism. Terms that critique racial hierarchy such as"White privilege" and "White supremacy" show small increases in attention. By contrast, the page for "White nationalism" shows a decrease in daily visits. All other data sources show larger decreases in interest in "White nationalism" (46-76% reduction, see Table A.6.)

These modest changes also indicate that engagement with antiracism during and since the George Floyd protests has not focused on White supremacy generally, but rather on specific facets of White supremacy and Black liberation. Pages for radical and relatively lesser known 1960s civil rights leaders Angela Davis and Stokely Carmichael had large increases in traffic, whereas Martin Luther King Jr.'s page shows only a slight increase. One very important historical term, "Slavery," does not have a large increase post-protest. This suggests that protests are increasing the visibility of terms that associated with BLM's radical vision.

3.2.6 Opponents of BLM generate less attention than BLM

Protest can trigger counter-protesters who dispute a movement's ideas or policy proposals. For this reason, it is important to examine negative discourse associated with Black Lives Matter and antiracism more generally. Here, I look at Twitter activity to measure anti-BLM backlash and compare the attention given to its opponents. I track #BlackLivesMatter and three major counter-movement hashtags (#AllLivesMatter,

#BlueLivesMatter, #WhiteLivesMatter) across time. Figure 3.6 shows the prevalence of these hashtags from the origin of #BlackLivesMatter following the acquittal of the killer of Trayvon Martin, George Zimmerman, on July 13, 2013, through 2020. These data allows me to address Question 7 and gain an understanding of how protests generate counter-movements.

The data show that counter-movement discourse happens in tandem with the use of the Black Lives Matter hashtag, which itself is prompted by street protest associated with police homicide. All three counter-movement hashtags emerge or mature during the second half of 2014, parallel to Black Lives Matter's emergence as a protest movement. #BlackLivesMatter consistently has greater volume, an order of magnitude or more. There are two major baseline shifts apart from the initial emergence of Black Lives Matter. The first begins with the homicide of Eric Garner and ends with the decisions by grand juries in late 2014 to not indict police officers Darren Wilson or Daniel Panteleo. A second major shift occurs after the George Floyd protests.

While #BlackLivesMatter and its opponents, #All- and #BlueLivesMatter, show fairly similar patterns, #WhiteLivesMatter, which is more overtly White supremacist, displays distinctive patterns. The use of #WhiteLivesMatter notably declines after the Unite the Right rally, whereas the others hold constant ($P<0.001$, see Table A.5). This may have happened because Twitter aggressively banned White nationalists in the 6 months following the rally. Spikes in #WhiteLivesMatter in the years between Unite the Right and George Floyd, such as the Shelbyville, TN "White Lives Matter" rally on October 29, 2017, did not shift the baseline. This hashtag enjoyed a much larger increase in use than the others during the George Floyd protests, though it tapers with a slightly sharper slope than #BlackLivesMatter ($P<0.001$, see Table A.5). I find that the #WhiteLivesMatter tweet volume was 2.22 orders of magnitude greater after the killing of

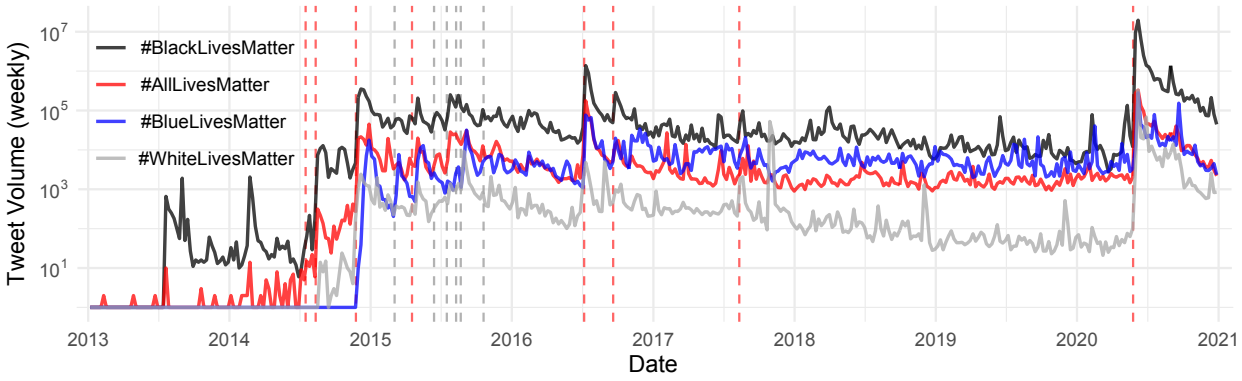

Figure 3.6: Weekly hashtag volume for #BlackLivesMatter and three opposing terms. Major and minor protest events are indicated by red and grey dashed lines respectively.

George Floyd (P<0.001). However, by the end of 2020, its volume was still more than half an order of magnitude less than #All- or #BlueLivesMatter (<1,000 tweets per day), which are themselves dwarfed by the volume of tweets containing #BlackLivesMatter (>10,000 tweets per day).

This analysis illustrates the importance of using different data sources. Google search data for "Black Lives Matter" (see Figure 3.2) shows a longer and more jagged, or episodic, emergence of public awareness of BLM. This is likely because Twitter may attract users who are more reactive to current events than the general public. This may also reflect underlying differences in how people use these digital platforms. On Twitter, one person may make dozens, even hundreds of tweets with a particular hashtag, but a single individual is unlikely to make more than a handful of Google queries. In this sense, the more protracted emergence of BLM on Google search, as opposed to Twitter, reflects its more egalitarian character.

These data on the use of counter-movement hashtags raises additional questions that can be answered with future research. Numerous scholars have noted that counter-movements often intensify when movements use counter-productive tactics, such as riots (Wasow, 2020), or they are perceived to be using such tactics in experimental settings (Ayoub et al., 2021). This raises the hypothesis that tactical choices may have a positive effect on the use of counter-movement discourse. Counter-movement actions can also complicate this dynamic. In some cases, counter-movements can be perceived to be intransigent and thus improve the standing of the original movement, as observed for LGBT+ activism (Fetner, 2008) and Occupy Wall Street activism (Milkman et al., 2013). In the BLM case, counter-movement actions, such as the Unite the Right assembly in Charlottesville, can turn violent and thus undermine their position. Future research can more thoroughly code BLM events according to tactic and investigate the relationship of tactical choice, public perceptions, and counter-movement behaviors.

3.3 Discussion

The purpose of this study is to measure and assess the cultural impact of protest. In summary, I focused on the association between BLM protests and search behaviors in platforms such as Google, Twitter, and Wikipedia. I also provided evidence from older forms of media such as books and national news organizations. I found large and consistent effects: large protests were followed by large increases in attention given to terms associated with Black Lives Matter. Furthermore, there is evidence that the attention was sustained in many cases. Antiracist discourse received much more attention after the wave of protests against George Floyd's murder and the attention given to antiracist terms remained much higher in December 2020 than in

the same period in the previous year. Taken together, this evidence indicates that BLM protests succeeded in drawing attention to antiracist theory in a large-scale and consistent way.

Here, it is important to situate these findings within scholarly discussions of social change. This study focuses on the relationship between protest and attention. Much of the analysis focuses on protest as a factor encouraging people to investigate and employ antiracist ideas. However, sociological research on the relationship between culture and political actions suggests an important contextualization of my findings. There is a recursive relationship between protest and attention: protest generates attention and attention generates protest. Thus, protest is part of a continuous cycle whereby changes in ideas facilitate political action and these actions can further cement ideas into the popular imagination.

These data refine the recursive model of culture and action. At times, protests can become a primary mover of public discourse when they are large, spontaneous reactions to catalytic events, such as the killing of a Black person by police. This is especially true of the large, multi-city contagious protest periods. It is important to emphasize why such political actions can be so important for changing how the public understands social problems. In the case of BLM, racialized police homicide, and police homicide more generally, have been a fact of American life for decades. What is new is the response of BLM organizers to consistently leverage these events to generate media attention, aided by new media technologies, such as social media and the mass availability of recording equipment for citizen journalism. This initial push by organizers sets off a cascading feedback loop of 1) dissemination of information about the initial event and earlier protests, 2) public attention, 3) new protests coordinated by BLM organizers, and 4) protest attendance by sympathizers. This cycle is enabled by prior changes in how people think about racial inequality and, at the same time, the ideas promoted by these protests may frame future discussions of race.

This research also contributes to ongoing discussions of how activists disseminate their ideas in a modern digital society. Computational social scientists have argued that movements operate in a larger "cultural environment," which includes social media and traditional news media (Bail, 2014). This line of research depicts the link between political action and cultural change in a fairly straightforward way: activists and their organizations use their resources to amplify the message they prefer in an attempt to have opinion leaders and the public adopt their view (Karpf, 2016; Van Laer and Van Aelst, 2010; Maghrabi and Salam, 2011; Tufekci, 2017). Particular organizations try to reframe issues by publicizing new policies, introducing new terms, and attracting attention to some issues over others. Often, activists will try to have highly prestigious elites or organizations promote their view in an attempt to legitimize their demands.

The pattern of BLM protest and vocabulary use indicates a different path to cultural change. BLM is a decentralized movement that does not rely on large organizations that lobby for its positions. Rather, it is a movement that has primarily mobilized followers through social media. Furthermore, the protests that have had the largest impact on the use of antiracist vocabulary are often the ones where footage of police violence was widely circulated to the public. This suggests that BLM is a movement whose cultural impact is initially established through the memorialization of specific events. Figure 3.4 is suggestive in this regard. Earlier waves of protests increased attention to victims of police violence, which then led to a rise in attention to broader ideas criticizing American society, such as White privilege and systemic racism. Thus, BLM affects American culture by using protests to focus attention first on individual victims and then drawing attention to larger policy issues.

The public's use of a new lexicon is complex. The rise of social movements often triggers dispute and "counter-movements" (Meyer and Staggenborg, 1996). Individuals who oppose a movement may try to mitigate the adoption of a movement's ideas through the introduction of their own distinct vocabulary. For example, Ince, Davis, and Rojas (2017) found that hashtag networks for BLM on Twitter frequently included "counter-movement" phrases such as "All Lives Matter." The analysis of counter-movement hashtags in the previous section deepens this point. Valence is a related issue. It is possible that movement opponents might use a term in a very different way than original intended. For example, BLM supporters may use the phrase "White supremacy" to denote patterns of prejudice, while White nationalists may use the same term as a rallying cry. I offer two observations with respect to this point. First, the data presented in the section on counter-movements suggests that such usage does occur though it is small in magnitude. The counter-BLM movement did appear after protest and, paradoxically, BLM may have provided an opportunity for a small community of White nationalists to propagate their message. Second, computational techniques such as neural network-based NLU and sentiment analysis can be used in future research to establish how much usage of antiracist terms comes from advocates and opponents.

It is also worth noting that people may have significantly different understandings of what policy reforms might be needed to implement a movement's demands. These differing policy responses reflect people's beliefs about the nature of a particular social problem such as racism. Dixon (2020) notes that people invoke "racial imaginaries," or visions of the nature of racial inequalities, when they interpret events such as the shooting of George Floyd. Some will see police violence as idiosyncratic and only a sign of "a few bad apples" while other see these events as evidence of a larger pattern of systemic racism. This

argument suggests that shifts in how people interpret social problems, such as police violence, set the stage for movements like BLM and how people use and interpret the antiracist vocabulary introduced by the movement.

3.3.1 Limitations and questions for future research

Finally, I review study limitations that can be addressed with future research. First, this study tracks aggregate use patterns. I do not collect information on the individuals who generate the data used in this study so it is not possible to link antiracist vocabulary to individual characteristics such as age, gender, socio-economic status, education, race or ideological positions. However, other researchers have explored the individual factors associated with supporting BLM or using antiracist discourse. Recent polling data suggests that approval of BLM and its policies is correlated with age, race, and political attitudes. In a recent poll of undergraduate students at a Midwestern university, Ilchi and Frank (2021) find that race and attitudes toward the police correlate with support for BLM. Updegrove et al. (2020) used data from a nationally representative sample of 2,114 people to show that older, conservative, and male respondents are less likely to support BLM. Arora and Stout (2019) use data from an experiment to show, among other things, that political party is a strong predictor of when a respondent will have a positive feeling toward a letter of support for BLM. None of these studies addresses digital behavior but they strongly suggest that the use of antiracist terminology would be strongly associated with age, political attitudes and other social and political characteristics, a hypothesis that can be addressed with future research. These studies can be brought into a dialogue with digital trace data to provide a more comprehensive and nuanced understanding of how BLM protests of the 2010s resulted in large scale cultural change.

Second, this research does not examine how antiracist words are used. For example, a person who supports Black Lives Matter is counted in the same way as a person who opposes the movement but uses their name in a social media post. Computational techniques such as sentiment analysis and neural network language models can be used to model how words are used. Third, this is a study of the consequences of the Black Lives Movement through 2020. I found evidence that antiracist vocabulary found greater usage after June 2020, but future political developments could reverse or mitigate that trend.

Third, I note that this study does not seek to establish a causal link between street protest and discursive change. Rather, this is a descriptive study showing that protest is associated with the heightened visibility of an antiracist lexicon, which allows people to create a space where such terms can be discussed. Eventually,

the amplification of antiracist discourse might facilitate political change. Subsequent research might search for naturally occurring random variations in protest in order to have the sufficient counter-factual cases needed for a causal identification.

Here I present two related questions for future research. This study does not examine the way that discursive change and agenda setting leads to the enactment of policies associated with BLM and antiracism such as ending qualified immunity for police officers, reparations, or diversity, equity, and inclusion (DEI) initiatives. Social scientists can track the adoption of these policies to test the hypothesis that a discursive shift precedes the institutionalization of social change. Social scientists can also study the cultural impact of BLM and antiracism on the youngest people in society. A consistent finding in public opinion research is that cohort replacement is a very important driver of social change. Already, multiple studies have begun to document the ways that BLM has changed the way youth think about inequality and how parents speak to their children about race (Underhill, 2018).

3.4 Conclusion

Often a movement's "success" is reflected in widespread acceptance of the movement's goals or institutional change such as legislation. Here, I document BLM's successful injection of the movement's framing into public discourse. These findings are relevant for researchers interested in social movements and long-term cultural change for two reasons. First, the introduction of antiracist discourse could reflect a pivot towards an alternative pathway to secure a more just future. The participants in the events of 2020 (and earlier BLM protests) have an advantage that previous generations of activists did not: they witnessed the shortcomings of civil rights discourse, diversity and inclusion discourse, and the liberal discourse of "reform." Today's activists are not rejecting older movement discourse so much as they are exposing the limitations of such framings. Second, the murder of George Floyd was a catalyst for an interracial coalition of actors and a key moment for organizers who are invested in social change. My exploration of BLM is framed by the killing of Eric Garner in 2014 and that of George Floyd 6 years later. These events look remarkably similar: middle-aged Black men recorded on video begging for their lives as they are choked to death by police in the street. Yet the collective response to these events is starkly different. In 2014 and 2015, the discursive impacts do not go far beyond the recognition of the existence of racialized police homicide. Progressively the response evolved to become more expansive, beyond police killings, even beyond policing, to the social structures that create and maintain the conditions of Black life in the U.S. This study demonstrates how

political organizers have leveraged these tragic events to produce a new collective understanding of society.

The broad discursive response to the George Floyd protests shows it is a mistake to characterize BLM as fundamentally, or exclusively, concerned with policing or even the carceral state. Positions taken by BLM organizers and rhetoricians in publications, interviews, and speeches make clear that policing is only one facet of Black emancipatory politics under the banner of Black Lives Matter. Legislative and Democratic platform debates have barely touched on the abundance of societal choices leading to the marginalization, exploitation, and disposal of Black life. The deliberate exclusion of Black families from the post-war growth of the middle class and creation of White suburbs (with White suburban tax-bases and schools) and Black urban ghettos (Freund, 2007), the withdrawal of already insufficient public funding of community and mental health care programs which began in the 1980s and continues today (Nelson, 2011; Hohle, 2015), and the dramatic growth of incarceration as the solution to social neglect of Black communities (Alexander and West, 2010) are almost entirely absent from popular media and political discourse, even when it purports to examine and remedy the social position of African Americans.

Even with this awareness, presenting this framing of BLM in the design and discussion of this study has been challenging for its authors. In part this is because it is easy and perhaps necessary to engage with the dominant mode of discourse. But it is also because the real problems BLM attempts to address, which include the problems of policing, pose a tremendous moral and social-structural challenge if we are to begin to resolve them. The narrative and most of the real outcomes of the 1960s civil rights movement have not come close to addressing these problems, and instead focused on the politically and structurally simpler problem of equal protection under the law and tokenized representation in White institutions. BLM is in part a response to limited social change and increased policing of Black communities coming out of the social movements of the 1960s. Beyond this, the past 50 years have seen new assaults on Black and other primarily urban minority communities as well as to the working class more generally with the weakening of funding for social programs and education under political control of both parties, which we generally refer to as neoliberalization (Hohle, 2015). I implore researchers focusing on BLM and other political actors who want to realize Black liberation to reorient toward this broader framing of BLM. Understanding BLM as a movement that arose in response to the limits of the civil rights era and amid rampant repression of Black communities challenges notions of progress held by many progressives. Addressing the problems raised by this framing may come at the cost of comfort, both emotional and material, for those who occupy positions of racialized privilege. Making this sacrifice offers no guarantee, but it is critical if we are to realize of the

incomplete goals of previous movements and work toward the restoration of Black communities and Black life.

Data Sources

To measure the use of antiracist vocabulary, I obtained data from multiple sources: Google Trends, Google Books, national news, Wikipedia, and Twitter. Each data platform provides unique information about how people are using different terms. The Supporting Information provides additional analyses incorporating multiple data sources when appropriate.

Google Search Google search data were retrieved using the Google Trends API. Google limits the resolution at which data can be downloaded. Daily search volumes were retrieved in 8-month chunks overlapping by 4 months. Google does not give access to absolute search volumes and uses a relative 0–100 scale where 100 is the maximum daily searches for the requested period. In the generalized additive models measurements are taken as a ratio, such that the relative measurements yield the same result as absolute ones.

Google N-grams Books Google N-grams data were downloaded from `http://storage.googleapis.com/books/ngrams/books/20200217/eng/totalcounts-1` and contain measurements through 2019.

Wikipedia Page Visits Wikipedia Page Visits data were downloaded using the Wikimedia REST API (`https://wikimedia.org/api/rest_v1/?doc`).

Daily News Coverage I searched keywords that related to Black Lives Matter through the Media Cloud API Roberts et al. (2021) to gather the number of publications that mentioned each keyword per day from Jan 1, 2013 to Dec 31, 2020. The source collection contains a total of 271 US national level news publications include 17 radio (e.g., NPR) and TV (e.g., CNN) broadcast, 114 digital native (e.g., Vox.com), 108 print native, (e.g., NYT), and 4 others (e.g., Reuters).

Twitter Daily counts of tweets were collected through the academic track of Twitter API version 2.0. The academic track allows researchers to access the full archive database of Twitter and collect the daily volume data of tweets that contained searched keywords. See details of the API documentation at `https://developer.twitter.com/en/docs/twitter-api/tweets/counts/introduction`.

BLM-related Protests The independent variable is obtained from multiple BLM protest data sets that researchers have collected to study the movement. The "Elephrame" data set is an open source data set collected since 2014 curated by Alisa Robinson with aid from site users. Data were scraped from `https://elephrame.com/textbook/BLM/chart` on January 8th 2021. The data set collected by Armed Conflict Location & Event Data Project (ACLED) uses automated searchers to collect news reports of U.S. protests (BLM and otherwise) since May 2020. Data were retrieved using the ACLED Data Export Tool (`https://acleddata.com/data-export-tool/`) on January 15th 2021.

CHAPTER 4

COMPLEXITY THEORY WITHOUT COMPLEX MODELS: THE 27 CLUB MYTH[1,2]

4.1 Introduction

The "27 Club" refers to the widespread legend that notable people, particularly musicians, are unusually likely to die at age 27. A 2011 inquiry in *The BMJ* showed this is not the case, dismissing the 27 Club as a myth. In this chapter I expand on this discourse by demonstrating that although the existence of the phenomenon cannot be empirically validated, it is real in its consequences. Using Wikipedia data, I show that while age 27 does not hold greater risk of mortality for notable persons, those who died at 27 are as a group exceptionally notable compared to those who died at other young ages. The 27 Club legend originated from a statistically improbable event circa 1970, wherein four superstar musicians died within the span of 2 years all at age 27. This coincidence captured the public imagination such that our fascination with the 27 Club brought itself into being, producing greater interest in those who died at age 27 than would have been otherwise. This demonstrates path dependence in cultural evolution, whereby an effectively random event evolves into a narrative that shapes otherwise unrelated events and thus the way we construct and interpret history.

This study is, from a data and methods perspective, a small version of Chapter 3, and could be thought of something like Chapter 3.5. Like the BLM study, I use Wikipedia data to study large scale cultural patterns. Here, I examine the relationship between notability and age, and show that a widespread myth, the "27 Club", has real consequences, although belief in the myth is erroneous.

From a methodological standpoint, this piece offers nothing new over the BLM study. In both, trace data are the substrate for a regression analysis, here, Bayesian regressions with spline terms. The main

[1]This chapter is based on as "Path dependence, stigmergy, and memetic reification in the formation of the 27 Club myth" with Patrick Kaminski accepted for publication by *PNAS* September 26 2024.

[2]A massive thank you Patrick for rubber-ducking me through this fun and blazingly fast paper. I am also particularly grateful to Kate Howell for the initial conversation that sparked this idea. Thanks also to Helge Marahrens, Christian Kipp, and anonymous reviewers for gracious and thoughtful comments toward improving the manuscript.

analysis of this chapter is less complicated than Chapter 3's, eschewing the need for close reading or a temporal analysis. Instead the temporal dimension of the study emerges by deduction through conceptual contributions from complexity theory and analogues in traditional social theory.

Typically it is assumed that only complex systems methodology can capture complexity in a complex system (see Rosen, 1987). This study shows that this is not the case. While the regression methods are indeed sophisticated (quantile regression, Bayesian modeling, non-parametric terms) they are decidedly not complex in name or in nature. Instead, the relatively simplistic model is used to verify the existence of a phenomenon, and I use verbal argument to reason how complex forces of chaos (path dependence), contagion (social transmission), and stigmergy (no direct analogue in traditional social theory) converged to produce the phenomenon. In truth, empiricists do this sort of deductive work frequently, but it is typically viewed as something suspicious and dirty if acknowledged at all. Humanists, philosophers and cultural studies scholars in particular, are completely comfortable relying on this sort of reasoning. While the case of the 27 Club is fairly uncomplicated, this study shows how empirics, theory, and deductive reasoning can harmonize to put forward "scientific" arguments.

4.1.1 Background

The legend of the 27 Club emerged in the early 1970s following the deaths of famous musicians Brian Jones, Jimi Hendrix, Janis Joplin, and Jim Morrison, each aged 27, within a span of two years (Sounes, 2013). The mystique of these deaths was strengthened by further uncanny parallels including participation in the counter-cultural music scene and festival circuit, the drug-related nature of the deaths, and the occurrence of the fourth and final death on the second anniversary of the first, as well as the apocryphal white lighter found on each musician (Evon, 2017). This event fostered a widespread belief that there is an increased risk of mortality at the age of 27 compared to other ages, particularly other similar young ages. Subsequently, the 27 Club expanded to include other notable cultural figures Jean-Michel Basquiat, Kurt Cobain, and Amy Winehouse, and even those who died before the term was coined, such as blues pioneer Robert Johnson in 1938. At present the English Wikipedia page for the 27 Club lists 87 individuals, and parallel pages exist in 50 other languages. No other year-age group has such a page on Wikipedia.

It is perhaps natural to assume the 27 Club is merely a myth and of no real significance. Supporting this view, a 2011 retrospective cohort study of British musicians conducted by Wolkewitz and colleagues found no evidence of increased risk of dying specifically at age 27 (Wolkewitz et al., 2011). Instead, the study

noted a generally higher risk of mortality throughout musicians' 20s and 30s, challenging the validity of the 27 Club narrative. There is no credible evidence or plausible medical mechanism supporting an increased likelihood of death at age 27.

Thomas & Thomas famously stated in their eponymous Thomas Theorem that, "If men define situations as real, they are real in their consequences" (Thomas and Thomas, 1938). Given the persistent legend of the 27 Club, I propose that the club is reified (made real) as a cultural phenomenon and measurable in its consequences even as empirical death rates do not support it. I argue that persons who die at the age of 27 receive outsized attention that can only be explained via a real 27 Club effect.

Thus, the reality of the 27 Club is decoupled from its myth. While the legend states that famous persons are more likely to die at 27, in actuality dying at 27 makes a person more famous than they would have been otherwise. The increased attention that members of the 27 Club phenomenon receive posthumously inflates our perception of the number of deaths at 27, at least among the uppermost echelons of notable persons. I demonstrate this decoupling and the true 27 Club effect with a series of Bayesian quantile regression models, which show a moderate and statistically significant boost to attention to persons who died at 27 around the 90th percentile of notable persons, and an even greater effect around the 99th percentile. Therefore, one might say, "To shine brightest in death, aim for 27."

My theory of the 27 Club phenomenon has three components: path dependence, stigmergy, and memetic reification. I will treat each of these briefly here, and elaborate on their relationships to the empirics and each other in the discussion.

The first concept, path dependence, refers to the highly improbable event that inspired the myth of the 27 Club. Path dependence is a theoretical tool for understanding historical developments in social systems (Mahoney, 2000). While individual events are often hard to predict, they may catalyze a causal chain of events which are predictable. In the case of the 27 Club, the original 4 deaths were a fluke, yet it is unsurprising that such a remarkable event with high public visibility would inspire a myth.

The second concept, stigmergy, is a mechanism of indirect coordination through the environment, whereby individual actions leave signs that trigger subsequent actions by the same or different agents (Van Dyke Parunak, 2005). This process facilitates complex, collective behaviors without centralized control. A classic example in human behavior are so-called "desire" paths through urban green space. As people repeatedly walk along paths, they leave traces that encourage others to follow the same routes, reinforcing the trails' formation and guiding the flow of traffic. Stigmergy is also a driver of recommendation algorithms:

as users designate a product or resource as useful, the system makes it easier to find. In this study of the 27 Club, stigmergy occurs through links to 27 Club members throughout Wikipedia, increasing the probability of page visits, and potentially inspiring subsequent authors to create new links to 27 Club members in Wikipedia and beyond. Thus is stigmergy a mechanism of contagion, spreading cultural knowledge through traces left in the environment.

By the final concept, memetic reification, I refer to the cultural evolution implicated by the Thomas Theorem. The 27 Club myth transformed into a real phenomenon through memetic processes, i.e., the transmission of culture. Real consequences for our shared understanding of the world, in this case, who among us reach fame and become part of the culture, were shaped by the transmission of a relatively small story, the myth of the 27 Club.

4.2 Results

I employ the Notable Persons data set curated in 2018 (Laouenan et al., 2022). The data comprise (nearly) all persons with Wikipedia pages at the time of collection. As a proxy for notability I use Wikipedia page visits from 2015–2018 as recorded in the Notable Persons data set. To remove spikes in attention in the period immediately following death, I restrict the data to those who died prior to 2015. Due to survival biases, which lead more notable persons to be more likely to persist in the collective memory, I further restrict the data to those who were born after 1900. Finally, I pulled exact dates of birth and death from each person's Wikidata page. The final data comprise 344,156 persons with pages across all languages. I chose not to restrict the analysis by language, as the 27 Club phenomenon is internationally known, with entries in 51 languages.

4.2.1 Musicians die younger, but there is no increased risk of death at 27.

Figure 4.1 shows the distribution of death ages for notable persons in Wikipedia. I give three distinct perspectives on mortality, dividing the data into all figures (N = 344,156), cultural figures (N = 125,198), and pop musicians (N = 14,517). Professions were determined using the Notable Persons occupational class variables.[3] Figure 4.1 indicates pop musicians tend to die younger than other notable persons. However, the discrepancy for musicians is partly obscured by the full set's fatter tail, which includes many who found

[3]Pop musicians were identified as those whose primary profession appeared more than 100 times in the data set and were likely to ascribed to pop musicians: singer, music, guitar, bandleader, saxophonist, trumpet, drummer, jazz, bassist. Contrast this with other musical designations e.g., opera, composer, pianist, conductor, violinist.

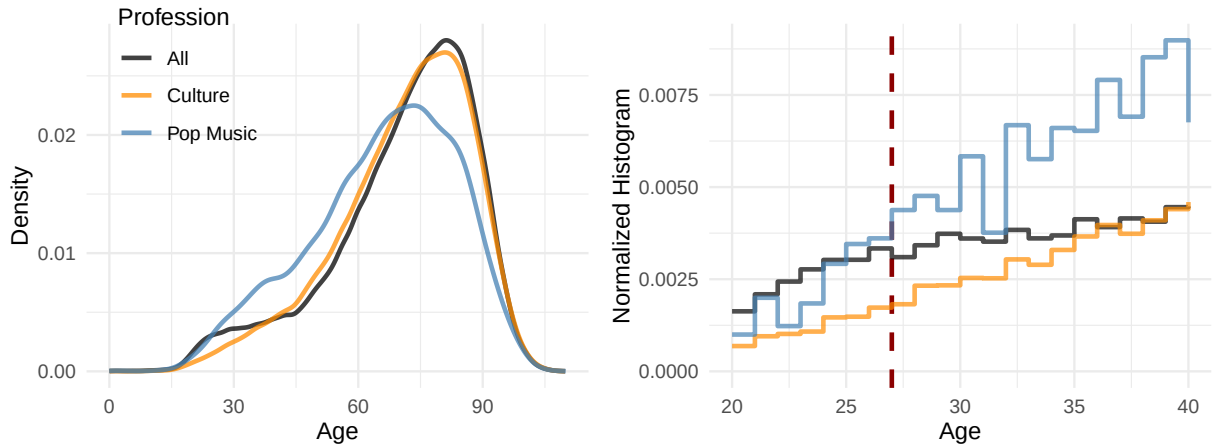

Figure 4.1: Death-age distributions among persons on Wikipedia born after 1900. Left: Kernel density for age of death in Wikipedia. Right: Normalized histogram of the same data for ages 20–40. Black: All notable persons. Orange: Persons whose primary profession is cultural. Blue: Persons whose primary profession is the production of popular music.

fame at younger ages, such as athletes and aristocrats. The distribution for cultural professionals shows elevated rates of early death for pop artists compared to others in similar fields, consistent with other studies (Kenny and Asher, 2016; Wolkewitz et al., 2011; Epstein and Epstein, 2013; Anisimov and Zharinov, 2014).

Figure 4.1 does not suggest an increased rate of death at age 27 for any category. The right panel shows a normalized histogram for death age, and 27 does not appear to deviate from the approximately linear trend for each professional class. The histogram does not show an elevated risk of death at age 27, consistent with intuition and Wolkewitz et al. (Wolkewitz et al., 2011).

4.2.2 The real 27 Club effect: Those who die at 27 are more famous than one would expect otherwise.

As the sociomedical hypothesis of higher mortality at age 27 is both implausible and unsupported, I turn to a purely sociocultural framing. This states that the those who die at age 27 are more famous than we would expect otherwise. Figure 4.2 shows notability arranged by rank for young deaths, those aged 25–40. A color gradient from dark (younger) to light (older) indicates notability increases with age. This is intuitive for two reasons. First, as demonstrated by Figure 4.1, mortality risk increases smoothly between ages 20 and 40. Within Wikipedia, fame is roughly log-normally distributed, so expected fame at each rank increases with the size of the pool. Second, achievement and network quality grow with age, increasing the probability of reaching exceptional notability (Hill and Dunbar, 2003; Badar et al., 2014).

The 27 Club, shown in green, visibly diverge from the tendency for notability to increase with age in

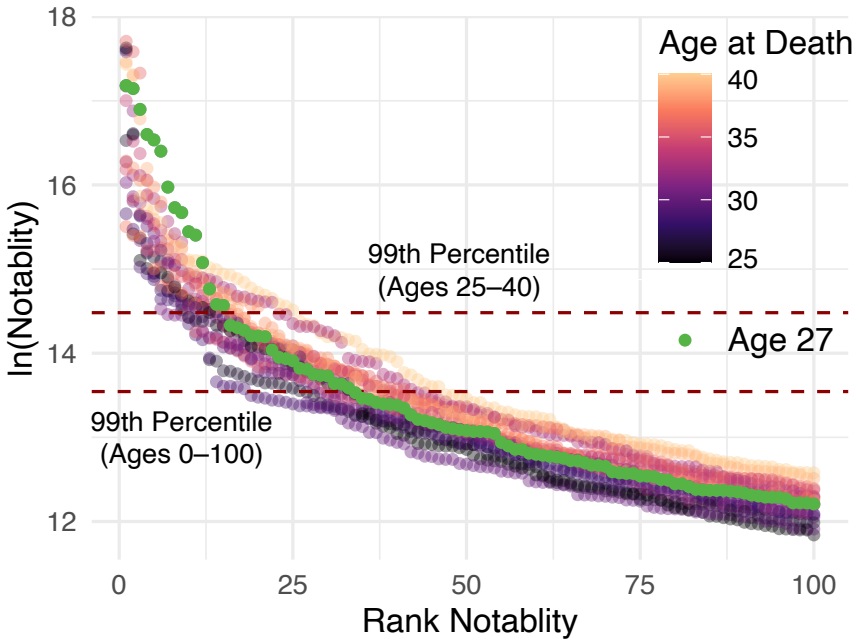

Figure 4.2: Notability (natural log) by rank within each death-age [25,40]. All 344,156 persons were stratified by age, then ordered and ranked by notability. This plot presents ln(notability) for the 100 most notable persons at each death-age, subsetted ages 25–40. Points are colored by age from dark (low) to light (high). Green points show notability of those who died at age 27. Dashed lines demarcate the 99th percentile for both the subset of ages plotted and the full data set.

Figure 4.3. For most of the range of rank, e.g., 25–100, the 27 year-old curve is at the midpoint of the distribution, around the early-mid 30s. This suggests a considerable 27 Club effect on posthumous fame. The effect is even starker in ranks 4–12, where the 27 Club floats above all other ages.

I developed a formal test of this 27 Club effect, visualized in Figure 4.3, by means of a series of Bayesian quantile regressions using the `brms` package in R (Bürkner, 2017). Whereas regression typically estimates the conditional mean of the dependent variable, quantile regression estimates the conditional median or any other specified quantile. This allows me to investigate the relationship between age of death and notability at various levels of fame. Fitting quantile regression at the 50th percentile, for instance, would estimate the expected level of notability for those of moderate fame for 27 year-olds as compared to all other ages. Fitting a regression to the same equation but specifying the 99th percentile would estimate how much more famous the most notable 27 year-olds compared to the all other ages are.

Models were computed using the `brms` package in R with Stan backend. Quantile regression was computed by estimating asymmetric Laplace likelihood, `asym_laplace`, for each quantile, τ. Age was measured as a pseudo-continuous variable, i.e., in exact days rather than exact years, in order to best approximate ac-

cumulation of notability outside a 27 Club effect. Cubic splines for age were fit using 5 knots. Early experiments with the number of knots, however, had only a marginal effect on the fits. I set naive priors for each term in the equation given by Equation 4.1. Betas were initialized with flat priors.

$$\begin{aligned} Q_{\text{fame}|X}(\tau) &= X\beta_\tau, \text{where } 20 \leq \text{age} \leq 40 \\ &= \beta_0 + \beta_1 27\text{Club} + s(\text{age}) + \epsilon. \end{aligned} \tag{4.1}$$

As we have intuitive and empirical support for a positive effect of age on notability, I estimate a continuous term for age in addition to a binary coefficient for the 27 Club. This regression effectively asks "How much more famous are the 27 year-olds than we would expect them to be given the otherwise smooth relationship between age and fame?" We should not necessarily expect the relationship between age and fame to be linear, so I fit a cubic spline to age allowing for some bend in this relationship.

I restrict the regression to ages 20–40 for a number of reasons. First, my 27 Club hypothesis proposes increased fame at 27 compared to similar ages. Second, the relationship between age and notability is more complex when estimated across the full range of ages and finding a good fit for the range of interest is easier when the spline is fit only in that range. Third, splines are less reliable at their ends, and 27 lies in a tail when fitting a fuller range of ages.

Although I visualize death distributions for different professional classes in Figure 4.1, the quantile regression model includes all professions. There are 21,931 persons aged 20–40 in the data set fitting the additional specifications enumerated in the first paragraph of the Results. This is the sample I analyze in the quantile regression.

The estimates of the Bayesian regression models are visualized in Figure 4.3. Each model fits its own coefficients predicting a particular quantile. The left panel shows the effect of age as estimated by the spline term. Figure 4.3 shows weakly positive, monotonic relationships between age and notablity with slight deviations from linearity. The right panel gives coefficients for the 27 Club, estimating how much more notable are those who died at 27 compared to other ages. The 50th and 80th percentiles do not estimate significant coefficients. However, the 90th through 97.5th show consistently positive effects around 0.3. The most famous persons show a much larger effect. The 99th percentile model estimates a full order of magnitude increase in notability for the 27 Club.

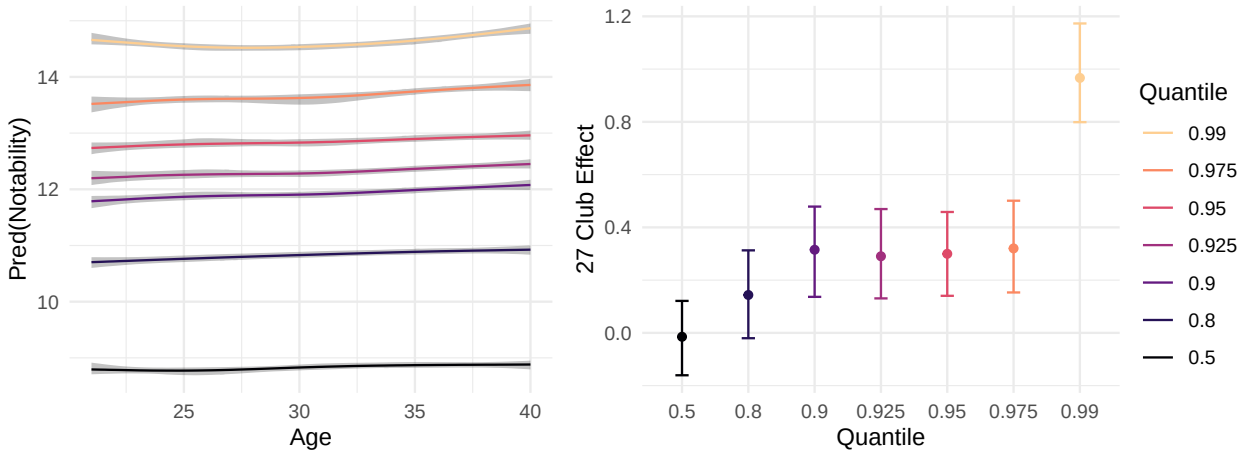

Figure 4.3: Bayesian quantile regression models for 7 quantiles. Left: Estimated ln(notability) for each quantile with the spline fit for age. Right: Coefficient for the 27 Club effect for each quantile with 95% CI.

Given the 27 Club Wikipedia page(s) containing links to noted club members, it is possible that the increased notability of the 27 year-olds on Wikipedia is entirely attributable to clicks originating from such pages. Were this the case, it would be dubious to state that death at 27 confers greater notability, and thus I could not support the claim that the myth has a real effect. In other words, if the 27 Club effect we find is caused by the 27 Club pages, we have learned something about Wikipedia, but fallen short of the Thomas Theorem: I haven't demonstrated that the 27 Club effect extends into the world beyond this one page.

To control for the effect of 27 Club Wikipedia pages, I make use of Wikipedia Clickstream data, which provide networked "referrer/resource" pair data for Wikipedia pages in 11 languages since 2018. For each of 944 persons who died at 27 in the data set, I calculate the percentage of Wikipedia page visits originating with the 27 Club page in each available language. I measured the proportion of notable persons page visits that originated from the 27 Club page across all 11 languages (Chinese, English, Farsi, French, German, Italian, Japanese, Polish, Portuguese, Russian, and Spanish) in the year 2018. I then adjust each page visit count by subtracting the proportion of 27 Club page clickstreams from the page visit count recorded in the Notable Persons data set. I find that this control only slightly alters the point estimates fit by the quantile regression models, modestly lowering them. A comparison of each model's fit on the two data variants (including vs. adjusted for 27 Club page clickstreams) is shown in Figure 4.5.

The lack of impact of the 27 Club page referrals to 27 Club members is unsurprising given that only a small proportion of those who died at 27 are mentioned on these pages. Of 944 notable persons in the data set, only 59 registered referrals from any of the eight 27 Club pages for which I have clickstream data.

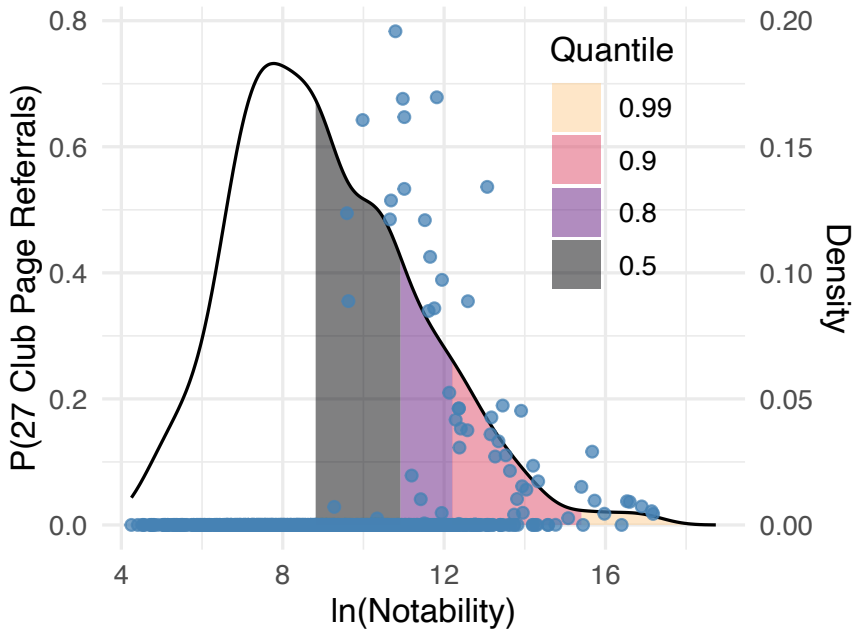

Figure 4.4: Scatterplot demonstrates the proportion of page visits via a 27 club page by notability (natural log) for all 944 persons who died at 27 in the data set. Kernel density shows the distribution of notability for 27 Club members. Shaded regions indicate the boundaries of several quantiles fit in the regression model in Figure 4.3.

Figure 4.4 shows that fewer than 20 moderately notable persons garner between 30% and 80% percent of their visits through 27 Club pages, while the more notable members of the 27 Club are under 20%. The most notable are, with the exception of Brian Jones (12%), well under 10%. Thus, where we see the strongest 27 Club effect, the 27 Club pages are least impactful. Still, most who died at age 27, even above the 80th percentile, attribute none of their visits to the 27 Club pages.

4.2.3 Stigmergic and non-stigmergic evidence of the 27 Club effect.

One component of my theory of the reification of the 27 Club myth is that the increased fame of 27 Club members is partly propelled by stigmergy. Recall stigmergy occurs when individual agents leave signals in the environment for subsequent agents to follow. Stigmergy is most obviously at play in the existence of the 27 Club page, but can be observed through other links to 27 Club members, both within Wikipedia and across the Web. On the other hand, page visits to 27 Club members that occur through search engines, and thus originate without a direct pathway, represent a measure of notability through non-stigmergic means.

The Wikipedia Clickstream data allow for multiple perspectives of where stigmergy is at play in the 27 Club phenomonenon. Figure 4.6 repeats the Bayesian quantile regression design using slightly different

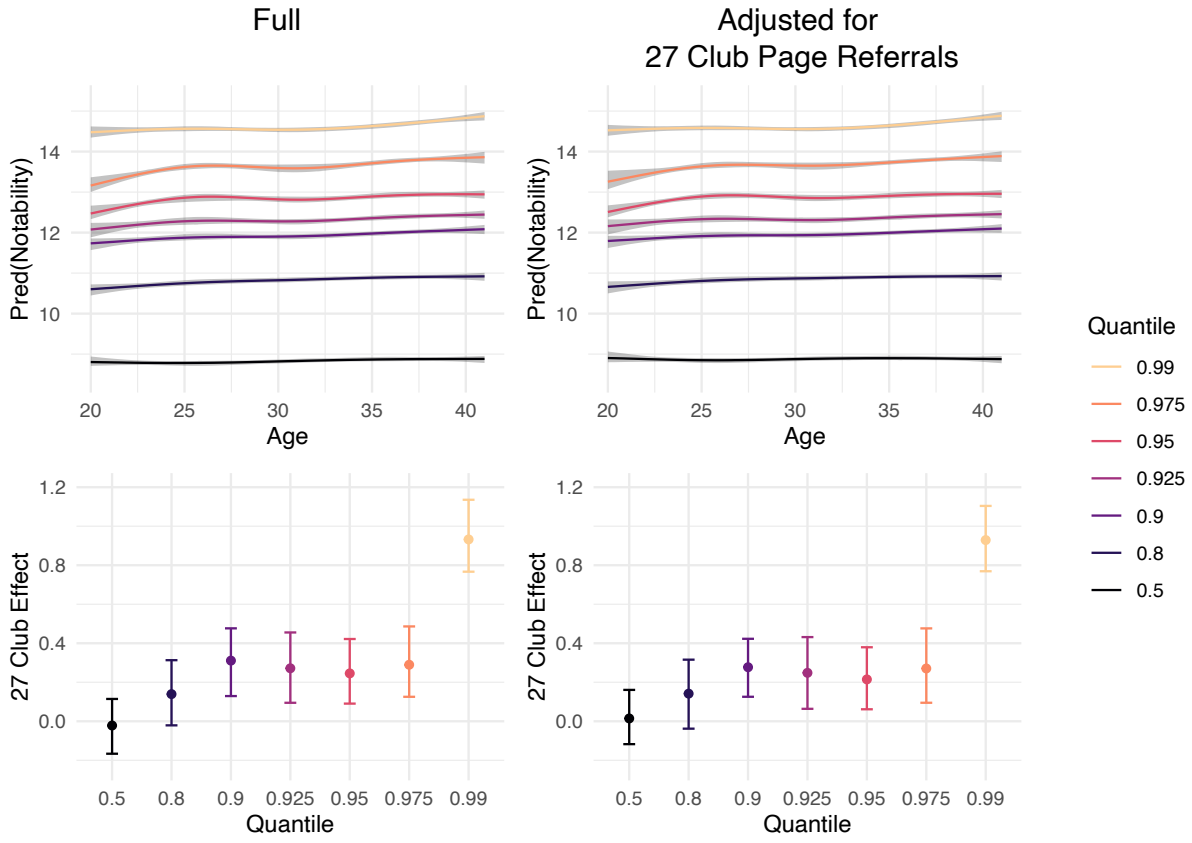

Figure 4.5: Estimates from Bayesian quantile regression models for 7 quantiles as in Figure 3. Left: Results from Figure 3. Right: Quantile regression results adjusting for referrals from Wikipedia 27 Club pages in 8 languages.

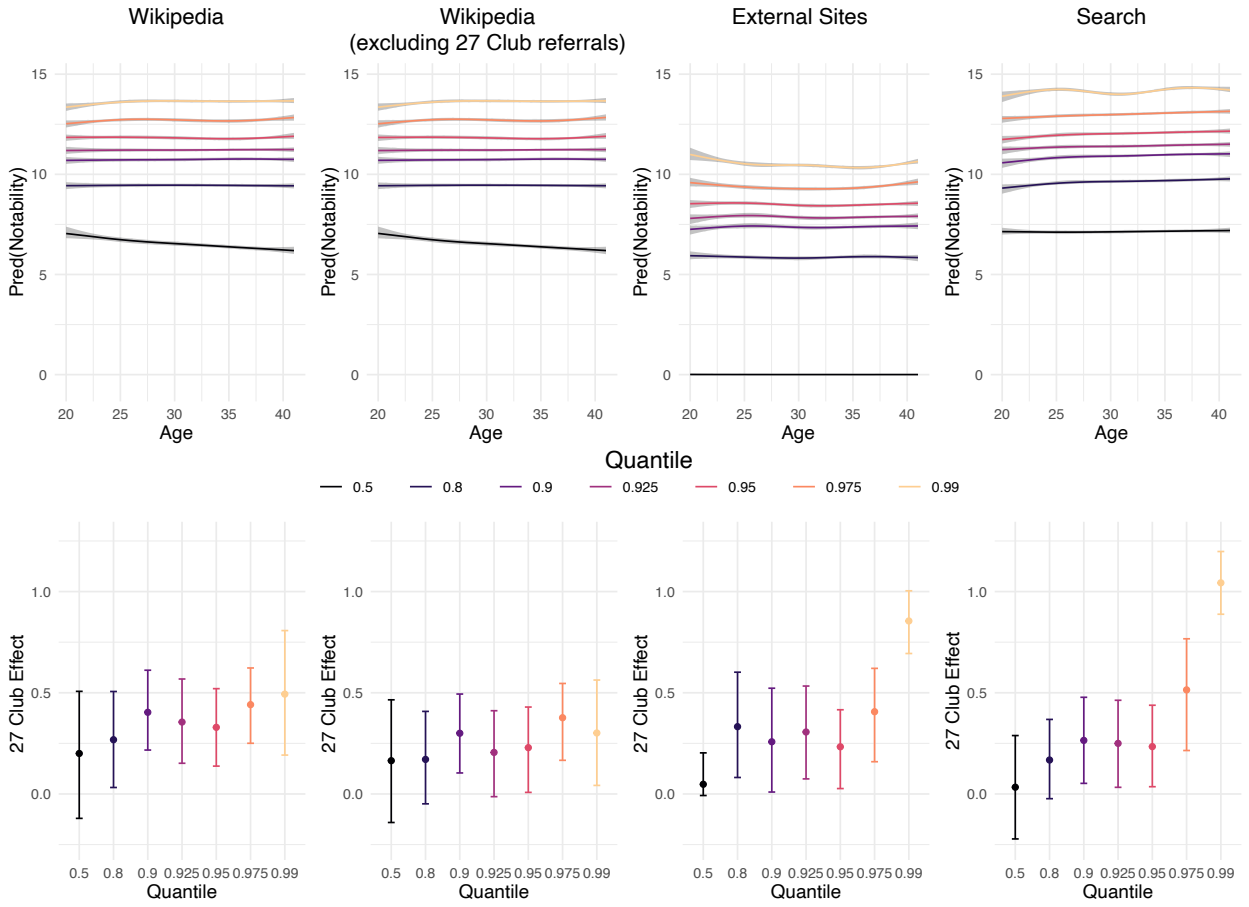

Figure 4.6: Estimates from quantile regression models on various measures of notability calculated from Wikipedia Clickstream data. As in Figure 4.3 the data include all 20-40 year olds in the data set. Top row: Estimated ln(notability) for each quantile with the spline fit for age. Bottom row: Coefficients for the 27 Club effect for each quantile with 95% CI.

flavors of the dependent variable, our measures of notability. Here the sample is the same as above, but the page views data comprise only the 11 languages in the Wikipedia Clickstream data set.

27 Club effects can be seen across all metrics of notability. Referrals from other Wikipedia pages measure stigmergic attention through Wikipedia links. Contrary to the main analysis on the full data set, we see that the 27 Club Wikipedia page accounts for a substantial portion of the 27 Club effect: when comparing notability including referrals from the 27 Club page to those excluding them, the coefficients diminish moderately. The stigmergic boost in notability of 27 Club members can also be see in referrals from external websites. While these may originate with articles referencing the 27 Club, they exclude the 27 Club Wikipedia page.

Non-stigmergic effects of the 27 Club are demonstrated by quantile regression of page visits originating

from search engines. Here we see yet the largest 27 Club effect in the uppermost echelons of notability and similar moderate effects at and above the 90th percentile. The large effect at the 99th percentile compensates for the considerable drop in page visits from Wikipedia when excluding 27 Club referrals.

These results demonstrate 27 Club effects across the Internet by both stigmergic and non-stigmergic means. While the 27 Club pages do confer considerable increases to notability of those listed, across Wikipedia there is evidence of a 27 Club effect beyond the 27 Club page. Further, we see the 27 Club phenomenon through search engine referrals. As many search queries are inspired by events, thoughts, and interactions offline, this provides strong evidence that a 27 Club effect exists in the cultural milieu outside the Web itself.

4.2.4 Path dependence in the 27 Club myth: The original 27 Club deaths were an unlikely event.

Finally, I argue that the origin of the 27 Club was an unlikely occurrence, and thus the effect we observe in the prior analysis is path dependent. To test this, I fit a Bayesian count model (zero-inflated Poisson) for multiple famous deaths at each year-age during a 2-year period. Again, we estimate a smooth term for age. I restrict the data to the 99.9th percentile, or 1.5 million annual visits. Brian Jones, the least famous of the original four, averaged 1.8 million visits. Zero-inflated Poisson models were computed using the `brms` package in R with Stan backend. I fit a smooth term for integer age in years using the default settings and default priors `brms` for Equation 4.2.

$$N(deaths) = \beta_0 + s(\text{age}) + \epsilon. \tag{4.2}$$

This model, visualized in Figure 4.7, estimates 3 deaths at 27 have a mean probability of 4.45×10^{-5}. 4 deaths are far less probable at 7.26×10^{-7} (95% CI = [1.63×10^{-6}, 2.61×10^{-7}]). Thus I estimate a 2-year period with 4 deaths at age 27 to be roughly a 1 in 100,000 event. This is a crude estimate, but it nevertheless gives a sense for the improbability of the 27 Club's origin, and in turn part of the event's mystique.

4.3 Discussion

This investigation shows that the 27 Club is a real phenomenon: those who died at 27 are more famous than we should expect by chance, and that benefit to notability increases with notability itself. While the Thomas Theorem explains the consequences of cultural beliefs, it does not account for their origins. Drawing on

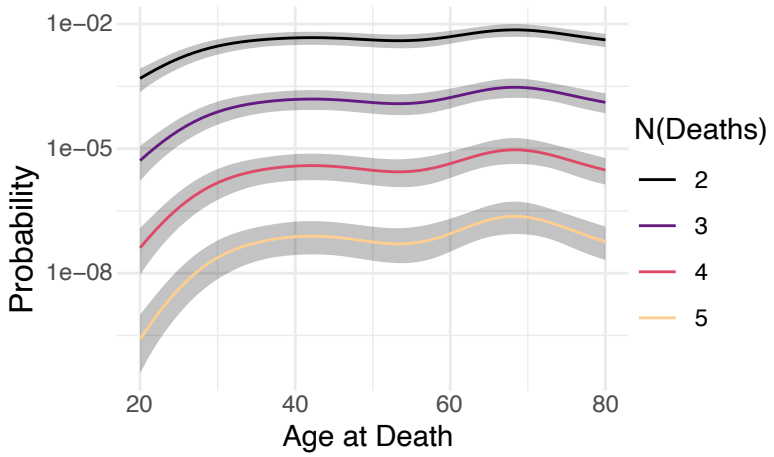

Figure 4.7: Probability of N deaths at a single year-age in a 2-year period with 95% CI.

sociological theories of diffusion (Boyd and Richerson, 1988; Christakis and Fowler, 2013; Centola, 2015; Sato, 2024), I characterize the process of mythmaking and its cultural consequence as "memetic reification". Social contagion theory accounts for the propagation of the belief or "meme" (Aunger, 2006), and the Thomas Theorem describes the downstream effects of the legend, its effect on reality, or reification. Further, the 27 Club effect is felt whether or not you are aware of it. The boost to 27 Club members' notability increases their visibility generally, forming a positive feedback loop. In this case, cumulative advantage occurs partly through stigmergy, whereby elevated visibility resulting from notability at earlier periods helps to grow and maintain fame during later periods, which can produce heavy-tailed distributions like we see in Wikipedia page visits (Mittal, 2013; Salganik et al., 2006). The stigmergic effect of the 27 Club is further inflated by the existence of the 27 Club Wikipedia page. The page's links to club members' pages provide a direct mechanism to boost our proxy for notability, Wikipedia page views, which undoubtedly raises awareness of members in our collective cultural imagination.

The final component of my theory of the formation of the 27 Club legend is path dependence. From a general systems perspective, path dependence occurs when a feature of the system that had no functional significance and/or resulted from chance predictably stabilizes or becomes emphasized over time (Mahoney, 2000; Page, 2006). Path dependence is found in historical as well as biological and cultural evolutionary systems. In cultural evolution, this entails the emergence of a schema from an initial contingent event and can only be attributed retrospectively to that event.

In the case of the 27 Club, an improbable ($\sim$1 in 100,000), but ultimately influential event precipitated a

cultural schema that allows people to attribute meaning and "explain" this unlikely event (DiMaggio, 1997). This schema is the 27 Club myth of elevated risk of death at 27. Initial belief in the myth is the result of over-generalization from the genuinely uncanny event. Early on, the myth spread only through narrative transmission. But gradually, a real 27 Club effect emerged: those who died at 27 truly were more visible. This was due in part to increased salience of their deaths (e.g., the 27 Club Wikipedia list), but also through downstream effects of inflated presence in the culture outside the context of 27 Club membership. Thus, as the 27 Club myth gave rise to a real, but distinct 27 Club effect, the effect itself serves to bolster the myth. While the elements of the origin of this particular phenomenon are especially tidy, the social forces and their interplay are more widely applicable in the development of not only folklore, but cultural patterns more generally.

CHAPTER 5

SIMULATING CULTURAL EVOLUTION: DYNAMICS OF COVERT SIGNALING[1,2]

5.1 Introduction

Covert identity signals permit the communication of group membership to ingroup members while avoiding potentially costly detection by members of other groups. If individuals are incentivized to detect others' group memberships, however, covert signals may not remain covert for very long. In this chapter I propose a theoretical extension to the literature on covert signaling in which conventionalized identity signals can become destabilized when learned by outgroup individuals, to be replaced by the emergence of new signaling conventions. I formalize this idea with both analytical and agent-based modeling of ingroup and outgroup individuals who learn about signals of group membership. Depending on the risk and associated cost of detection by the outgroup, the model yields three dynamic classes: saturation, where all identity signals become stable conventions and never go extinct; cycling, in which new signals emerge to replace old ones as they are learned by the outgroup; and suppression, in which informative identity signals never emerge. This analysis has implications for understanding identity signaling, the emergence of conventions, coded speech, and the ebb and flow of fashion cycles.

The studies in previous three chapters have largely used computational tools or computational data that feed into traditional methods of analysis. In Chapter 2 traditional analysis was qualitative: close reading. In Chapters 3 and 4 traditional analysis was quantitative: frequentist and Bayesian statistics. In Chapters 2 and 3, I take semantic data and produce analysis where meaning and concepts are central to the analysis. In this chapter I eschew traditional methods for an approach that employs only methods originating from complex systems to produce an ad hoc model of communication. These methods, dynamical systems modeling

[1]This chapter is based on "Dynamics of covert signaling: Modeling the emergence and extinction of identity signals" with Paul Smaldino accepted for publication by *Psychological Review* September 16 2024.

[2]Much gratitude to Paul for persevering through multiple (rather frustrating) review processes. It shouldn't be this hard to get ABMs published, but I'm happier doing it with you! Thanks also to Veronica Capelli, Mirta Galesic, Patrick Kaminski, Helena Miton, Tamara van der Does, Mikkel Werling, and Harry Yan for helpful comments that strengthened this manuscript. This work was supported by ARO grant W911NF-20-1-0220 to Paul. E. Smaldino and by the Santa Fe Institute.

(physics-style modeling) and agent-based modeling, a genre of algorithmic computer simulation, allow the probing of social phenomena without any data, or, more accurately, data are the outputs not the inputs of the method. Thus, in this study I examine a cultural system and abstract meaning away entirely to examine a process of the emergence of meaning.

In this study I developed a model of these covert signaling dynamics. Ingroup agents signal to each other using arbitrary symbols, here simply ordinal numbers, and learn to associate particular signals with their ingroup identity through successful interactions with other ingroup members. Outgroup members also interact with the ingroup, and similarly learn to associate ingroup identity with the ingroup, after which the ingroup abandons the signal due to punishment from the outgroup. The cycle starts anew with a different signal finding salience among the ingroup.

The study of White Nationalists in Chapter 2 contains much such coded political signalling. Other examples include political dissidents on internet fora such as Winnie the Pooh Xi Jinping memes (Hearn, 2020) and gay haircuts (Hayfield et al., 2013). When a signal that effectively communicates ingroup identity to other ingroup members becomes known to the outgroup, ingroup members must abandon it to avoid sanctions, develop new signals in their place to continue clandestine coordination. The model also can be understood to represent Simmel's fashion cycles (1904), in which the fashions shift so that elites can remain distinct from non-elites who learn to mimic elite trends.

The primary virtue of this style of modeling is that it does not require real data. While a good model derived from observations of a real system would certainly be preferable, covert signaling dynamics are especially challenging to study because draws on so much what makes culture difficult. Cultural signals are difficult to identify generally because they are numerous and highly contextual. In this case there is even greater difficulty because the signalers are attempting to avoid detection. Further, because the signals are changing over time and diffusing through the ingroup, we require high resolution data. Ad hoc modeling allows us to avoid both of these challenges by 1) abstracting away content, and 2) controlling all aspects of the system to give the desired resolution. One further strength of ad hoc modeling is that we can modulate the parameter values to give an full picture of the kinds of dynamics that can emerge. By contrast a single data set is likely to capture a small set of parameter values. Nevertheless, because it may be difficult or impossible to observe particular parameter combinations, and, worse still, parameter values may be fundamentally unquantifiable in the real system, ad hoc modeling often produces results which are solipsistic, and it is difficult to know how to integrate such knowledge with a broader field of work that is derived from real

systems, whether through observational or experimental studies.

5.1.1 Background

When interacting with other people they don't know very well, people often broadcast signals to establish the sort of person they are or aren't. These identity signals can be linguistic, sartorial, or even behavioral: they serve to inform audiences about what the sort of person the signaler is or is not (Berger and Heath, 2008; Smaldino, 2019). The identities being signaled may concern the signaler's role in society, their membership in a particular group, or even their individual behavioral characteristics (Burke and Stets, 2009). For example, a person may communicate their identity as a mother (role), a democratic socialist (group membership), or someone with a morbid sense of humor (individual behavior). Identities serve functions far beyond establishing the goals and psychological well-being of those that hold them, though these are surely important. Identity signals are used instrumentally to facilitate social assortment (Smaldino, 2019, 2022). They provide audiences with information with which to evaluate the likelihood of success of cooperative interactions (including economic or romantic interactions) and about potential dangers of direct engagement. For example, numerous studies have shown that altruistic behaviors are preferentially directed toward individuals perceived as ingroup members (Chen and Li, 2009; Henrich and Muthukrishna, 2021). The signaling function of how we present ourselves in public has long been appreciated by social scientists (Goffman, 1978; Barth, 1969; Donath, 1999; Berger and Heath, 2008; Wimmer, 2008).

At their core, identity signals serve a key social function by enabling individuals to rapidly characterize others as similar or dissimilar. In the politically polarized United States, signals have emerged to identify individuals as belonging to the political right or left (Urbatsch, 2014; Sloman et al., 2021; Powell et al., 2023). Some of these signals are directly connected to the signaler's political views, such as bumper stickers declaring support for particular candidates or polarizing social issues (e.g., "Abortion Is Healthcare" or "Gun Control Means Using Both Hands"). Using one of these signals requires minimal common knowledge between the signaler and receiver, because the signal contains explicit information about its intended meaning. Other signals, however, are seemingly arbitrary and become conventionalized as identity signals only over time. Some of these may reflect the phenomenon of "lifestyle clustering" so that stereotypes about "latte-drinking liberals" and "bird-hunting conservatives" can emerge via the tendencies of liberals to live in cosmopolitan urban centers more likely to have upscale coffee shops and of conservatives to live more in rural areas where hunting is more accessible (McPherson, 2004; DellaPosta et al., 2015). The correlation

between identity and lifestyle need not be strong, however, for signals to become conventionalized. Signaling conventions can emerge entirely through the amplification of small differences when assortment with similar individuals is beneficial. Numerous behavioral experiments and formal models have shown how incentives for coordination can facilitate strong correlations between observable signals and unobservable characteristics when no such associations existed in the initial population (Boyd and Richerson, 1987; Nettle and Dunbar, 1997; McElreath et al., 2003; Castro and Toro, 2007; Efferson et al., 2008; Puglisi et al., 2008; Cohen and Haun, 2013; Centola and Baronchelli, 2015; Bell and Paegle, 2021; Guilbeault et al., 2021).

If successfully signaling to similar others is incentivized and, as is usually assumed, signaling to dissimilar others is not penalized, then signaling conventions can not only emerge, but stabilize. In other words, a prediction that follows from the logic of most models of convention is that, once established, a convention will remain conventional. However, this prediction is not aligned with many cases of signaling conventions, in which signal trends rise and then wane in popularity (Berger and Heath, 2008). Fashions may simply go out of style. However, another reason is that signaling one's identity to dissimilar others may in fact be costly. If revealing one's otherwise-hidden identity to members of an outgroup entails costs—such as those faced by members of certain ethnic minorities and religious groups, political dissidents, and LGBTQ+ individuals—then overt identification may not be worth the risk. In such cases, covert identity signals may arise.

Covert signals are accurately received by their intended audience but obscured when received by others (Smaldino et al., 2018; Smaldino and Turner, 2022). They allow individuals who share social traits to recognize one another while simultaneously allowing signalers to avoid being recognized as dissimilar by those not "in the know." Political dog whistling is perhaps the most widely known example of covert signaling, in which speakers will make references that are interpreted as innocuous by most listeners but signal more controversial commitments to insiders (Henderson and McCready, 2017). For example, former US President George W. Bush regularly decried the 1857 Supreme Court decision that denied the freed slave Dred Scott's right to file suit, tapping into the connections that conservative Evangelicals at the time made between that decision and the 1973 Roe vs. Wade decision that until recently upheld the right to legal abortion (Kirkpatrick, 2004). While some dog whistles are identified after the fact (as in the case just mentioned), their prevalence indicates that many if not most go largely undetected, even if little research has investigated the efficacy of dog whistles in remaining covert. More quotidian examples of covert signaling abound concerning the ways people implement fashion, humor, and other semiotic tools to subtly indicate identity

(Berger and Ward, 2010; Flamson and Bryant, 2013; Fischer, 2015). These signals are probably less easily detected than more overt signals, but trade clarity for the benefits of encryption or plausible deniability (Lee and Pinker, 2010). Covert signaling may be particularly important to members of persecuted minorities, such as LGBTQ+ individuals or political dissidents, who have strong incentives to assort with one another but also to avoid detection by nonmembers.

Modeling work has formalized this idea (Smaldino et al., 2018; Smaldino and Turner, 2022), indicating that covert signals should be favored over overt identity signals when being revealed as dissimilar is costly and when individuals cannot count on being able to partner only with those they prefer. These conditions are more likely to be met in more diverse societies and among those with minority-group status. A recent empirical study provides explicit support for the theory in the context of political identity signaling online, showing that Twitter users with more heterogeneous follower networks tweeted more covertly and that participants in a behavioral experiment strategically selected more covert signals when their audience consisted of more outgroup members (van der Does et al., 2022).

Covert signals work because they are known to insiders but not to outsiders. At minimum, they must be substantially less reliable as signals of identity when received by outsiders. However, the information content of a signal is not fixed. As noted above, signaling conventions can emerge dynamically as people learn to associate particular signals with particular identities. When the incentives of the signalers and the receivers are sufficiently aligned, these conventions can become stabilized and even institutionalized. When audiences are antagonistic, deceptive signals are incentivized (Crawford and Sobel, 1982). Covert signals occupy an in-between space (Smaldino and Turner, 2022), conveying honest messages to ingroup members (whose interests are aligned with those of the signaler) while deceiving outgroup members (whose interests are not aligned). Outgroup audience members may have incentives to avoid being deceived, however, and to correctly identify outgroup individuals despite their efforts to signal covertly. The police may wish to arrest dissidents, and an employer may wish to avoid hiring someone with divergent or unconventional views. The effectiveness of covert signals may therefore be more ephemeral when compared with overt signals. Once a particular signal has outlived its usefulness, new signals can arise to take its place.

The presence of covert signals provides the conditions for dynamic cycles of signaling conventions. Previous models of covert signaling focused on competition between strategies of covert vs. overt signaling, and did not explicitly consider the dynamic usage of specific signals, though prior work has speculated about the possibility of such cycles (Smaldino et al., 2018; Smaldino and Turner, 2022; van der Does et al., 2022).

Here, I explore this idea more extensively.

My proposal works as follows. In any sufficiently large population in which prior knowledge of interaction partners is not guaranteed, arbitrary signals can become reliable markers of identity when individuals learn to associate particular signals with particular identities and choose their own signals accordingly. If accidentally revealing one's identity to an outgroup individual is either very unlikely or does not carry particularly high costs, a signal can become conventionalized as a stable marker of identity despite being known to both ingroup and outgroup audiences, i.e., it becomes an overt signal. If, however, becoming known to outgroup audiences carries sufficient risk, a signal may lose its value once it is in regular use by an ingroup, as the outgroup comes to associate the signal with the ingroup. In this case, the usage frequency of a dominant covert signal may decrease, and a new covert signal may rise to dominance. This dynamic should repeat indefinitely as long as other inter- and intra-group relationships remain relatively constant. In extreme cases where detection by the outgroup is both sufficiently likely and sufficiently costly, it is possible that no reliable covert signals will ever emerge, and group members will have to rely on other means to assort. In the subsequent sections of this paper, I demonstrate the plausibility of the proposal, and examine conditions for the emergence of stable signals, cycles, and the total absence of reliable signals.

This proposal for the emergence of signaling cycles follows a logic similar to mechanisms generating cyclical group dynamics in other systems. In the social sciences, perhaps the best known is Simmel's (1904) theory of fashion. He proposed that the function of many fashion trends, which clearly serve little utilitarian purpose, is to distinguish members of the elite from members of the lower social classes. Members of the lower classes, however, are incentivized to appear more upper class, and therefore strive to copy the fashions of the elite. This in turn incentivizes the elite classes to continually innovate new fashion trends so as to stay above the rabble. Although this idea cannot explain all fashion cycles, later work has expanded upon Simmel's theory, finding some empirical support (Miller et al., 1993; Krawczyk et al., 2014) and using formal modeling to explore specific conditions for the emergence and character of fashion cycles (Pedone and Conte, 2001; Acerbi et al., 2012; Di Giovinazzo and Naimzada, 2015). My proposal is similar to Simmel's, but differs by virtue of the differential effect signals have on ingroup and outgroup audiences and in the use of a particular signal by only one group. I focus on scenarios in which members of a disadvantaged or persecuted group are trying to identify each other while avoiding detection by a hostile outgroup—this latter group is trying to punish rather than imitate them.

Cyclical dynamics can arise in any coupled two-component system in which the first component acti-

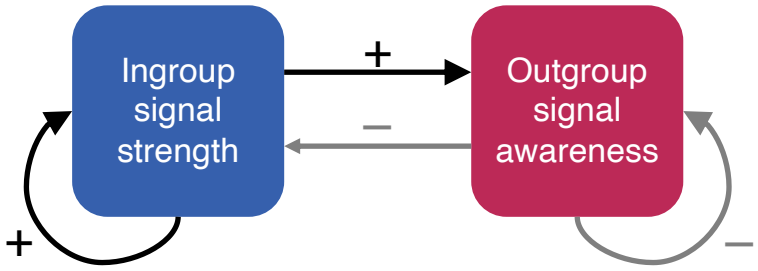

Figure 5.1: Cyclical dynamics can arise in any coupled two-component system in which the first component activates growth in both itself and the other component, and the second component inhibits growth in both itself and the other component. Here I illustrate how cycles can emerge in the usage of an ingroup signal through exposure to a hostile outgroup.

vates growth in both itself and the other component, and the second component inhibits growth in both itself and the other component (Figure 5.1). Predator-prey systems can famously exhibit these cycles, typified by the Hudson's Bay Company data on Canadian lynx and snowshoe hare populations and formalized by the classic Lotka-Volterra model (Smaldino, 2023). In this model, the prey population grows in the absence of predators and also stimulates growth in the predator population, while the predator population shrinks in the absence of prey and also inhibits growth in the prey population. When the two populations are present simultaneously, coupled cycles of growth and decline emerge, though some parameter combinations can also lead to the collapse of one or both populations. A range of other systems exhibit cycles for similar reasons, including endoparasites (Otto and Day, 2011), neural firing rates (Stiefel and Ermentrout, 2016), and even the rise and fall of empires (Turchin, 2003).

Below, I present two formal investigations of the verbal theory presented above. In the next section I present a relatively simple analytical model, adapted from a "matching alleles" host-pathogen model in which a host evolves genetic resistance to multiple pathogens. Individuals in our model learn to adopt or disadopt particular signals through feedback from interactions with ingroup and outgroup individuals. Successful coordination with ingroup partners increases the likelihood of using a particular signal again, while discovery by outgroup agents decreases that same likelihood. This model is able to capture the broad strokes of our verbal theory, and serves to demonstrate how the dynamics we propose may arise. However, the analytical model is also limited in its ability to capture important features of human cognition and communication, and so in the following section we then turn to an agent-based model, in which agents use reinforcement learning to make decisions concerning the costs and benefits associated with the use of various signals. The convergent findings from this second model provide robustness for our verbal theory.

5.2 The analytical model

Here I present an analytical model based on coupled differential equations. This model captures several core elements of the theory discussed in this paper, and also produces patterns of stability, cycling, and noise that are qualitatively similar to the agent-based model presented in the main text. Nevertheless, the complexity of the theory means that this model omits some key elements, such as explicit agent learning and the ability of multiple signals to be simultaneously expressed. In the interest of robustness and completeness, I present this model here.

Consider a signaling system in which there are a set of K possible signals, any of which can be used by an ingroup to effectively assort. The strength of a signal i among the ingroup, x_i, is its proclivity to be used by members of the ingroup, and is therefore equivalent to its frequency of use in that population. In the absence of an outgroup, popular signals will increase in strength more rapidly than unpopular ones. However, popular signals are also likely to be learned by the outgroup. I designate the strength of awareness of a signal i among the outgroup as y_i. Awareness among the outgroup decreases the utility and therefore the strength of the signal among the ingroup. However, when a signal is not in high usage, outgroup awareness will diminish to baseline levels. We can represent this dynamic as a system of coupled differential equations. The strength of ingroup signals changes as follows:

$$\dot{x}_i = rx_i \left(1 - \sum_j x_j\right) - cpy_i \frac{x_i}{1 + x_i}, \tag{5.1}$$

where r is the reward for successfully using the signal among the ingroup, and the parenthetical in the first term represents the fact that the strengths of all signals must sum to one; c is the contact rate between in- and outgroup members, p is the punishment of ingroup members from detection by the outgroup, and the fraction in the second term indicates diminishing marginal returns to punishment.

I similarly represent the change in outgroup signal awareness as follows:

$$\dot{y}_i = c\rho y_i(1 - y_i)\frac{x_i}{1 + x_i} - \delta y_i + \epsilon, \tag{5.2}$$

where ρ is the rate at which outgroup individuals learn to recognize ingroup signals, which I model as a logistic function; δ is the intrinsic decay rate of signal knowledge among the outgroup, and ϵ is the baseline

awareness of a signal among the outgroup.

For simplicity, I present model explorations with the following default parameter values: $K = 3$, $r = 1$, $\delta = 0.05$, $\rho = 0.4$, $\epsilon = 0.001$. R code to fully explore the model through numerical simulation is provided at `https://osf.io/4vcug/?view_only=14c6e87c32f146a6910a24e7dd079191`. Our focus is on the two parameters most central to our verbal theory: the punishment to ingroup individuals from outgroup detection, p, and the relative rate of contact between the ingroup and the outgroup, c. These parameters are at the core of our verbal theory described above. When p is very low, detection by the outgroup should matter little, and stable signaling conventions should emerge. As p increases, dominant signals should be replaced with increasing frequency until persistent dominance is either transient or impossible. Example dynamics for both ingroup signal strength and outgroup signal awareness are shown in Figure 5.2. Similarly, lower contact rate (c) reduces the expected cost of punishment, so that a stable, dominant signal can persist more easily for a given level of punishment, p. The left panel of Figure 5.3 shows the probability of a new dominant ingroup signal (peak) arising per unit time, estimated over the course of a simulation run lasting 10,000 time steps. Contact rates of $c > 1$ imply that ingroup individuals are more likely to interact with members of the outgroup than members of their own group. The middle panel of Figure 5.3 extends this analysis for much larger values of p for $c = 1$, showing a critical transition in which the number of peaks stops increasing with p and instead decreases sharply. This is when the overall strength of the dominant signal starts to become indistinguishable from that of the other signals. This corresponds to a suppression of effective identity signaling among ingroup members. This transition from cycling to suppression happens somewhat gradually as p increases, with the mean strength of the dominant signal decreasing gradually starting with much lower values of p (Figure 5.3, right panel).

This mathematical model captures several key features of our theory. We observe the persistence of a single dominant ingroup signal when the expected cost of outgroup interactions are low (driven by low contact rate and/or cost of punishment for detection), cycling between different signals with increasing frequency when the expected cost increases sufficiently, and the suppression of any dominant signals once the expected cost of signaling exceeds some threshold. However, this model also has several limitations. First, the model assumes that all individuals have identical knowledge of signal prevalence, and therefore identical proclivities to use those signals. This means that ingroup and outgroup knowledge *necessarily* rise and fall together. Second, the model requires that individuals' knowledge of signals must exactly track the relative frequency at which they use those signals. This means that increased knowledge of one signal automatically

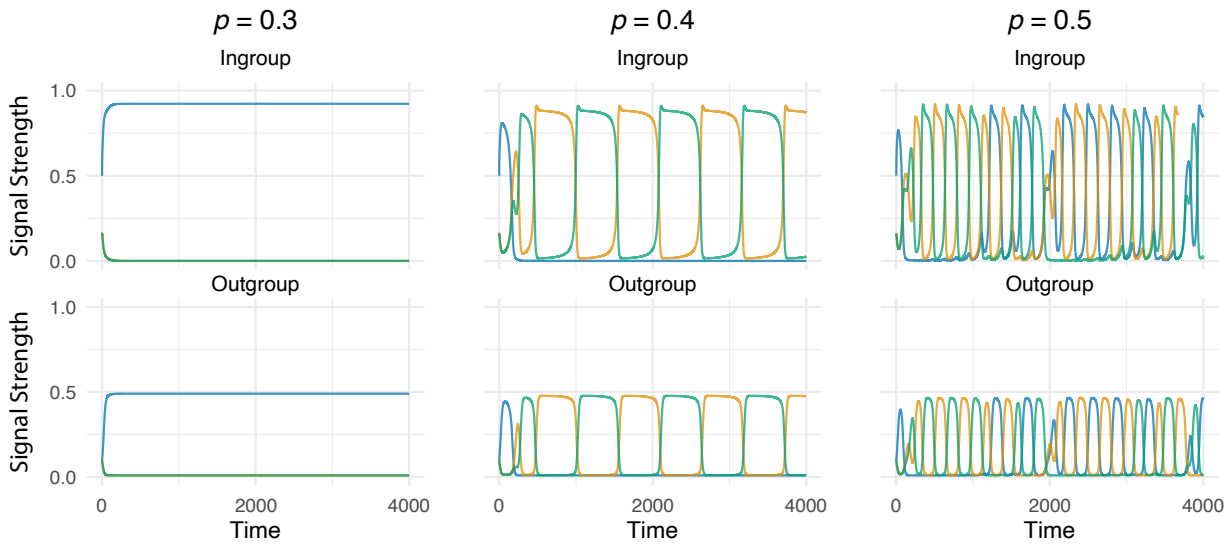

Figure 5.2: Example dynamics for ingroup signal strengths (top) and outgroup signal awareness (bottom) for different values of punishment, p. Here $c = 1$.

implies decreased knowledge of other signals. Third, the model does not allow for the emergence of novel signals. In reality, when a signal is driven out of use, human signalers should be able to replace it with a new signal that is (initially) unknown to the outgroup. Finally, the analytical model represents human learning as a process of selection at the population level. While this sort of abstraction is useful, it must still be tested against more individual-level cognitive mechanisms for learning to verify that the behavior of individual learners can in fact be accurately represented by the mean-field approach described above.

5.3 The agent-based model

Consider a population of agents that interact with others agents both from their own ingroup and from an outgroup. As in the mathematical model, I focus on one particular group and its members' relationships with both ingroup and outgroup partners. Each interaction is an opportunity for identity signaling in which they can potentially identify a fellow ingroup member. Signals are arbitrary, and therefore the association between a particular signal and its meaning confirming the sender as an ingroup member must be learned. When signals become established as reliable indicators of ingroup status, they help members find each other and receive the benefits of cooperation and coordination. When an agent receives a signal it believes communicates ingroup status, it may take a risk and overtly declares its identity status to its partner in the hopes of initiating a successful coordination. This sort of declaration is necessary to confirm similarity and receive benefits. The use of arbitrary signals is both methodologically convenient and theoretically motivated. From

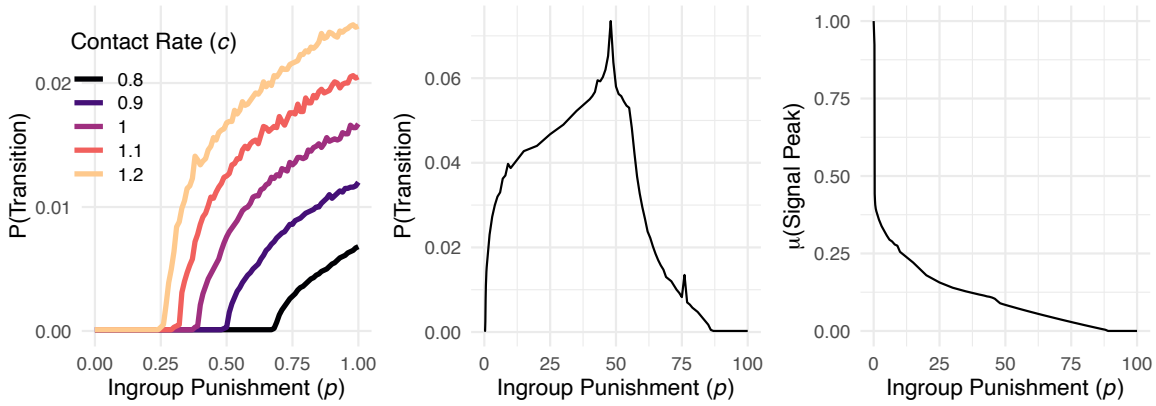

Figure 5.3: Summary results from numerical simulations of the analytical model. Left: Probability of a new dominant ingroup signal, P(Transition), as a function of p, for several values of c. For low values of p and c, there is only one stable peak. Middle: P(Transition) for much larger values of p. As p approaches a critical value just below 50, the rate of cycling stops increasing and begins to decrease. Right: Cycle amplitude diminishes as p increases. When p gets very large, all signals are suppressed. In both the middle and right panels $c = 1$.

a practical standpoint, the model represents the simplest possible system, and so arbitrary signals are used in principle, making them the standard for both signaling and evolutionary (genetic/phenotypic) models. Work in semiotic theory also shows that many symbols are fairly arbitrary and gain meaning only in reference to their communicative association with aspects of reality and with other symbols (Tylén et al., 2013). In this sense the model's arbitrary signals are realistic, as meaning emerges through association of symbols with group identity. That said, many signals operate through contextual and referential indicators, and these aspects are not captured by my model.

Outgroup individuals also learn to associate certain signals with group membership. If an outgroup member receives a signal from an ingroup member and chooses to identify that agent as a member of the ingroup, the outgroup agent learns that the focal agent's signal is associated with ingroup membership. If a subsequent ingroup member uses this signal with that outgroup member, the outgroup member is more likely to recognize its ingroup status and identify the agent as a member of the ingroup. Depending on whether this outcome is costly, the ingroup agent may become less likely to use the same signal in the future. Note that for convenience, I consistently refer to the group that is motivated to signal covertly as the ingroup and to the other group as the outgroup, though of course in reality individuals will typically conceptualize members of their own group as "ingroup" and members of other groups as "outgroup."

The model dynamics proceed in discrete time steps, each of which consists of five phases: an ingroup

coordination phase, the first signal repertoire update phase, an outgroup detection phase, the second signal repertoire update phase and a signal extinction/emergence phase. A simplified schematic is given by Figure 5.4, which omits the extinction/emergence phase. I code distinct phases for ingroup and outgroup signaling largely to simplify the code and model description. In a real system the processes in all these phases would occur concurrently. Simulations run with the two phases interspersed (not reported here) confirmed my prediction that this modeling decision did not qualitatively alter the model dynamics or outcomes. Due to the complexity of the model, I describe the initialization conditions as I introduce each parameter. A full list of model parameters and their default values is shown in Table 1. Python code for the model dynamics and analysis is available at `https://osf.io/4vcug/?view_only=14c6e87c32f146a6910a24e7dd079191`.

5.3.1 Ingroup coordination interaction phase

I consider a population of N_{IN} agents, all members of a group, which I will refer to as the ingroup. Each ingroup agent i is characterized by a *repertoire* of K identity signals, defined as a vector of signal weights $\mathbf{S}_i^{\text{IN}} = \{s_{i1}^{\text{IN}}, s_{i2}^{\text{IN}}, \ldots, s_{iK}^{\text{IN}}\}$. Each signal weight s_{ik}^{IN} represents the informational value of signal k to a particular agent i. Each signal in the repertoire of each ingroup agent i is bounded in [0, 1] so that the theoretical maximum of $\sum_k^K s_{ik}^{\text{IN}} = K$. Upon initialization of the model, every ingroup agent is instantiated with an identical signaling repertoire, such that $\forall(i,k), s_{ik}^{\text{IN}} = 0.1$.

At each time step, each ingroup agent is paired with a partner, randomly chosen from among the remaining ingroup agents such that each agent has exactly one ingroup partner. Each agent chooses a signal to send to its partner from a normalized vector, $\mathbf{S'}_i^{\text{IN}} = \frac{s_{ik}^{\text{IN}}}{\sum_k^K s_{ik}^{\text{IN}}}$, such that a signal k is chosen by agent i with a probability equal to its relative informational value, s'^{IN}_{ik}. Based on their respective partners' signals, both agents independently decide whether to overtly declare their identity. This decision is made with a probability equal to the absolute informational value, s_{ik}^{IN} that a signal's receiver had previously assigned to their partner's signal. If neither agent declares their identity, they are presumed not to recognize one another as ingroup members and so nothing happens. If either partner does declare their identity, then each agent learns to associate both the sent and received signals with the ingroup identity, and therefore increases the informational value of those signals by ρ_{IN}, such that $s_k^{\text{IN}} \leftarrow \rho_{\text{IN}} + s_k^{\text{IN}}$. Although this phase consists entirely of interactions between ingroup members, the entire set of phases are intended to represent a mixture of relatively concurrent interactions, and so agents have no *a priori* reason to assume their interaction partners are fellow ingroup members. This is a minimal assumption given that the probability of choosing a signal

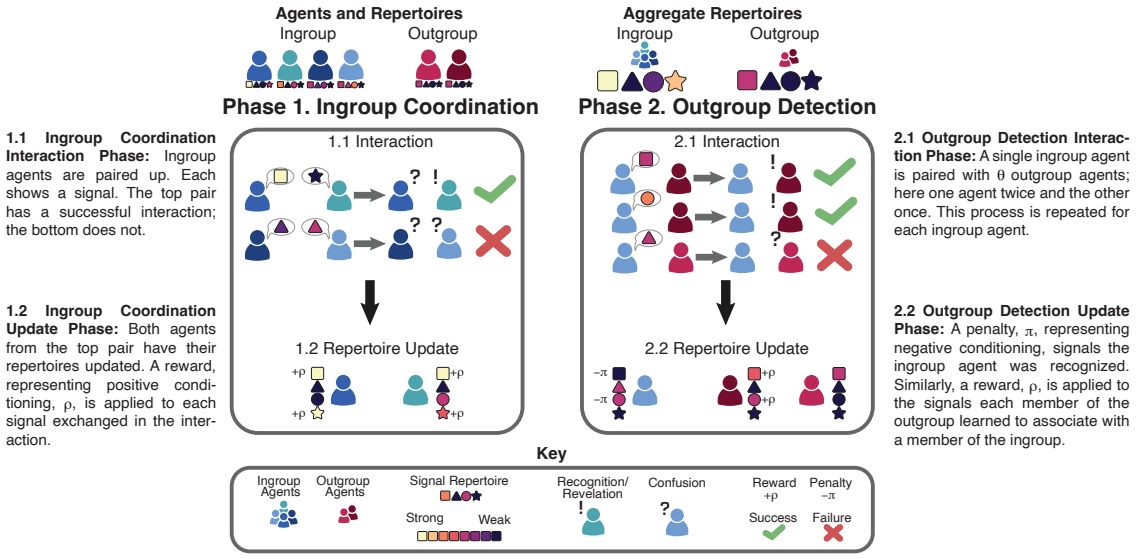

Figure 5.4: A schematic of the agent-based model representing a single time step. Signaling repertoires are denoted by the array of colored shapes. Each agent has its own signaling repertoire, distinguished by the different colors for each shape, indicating the informational value of each signal for that particular agent at that time. Stronger signals are lighter, weaker are darker. Aggregate signaling repertoires, the primary outcome of the model, are also pictured, representing informational values averaged across all agents. Not pictured is the extinction/emergence phase where signals that have risen and then fallen to particular thresholds in the ingroup population are reset to their initial values in the ingroup and outgroup. This represents the emergence of a novel signal at the location in the signal repertoires that was occupied by a newly extinct signal.

should be an increasing function of the signal's perceived informational value for other members of one's ingroup.

5.3.2 Ingroup coordination repertoire update phase

Following the coordination phase, the signaling repertoire of each ingroup agent is updated to account for new information acquired. First, each agent's signal repertoire is updated to account for any rewards accumulated during the coordination phase, as described above. Then, each signal is truncated such that any value great than 1 is set to 1 and any value less than 0 is set to 0.

Without truncation, values greater than one would tend towards positive infinity, and values less than 0 would tend towards negative infinity. Indefinitely increasing or decreasing signal values could result in runaway feedback loops, which would make model evaluation more challenging. However, there are theoretical reasons for constraining the informational value of the signals. From the perspective of signal interpretation, identity signals represent the probability of identity, and are thus bounded in $[0, 1]$, so that signals that are never used have a value of 0, and signals that reliably indicate ingroup membership have

a value of 1. From the perspective of signal generation, relative signal salience should match interpretive probabilities. If two signals are equivalent predictors of identity, and the agent may select only one, there is no reason to prefer one to the other.

5.3.3 Outgroup detection interaction phase

After the ingroup agents interact amongst themselves and update their signal repertoires, they enter another round of signaling. Here the ingroup agents are partnered with members of an outgroup. The outgroup consists of N_{OUT} agents, each of which also tracks the information value of each signal that may be used by the ingroup. Each outgroup agent j is therefore characterized by a signal repertoire of length K, $\mathbf{S}_j^{\text{OUT}}$. Instead of using these signals amongst themselves, outgroup agents use their information to identify members of the ingroup. The strength of a particular signal for a given outgroup agent is the probability that they will correctly identify another agent using that as a member of the ingroup. I assume the outgroup is initially naïve about all signals, and so we initialize each signal in an outgroup agent's repertoire to a strength of 0.05.

At each time step each ingroup agent interacts with up to θ random outgroup agents, where $\theta \geq 0$. If θ is not an integer, the ingroup agent calculates a probability $p = \theta - \text{floor}(\theta)$, and interacts with $\text{ceil}(\theta)$ with probability p and $\text{floor}(\theta)$ with probability $1 - p$. For example, if $\theta = 3.2$, the agent will interact with four outgroup agents with probability 0.2, and three outgroup agents otherwise. We can therefore explore cases where the ingroup interacts mostly among themselves ($\theta < 1$), and mostly with the outgroup ($\theta > 1$), as might be the case for a minority group.

During each interaction, ingroup agents select their signal as before. Each ingroup agent i chooses a signal k from its repertoire with a probability equal to the signal's normalized informational value, s'^{IN}_{ik}. The outgroup agent j then recognizes the selected signal with a probability equal to the value of the same signal in its own repertoire, s^{OUT}_{jk}. If it identifies the agent as a member of the ingroup, the outgroup agent strengthens the signal's association with the ingroup, and adds ρ_{OUT} to its representation of the signal's information value. Additionally, the outgroup agent may behave in a manner to sanction the ingroup member. The member of the ingroup receives a penalty π that is subtracted from the signal's informational value, which may disincentivize the ingroup agent from using the same signal again in future interactions. This is analogous to real-world sanctions for revealing a marginalized identity, ranging from a cold shoulder, verbal abuse, or an employer reprimand to physical violence, political imprisonment, etc.

5.3.4 Outgroup detection repertoire update phase

During this stage, signal weights are updated as a result of interactions between the ingroup and the outgroup. Both the ingroup and outgroup update their signal repertoires according to the same process, which is identical to the process for the ingroup coordination update phase. First, each agent's signal repertoire is updated to account for any rewards accumulated during the coordination phase, as described above. Then, for each agent, each signal is truncated such that any value great than 1 is set to 1 and any value less than 0 is set to 0.

5.3.5 Signal extinction/emergence phase

In real cases of identity signaling, the number of novel signals that a population may develop is effectively infinite. However, the constraints of my model demand we simplify to a small number of arbitrary signals, K. We simulate the emergence of genuinely novel signals by imposing extinction on a signal when its average informational value among ingroup agents decreases below a threshold, which we set to 0.05. When the extinction threshold is reached, the signal is reset to its initial value of 0.1 among the ingroup. Importantly, the outgroup signal knowledge is also reset to its initial value, 0.05, as the arbitrary ordinal position in the signal array now represents a new signal, of which the outgroup has minimal knowledge. This results in differing cycles of signal renewal in the ingroup and outgroup. The ingroup signals decline gradually as a result of punishment by the outgroup, whereas outgroup signals have sharp vertical drops, as seen in Figure 5.7.

Importantly, prior to extinction, a signal's average informational value among the ingroup must first exceed a "prevalence" threshold, which we set at 0.15, indicating it is widely acknowledged as an identity marker among the ingroup. Only after surpassing this level can a subsequent drop below the extinction threshold initiate a reset. Setting the prevalence threshold too low results in signal extinction before becoming recognized as an identity marker, usually as a result of random noise (drift).

This mechanism of signal renewal approximates recurrent novelty in real identity signaling. Because outgroup knowledge is reset when an ingroup signal goes extinct, that signal should subsequently be interpreted as a completely new signal despite being indexed by the same number. So, even though K is small and finite, the model effectively generates new signals continually. In fact, even $K = 1$ is capable of cycling in the model. However we set $K = 3$ in my analysis in order to make cycling easier to interpret for the

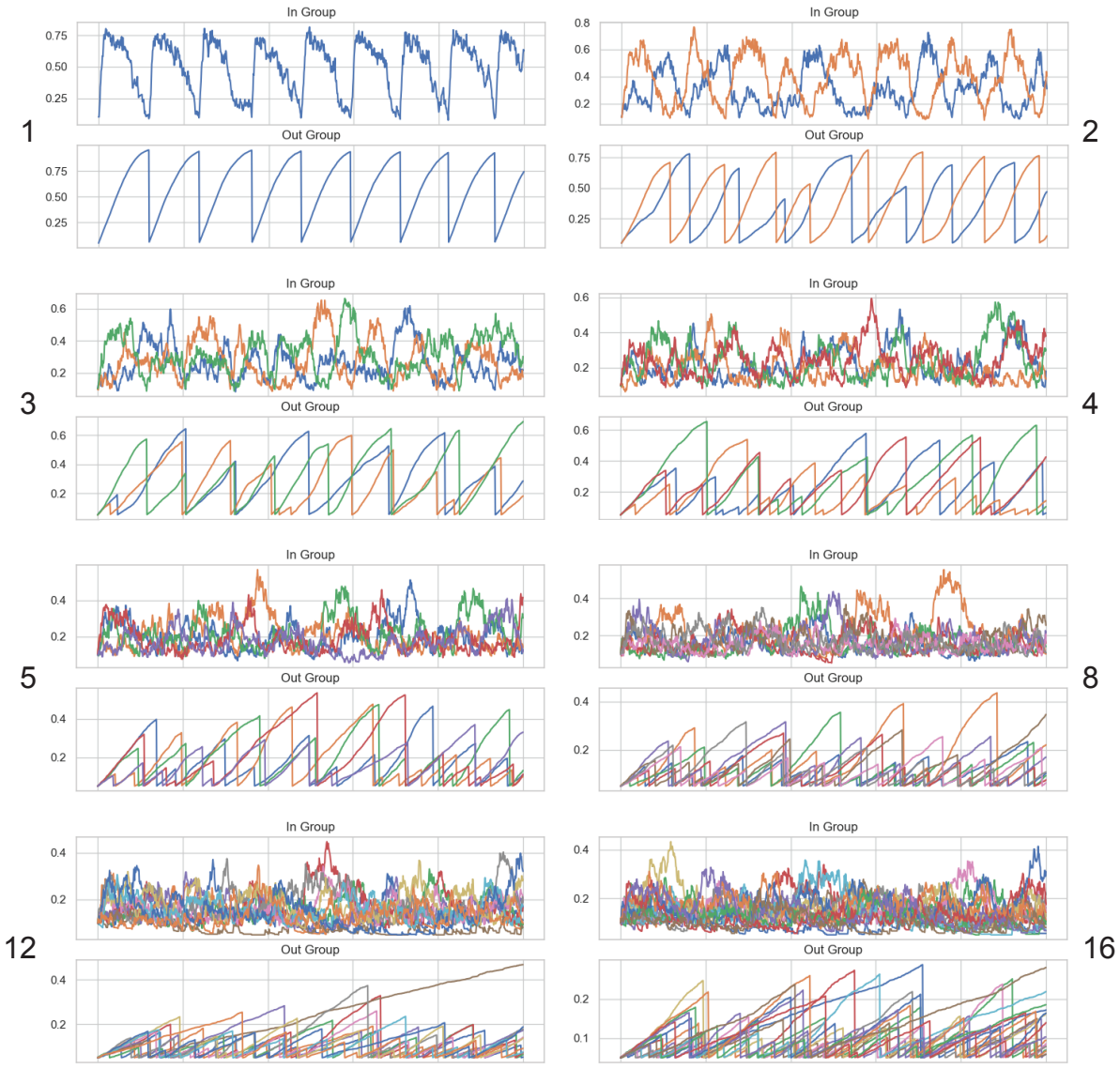

Figure 5.5: Signaling dynamics sweeping K. All other values set to the defaults given by Table 5.1

reader. Figure 5.5 demonstrates cycling signaling dynamics with between one and 16 signals.

While signal renewal is best understood as novel signal emergence, one could also interpret a new signal as the return of a signal which has been out of use long enough to lose salience with the outgroup to return as an effective identity signal among the ingroup. It may be that this signal never truly left the vocabulary of the ingroup, and was only removed from contexts where the audience identity is uncertain; thus, formerly covert signals may be relegated to "safe spaces", and reappear after a long period of dormancy, especially after generational turnover. While my design doesn't model such mechanisms directly, the process is effectively

Table 5.1: Parameters for the signaling model.

Parameter	Symbol	Variable[a]	Default Value
Ingroup size	N_{IN}	No	50
Outgroup size	N_{OUT}	No	500
Number of signals	K	No	3
Length of simulation runs	T	No	1000
Ingroup reward	ρ_{IN}	Yes	0.5
Outgroup reward	ρ_{OUT}	Yes	0.2
Ingroup penalty	π	Yes	0.3
Out/Ingroup interaction ratio	θ	Yes	1
Ingroup initial signal value		No	0.1
Outgroup initial signal value		No	0.05
Ingroup signal prevalence threshold		No	0.15
Ingroup signal extinction threshold		No	0.05

[a] Parameter value changes between model runs in this study.

approximated by the extinction mechanism.

5.3.6 Outcome measures

We ran each simulation for $T = 1000$ time steps, performing 50 runs for each combination of parameters. This number of runs is justified both because there was very little variation in classification between runs using any particular combination of parameters, and because 1000 time steps was more than enough time to observe the model dynamics settle into stable, long-term behavior patterns (see Figure 5.7). In some cases (very infrequently) observed but not explored, cycles may be too wide to appear in 1000 time steps. This occurs when agent learning is low but balanced with ingroup punishment. We are content knowing that such cases exist and how one might locate them model in parameter space. These cases are not only marginal, but do not exhibit fundamentally different dynamics than those explored here; only the rate of the dynamics differs.

The primary outcome measure of interest is the informational value of each signal (i.e., the signaling repertoires) for each agent over time. Specifically, we are interested in the conditions under which the all signals become effective and sustained identity markers, those under which cyclical dynamics would emerge, and those in which no signal ever attains (or retains) widespread use. Ultimately, it is which of these three categorical outcomes (saturation, cycling, or suppression) that we are really interested in. To categorize the model outcomes into one of these three classes, I relied on visual inspection of the model

outcomes, i.e., time series of aggregate ingroup and outgroup signal informational values.

I also trained a random forest model on 11 time series features to automatically identify the three classes of dynamics. I manually classified 100 randomly selected parameter combinations (10 runs per combination) from across the ranges defined in Table 5.1. Holding out 20% of parameter settings, cross-validation of random forest models achieved average accuracy of 0.96. This indicates that the criteria used for model outcome classification can be reliably automated.

The random forest model was trained on a set of time series features. I manually classified 100 randomly selected parameter combinations (10 runs per combination) from across the ranges defined in Table 5.1. Holding out 20% of parameter settings, cross-validation of random forest models achieved average accuracy of 0.96. Figure 5.6 reports the results of Shapley decomposition, an analytical tool for estimating the contribution of features to the output of machine learning models.

The human-defined features were selected to mimic the authors' intuitive process for identifying the dynamical class exhibited by a particular run of the model. Most of these features are derived from the mean (μ) and standard deviation (σ) of each signal's informational value summed and normalized across all agents across all time steps within a model run. This is calculated separately for the ingroup and outgroup. Within a single model run, I use the following 5 features calculated for both ingroup and outgroup to train the random forest model and 1 feature calculated for only the ingroup for a total of 11 features: (1,2) the mean signal value across all signals, $\mu(\mu)$; (3,4) the mean signal standard deviation, mean(σ); (5,6) the mean signal value of a signal peak, μ(Signal Peak); (7,8) the standard deviation of all signal peak values, σ(Signal Peak); (9,10) the probability that a signal goes extinct per time step (definitionally the same for in- and outgroup) P(Extinction); and (11) the proportion of ingroup interactions which successfully resulted in rewards for the agents, % Ingroup Success.

5.4 Results

5.4.1 Characterizing model outcomes

Visual inspection of the model output led us to consistently identify each model run in terms of one of three classes of behavior: *saturation, cycling*, or *suppression*. Figure 5.7 shows example runs in which each of these three classes of dynamics arose. Each column shows three example runs, with each example involving two graphs: the signal strengths for the ingroup (top) and outgroup (bottom). Each colored line represents the average weight a particular signal among either the ingroup or outgroup agents.

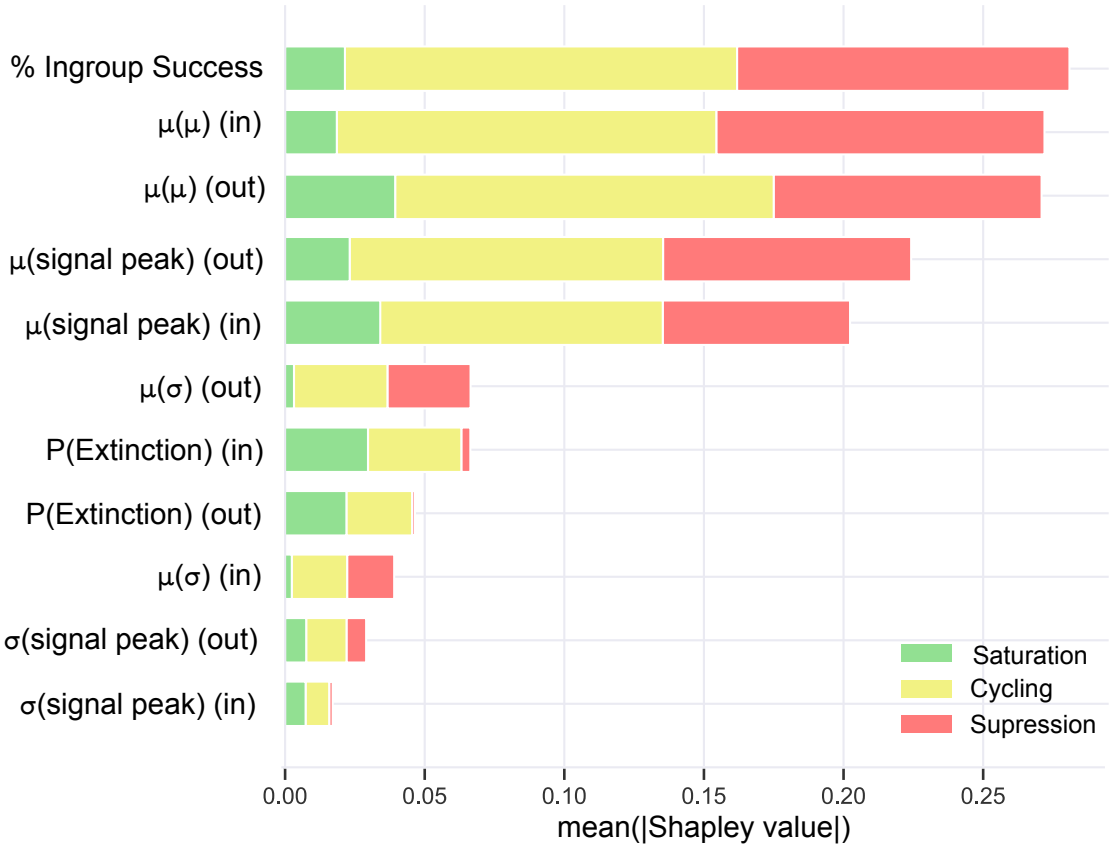

Figure 5.6: Shapley decomposition of the random forest model. Length of each bar shows the contribution of the feature to determining the probability of each class of model dynamics. A long bar in a particular color indicates that the feature is highly informative in determining whether it belongs to the class associated with that color. A short bar indicates that the feature is not very informative.

In instances of saturation, all signals become effective identity markers and remain so, despite occasional penalties from being identified by outgroup agents. This differs from the analytical model, which prevents multiple signals from becomingly strongly associated with the identity concurrently. In this regard the agent-based model is considerably more realistic, as groups rarely rely on a signal signal to communicate group membership.

When identification by the outgroup becomes costlier, the dominant signal may be sufficiently disincentivized once the outgroup learns to associate it with the ingroup, and so ingroup agents begin to use alternative signals. Through reinforcement learning, they converge on a new conventional signal with which to identify other ingroup agents. Once conventionalized, however, the new dominant signal becomes a new target for outgroup learning. When the outgroup learns the new ingroup signal, ingroup agents are once again forced to abandon it. A new signal emerges as dominant, the cycle begins anew. Note that in the

model, when the ingroup abandons a signal, the outgroup knowledge of that signal is also reset, effectively generating a novel signal about which the ingroup and outgroup have minimal knowledge. Thus, despite the finite and small number of signals in the model, the model should be interpreted as capable of generating infinitely novel signals across a run, and also of recycling an old signal which has lost its salience in both populations, as often happens in fashion cycles. In other words, the parameter K represents not the total the number of possible signals, but rather the maximum number of signals used *simultaneously* at any given time.

The final class occurs in cases where the outgroup learns quickly and when being identified as a member of the ingroup is severely punished. Here, no signal ever dominates or becomes informative as a group marker. Rather we see a pattern of suppression. In this circumstance all signals are equally likely and none carries identity information.

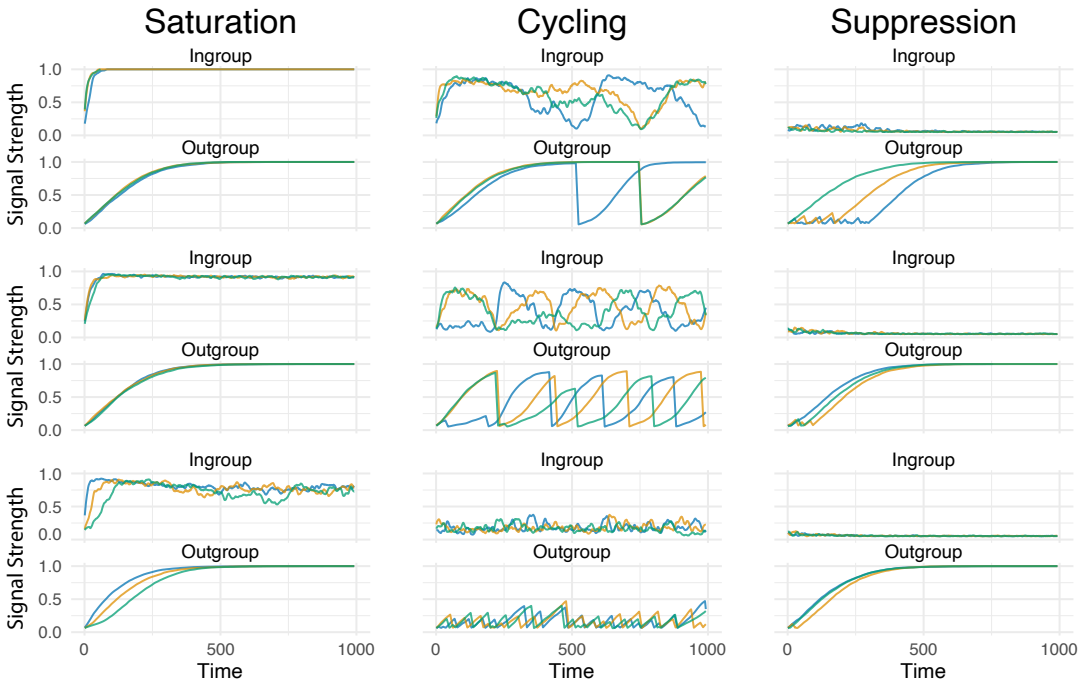

Figure 5.7: The three classes of model dynamics. Examples were selected by nonrandomly sampling runs of a parameter sweep of ingroup punishment (π). Time series occur over 1000 time steps, and data are smoothed using a moving average with a 10-time step window. The remaining parameters used in these runs are given by the Default Parameters column of Table 5.1.

I also defined three features of the model runs that correspond with theoretically relevant features of covert signaling dynamics. The first, P(Extinction), is the probability that a signal that has become prevalent in the ingroup population goes extinct and is replaced by the emergence of new signal. This corresponds to

the complete "life cycle" of a covert signal. I also measure μ(Signal Peak), the mean height of the highest value a dominant signal reaches before it is replaced by a new dominant signal. The final metric is P(Ingroup Success), the probability that an ingroup signaling interaction results in successful coordination. Each of these features is included in training the random forest classifier employed in some of the analyses.

Readers may notice parallels between these measures and the results given in Figure 5.3. μ(Signal Peak) is calculated and reported for both the agential and analytical models. P(Extinction) is analogous to P(Transition) in the analytical model; P(Transition) is also measurable here, but P(Extinction) is a more theoretical meaningful metric, as it corresponds to signal abandonment and novel signal emergence, rather than resurgence of an existing signal. P(Ingroup Success) has no analogue in the analytical model, as signaling behavior is effectively abstracted away and changes in frequency are based entirely on "signal population" sizes.

5.4.2 Characterizing the parameter space

With our analytical methods established, we can say more about the relationships among the model parameters in producing each the three categorical model outcomes. The model produces three distinct classes of behavior: saturation, where all identity signals become stable conventions and never go extinct; cycling, in which new signals emerge to replace old ones as they are learned by the outgroup; and suppression, in which informative identity signals never emerge. Cycling exists in the regions of parameter space between saturation and suppression. We can draw an analogy to phases of matter. Saturation is gas. Then, as the (social) pressure increases, we observe transitions to cycling (liquid) followed by suppression (solid), so that the state of matter changes in response to the expected costs of outgroup interaction. Holding all other parameters constant, as we sweep a particular parameter, the dynamics are driven towards or away from the neighboring class of dynamics. For example, parameter settings that produce saturation will be driven towards cycling as the expected cost of interacting with the outgroup is increased, and past cycling into suppression if it is increased further. Similarly, parameter settings that produce cycling will be driven into saturation as the ingroup reward is increased, and into suppression as the reward is decreased. The tendency for each parameter to drive the system toward saturation or suppression as it is increased is given by the "Dynamic Tendency" column of Table 5.2. Generally, saturation occurs under conditions that favor ingroup coordination or disfavor outgroup detection, and suppression occurs under conditions that favor outgroup

detection or disfavor ingroup coordination. Put another way, suppression occurs under conditions where we expect a high cost of outgroup interaction relative to the benefit of ingroup interaction. Cycling occurs in the ranges where the expected payoffs from outgroup and ingroup interactions are more balanced. Figure 5.8 shows that for the default parameter values (given by the rightmost column of Table 5.1), most parameter settings in the sweeps result in either suppression or cycling.

Table 5.2: Four key model parameters and the effect of increasing them. Increasing each parameter drives the model toward either saturation or suppression (given by the "Class Tendency" column) when all other parameters held constant. Between these two equilibria lies cycling dynamics.

Parameter		Class Tendency
ρ_{IN}	Ingroup reward	Saturation
ρ_{OUT}	Outgroup reward	Suppression
π	Ingroup punishment	Suppression
θ	Out/Ingroup interaction ratio	Suppression

While the model has many parameters, I focus initially on four, given by Table 5.2. For two of these, predictions exist based on prior models of covert signaling (Smaldino et al., 2018; Smaldino and Turner, 2022). These models predict that, compared with a strategy of overt signaling (in which a stable dominant signal is plausible), covert signals (which are costly when detected by outgroup individuals) will be favored when the risk of detection by the outgroup is large and the cost of such detection is high. In the current model, the ingroup penalty, π, represents the cost of detection, while the ratio of outgroup-to-ingroup interactions, θ, represents the overall risk of detection by the outgroup.

Figure 5.8 shows the outcome variables P(Transition), μ(Signal Peak), P(Ingroup Success), and P(Outgroup Success) as we sweep across the parameters of interest (Table 5.2). Colored bands indicate which class of dynamics the model is exhibiting at each parameter setting. Figure 5.8 illustrates the tendency for each parameter to lead to one or the other pole of dynamics (saturation or suppression). Increasing ingroup reward decreases the frequency of cycling, though under these parameter settings it does not reach full saturation. Increasing the other parameters pushes the dynamics toward suppression. More specifically, ingroup punishment, which represents negative reinforcement, formalizes the outgroup's ability to sanction the ingroup when recognized. Without any such sanctioning, the initially-dominant signal can persist; at least some ingroup punishment is required to produce cycling. However, if the ingroup punishment becomes too great, cycling dynamics will give way to suppression. This is also the case for the ratio of outgroup to ingroup interactions, which represents the relative ability of the ingroup to assort preferentially among themselves

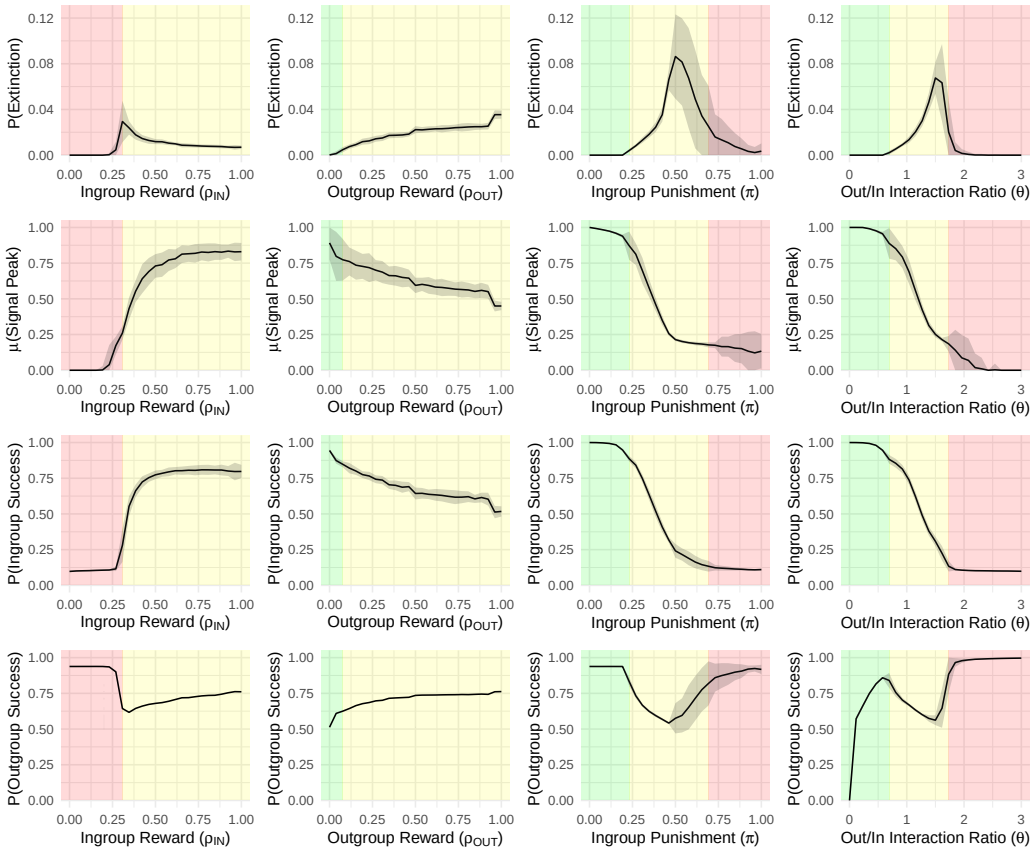

Figure 5.8: Sweeps of theoretically significant parameters holding all others constant. Each plot shows the relationship between a parameter and a feature of model dynamics. 95% confidence intervals are demarcated by the shaded gray band on either side of the line. Dynamic classes are shown by the shaded rectangles: green, saturation; yellow, cycling; red, suppression. Each column corresponds to 1 of the 4 parameters of interest. The *x*-axes show the parameters by which they differ: ingroup reward (ρ_{IN}), outgroup reward (ρ_{OUT}), ingroup punishment (π), and outgroup/ingroup interaction ratio (θ). Rows correspond to four outcome features. Top: P(Extinction), the probability that a signal goes extinct and restarts, corresponding to the period of the cycles. Upper Middle: μ(Signal Peak), the average knowledge of a signal among the ingroup, corresponding to the amplitude of the cycles. Lower Middle: P(Ingroup Success), the proportion of ingroup interactions which resulted in one or both of the ingroup agents recognizing their partner's signal and revealing their own identity. Bottom: P(Outgroup Success), the proportion of ingroup-outgroup interactions which resulted in the outgroup member identifying the ingroup member.

(even if they don't always know it), and avoid too many potential encounters with the outgroup. Decreasing the relative number of outgroup interactions produces saturation: the ingroup is able to coordinate and is rarely sanctioned. As the rate of interaction with the outgroup increases, we see cycling dynamics: the outgroup punishment forces the ingroup to periodically abandon their dominant signal and learn to coordinate around a new signal. Note P(Outgroup Success) mirrors P(Extinction), illustrating how cycling is an adaptive response to moderately costly detection by the outgroup. When outgroup interaction grows too frequent, however, the ingroup cannot effectively coordinate at all as the sanctions wipe out any progress

toward converging on a signal.

The simulations represented in Figure 5.8 were initialized with all ingroup and outgroup signals set to a low value, 0.1 and 0.05 respectively. Given that cycling emerges under conditions intermediate between saturation and suppression, one might wonder whether it is a merely transitory state and not actually stable, tending eventually toward saturation or suppression. To show that cycling is *indeed* a stable state, we ran two additional sets of parameter sweeps from initial conditions corresponding to saturation- and suppression-like states. Figure 5.9 demonstrates the dynamics from initial saturation-like conditions, in which all ingroup signals were initialized to 1 instead of 0.1. This figure shows nearly identical patterns to Figure 5.8, indicating that within parameter settings that lead to cycling, cycling is a stable attractor from both neutral and saturation-like regimes. These results differ, however, in the wide confidence interval indicating that sometimes 1000 time steps are sufficient to shift the signals away from 1, and sometimes they are not.

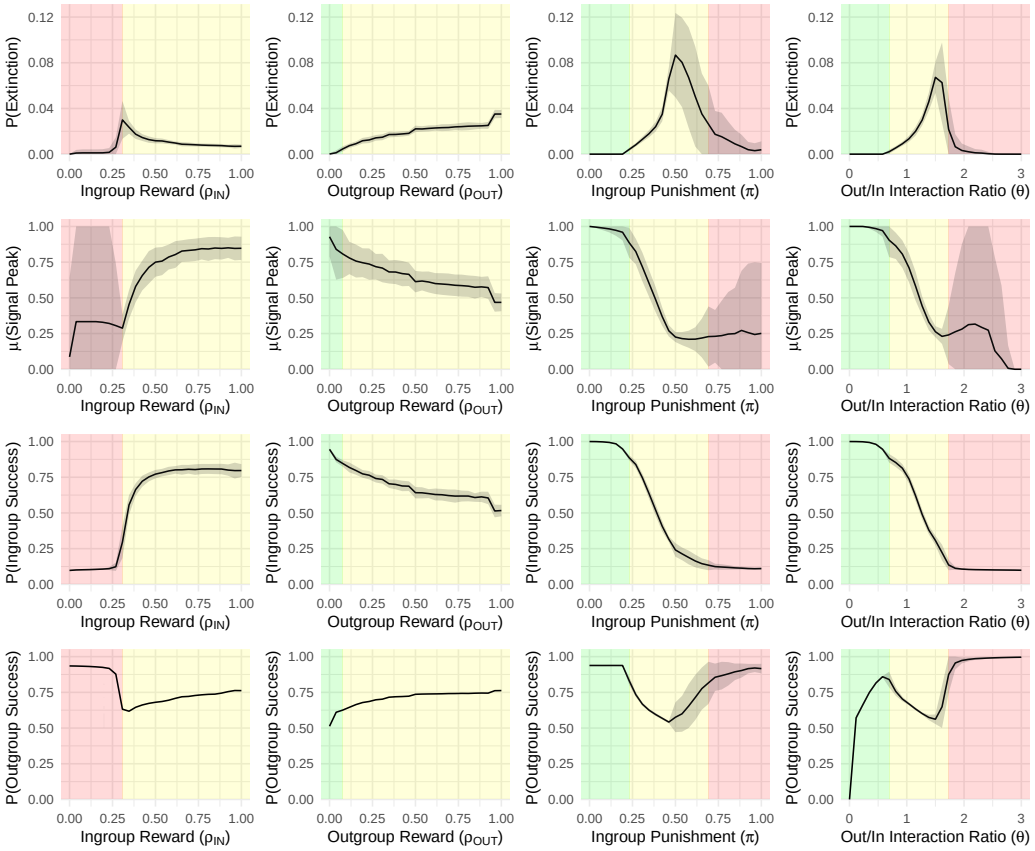

Figure 5.9: Sweeps of theoretically significant parameters holding all others constant as in Figure 5.8. Unlike Figure 5.8, which initializes with suppression-like conditions, wherein all ingroup signals are 0.1, these simulations initialize with saturation-like conditions, wherein all ingroup signals are 1. The patterns observed are nearly identical to Figure 5.8, differing only for the suppression regime for μ(Signal Peak).

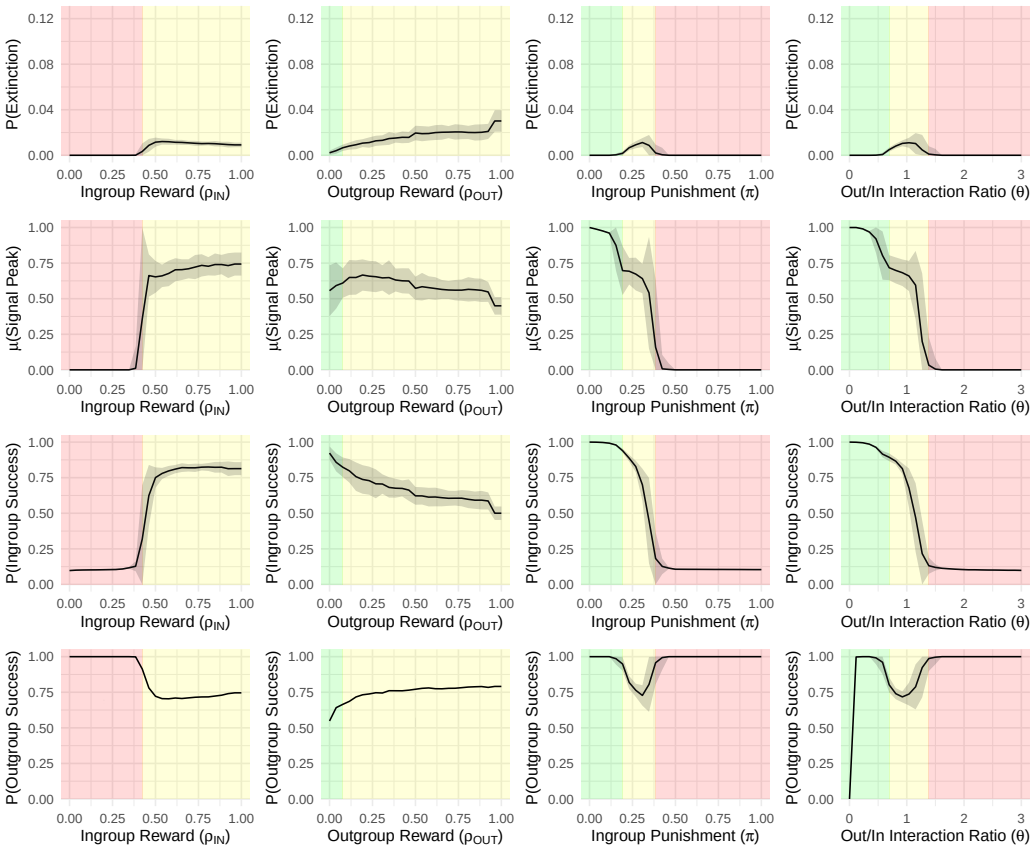

Figure 5.10: Sweeps of theoretically significant parameters holding all others constant as in Figure 5.8. Unlike Figure 5.8, which initializes with neutral conditions, wherein all ingroup signals are 0.1, these simulations initialize with suppression-like conditions, wherein all outgroup signals are 1. The patterns observed are similar to Figures 5.8 and 5.9, but differ notably due to artifacts of model design.

Similarly, to show that cycling is a stable attractor from a suppression-like state we executed a set of model runs that initialized the outgroup signals to 1 and the ingroup signals to 0.1. These results are depicted in Figure 5.10 and are likewise similar to the results in Figure 5.8, though they differ slightly as some parameter combinations that typically result in cycling dynamics remained suppressive. The discrepancy is due to artifacts of model design. First, P(Extinction), or rate of cycling, is lower throughout the regions of parameter space that produce cycling. Second, the region of parameter space that produced cycling in companion plots has partly given way to suppression. The cause of both discrepancies is similar and stems from the mechanism for signal extinction and novel signal emergence. Both in a real system and our model the ingroup will develop a novel signals unable to use existing ones. However, in our model novel signals emerge only after reaching a prevalence threshold in the total population, here set to an average informational value of 0.15. In some cases, ingroup reward is not sufficient to overcome ingroup penalty at the interaction

rate to reach that 0.15 threshold, and thus no signals can be renewed and the system remains in a state of suppression. Similarly, the rate of extinction is low even when cycling because it takes a substantial portion of the 1000 time steps to reach the threshold and enter cycling. However, these are both consequences of the model design and number of time steps; cycling is a stable attractor from this suppression-like state.

Despite these discrepancies, the results of Figure 5.10 indicate cycling appears to be a stable attractor from a suppression-like state. Notably, the suppression-like state used for the initial conditions in these model runs is highly artificial. When the model is run with parameter settings that produce cycling from conditions that are not close to suppression (i.e., at least one signal is not well-known to the outgroup), the model is extremely unlikely to reach a suppression-like state by random drift. Moreover, this scenario—in which no initial signals are unknown to the outgroup, as opposed to arriving at suppression from some other state—is outside of the range of conditions intended to be captured by the model. In reality, we imagine that members of a group could always put forth a novel signal for consideration as a group marker.

An alternate perspective is given by Figure 5.11, which examines the effects of interacting parameters. Again we sweep the four parameters of greatest interest, this time at four regular intervals in $[0.25, 1]$, giving 256 (4^4) parameter combinations (10 runs each). Each cell of figure 5.11 indicates the random forest classifier's average predicted probability that it is observing each class. The predicted probability should be thought of not as the probability of observing each class, as the dynamics are consistent across runs at a particular parameter setting, but rather the classifier's "confidence" that it is observing a particular class of dynamics. Figure 5.11 demonstrates that parameters interact additively. There is a generally continuous gradient from saturation (green), through cycling (yellow), to suppression (red). As three of the parameters—outgroup reward, ingroup punishment, and interaction rate—tend toward suppression as they increase, the gradient appears to flow toward suppression in an upward and rightward linear vector. Within each 4×4 subgrid, truly continuous gradients flow upward and *leftward* as ingroup reward tends toward saturation as it increases.

5.5 Discussion

Humans living in diverse societies require conventionalized signals to help them recognize and assort with similar partners, as well as to avoid or even (in some cases) aggress against dissimilar others. When group boundaries are clearly defined and delineate the structure of social interactions, the identity signals that

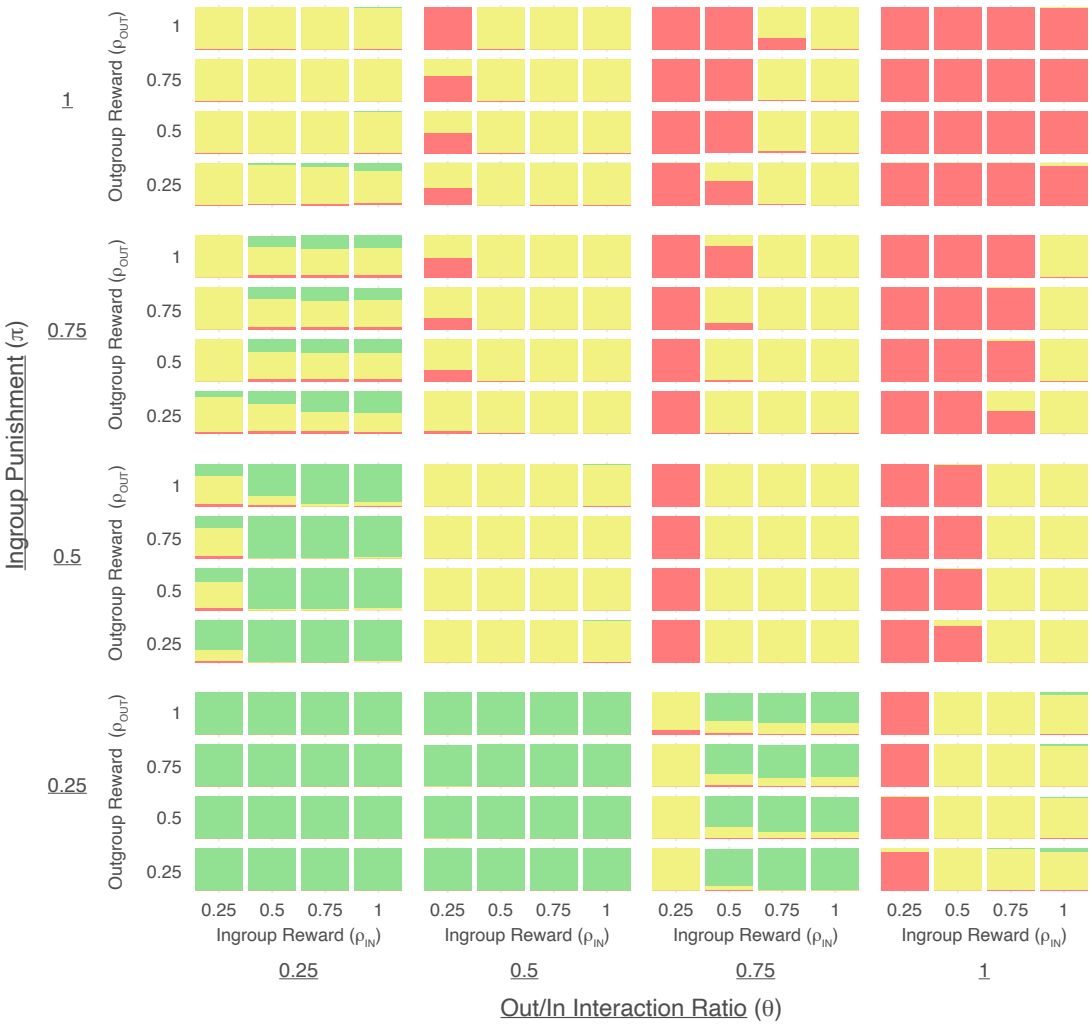

Figure 5.11: Sampling parameter space across theoretically significant parameters at regular intervals, holding all other parameters at the values given in Table 5.1. In total 256 (4^4) parameter combinations are visualized. Each colored bar shows random forest model confidence at a given parameter setting in observing each of the three dynamic classes: green, saturation; yellow, cycling; red, suppression. Inner x- and y-axes indicate the value of ingroup reward and outgroup reward respectively. Outer x- and y-axes indicate the value of out/in interaction ratio and ingroup punishment respectively.

emerge can persist for as long as the groups do. When groups mix, however—as they typically do in most places today—dividing lines between groups can cause problems. Such a scenario can select for covert signaling strategies, in which individuals are more likely to use encrypted or obfuscated identity signals that are less likely to correctly interpreted by outsiders (Smaldino et al., 2018; Smaldino and Turner, 2022). Previous work on covert identity signaling has been largely agnostic about the specific signals individuals might be used, or for how long a particular signal might effectively remain covert if outgroup individuals are incentivized to recognize members of other groups. The model demonstrates how identification by an

outgroup, if sufficiently likely and sufficiently costly, can cause covert identity signals to have natural life cycles in which they first emerge as ingroup signaling conventions but are then discarded as they become known to outsiders, replaced by new signals.

The dynamic I describe is reminiscent of the Red Queen hypothesis from evolutionary biology (Van Valen, 1973). There, a species at risk of predation, parasitism, or competition from other species must constantly evolve new adaptations just to maintain constant fitness levels, in response to the continuous evolution of adaptations to compete or predate on the part of the opposing species. My proposal is that, similarly, groups with incentives to remain even partially obscured to outsiders will need to constantly create new identity signals in response to the incentives faced by outsiders to recognize signals identifying other groups. In the model, particular signals appear to cycle in and out of fashion, but this need not be the case—indeed, my model does not require this interpretation. Instead, we should imagine each signal to merely be an index or pointer, for which the signal being represented is replaced by a novel one whenever the informational value of a particular index becomes sufficiently small.

I further show that if identification by the outgroup is *too* likely and *too* costly, conventionalized identity signals may *never* emerge for public display. In this scenario, members of the ingroup are quite bad at identifying one another in public, because the risk of being identified by unfriendly outsiders is too great. This is perhaps akin to the situation faced by political dissidents under a brutal totalitarian regime, or even by xenophobic extremists in a well-functioning cosmopolitan society. More mundanely, it may also represent scenarios in which the benefit of coordinating on a particular idiosyncratic identity may be outweighed by the social costs of outing oneself as different. The model assumes that agents must signal, and I find that under substantial threat they simply use signals that are minimally informative. In the real world, one could propose that people simply avoid signaling their identity, but in practice this is hard to do. It is nearly impossible to avoid transmitting any identity information during a social interaction (Moffett, 2019). The best one can do in these cases may be to avoid transmitting information that can *reliably* be linked to a persecuted group identity, such as by using ambiguous or generic speech, or even outright deception. This is in line with the model results.

My analysis identifies the importance of learning rates to the dynamics of identity signal use, and shows that these rates contribute additively, along with expected costs of interactions with outgroup, to whether signal dynamics exhibit stable saturation, cycling, or suppression (see Figures 5.8 and 5.11). Sufficient reward from successfully using signals to identify ingroup is required to conventionalize particular signals

as group markers. However, the rate at which the outgroup acquires knowledge of the ingroup's identity signals must be sufficiently slow as to allow time for that conventionalization to usefully take hold. Empirical determination of these learning rates, which must be determined relative to rates of outgroup interaction and punishment, is likely to be difficult. Complicating matters is the fact that in the model, learning rates could be set independently from the incentives of either successful ingroup coordination or harmful outgroup punishment, whereas in reality, these factors are likely to be entangled. It is likely that learning rates will often be optimized to meet the strategic needs of group members, conditional upon other constraints of behavior, cognition, and physiology.

The theory of covert signaling (Smaldino et al., 2018; Smaldino and Turner, 2022; van der Does et al., 2022) indicates that speech or other communicative acts that deliberately or at least instrumentally (and so potentially unconsciously) obscure identity information from non-insiders should be more common when ingroup members are relatively uncommon and the cost of being identified as such by outgroup individuals is nontrivial. These results indicate that individuals must do more than simply attune themselves to signals that are intrinsically overt or covert. Rather, the results imply several consequences for individuals living in diverse populations in which covert signals are the best strategic choice for identity signaling. Individuals must of course possess, consciously or not, some understanding of the informational value of both sent and received identity signals (Skyrms, 2010; Bergstrom and Rosvall, 2011), so that they correctly present themselves to others and correctly identify coalitional commitments and behavioral tendencies in others. Individuals must of course also possess appropriate strategic caution to favor covert or encrypted signals in environments in which both friendly and unfriendly audience members are present (Loury, 1994; Smaldino et al., 2018; Smaldino and Turner, 2022). But individuals for whom covert signaling is strategic must also learn and continuously re-learn (1) the likely informational content of potential signals so as to accurately present themselves, (2) the information content of particular signals used by others so as to accurately identify them, and (3) the likely risks of being discovered by outgroup listeners condition on using particular signals. The key point is that each of these estimates (1–3) may be continuously changing. Understanding the cognitive mechanisms, and their associated accuracy, behind these estimates is an important task for research on the psychology of social cognition and intergroup interactions.

Previous work described overt and covert signals as if they were distinct classes of symbolic communication, with distinct lexicons (Smaldino et al., 2018; Smaldino and Turner, 2022). The present study illustrates how this need not be the case. Rather, the property of a signal as covert or overt results from a dynamic

process that emerges from the changing state of collective knowledge. An initially arbitrary signal may become a covert signal of identity within an ingroup through repeated association while remaining obscure to outsiders, only to later become an overt signal once sufficiently common use allows those outsiders to determine its meaning as an identity signal. At this point, the signal may either remain an oft-used identity signal if the costs of identification by outsiders is negligible, or be abandoned if covertness is required.

This work presented here is agnostic about some of the psychological process that determine or moderate the model parameters. For example, the reward for successfully using identity signals for assortment may stem from emotional responses, monetary gains, or the maintenance of partnerships or coalitions. Any of these could be used by learning mechanisms to reinforce some signaling behaviors and to suppress others. I view this agnosticism as a strength of the model, as it is similar to the phenotypic gambit common in behavior ecology (Grafen, 1991), which focuses on how behavioral strategies create adaptive value without worrying overly about the mechanisms that generate those behaviors (temporarily at least). This approach is no less fruitful when applied to human behaviors that may evolve culturally (Smith and Winterhalder, 1992), even if behaviors must eventually be understood in terms of their generating mechanisms (Heyes, 2016). Indeed, the present work can be viewed in the context of other attempts to understand dynamic patterns of social and cultural change using formal modeling approaches (e.g., Turchin, 2003, 2011), along with more recent efforts to better integrate cultural evolution with cognitive science (e.g., Heyes, 2018).

In general, although my model does embrace some of the important complexity present in the real world, it is still a drastically simplified representation of human behavior. This kind of simplification is necessary for the development of robust formal theories of social behavior, particularly in domains where the number of formal models is still relatively small (Smaldino, 2017, 2023). Nevertheless, it is important to acknowledge how some of the model assumptions constrain my ability to generalize. I assume that the benefits to successful assortment manifest simply as positive reinforcement for the use of a particular signal, while the costs of being identified by the outgroup manifest simply as negative reinforcement for the use of that signal, with both of these identically implemented across all agents. This is a plausible dynamic that is consistent with well-mixed models of weak selection commonly used to model both genetic and cultural evolution (Mullon and Lehmann, 2014; Rodrigues and Kokko, 2016). In doing so, I ignore issues like individual differences in signaling strategies and power dynamics, as well as the potential for the population structure to evolve if punished individuals are removed from the population. Exploring these and other limitations are important avenues for future modeling work.

This work highlights the need for more empirical work to investigate the dynamics of identity signals, particularly those that may be covert and used only within particular groups. Studying such signals is difficult, because covert signals are not easily identified by non-group members (van der Does et al., 2022). This difficulty is further complicated if the use of a particular signal as an identity marker and the level of covertness attributable to that signal both change over time. My model yields qualitative but testable hypotheses about the nature of these signaling dynamics. We expect to see clear relationships between the expected costs of detection by outgroup individuals and the character and lifespan of the identity signals used by ingroup individuals to coordinate. These predictions should aid future empirical investigation. Given the importance of considering identity for understanding social processes, such investigations are warranted.

CHAPTER 6

COMPLEX TEXTUAL INTERPRETATION AT SCALE: QUALITATIVE CODING USING LLMS[1,2]

6.1 Introduction

Qualitative coding, or content analysis, extracts meaning from text to discern quantitative patterns across a corpus of texts. Recently, advances in the interpretive abilities of large language models (LLMs) offer potential for automating the coding process (applying category labels to texts), thereby enabling human researchers to concentrate on more creative research aspects, while delegating these interpretive tasks to AI. The case study in detailed in this chapter comprises a set of socio-historical codes on dense, paragraph-long passages representative of a humanistic study. I show that GPT-4 is capable of human-equivalent interpretations, whereas GPT-3.5 is not. Compared to our human-derived gold standard, GPT-4 delivers excellent intercoder reliability (Cohen's $\kappa \geq 0.79$) for 3 of 9 codes, and substantial reliability ($\kappa \geq 0.6$) for 8 of 9 codes. In contrast, GPT-3.5 greatly underperforms for all codes ($mean(\kappa) = 0.34$; $max(\kappa) = 0.55$). Importantly, I find that coding fidelity improves considerably when the LLM is prompted to give rationale justifying its coding decisions (chain-of-thought reasoning). I present these and other findings along with a set of best practices for adapting traditional codebooks for LLMs. Our results indicate that for certain codebooks, state-of-the-art LLMs are already adept at large-scale content analysis. Furthermore, they suggest the next generation of models will likely render AI coding a viable option for a majority of codebooks.

The first and final studies make for poetic bookends to this dissertation. In Chapter 2 I expand upon Nelson's Computational Grounded Theory (2017) arguing that computational tools for text analysis are typically better understood as data organizers, drawing structure from unstructured data. This is largely because our

[1]This chapter is based on "Scalable qualitative coding with LLMs: Chain-of-thought reasoning matches human performance in some hermeneutic tasks", which was published to arXiv on February 6th 2024.

[2]I am especially grateful to Tania Ravaei for collaborating on codebook development. Thanks also Harry Yan, Pat Wall, Patrick Kaminski, Adam Fisch, Alicia Chen, and Francisco Muñoz for their helpful comments toward improving this manuscript.

tools for text analysis at the time were considerably less sophisticated than qualitative researchers'. Thus I advocate for simpler approaches to text organizing than topic modeling that allow for a wide range of downstream quantitative and qualitative analyses.

In the interceding years, we have leapt past natural language processing and achieved true, if flawed, natural language understanding. The aspirations of Computational Grounded Theory are all but realized. LLMs are zero-shot learners and as of 2024 capable of incredibly sophisticated literary analysis. These models are adequate and frequently impressive qualitative coders, and an effective replacement for coding teams.

However, I still stand by most of what I wrote in the White nationalist study. An LLM would not be able to so thoroughly draw the boundary around the discourse of White supremacy. The current state of language models leaves them somewhere between very dumb and fast NLP models and very smart and slow human researchers. The brute force approach of relative frequencies between a target and relevant reference group is almost certainly not recoverable through even a language model fine-tuned on the data. Nor is a language model as capable of close reading through the data set, which would require memory, iterative observation, hypothesis refinement, and a multilevel scientific imperative to understand the social world.

LLMs still cannot replace us, even as they cannot replace some of our crudest informational tools. And neither should they. An LLM is not capable of performing regression from reading a dataset anymore than a human is (though undoubtedly this could be trained into the underlying network). But this would be a waste of time and compute when we have perfectly good existing tools for regression. This is equally true for existing methods of natural language processing.[3] If deep learning-based language models continue to improve at the rate they have since word2vec (2013) or since transformers (GPT-1, 2018), we will face difficult questions about the value of human experience and in particular subjectivity in science. The humanities have largely settled this for themselves, though genuinely superhuman intelligence would also complicate their position, which was already greatly threatened. There is reason to be cautious and perhaps fearful of these tools, even from the narrow perspective of social science and humanities research. Nevertheless, there is also tremendous potential for new and faster understanding, particularly of complex systems, which have thus far proven resistant to most of our existing analytical tools.

[3]Though probably not LDA-based topic modeling, which was already supplanted by methods that leveraged neural networks, such as BERTTopic.

6.2 Background

Text categorization, commonly referred to as content analysis and qualitative coding in the social sciences, plays an important role in scholarly research and industrial applications. This process traditionally relies on human expertise to interpret the nuanced and often complex meanings embedded in texts (Strauss, 1967; Saldaña, 2009). The difficulty lies in the multifaceted nature of meaning and the challenge of fitting real-world complexity into discrete categories, even for skilled readers. Historically, these challenges have positioned text categorization as a task unsuitable for machine learning approaches (Malik, 2020), despite robust attempts (Nelson, 2017; Dhar et al., 2021).

Recent developments in artificial intelligence, notably the advent of transformers with billions of parameters known as large language models (LLMs), have begun to challenge this notion. These models demonstrate increasing capabilities in knowledge, interpretation, reasoning, and creativity expressed in natural language, approaching or even surpassing human performance (Bubeck et al., 2023; Romera-Paredes et al., 2023; Team et al., 2023). The processing speed of artificial intelligence opens up the possibility of categorizing vast quantities of text, far exceeding the limitations of human coding teams restricted to smaller samples. Yet, this opportunity raises a critical question: how can we ensure and maintain the accuracy of machine categorization at a level comparable to human standards?

This study provides the strongest evidence to date that machines are capable of human-quality interpretations of text for the purposes of qualitative coding. Additionally, this report serves as a practical guide to employing LLMs in text categorization and as a reference for those encountering machine-assisted qualitative coding in empirical research. I contribute to the growing body of work that builds confidence in the rigor of LLM-based text categorization (Xiao et al., 2023; Chew et al., 2023; Dai et al., 2023; Tai et al., 2023), a field that will expand as these models continue to evolve. This report emphasizes the redesign of codebooks—comprising category descriptions and coding instructions—specifically for LLMs. I demonstrate how the structure of prompts, the specific requests made to the generative model for categorizing passages, significantly impacts coding fidelity. Even as these models continue to rapidly improve, I expect most of the principles of prompt design I report will remain useful and informative as methodologists explore new models and empiricists automate their coding workflows. These results are presented through narratives detailing my approach and highlighting potential challenges and demonstrated by LLM-generated analyses compared to a human-derived gold standard. A summary of best practices for content analysis with

an LLM is also presented in tabular format for quick reference.

Key findings of the study include:

- GPT-4 exhibits human-equivalent performance with zero-shot prompts. 8 of 9 tasks exceed the 0.6 threshold for substantial agreement using Cohen's κ. 3 of 9 tasks exceed the 0.75 threshold for excellent agreement.
- GPT-3.5, when given the same prompts, has an average intercoder reliability of 0.34 across all codes.
- Codebooks designed for human coders need reworking for LLM application, requiring iterative manual testing to refine phrasing and improve model comprehension.
- Agreement improves when the LLM provides rationale for code assignments: $\mu(\kappa) = 0.68$ vs. $\mu(\kappa) = 0.59$.
- Agreement improves when presenting each code as a separate prompt, rather than the codebook as a whole: $\mu(\kappa) = 0.68$ vs. $\mu(\kappa) = 0.60$.

6.2.1 Automating Content Analysis: Past and Present

Prior work on automating content analysis entailed training machine learning models on large quantities of text. Supervised models, typically some form of linear regression, learn to associate text features with user-specified categories (Kadhim, 2019). This process captures half the traditional human-coded process by using human-derived codes and examples, but fails to leverage abstract code descriptions found in a codebook, as well as requiring large quantities of human-annotated data. Unsupervised models, such as LDA (Blei et al., 2003; Jelodar et al., 2019) or BERTTopic (Grootendorst, 2022), develop their own categorizations from unlabeled training sets. This process does not require time-intensive labeling, but rarely captures the specific categories that the researcher intends to target.

The latest generation of LLMs (e.g., GPT (Brown et al., 2020), LLaMa (Touvron et al., 2023), Mistral (MistralAI, 2023), Claude (Anthropic, 2023)) differ notably from previous machine learning models in that they can perform new tasks specified through natural language prompts. A user can specify a task that the model was not trained on, give few (single digit) or no examples, and the model will return output conforming to the specifications. Demonstrated successes include computer code generation (Nowakowski and Keller, 2024), creative writing (Gómez-Rodríguez and Williams, 2023), and quantitative reasoning

(Bischoff, 2024). We are only beginning to understand and expand upon the limitations of these models. By converting natural language requests into highly intelligent output across vast and indeterminate domains, LLMs lower the technical barriers to machine learning by making its application more naturalistic and eliminating the need for large training data. Beyond this, LLM's capacities in many domains far exceed the specialized machine learning models that preceded them, suggesting that for many applications, including scholarly inquiry, artificial intelligence is overwhelmingly more accessible and capable in 2024 than it was just two years prior.

Early studies of content analysis with LLMs are encouraging. Xiao et al. (Xiao et al., 2023) demonstrate moderate success, Cohen's $\kappa = 0.61$ and $\kappa = 0.38$, in two linguistic tasks using GPT-3. Chew et al. (Chew et al., 2023) report high success on many of 19 tasks across three datasets, and results that are indistinguishable from random for others. It is difficult to evaluate their results due to the choice of Gwet's AC1, which is biased toward agreement on negative codings rather than positive, whereas most standard measures of intercoder reliability do the opposite (Vach and Gerke, 2023). However, I commend Chew et al.'s approach to adapting codebooks for LLMs, which is communicated with great detail and clarity. A survey of 20 empirical pieces reports "mixed-results" of using GPT-3 to automate "text annotation," a term that ties their framing to "data annotation", labeling data for in machine learning (Ollion et al., 2023), rather than content analysis in the tradition of grounded theory (Strauss and Corbin, 1997).

I present here three advances to these studies. 1) I report the first methodological account of automating qualitative coding using GPT-4, which, along with other recent models, greatly improves upon many of GPT-3's capabilities (OpenAI, 2023; Yuan et al., 2024; MistralAI, 2023). 2) I provide the first conclusive evidence that LLMs are capable of human-equivalent performance in qualitative coding, and do so on larger passages of text, wherein meaning is often woven through multiple interrelated clauses. 3) I demonstrate that GPT is better at interpreting text when it is tasked with justifying its coding decisions (chain-of-thought prompting) rather than applying codes without an accompanying explanation.

6.2.2 Case study: W.E.B. Du Bois's characterization in news media

In order to present a realistic challenge of using an LLM to do qualitative coding, I make a case study of my own work. I adapted a codebook written by the authors to understand how the scholar and activist W.E.B. Du Bois has been characterized in news media over time. The codebook is composed of 9 codes in 3 categories. Due to multiple layers of agency (who is doing what) and voice (who is saying what), the tasks

are difficult even for human interpreters. Applying the codes is also complicated because it can be difficult to differentiate Du Bois's scholarship from his political activism, as Du Bois's theoretical contributions have profound implications for understanding race and the social-historical position of Black persons in the United States and beyond, making them powerful activist tools. I am particularly interested in understanding how different facets of Du Bois's activities contributed to his canonization in the public imagination as the preeminent figure for understanding Black political struggle. Table 6.1 gives the codes in brief. Complete examples of the original human and modified-for-GPT codebook are included in the appendix.

The training and test data for this study were random samples of passages from New York Times articles (1970–2023) that mention W.E.B. Du Bois. 232 passages were automatically extracted as concurrent paragraphs containing "Du Bois". The average number of words was 94 ($\sigma = 70$), and the average number of sentences was 3.75 ($\sigma = 2.88$). To give a better sense of the size of the passages, this paragraph has 76 words across 4 sentences.

Table 6.1: Categories and descriptions for 9 codes.

Characterization of Du Bois	
Scholar	Describes Du Bois as a scholar or intellectual.
Activist	Refers to Du Bois's political or social activism.
General Themes	
Monumental Memorialization	Refers to an enduring cultural object named after Du Bois.
Mention of Scholarly Work	Mentions or quotes specific academic works by Du Bois.
Social/Political Advocacy	Mentions or implies social or political activism, advocacy, or critique.
Canonization Processes	
Coalition Building	Refers to Du Bois's activities with activist or academic organizations.
Out of the Mouth of Academics	Describes an academic organization engaging with Du Bois's legacy.
Out of the Mouth of Activists	Describes an activist organization engaging with Du Bois's legacy.
Collective Synecdoche	Mentions Du Bois alongside other figures in order to represent some facet of a culture, era, or ideology.

6.3 Results

6.3.1 Adapting a codebook for an LLM

Initially, I developed a codebook for human coders using standard methods. This process involved exploratory reading to define and refine codes. Codes were derived to probe particular substantive hypotheses, some of which preceded exploratory reading, and others which resulted from it. A collaborator and I then applied these codes iteratively, adjusting them as needed until I achieved high intercoder reliability with a test set. I adapted these code descriptions for use with a large language model, evaluating the LLM's performance on a training set of text passages. Where I found ambiguities or deficiencies in the model's interpretation, I refined the code descriptions accordingly. This iterative process of definition, evaluation, and refinement follows Nelson's Computational Grounded Theory paradigm (Nelson, 2017), the core of which is common to all qualitative code development processes whether or not coding is automated (Strauss and Corbin, 1997).

My experience modifying the code descriptions yielded several key insights related in the following paragraphs. I encourage readers interested in a fuller account of this process to read Chew et al.'s study (Chew et al., 2023) describing their process of LLM-Assisted Content Analysis (LACA), which relates a process similar to my own in greater detail.

LLM-generated rationale are essential for evaluating performance. In adapting the codebook, I wanted to understand not just which codes the model struggled to interpret correctly, but what aspects of the code the model failed to capture. To achieve this, I structured the prompts to require GPT to justify its decision to apply or not apply each code. These rationale were invaluable. They often highlighted parts of the code description that were ambiguous or imprecisely defined, leading the model to misinterpret them. Whenever a rationale repeatedly pointed to such an issue, I revised the corresponding code. I then retested the passage to check that the code was correctly applied and the rationale aligned with the intended interpretation of the code. Sometimes a revision would not improve the interpretations for the passages in question; other times it would fix the interpretations for those passages, but would introduce new problems in passages which were previously coded correctly.

Figure 6.1 demonstrates an effective method of prompting GPT to provide rationale for its code selections. The initial instruction is given by the Justification section of the prompt, and solicited again in the

Role Assignment

```
You are tasked with applying qualitative codes to articles, book reviews, and
opinion pieces referencing W.E.B. Du Bois. The purpose of this task is to track
how Du Bois is represented in news media over time.
```

Code Definition

```
Below I will explain how to apply the code:

Title: Monumental Memorialization
Description: Apply when an enduring cultural object is named after Du Bois.
Such objects include prizes/awards, named professorships, buildings or rooms,
geographical features, institutes, schools, or activist organizations. Do not
apply when Du Bois is mentioned in the title of a book or theater production.
```

Justification

```
When you evaluate the passage, provide a justification of why you did or did
not apply the code.
```

Decision / Formatting

```
Then list the code in the following fashion if you applied the code:

Justification: [insert 2-3 sentence rationale for applying the code here]

Codes Applied:
- Monumental Memorialization

Otherwise you can format it like this:

Justification: [insert 2-3 sentence rationale for not applying the code here]

Codes Applied:
- None

Do not write anything in your reply after listing the "Codes Applied:"
```

Figure 6.1: The chain-of-thought prompt sequence.

Decision/Formatting box.

LLMs require more precise descriptions than do human readers. Human coders do not rely solely on a written codebook. Their interpretation of the codes is enriched through the codebook development process, discussions with fellow coders, and supplementary oral instructions. An LLM lacks this interactive and historical context and must interpret codes entirely from written descriptions. My work modifying the codebook for GPT revealed information that, while implicitly understood by the code developers, wasn't explicitly stated in the code descriptions. This process not only aided in refining the codebook for automated coding, but also improved my own understandings of the codes. This ultimately led to clearer definition of the codes, thereby enhancing future manual coding processes as well. Figure 6.2 demonstrates how the Monumental Memorialization and Social/Political Advocacy codes were redefined to improve GPT's comprehension.

Often, I encountered cases where ambiguous phrasing was obvious to humans, but challenging for the

LLM. The codebook contains two codes that relate to Du Bois's reputation among academics and activists. These codes are meant to evaluate whether Du Bois appears in a news story because either an academic or activist mentioned him. Initially, I titled this code "Academic Repute," which worked well for human coders. GPT, however, consistently misinterpretted this code as pertaining to Du Bois's esteem *as* an academic, rather than *among* or *by* academics (the meaning of "among" remains ambiguous even here). I tried numerous iterations this code without success. Nevertheless, altering the title of the code to the far more literal "Out of the Mouth of Academics" dramatically improved performance, even when paired with the original code description. In another case, the code titled "Social/Political Activism" was revised to "Social/Political Advocacy" (Figure 6.2 C) because GPT did not consider social critique to be a form of activism, even when it was specifically instructed to.

I found that words indicating how much the model should draw on context or its own outside knowledge had large impacts on the model's outputs, often to the desired effect. In particular, instructing the model to restrict itself to "explicit" meanings, or to draw on "implicit" meanings, often helped the model with part of a code description it had struggled with. Figure 6.2 B and D demonstrate the addition of such verbiage to control scope.

Both mandatory (do) and prohibitory (do not) phrasing were observed by the model, though mandatory phrasing seemed more successful, a finding reported by other researchers (Bsharat et al., 2023). The ordering of directives also impacted how likely the model was to follow them. I found that moving a phrase that was ignored in the coding rationale toward the front of the definition made the model more likely to follow its specifications, as in Figure 6.2 A. When a very specific problem was observed repeatedly, it was sometimes necessary to add a directive to correct it, as in Figure 6.2 H.

Prompting for machine-readable output. To fully automate the coding process, model output must be reliably readable by a computer. The LLM generates text, which must be interpreted by another script into a data structure, such as a table, for further analysis. Instructing exactly how to format the output produced machine-readable results with GPT-4 and GPT-3.5. Critically, this involves specifying a tag that the interpreting script locates, after which follows a reliably formatted list of codes. The Decision/Formatting component Figure 6.1 illustrates how to constrain model output and produce consistent results across queries. Additionally, because GPT tends to be excessively verbose and summarize its output, particularly at higher temperatures, I informed the model that I do not want any output to follow the code list.

Original: Activist

Apply when Du Bois is developing activist organizations, giving public speeches, participating in meetings with politicians and organizers, running for office, promoting a candidate, organization or initiative. **Also apply when Du Bois is *explicitly* described as an activist or leader.** (A)

Redefined: Activist

(A) **Apply this code when Du Bois is *explicitly* called an "activist" or "leader",** or **when his political or social activism is either *explicitly noted or clearly implied through context.*** (B) Examples include being mentioned in the context of leadership, activism, developing activist organizations, giving public speeches, participating in meetings with politicians and organizers, running for office, or promoting a candidate, organization, or initiative.

Original: Social/Political Activism (C)

(D) **Refers to any form of social or political participation in promoting change in society. Can be used in conjunction with Du Bois' scholarly work, specifically when Du Bois's ideas are used to frame Black political struggle. Can be used** (E) (F) **to describe Du Bois' political work or those who invoke Du Bois in service of their activism.** (G)

Redefined: Social/Political Advocacy (C)

(D) **This code applies when a passage mentions or *implies* any form of social or political activism, advocacy, critique, or discourse, including discussions about current or historical social problems. This includes not only direct activism of Du Bois and others, but also the framing and challenging of social norms,** (E) **historical narratives, and racial or cultural identities. Apply this code when Du Bois's work, persona, or ideas are invoked in discussions that critically engage with Black identity, positionality, or broader systemic circumstances of Black people. Adjacency to other activists, such as inclusion in a list, is insufficient; advocacy must be *explicitly mentioned* in the passage.** (F) (G) (H)

Figure 6.2: Two examples of prompt redefinition. Colored, alphabetically labeled blocks of text show alterations derived through iterative code refinement. Italics draw attention to direction to constrain interpretive scope to implicit or explicit information.

6.3.2 Selecting a model and writing prompts for optimal performance

Once the code descriptions have been revised for LLM text categorization, numerous other decisions remain about how to prompt a model to execute the content analysis. I present these as a separate step for the sake of clarity, but in reality, I developed my approach iteratively and in tandem with revising the code descriptions. I hope future methodologists and empiricists will benefit from what I learned during this process, and that less exploration of these components will be necessary so practitioners can focus on application or exploration of calibrations not explored here. I summarize all my recommendations for qualitative coding with an LLM in Table 6.3.

There is large and growing body of academic and nonacademic literature on prompt engineering: con-

structing user-defined input to elicit the best model output. In fact, the codebook adaptation in the previous section was in large part an exercise in prompt engineering. However, in this second section, prompt engineering refers more to the broader context of task description than the code definitions. In this section I report how different prompts influence the quality of machine categorization. Additionally I compare performance when the LLM is tasked with assigning each code independently to when the model is given the full codebook and assigned with coding all 9 codes as a single task. I refer to these as the "Per Code" and "Full Codebook" approach respectively.

Studies have shown that LLM decision-making improves when the model is prompted to account for its decisions (Wei et al., 2022; Madaan and Yazdanbakhsh, 2022). This is generally known as chain-of-thought (CoT) prompting or reasoning, and refers to breaking down tasks into specific components, one or more of which involve planning for future steps or reflection on previous ones. The prompts, which can be viewed in full in the appendix, apply chain-of-thought prompting by including 1) a role assignment step, informing the machine of its purpose, 2) a task description step, specifying the code definition, 3) a justification step, instructing the model to provide a rationale for its decision, and 4) a decision step, wherein the model delivers its ultimate analysis in a consistent, machine-readable format. An example of the chain-of-thought prompt sequence is given by Figure 6.1.

I use zero-shot prompts throughout this study. Zero-shot refers to providing the model only the task description, without giving examples of correctly executed responses. Xiao et al. found few-shot prompting improves coding and performance on other tasks (Xiao et al., 2023), whereas Chew et al. largely employed zero-shot prompts (Chew et al., 2023). This case study involves evaluating paragraph-long passages rather than single clauses. I found that information in the examples was drawn upon by the LLM and interfered with its coding decisions. I also found that in Full Codebook prompts, giving examples greatly expanded the prompt, negatively impacting results. When content is more literary or historical, zero-shot prompts are probably preferred, but that most coding tasks will benefit from few-shot prompting as demonstrated by the results of many other studies across domains.

Performance comparisons are relative to the human-derived gold standard hidden from the LLM at all stages of development. I used the default settings for the GPT API where temperature is set to 0 and nucleus sampling (`top_p`) is set to 1. I specified the task description as a "system prompt", and provide each passage as a "user prompt". A system prompt gives the LLM its purpose, clearly specifying the task it is meant to address, whereas a user prompt provides the input to which the model responds by generating output. I did

not investigate whether intercoder agreement suffers with the default system prompt, while combining the task description and passage as a user prompt.

Table 6.2: Intercoder reliability (Cohen's κ) for all codes on 111 gold standard passages. Best overall performance is shown in bold. Italics indicate the highest intercoder reliability between pairs with and without prompting for rationale (CoT vs. No CoT); if the pair are equivalent neither is italicized. Two values are considered equivalent if their difference does not exceed 0.02.

		GPT-4				**GPT-3.5**	
		Per Code		**Full Codebook**		**Per Code**	
Code	**Count**	**CoT**	**No CoT**	**CoT**	**No CoT**	**CoT**	**No CoT**
Scholar	27	**0.61**	0.52	**0.59**	0.42	*0.29*	0.21
Activist	23	**0.81**	0.65	*0.67*	0.62	*0.39*	0.32
Monumental Memorialization	13	**1.00**	0.91	*0.75*	0.48	0.29	0.31
Mention of Scholarly Work	24	**0.71**	**0.69**	*0.52*	0.44	0.33	*0.39*
Social/Political Advocacy	51	**0.64**	0.60	0.60	0.60	*0.55*	0.51
Coalition Building	9	**0.60**	0.44	*0.43*	0.13	*0.33*	0.17
Out of the Mouth of Academics	30	**0.63**	**0.65**	**0.65**	0.62	*0.37*	0.33
Out of the Mouth of Activists	11	*0.30*	0.09	**0.34**	0.18	*0.21*	0.09
Collective Synecdoche	26	**0.79**	0.78	**0.81**	0.71	0.27	0.27
Mean	24	**0.68**	0.59	*0.60*	0.46	*0.34*	0.29

GPT-4 greatly outperforms GPT-3.5. I found that GPT-4 approaches human performance for 3 codes: Activist: κ = 0.81; Monumental Memorialization: κ = 1.00; Collective Synecdoche κ = 0.79. GPT-4 prompted for rationale provides considerably higher quality code assignments than GPT-3.5, except in the case of the Out of the Mouth of Activists code, which no configuration handled well. It is especially notable that GPT-4 and GPT-3.5 differed in their most accurately interpreted codes. In the 3 tasks GPT-4 executed best, GPT-3.5's performance was slightly below its own average, $mean(\kappa_{all})$ = 0.34 vs. $mean(\kappa)$ = 0.32.

Coding fidelity improves when codes are presented as individual tasks. I adapted the codebook by presenting the entire codebook to GPT along with task instructions. However, I found in testing that performance improved when GPT was given each task independently. This "per code" approach was taken by one recent study exploring content analysis with non-mutually exclusive codes (permitting multiply coded

passages) (Chew et al., 2023), but not another, which tested only two codes (Xiao et al., 2023). Table 6.2 compares the GPT-4 performance when presented individual tasks for each code ("Per Code") and when presented all tasks in a single prompt ("Full Codebook"). I found that for the 3 human-equivalent tasks (Activist, Monumental Memorialization, and Collective Synecdoche) the Per Code performance far exceeded the Full Codebook for 2 tasks, and was comparable for 1. For 2 other tasks, Mention of Scholarly Work and Coalition Building, I found that the Per Code configuration produced considerably higher agreement, whereas Full Codebook performed comparably to the Per Code in the remaining 4 tasks.

Coding fidelity improves when the model is prompted to justify its coding decisions. Consistent with other experiments with chain-of-thought (CoT) reasoning in LLMs, I found that coding agreement benefited strongly from prompting the model to explain itself (Chu et al., 2023). Table 6.2 shows the effect of prompting for rationale on three pairs of conditions: Per Code GPT-4, Full Codebook GPT-4, and Per Code GPT-3.5. I found that across all codes and conditions, with one exception, CoT prompting produces higher or equivalent intercoder reliability with the gold standard. Using GPT-4, average Per Code agreement improved from 0.59 to 0.68, and average Full Codebook agreement improved from 0.46 to 0.60. Moreover, a majority of pairs showed substantial improvement when the codes were assigned after providing reasoning for coding decisions.

6.4 Discussion

Determining appropriate domains for LLM-assisted qualitative coding. Previous methods of automated text categorization, both supervised and unsupervised, rarely met the standards of traditional social scientists and humanists, and were instead generally employed by data scientists. Capturing meaning, particularly complex meaning, through machine learning has largely been an elusive goal (Malik, 2020). Despite my own former skepticism, I predict that LLMs will be capable of applying most qualitative codebooks within the year. However, my results show that even within the scope of a single codebook, interpretation quality varies. Thus, different disciplines and domains should expect model success and the ease of transitioning a codebook to vary considerably. I suspect that more humanistic and "softer" scientific approaches will (continue to) be more resistant to machine interpretation than problems posed by scholars who identify with "harder" sciences, to say nothing of their ability to convince their peers of its validity. I do not oppose developing evaluation benchmarks for qualitative coding to assess which models are adept at what

Table 6.3: Principles of prompting an LLM for qualitative coding.

Task Instructions	
Prompt for Rationale	Model fidelity improves when instructed to justify its coding decisions.
One Task Per Code	Model fidelity improves when given each code as a separate task.
Brevity	Shorter task descriptions are more likely to be faithfully executed by the model.
Structured Output	Instruct the model to format its output to ensure uniform responses.
Code Definitions	
Word Choice	A single high-content word can be changed to align with the LLM's built-in ontology.
Clause Order	Clauses are more likely to be observed when introduced earlier in the code description.
Mandates/Prohibitions	Both can be effective, but it is easier to get the model to "do" than "do not".
Code Titles	Altering the code title can have a large effect even without altering the definition.
Interpretation Scope	Use words like "implicit" and "explicit" when interpretation is too limited or expansive.
Chain-of-Thought Prompt Sequence	
1. Role Assignment	Supply the model its purpose, e.g., "You will be applying category labels to passages."
2. Code Definition	Provide the code title(s) and description(s).
3. Justification	Request that the model provide evidence of its reasoning.
4. Decision	Instruct the model to list the codes that apply to the passage in a consistent format.

variety of task, but neither do I advocate it; meaning is manifold and emergent, and much of its beauty derives from its resistance to reduction and definition. Instead, I suggest those who wish to employ an LLM to perform content analysis survey similar attempts and simply experiment on their own. The process of discovering triumphs, workarounds, and limitations of working with these models was not only fascinating, but tremendously fun.

Practical aspects of transitioning to content analysis with LLMs. While artificial intelligence potentially opens up much larger datasets to qualitative scholars, there is still a considerable technical barrier to automating content analysis. Development of an LLM adapted codebook is feasible for anyone regardless of technical skill by interacting with an LLM through chat-like Web platforms provided by proprietary model developers. However, systematically testing prompts or applying a completed codebook to the full dataset requires moderate skill in writing scripts in a language such as Python. Rather than suggest that all scholars

become programmers, I encourage researchers to develop partnerships with students or community members seeking programming or research experience as a form of project-based education. Conversely, data scientists should pursue partnerships with traditional social scientists and humanists, who are often better positioned to develop coding schema to flush out complex meanings embedded in text, which are now more tractable to machine learning.

Handling passages where model interpretation is poor. Overwhelmingly, GPT-4's interpretations were accurate and human-like. However, I found repeatedly that GPT-4, like a human reader, struggled with edge cases, especially where implicit information was required to make a judgment. I am encouraged by this finding, and argue that with automated analysis, fidelity is less important than it is with humans. Because statistical power increases with the number of observations, noise is more tolerable in machine-applied codes, as automated coding potentially increases sample size by orders of magnitude. Notably, this assumes that error is restricted to edge cases and is not otherwise systematically biased. I also advise against automated coding where datasets are small, as in interviews, where it is likely as efficient to code entirely by hand. As models improve and can provide confidence estimates for their statements (Lin et al., 2023; Zhou et al., 2023; Chen, Jiefeng and Yoon, Jinsung, 2024), ML content analysis workflows should include manual review of passages with uncertain code assignments. Anecdotally, I found that GPT-4 could intelligently reflect on its responses when prompted to do so. When presented the output of another model instance, GPT-4, acting as an untrained "critic" model (Paul et al., 2023), was often able to identify when it had encountered an edge case without prompting, as well as recognize and revise obvious mistakes. My experiences suggest that a human-in-the-loop tag-for-manual-review workflow or a two-step automated reflect-and-revise workflow may already be feasible with GPT-4 and similar models.

6.5 Conclusion

These results using state-of-the-art models lead me to recommend that scholars who do much qualitative coding consider automated coding with LLMs a potentially viable option today. I especially encourage skeptics to probe these tools' capacities, as it is useful to know their limitations. Over the next year, models such as those incorporating memory (Shinn et al., 2023), (multi-)agential models that dialogue and revise prior to rendering output (Yao et al., 2022; Xu et al., 2023), and architectures that can handle larger inputs (Gu and Dao, 2023), will almost certainly greatly improve upon GPT-4's current abilities. When those

models are made available, researchers who have already experimented with LLMs will be best positioned to make use of the new tools. The efficiency of automation is compelling, but I am most enthusiastic about the ability to probe much larger datasets than ever before, potentially illuminating patterns too rare or too fuzzy to detect with a sample numbering in the tens or hundreds rather than thousands or beyond.

CHAPTER 7
CONCLUSION

The overarching goal of this manuscript is to demonstrate the virtues of computational approaches to the study of culture. The five studies detailed herein show that computation can enter into inquiry at any of three points: data generation, data processing, and data analysis. Computational tools can enable cultural understanding that is inaccessible to other approaches, in part because they are frequently designed to capture complexity.

At the same time, I have endeavored to advance computation not as an analytical paradigm or disciplinary approach in the form of "computational social science". Rather computation offers sets of tools and data which can and should be integrated into existing paradigms. The mixed methods approaches I highlight in my own work were not developed as a pedagogical exercise. Rather, I arrived at each approach by exploring the data, iterating on hypotheses and methods for testing, and filling explanatory gaps exposed by both successful and failed tests. This process, powered by my open methodological commitments and diverse collaborators, naturally led to mixed methods work. Sometimes this work centers computation, but just as often it does not.

7.1 Where computation fits in cultural research

Throughout this dissertation I exploit what I call "computational data", which require computational data processing, but then perform more or less traditional qualitative or quantitative analyses on these data. The data acquisition in Chapter 2 (short, unstructured text in the form of tweets), Chapter 3 (Google Trends, Wikipedia Pageviews, tweets, news media mentions), and Chapter 4 (Wikipedia Pageviews) could only have been performed by a researcher with computer coding experience. Further, the selected scope of the data, not the sources but rather the particular set of observations from the sources, typically required careful consideration upstream of the process of data acquisition. In the case of the White nationalists in Chapter

2 this involved close reading much of their discourse prior to collecting data to determine an effective seed group for snowball sampling. Similarly the study BLM discursive change in Chapter 3 required close reading of Black emancipatory discourse online and in seminal documents by movement participants to determine a set of relevant terms representing BLM's political agenda.

The strictly methodological study in Chapter 6 shows how computation can be used to replicate one component of a widely used qualitative method. By employing a large language model in applying a codebook, we retain the most important parts of qualitative analysis, human creative and interpretative capacities. Nevertheless we exploit the power of an incredibly complex model of networked meaning, the LLM, and the speed of computation to overcome a bottleneck in the analytical pipeline. LLM-assisted coding replaces the most time consuming and least creative step in the qualitative coding process, applying the codes, with a machine that reads and interprets text far faster than a human can and never falters due to fatigue or distraction. This opens up large computational data sets to robust qualitative coding for the first time. The LLM-assisted approach differs from prior machine learning approaches to text analysis which were incapable of drawing on complex networked meaning to classify text.

One of my primary goals in this dissertation was to illustrate how computational tools can facilitate cultural work. In particular I drew distinctions between three classes of computation that are rarely explicitly demarcated. These are computational data: large digitized archives, crowdsourced data, and simulated data; computational data acquisition and processing: the use of scripts to download and scrape computational data, and process and organize those data into a more tractable form; and computational analysis: the use of computational models, such as network modeling or machine learning to do inference. Of these computational analysis is the least useful for the majority of practitioners. In large part the value of computation in the analysis of culture and social systems more generally derives from being able to exploit large data sets with with rich data, including natural human behavior, such as social media trace data and search queries. These data may be analyzed by computational or traditional methods, and traditional analytical methods are both broader than computational ones and much more widely practiced. But these data are not accessible to traditional researchers. Both acquiring them and processing them typically require the ability to write code from scratch; in the case of both this coding may be trivially simple or devilishly complicated, though the general rule is that data procurement is less fraught. Thus, while these data have proliferated, computer and data scientists have produced much more research using computational data, despite traditional researchers being better positioned to ask meaningful questions of them. Interdisciplinary collaboration will continue

to lead to both better research, and the diffusion of extradisciplinary paradigms into both computer/data scientific and disciplinary social scientific fields.

7.2 Where complexity fits in cultural research

Throughout these studies I pair computational data with traditional analyses for both epistemic and rhetorical purposes. As I discuss in the Introduction, complexity of cultural processes often renders them resistant to quantification. As such, qualitative analysis is often the best approach to examining culture. In Chapter 2 I employ computational data processing to produce a lexicon which is easily analyzed by close reading, leveraging the computation to make the scale of discourse with over 100,000 participants tractable to qualitative analysis. Similarly, I use close reading to derive a set of terms representing multiple facets of BLM's political agenda, capturing part of the complexity of diverse and interrelated concepts. However, instead of using a complex model of diffusion or language representation, I leverage the simplicity and clarity of frequentist statistics to make arguments that illuminate complexity in the conceptual and temporal patterns of discourse. Moreover, complexity in the hypothesis design often demands that my approach to frequentist statistics is more creative, such as performing multiple regression analyses with the same sample, but slightly different variations of the dependent variable as in Chapters 3 and 4. Similarly I use binary variables to represent different epochs in the development of the BLM discourse and visualize these coefficients chronologically to capture temporal progression in a generalized additive regression model that otherwise does not generate interpretable representations of time. The choice of standard methods is not merely determined by the demands of my inquiry, but also by my desire to produce knowledge which is interpretable and credible to a wide audience. By contrast, Chapter 5 employs a fully computational agent-based modeling approach, which originates with complex systems, not traditional social science, and is difficult for sociologists or psychologists to recognize as a piece of social theory.

Framing culture through the verbiage of complex systems is not an assertion of complexity science's sole or superior claim to the study of complex systems. Traditional social scientists as well as humanists have developed theoretical and methodological approaches to studying complexity while remaining completely unaware of complexity theory or the computational methods which are especially good at analyzing complex systems. However, complexity theory can augment existing theory without supplanting it. Perhaps its most significant contribution is to articulate in the language of mathematics why quantitative methods

typically struggle to model cultural processes and in part justify the preference of many cultural scholars for qualitative methods. Properly problematized it becomes easier to address complexity through quantitative analysis, particular through methods developed specifically to analyze complex systems.

I took different approaches to complexity *qua* complexity throughout the studies in this dissertation. Chapters 2 and 3 rely on deep reading to extract complex meaning from large discursive systems, and do employ modling techniques from complex systems. Chapter 4 also does not draw on a complex model, and instead uses verbal argument to demonstrate how concepts from complexity are at play in the evolution of a complex system. Chapter 5 uses two separate techniques from complexity science to model a complex system, dynamical systems and agent-based modeling, though I failed to represent the evolution of covert signaling as a system of differential equations, which I believe is because the system contains too much interaction for such a representation. Finally, Chapter 6 exploits a massively complex model in the form of an LLM, but rather than use it as a model of a complex system we wish to study, I leverage its complex network of meanings in the same way a researcher uses their own internal network of meanings to interpret text. As such LLMs and other deep learning models are tragically poor models of the world because while they may capture real world complexity, they typically offer no explanatory power, as we cannot interpret how the components of the system interact, or even what the components of the system are.

Complexity undoubtedly underlies cultural systems and largely accounts for the difficulties researchers face in understanding them. I am not the first to note this, and various disciplines have developed their own articulations of the problems of complexity in culture from humanities to social sciences. I have endeavored to show that formulating this problem in the language of complex systems science largely explains why cultural systems challenge our analytical and epistemic frameworks, particularly those that rely on quantification. However, I hope that my work demonstrates that there are many potential ways to capitalize on the insights of complex systems science, ranging from concepts to data processing tools and techniques to analytical models. Such a framing of complexity offers something to qualitative, quantitative, and computational researchers alike.

7.3 Some Limitations

Despite the tremendous affordances of computational work, there are still serious limitations. Some of these limitations are general to many forms of research, but manifest in particular ways. Others are very

specific to computational work. For instance, up until very recently, with the development of GPT-4-like LLMs, automated qualitative coding was possible, but severely limiting in the kinds of questions that could be asked. Because machine learning methods were crude, we were only able to extract crude categories from text, which does not allow for questions that involve complex meaning. Even today, it is difficult to interpret something like Twitter discourse through automated means, because much of the sense in posts is contextual, drawing meaning from interaction, implicit reference to other posts, and user information such as handle or profile image.

Generally speaking, often the data we can recover from online sources lacks some of the features necessary to address our hypotheses. When I started the White nationalist project, my goal was to capture the process of "becoming" a White nationalist through developing ties and facility with the discourse. However, most users did not have enough posts for a reliable individual-level analysis. Those that did were limited to 3,200 hundred messages allowed by the Twitter API at that time, and often that meant that posts from a time before they were White nationalists were inaccessible. Given the limits of the data, I was forced to pivot to population-level hypotheses and abandon my work on adult socialization. In the case of the 27 Club study, I was only able to control for the effects of the 27 Club pages because other users had tracked the clickstream data. Without these additional data, which I found in a different source and had not discovered when I submitted the first draft for publication, the results of my work would have been almost meaningless. Computational data are often highly desirable because they allow us the forgo data collection, meaning we can do our work more cheaply and quickly, and exploit data that have been recorded for reasons other than research. But one of the consequences of this is that we are incapable of transcending the limitations of these data. We cannot run new experiments or surveys. We can use only what others have provided.

Complexity is inherently challenging to study, and in some cases, may be provably unknowable. Even when it is knowable in principle, a system may be too small to provide enough observations to deduce complex relationships. This is not a consequence of sampling restrictions, but rather a limit of population size. Human populations may be too small or insufficiently varied to observe how each component of our social identities, life histories, or genetics interact to shape us. The greater the number of variables, complexity of interaction, and error (some combination of unobserved effects and perhaps genuine stochasticity) the larger the system needs to be to in order to effectively model it. In the cases where we do have enough observations and variation to learn something about interactions within the system, often computational (i.e., machine learning) models are only able to make predictions about outcomes, but cannot render interpretable

explanations of the mechanisms behind their predictions or allow us to understand how we might intervene to produce more desirable outcomes.

7.4 In Summa

It may seem odd to end on a note that amounts to "take only what you find useful". I believe the approaches to research demonstrated in this dissertation to be underused among social scientists given the explanatory power they offer. But I do not aspire to convert traditional researchers to complexity scientists; even if this were my goal, switching paradigms mid-career is a tall order. Neither is it my intention to advocate supplanting traditional scholarship with complexity theory or computational methods. Rather it is to work to harmonize the different perspectives in such a way that strengths of each are drawn upon, particularly to compensate for the others' weaknesses. Almost always this will involve multidisciplinary collaboration. Research will be best served by integration with standard approaches rather than submission to or outright rejection of them.

Bibliography

Acerbi, A., Ghirlanda, S., and Enquist, M. (2012). The logic of fashion cycles. *PLOS ONE*, 7(3):e32541.

Adams, J. and Roscigno, V. J. (2005). White supremacists, oppositional culture and the World Wide Web. *Social Forces*, 84(2):759–778.

Alexander, M. and West, C. (2010). *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*. New Press.

Anisimov, V. and Zharinov, G. (2014). Lifespan and longevity among representatives of creative professions. *Advances in Gerontology*, 4:83–94.

Anthropic (2023). Claude 2. `https://www.anthropic.com/index/claude-2`. Accessed: 2024-01-18.

Arora, M. and Stout, C. T. (2019). Letters for black lives: Co-ethnic mobilization and support for the Black Lives Matter movement. *Political Research Quarterly*, 72(2):389–402.

Aunger, R. (2006). An agnostic view of memes. In *Social Information Transmission and Human Biology*, pages 89–96. CRC Press.

Ayoub, P. M., Page, D., and Whitt, S. (2021). Pride amid prejudice: The influence of LGBT+ rights activism in a socially conservative society. *American Political Science Review*, 115(2):467–485.

Badar, K., M. Hite, J., and F. Badir, Y. (2014). The moderating roles of academic age and institutional sector on the relationship between co-authorship network centrality and academic research performance. *Aslib Journal of Information Management*, 66(1):38–53.

Bail, C. A. (2014). The cultural environment: Measuring culture with big data. *Theory and Society*, 43:465–482.

Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J. J., and Vespignani, A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51):21484–21489.

Barth, F. (1969). Introduction. In Barth, F., editor, *Ethnic Groups and Boundaries*, pages 9–38. Little, Brown, and Company, New York.

Bartley, T., Koos, S., Samel, H., Setrini, G., and Summers, N. (2015). *Looking behind the label: Global industries and the conscientious consumer*. Indiana University Press.

Beck, C. J. (2008). The contribution of social movement theory to understanding terrorism. *Sociology Compass*, 2(5):1565–1581.

Bedau, M. A. (1997). Weak emergence. *Philosophical Perspectives*, 11:375–399.

Bell, A. V. and Paegle, A. (2021). Ethnic markers and how to find them. *Human Nature*, 32(2):470–481.

Benford, R. D. and Snow, D. A. (2000). Framing processes and social movements: An overview and assessment. *Annual Review of Sociology*, 26(1):611–639.

Berbrier, M. (2000). The victim ideology of white supremacists and white separatists in the united states. *Sociological Focus*, 33(2):175–191.

Berger, J. and 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.

Berger, J. and Ward, M. (2010). Subtle signals of inconspicuous consumption. *Journal of Consumer Research*, 37(4):555–569.

Bergstrom, C. T. and Rosvall, M. (2011). The transmission sense of information. *Biology & Philosophy*, 26:159–176.

Berlet, C. and Vysotsky, S. (2006). Overview of US white supremacist groups. *Journal of Political and Military Sociology*, 34(1):11–48.

Bischoff, M. (2024). AI matches the abilities of the best Math Olympians. https://www.scientificamerican.com/article/ai-matches-the-abilities-of-the-best-math-olympians/.

Blee, K. M. (1996). Becoming a racist: Women in contemporary Ku Klux Klan and neo-Nazi groups. *Gender & Society*, 10(6):680–702.

Blee, K. M. (2017). How the study of white supremacism is helped and hindered by social movement research. *Mobilization*, 22(1):1–15.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(1):993–1022.

Blumer, H. (1986). *Symbolic interactionism: Perspective and method*. University of California Press.

Bonilla-Silva, E. (2006). *Racism without racists: Color-blind racism and the persistence of racial inequality in the United States*. Rowman & Littlefield Publishers.

Boyd, R. and Richerson, P. J. (1987). The evolution of ethnic markers. *Cultural Anthropology*, 2(1):65–79.

Boyd, R. and Richerson, P. J. (1988). *Culture and the evolutionary process*. University of Chicago Press.

Boyles, A. S. (2019). *You Can't Stop the Revolution: Community Disorder and Social Ties in Post-Ferguson America*. University of California Press.

Brown, M., Ray, R., Summers, E., and Fraistat, N. (2017). #SayHerName: A case study of intersectional social media activism. *Ethnic and Racial Studies*, 40(11):1831–1846.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Bsharat, S. M., Myrzakhan, A., and Shen, Z. (2023). Principled instructions are all you need for questioning LLaMA-1/2, GPT-3.5/4. *arXiv preprint arXiv:2312.16171*.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Burke, P. J. and Stets, J. E. (2009). *Identity Theory*. Oxford University Press.

Burris, V., Smith, E., and Strahm, A. (2000). White supremacist networks on the internet. *Sociological Focus*, 33(2):215–235.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28.

Caiani, M. and Kröll, P. (2015). The transnationalization of the extreme right and the use of the Internet. *International Journal of Comparative and Applied Criminal Justice*, 39(4):331–351.

Campbell, T. (2024). Black lives matter's effect on police lethal use of force. *Journal of Urban Economics*, 141:103587.

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal Of The Royal Statistical Society Series B: Statistical Methodology*, 57(3):473–484.

Castro, L. and Toro, M. A. (2007). Mutual benefit cooperation and ethnic cultural diversity. *Theoretical Population Biology*, 71(3):392–399.

Centola, D. (2015). The social origins of networks and diffusion. *American Journal of Sociology*, 120(5):1295–1338.

Centola, D. and Baronchelli, A. (2015). The spontaneous emergence of conventions: An experimental study of cultural evolution. *Proceedings of the National Academy of Sciences*, 112(7):1989–1994.

Chakraborti, M., Atkisson, C., Stănciulescu, Ş., Filkov, V., and Frey, S. (2024). Do we run how we say we run? Formalization and practice of governance in OSS communities. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–26.

Chen, Y. and Li, S. X. (2009). Group identity and social preferences. *American Economic Review*, 99(1):431–457.

Chen, Jiefeng and Yoon, Jinsung (2024). Introducing ASPIRE for selective prediction in LLMs. `https://blog.research.google/2024/01/introducing-aspire-for-selective.html?m=1`. Accessed: 2024-01-20.

Chetty, N. and Alathur, S. (2018). Hate speech review in the context of online social networks. *Aggression and violent behavior*, 40:108–118.

Chew, R., Bollenbacher, J., Wenger, M., Speer, J., and Kim, A. (2023). LLM-assisted content analysis: Using large language models to support deductive coding. *arXiv preprint arXiv:2306.14924*.

Christakis, N. A. and Fowler, J. H. (2013). Social contagion theory: examining dynamic social networks and human behavior. *Statistics in Medicine*, 32(4):556–577.

Chu, Z., Chen, J., Chen, Q., Yu, W., He, T., Wang, H., Peng, W., Liu, M., Qin, B., and Liu, T. (2023). A survey of chain of thought reasoning: Advances, frontiers and future. *arXiv preprint arXiv:2309.15402*.

Clark, A. D., Dantzler, P. A., and Nickels, A. E. (2018). Black lives matter:(re) framing the next wave of black liberation. In *Research in social movements, conflicts and change*. Emerald Publishing Limited.

Cohen, E. and Haun, D. (2013). The development of tag-based cooperation via a socially acquired trait. *Evolution and Human Behavior*, 34(3):230–235.

Collins, J. (1993). Genericity in the nineties: Eclectic irony and the new sincerity. In *Film theory goes to the movies*, pages 242–63. Routledge.

Crawford, V. P. and Sobel, J. (1982). Strategic information transmission. *Econometrica*, 50(6):1431–1451.

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*.

Dai, S.-C., Xiong, A., and Ku, L.-W. (2023). LLM-in-the-loop: Leveraging large language model for thematic analysis. *arXiv preprint arXiv:2310.15100*.

Danon, L., Díaz-Guilera, A., Duch, J., and Arenas, A. (2005). Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(09):9008–9017.

De Koster, W. and Houtman, D. (2008). 'stormfront is like a second home to me': On virtual community formation by right-wing extremists. *Information, Communication & Society*, 11(8):1155–1176.

DeCook, J. R. (2020). Trust me, I'm trolling: Irony and the Alt-Right's political aesthetic. *M/C Journal*, 23(3).

DellaPosta, D., Shi, Y., and Macy, M. (2015). Why do liberals drink lattes? *American Journal of Sociology*, 120(5):1473–1511.

Dhar, A., Mukherjee, H., Dash, N. S., and Roy, K. (2021). Text categorization: past and present. *Artificial Intelligence Review*, 54:3007–3054.

Di Giovinazzo, V. and Naimzada, A. (2015). A model of fashion: Endogenous preferences in social interaction. *Economic Modelling*, 47:12–17.

DiMaggio, P. (1997). Culture and cognition. *Annual Review of Sociology*, 23(1):263–287.

DiMaggio, P. J. and Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review*, 48(2):147–160.

Dixon, P. J. and Dundes, L. (2020). Exceptional injustice: Facebook as a reflection of race-and gender-based narratives following the death of george floyd. *Social Sciences*, 9(12):231.

Donath, J. S. (1999). Identity and deception in the virtual community. In Kollock, P. and Smith, M., editors, *Communities in Cyberspace*, pages 29–59. Routledge.

Douglas, K. M. (2007). Psychology, discrimination and hate groups online. *The Oxford Handbook of Internet Psychology*, pages 155–164.

Dunivin, Z. and Smaldino, P. E. (2023). Dynamics of covert signaling: Modeling the emergence and extinction of identity signals. *PsyArXiv*, 17.

Dunivin, Z., Zadunayski, L., Baskota, U., Siek, K., and Mankoff, J. (2020). Gender, soft skills, and patient experience in online physician reviews: A large-scale text analysis. *Journal of Medical Internet Research*, 22(7):e14455.

Dunivin, Z. O. (2024a). A lexical approach to locating symbolic boundaries around cultural identities on social media. *Unpublished*.

Dunivin, Z. O. (2024b). Scalable qualitative coding with LLMs: Chain-of-thought reasoning matches human performance in some hermeneutic tasks. *arXiv preprint arXiv:2401.15170*.

Dunivin, Z. O. and Kaminski, P. (2024). Path dependence, stigmergy, and memetic reification in the formation of the 27 club myth. *PNAS*, 121.

Dunivin, Z. O. and Lanigan, A. (2024). White genocide, ethnocrisis, and far-right identity politics: Constructing white nationalist identity through twitter discourse. *Unpublished*.

Dunivin, Z. O., Yan, H. Y., Ince, J., and Rojas, F. (2022). Black lives matter protests shift public discourse. *Proceedings of the National Academy of Sciences*, 119(10):e2117320119.

Eberhard, M. J. W. (1975). The evolution of social behavior by kin selection. *The Quarterly Review of Biology*, 50(1):1–33.

Efferson, C., Lalive, R., and Fehr, E. (2008). The coevolution of cultural groups and ingroup favoritism. *Science*, 321(5897):1844–1849.

Epstein, C. and Epstein, R. (2013). Death in The New York Times: The price of fame is a faster flame. *QJM: An International Journal of Medicine*, 106(6):517–521.

Evon, D. (2017). Did several musicians die with white BIC lighters in their pockets? *Snopes*. Retrieved 13 May 2024.

Feagin, J. (2013). *Systemic racism: A theory of oppression*. Routledge.

Ferber, A. L. (1999). *White Man Falling: Race, Gender, and White Supremacy*. Rowman & Littlefield Publishers.

Ferrell, J. (1999). Cultural criminology. *Annual Review of Sociology*, 25(1):395–418.

Fetner, T. (2008). *How the religious right shaped lesbian and gay activism*, volume 31. University of Minnesota Press.

Fischer, H. (2015). *Gay Semiotics: A Photographic Study of Visual Coding Among Homosexual Men*. Cherry and Martin.

Flamson, T. J. and Bryant, G. A. (2013). Signals of humor: Encryption and laughter in social interaction. In Dynel, M., editor, *Developments in Linguistic Humour Theory*, volume 1, pages 49–73. John Benjamins Publishing, Amsterdam.

Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4):1–30.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3):75–174.

Foster, K. R., Wenseleers, T., and Ratnieks, F. L. (2006). Kin selection is the key to altruism. *Trends in Ecology & Evolution*, 21(2):57–60.

Fox, K. J. (1987). Real punks and pretenders: The social organization of a counterculture. *Journal of Contemporary Ethnography*, 16(3):344–370.

Freund, D. M. (2007). *Colored Property: State Policy and White Racial Politics in Suburban America*. University of Chicago Press.

Futrell, R., Simi, P., and Gottschalk, S. (2006). Understanding music in movements: The white power music scene. *The Sociological Quarterly*, 47(2):275–304.

Geertz, C. (1973). *The interpretation of cultures*. Basic Books.

Gell-Mann, M. (1994). Complex adaptive systems. In *Santa Fe Institue Studies in the Sciences of Complexity Proceedings*, volume 19, pages 17–45. Addison-Wesley.

Gibson, M. A. (2001). Immigrant adaptation and patterns of acculturation. *Human Development*, 44(1):19–23.

Gilbert, L. S. (2002). Going the distance: "Closeness" in qualitative data analysis software. *International Journal of Social Research Methodology*, 5(3):215–228.

Gillion, D. Q. (2013). *The political power of protest: minority activism and shifts in public policy*. Cambridge University Press.

Gillion, D. Q. (2016). *Governing with words: The political dialogue on race, public policy, and inequality in America*. Cambridge University Press.

Giugni, M. (2007). Useless protest? a time-series analysis of the policy outcomes of ecology, antinuclear, and peace movements in the united states, 1977-1995. *Mobilization: An International Quarterly*, 12(1):53–77.

Goffman, E. (1978). *The Presentation of Self in Everyday Life*. Harmondsworth.

Goldberg, A. and Stein, S. K. (2018). Beyond social contagion: Associative diffusion and the emergence of cultural variation. *American Sociological Review*, 83(5):897–932.

Gómez-Rodríguez, C. and Williams, P. (2023). A confederacy of models: A comprehensive evaluation of LLMs on creative writing. *arXiv preprint arXiv:2310.08433*.

Grafen, A. (1991). Modelling in behavioural ecology. In Krebs, J. and Davies, N., editors, *Behavioural Ecology, 3rd edition*, pages 5–31. Blackwell Scientific Publications.

Gray, P. W. (2018). 'The fire rises': Identity, the alt-right and intersectionality. *Journal of Political Ideologies*, 23(2):141–156.

Grindstaff, L., Lo, M.-C. M., and Hall, J. R. (2010). *Handbook of cultural sociology*. Routledge.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Gu, A. and Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Guilbeault, D., Baronchelli, A., and Centola, D. (2021). Experimental evidence for scale-induced category convergence across populations. *Nature Communications*, 12(1):1–7.

Guiso, L., Sapienza, P., and Zingales, L. (2006). Does culture affect economic outcomes? *Journal of Economic Perspectives*, 20(2):23–48.

Hayfield, N., Clarke, V., Halliwell, E., and Malson, H. (2013). Visible lesbians and invisible bisexuals: Appearance and visual identities among bisexual women. *Women's Studies International Forum*, 40:172–182.

Hearn, K. (2020). Peppa pig is gangsta: China's challenging memes. In *Tracing Behind the Image*, pages 73–85. Brill.

Heims, S. J. (1991). *The cybernetics group*. The MIT Press.

Helbing, D., Farkas, I., and Vicsek, T. (2000). Simulating dynamical features of escape panic. *Nature*, 407(6803):487–490.

Henderson, R. and McCready, E. (2017). How dogwhistles work. In *JSAI International Symposium on Artificial Intelligence*, pages 231–240. Springer.

Henrich, J. and Muthukrishna, M. (2021). The origins and psychology of human cooperation. *Annual Review of Psychology*, 72:207–240.

Heyes, C. (2016). Blackboxing: social learning strategies and cultural evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1693):20150369.

Heyes, C. (2018). Enquire within: Cultural evolution and cognitive science. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1743):20170051.

Hill, R. A. and Dunbar, R. I. (2003). Social network size in humans. *Human Nature*, 14(1):53–72.

Hine, G., Onaolapo, J., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Samaras, R., Stringhini, G., and Blackburn, J. (2017). Kek, cucks, and God Emperor Trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1).

Hoffmann, L. (2016). *Postirony: The Nonfictional Literature of David Foster Wallace and Dave Eggers*. transcript Verlag.

Hohle, R. (2015). *Race and the origins of American neoliberalism*. Routledge.

Holland, J. H. (1992). Complex adaptive systems. *Daedalus*, 121(1):17–30.

Holland, J. H. and Miller, J. H. (1991). Artificial adaptive agents in economic theory. *The American Economic Review*, 81(2):365–370.

Hong, L. and Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46):16385–16389.

Hugh, M. (2017). MCMC and Bayesian modeling.

Ilchi, O. S. and Frank, J. (2021). Supporting the message, not the messenger: The correlates of attitudes towards black lives matter. *American journal of criminal justice*, 46(2):377–398.

Ince, J., Rojas, F., and Davis, C. A. (2017). The social media response to Black Lives Matter: How Twitter users interact with Black Lives Matter through hashtag use. *Ethnic and Racial Studies*, 40(11):1814–1830.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., and Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications*, 78:15169–15211.

Jorgensen, D. L. (1989). *Participant Observation: A Methodology for Human Studies*, volume 15. SAGE.

Kadhim, A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review*, 52(1):273–292.

Karpf, D. (2016). *Analytic activism: Digital listening and the new political strategy*. Oxford University Press.

Kenny, D. T. and Asher, A. (2016). Life expectancy and cause of death in popular musicians: is the popular musician lifestyle the road to ruin? *Medical Problems of Performing Artists*, 31(1):37–44.

King, B. G. and Soule, S. A. (2007). Social movements as extra-institutional entrepreneurs: The effect of protests on stock price returns. *Administrative Science Quarterly*, 52(3):413–442.

Kirkpatrick, D. D. (2004). Speaking in the tongue of evangelicals. *New York Times*, 17 October 2004.

Kiser, E. and Hechter, M. (1998). The debate on historical sociology: Rational choice theory and its critics. *American Journal of Sociology*, 104(3):785–816.

Klein Teeselink, B. and Melios, G. (2021). Weather to protest: The effect of black lives matter protests on the 2020 presidential election. *Available at SSRN 3809877*, 0.

Kleinman, S., Stenross, B., and McMahon, M. (1994). Privileging fieldwork over interviews: Consequences for identity and practice. *Symbolic Interaction*, 17(1):37–50.

Konstantinou, L. (2009). *Wipe that smirk off your face: Postironic literature and the politics of character*. Stanford University.

Kozlowski, A. C., Taddy, M., and Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949.

Krawczyk, M. J., Dydejczyk, A., and Kułakowski, K. (2014). The Simmel effect and babies' names. *Physica A: Statistical Mechanics and Its Applications*, 395:384–391.

Kreuter, M. W. and McClure, S. M. (2004). The role of culture in health communication. *Annual Review of Public Health*, 25(1):439–455.

Kudesia, R. S. (2021). Emergent strategy from spontaneous anger: Crowd dynamics in the first 48 hours of the ferguson shooting. *Organization Science*, 32(5):1–25.

Lamont, M. and Molnár, V. (2002). The study of boundaries in the social sciences. *Annual Review of Sociology*, 28(1):167–195.

Lamont, M. and Molnár, V. (2002). The study of boundaries in the social sciences. *Annual Review of Sociology*, 28(1):167–195.

Laouenan, M., Bhargava, P., Eyméoud, J.-B., Gergaud, O., Plique, G., and Wasmer, E. (2022). A cross-verified database of notable people, 3500BC-2018AD. *Scientific Data*, 9(1):290.

Laumann, E. O., Marsden, P. V., and Prensky, D. (1989). The boundary specification problem in network analysis. In *Research Methods in Social Network Analysis*, pages 61–87. Routledge.

Lee, J. J. and Pinker, S. (2010). Rationales for indirect speech: The theory of the strategic speaker. *Psychological Review*, 117(3):785.

Lee, T. (2002). *Mobilizing public opinion: Black insurgency and racial attitudes in the civil rights era*. University of Chicago Press.

Lee, Y.-J. and Roth, W.-M. (2004). Making a scientist: Discursive "doing" of identity and self-presentation during research interviews. *Forum: Qualitative Sozialforschung*, 5(1).

Lin, Z., Trivedi, S., and Sun, J. (2023). Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*.

Liu, X. and Duyn, J. H. (2013). Time-varying functional network information extracted from brief instances of spontaneous brain activity. *Proceedings of the National Academy of Sciences*, 110(11):4392–4397.

Loury, G. C. (1994). Self-censorship in public discourse: A theory of “political correctness" and related phenomena. *Rationality and Society*, 6(4):428–461.

Luhmann, N. (1990). *Essays on self-reference*. Columbia University Press.

MacDonald, K. B. (1994). *A People That Shall Dwell Alone: Judaism as a Group Evolutionary Strategy*. Praeger.

Madaan, A. and Yazdanbakhsh, A. (2022). Text and patterns: For effective chain of thought, it takes two to tango. *arXiv preprint arXiv:2209.07686*.

Maghrabi, R. and Salam, A. F. (2011). Social media, social movement and political change: The case of 2011 cairo revolt. *ICIS 2011 Proceedings*, 8.

Mahoney, J. (2000). Path dependence in historical sociology. *Theory and Society*, 29(4):507–548.

Malik, M. M. (2020). A hierarchy of limitations in machine learning. *arXiv preprint arXiv:2002.05193*.

Mann, L. (1993). Protest movements as a source of social change. *Australian Psychologist*, 28(2):69–73.

McAdam, D. and Su, Y. (2002). The war at home: Antiwar protests and congressional voting, 1965 to 1973. *American Sociological Review*, 67(5):696–721.

McCammon, H. J., Muse, C. S., Newman, H. D., and Terrell, T. M. (2007). Movement framing and discursive opportunity structures: The political successes of the us women’s jury movements. *American Sociological Review*, 72(5):725–749.

McCurdy, P. and Uldam, J. (2014). Connecting participant observation positions: Toward a reflexive framework for studying social movements. *Field Methods*, 26(1):40–55.

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

McPherson, M. (2004). A Blau space primer: Prolegomenon to an ecology of affiliation. *Industrial and Corporate Change*, 13(1):263–280.

Merrin, W. (2019). President troll: Trump, 4chan and memetic warfare. In *Trump’s media war*, pages 201–226. Springer.

Merton, R. K. (1948). The self-fulfilling prophecy. *The Antioch Review*, 8(2):193–210.

Meyer, D. S. and Staggenborg, S. (1996). Movements, countermovements, and the structure of political opportunity. *American journal of sociology*, 101(6):1628–1660.

Milkman, R., Luce, S., and Lewis, P. W. (2013). *Changing the subject: A bottom-up account of Occupy Wall Street in New York City*. CUNY, The Murphy Institute.

Miller, C. M., McIntyre, S. H., and Mantrala, M. K. (1993). Toward formalizing fashion theory. *Journal of Marketing Research*, 30(2):142–157.

Miller, J. H. and Page, S. E. (2009). *Complex adaptive systems: An introduction to computational models of social life: an introduction to computational models of social life*. Princeton University Press.

Mingers, J. (2002). Can social systems be autopoietic? Assessing Luhmann's social theory. *The Sociological Review*, 50(2):278–299.

MistralAI (2023). Mixtral of experts: A high quality sparse mixture-of-experts. `https://mistral.ai/news/mixtral-of-experts`. Accessed: 2024-01-13.

Mittal, S. (2013). Emergence in stigmergic and complex adaptive systems: A formal discrete event systems perspective. *Cognitive Systems Research*, 21:22–39.

Moffett, M. W. (2019). *The Human Swarm: How Our Societies Arise, Thrive, and Fall*. Basic Books.

Monroe, B. L., Colaresi, M. P., and Quinn, K. M. (2008). Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Mullon, C. and Lehmann, L. (2014). The robustness of the weak selection approximation for the evolution of altruism against strong selection. *Journal of Evolutionary Biology*, 27(10):2272–2282.

Murthy, D. (2008). Digital ethnography: An examination of the use of new technologies for social research. *Sociology*, 42(5):837–855.

Murthy, D. (2012a). Towards a sociological understanding of social media: Theorizing twitter. *Sociology*, 46(6):1059–1073.

Murthy, D. (2012b). Towards a sociological understanding of social media: Theorizing Twitter. *Sociology*, 46(6):1059–1073.

Nagle, A. (2017). *Kill all normies: Online culture wars from 4chan and Tumblr to Trump and the alt-right*. John Hunt Publishing.

Nelson, A. (2011). *Body and soul: The Black Panther Party and the fight against medical discrimination*. U of Minnesota Press.

Nelson, L. K. (2017). Computational Grounded Theory: A methodological framework. *Sociological Methods & Research*, 49(1):3–42.

Nelson, L. K. (2020). Computational grounded theory: A methodological framework. *Sociological Methods & Research*, 49(1):3–42.

Nettle, D. and Dunbar, R. (1997). Social markers and the evolution of reciprocal exchange. *Current Anthropology*, 38:93–99.

Noel, H. (2014). *Political ideologies and political parties in America*. Cambridge University Press.

Nowakowski, J. and Keller, J. (2024). AI-powered patching: The future of automated vulnerability fixes. *Google Security Engineering Technical Report*.

Olie, R. (1994). Shades of culture and institutions in international mergers. *Organization Studies*, 15(3):381–405.

Ollion, E., Shen, R., Macanovic, A., and Chatelain, A. (2023). ChatGPT for text annotation? Mind the hype! *SocArXiv preprint doi:10.31235/osf.io/x58kn*.

OpenAI (2023). GPT-4. https://openai.com/research/gpt-4. Accessed: 2024-01-18.

Otto, S. P. and Day, T. (2011). *A Biologist's Guide to Mathematical Modeling in Ecology and Evolution*. Princeton University Press.

Page, S. E. (2006). Path dependence. *Quarterly Journal of Political Science*, 1(1):87–115.

Paul, D., Ismayilzada, M., Peyrard, M., Borges, B., Bosselut, A., West, R., and Faltings, B. (2023). REFINER: Reasoning feedback on intermediate representations. *arXiv preprint arXiv:2304.01904*.

Pedone, R. and Conte, R. (2001). Dynamics of status symbols and social complexity. *Social Science Computer Review*, 19(3):249–262.

Peel, L., Larremore, D. B., and Clauset, A. (2017). The ground truth about metadata and community detection in networks. *Science Advances*, 3(5).

Pepper, S. C. (1926). Emergence. *The Journal of Philosophy*, 23(9):241–245.

Perry, B. L., Yang, K. C., Kaminski, P., Odabas, M., Park, J., Martel, M., Oser, C. B., Freeman, P. R., Ahn, Y.-Y., and Talbert, J. (2019). Co-prescription network reveals social dynamics of opioid doctor shopping. *PloS one*, 14(10):e0223849.

Powell, M., Kim, A. D., and Smaldino, P. E. (2023). Hashtags as signals of political identity: #BlackLivesMatter and #AllLivesMatter. *PLOS ONE*, 18(6):e0286524.

Puglisi, A., Baronchelli, A., and Loreto, V. (2008). Cultural route to the emergence of linguistic categories. *Proceedings of the National Academy of Sciences*, 105(23):7936–7940.

Roberts, H., Bhargava, R., Valiukas, L., Jen, D., Malik, M. M., Bishop, C., Ndulue, E., Dave, A., Clark, J., Etling, B., et al. (2021). Media cloud: Massive open source collection of global news on the open web. *arXiv preprint arXiv:2104.03702*, 0.

Rodrigues, A. M. and Kokko, H. (2016). Models of social evolution: Can we do better to predict "who helps whom to achieve what"'? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371.

Rojas, F. (2010). *From black power to black studies: How a radical social movement became an academic discipline*. JHU Press.

Romera-Paredes, B., Barekatain, M., Novikov, A., Balog, M., Kumar, M. P., Dupont, E., Ruiz, F. J., Ellenberg, J. S., Wang, P., Fawzi, O., et al. (2023). Mathematical discoveries from program search with large language models. *Nature*, 625:1–3.

Rosen, R. (1987). On complex systems. *European Journal of Operational Research*, 30(2):129–134.

Saini, A. (2019). *Superior: The Return of Race Science*. Beacon Press.

Salant, T. and Lauderdale, D. S. (2003). Measuring culture: A critical review of acculturation and health in Asian immigrant populations. *Social Science & Medicine*, 57(1):71–90.

Saldaña, J. (2009). *The coding manual for qualitative researchers*. SAGE Publications.

Salganik, M. J., Dodds, P. S., and Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–856.

Sato, Y. (2024). Sociological meaning of contagion. In *Sociological Foundations of Computational Social Science*, pages 91–100. Springer.

Sawyer, J. and Gampa, A. (2018). Implicit and explicit racial attitudes changed during black lives matter. *Personality and Social Psychology Bulletin*, 44(7):1039–1059.

Schudson, M. (1989). How culture works: Perspectives from media studies on the efficacy of symbols. *Theory and Society*, pages 153–180.

Shinn, N., Labash, B., and Gopinath, A. (2023). Reflexion: An autonomous agent with dynamic memory and self-reflection. *arXiv preprint arXiv:2303.11366*.

Shweder, R. A. (1999). Why cultural psychology? *Ethos*, 27(1):62–73.

Simi, P. and Futrell, R. (2006). Cyberculture and the endurance of white power activism. *Journal of Political and Military Sociology*, 34(1):115–142.

Simmel, G. (1904). Fashion. *International Quarterly*, 10:130—155.

Simon, H. A. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society*, 106(6):467–482.

Simon, H. A. (1990). Bounded rationality. In Eatwell, J., Milgate, M., and Newman, P., editors, *Utility and Probability*, pages 15–18. Palgrave Macmillan UK, London.

Singer, P., Lemmerich, F., West, R., Zia, L., Wulczyn, E., Strohmaier, M., and Leskovec, J. (2017). Why we read wikipedia. In *Proceedings of the 26th international conference on world wide web*, pages 1591–1600.

Skyrms, B. (2010). *Signals: Evolution, Learning, and Information*. Oxford University Press.

Sloman, S. J., Oppenheimer, D. M., and DeDeo, S. (2021). Can we detect conditioned variation in political speech? Two kinds of discussion and types of conversation. *PLOS ONE*, 16(2):e0246689.

Smaldino, P. (2023). *Modeling social behavior: Mathematical and agent-based models of social dynamics and cultural evolution*. Princeton University Press.

Smaldino, P. E. (2017). Models are stupid, and we need more of them. In Vallacher, R. R., Nowak, A., and Read, S. J., editors, *Computational Social Psychology*, pages 311–331. Routledge.

Smaldino, P. E. (2019). Social identity and cooperation in cultural evolution. *Behavioural Processes*, 161:108–116.

Smaldino, P. E. (2022). Models of identity signaling. *Current Directions in Psychological Science*, 31:231–237.

Smaldino, P. E., Flamson, T. J., and McElreath, R. (2018). The evolution of covert signaling. *Scientific Reports*, 8(1):1–10.

Smaldino, P. E. and Turner, M. A. (2022). Covert signaling is an adaptive communication strategy in diverse populations. *Psychological Review*, 129(4):812–829.

Smalls, K. A. (2018). Languages of liberation: Digital discourses of emphatic blackness. In *Language and Social Justice in Practice*, pages 52–60. Routledge.

Smith, D. E. (2005). *Institutional ethnography: A sociology for people*. Rowman Altamira.

Smith, E. A. and Winterhalder, B. (1992). Natural selection and decision-making: Some fundamental principles. In *Evolutionary Ecology and Human Behavior*, pages 25–60. De Gruyter.

Sounes, H. (2013). *27: A History of the 27 Club Through the Lives of Brian Jones, Jimi Hendrix, Janis Joplin, Jim Morrison, Kurt Cobain, and Amy Winehouse*. Da Capo Press.

Stiefel, K. M. and Ermentrout, G. B. (2016). Neurons as oscillators. *Journal of Neurophysiology*, 116(6):2950–2960.

Strauss, A. and Corbin, J. M. (1997). *Grounded theory in practice*. SAGE Publications.

Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Aldine.

Summers, N. (2016). Ethical consumerism in global perspective: A multilevel analysis of the interactions between individual-level predictors and country-level affluence. *Social Problems*, 63(3):303–328.

Swidler, A. (1986). Culture in action: Symbols and strategies. *American Sociological Review*, pages 273–286.

Tabilo Alvarez, J. and Ramírez-Correa, P. (2023). A brief review of systems, cybernetics, and complexity. *Complexity*, 2023(1):8205320.

Tai, R. H., Bentley, L. R., Xia, X., Sitt, J. M., Fankhauser, S. C., Chicas-Mosier, A. M., and Monteith, B. G. (2023). An examination of the use of large language models to aid analysis of textual data. *bioRxiv preprint bioRxiv:2023.07.17.549361*.

Tavory, I. (2010). Of yarmulkes and categories: Delegating boundaries and the phenomenology of interactional expectation. *Theory and Society*, 39(1):49–68.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. (2023). Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Thomas, W. I. and Thomas, D. S. (1938). *The child in America*. Alfred A. Knopf.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). LLaMa: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tsesis, A. (2002). *Destructive messages: How hate speech paves the way for harmful social movements*. NYU Press.

Tufekci, Z. (2017). *Twitter and tear gas*. Yale University Press.

Turchin, P. (2003). *Historical Dynamics: Why States Rise and Fall*. Princeton University Press.

Turchin, P. (2011). Toward cliodynamics: An analytical, predictive science of history. *Cliodynamics*, 2:167––186.

Tylén, K., Fusaroli, R., Bundgaard, P. F., and Østergaard, S. (2013). Making sense together: A dynamical account of linguistic meaning-making. *Semiotica*, 194:39–62.

Underhill, M. R. (2018). Parenting during Ferguson: Making sense of white parents' silence. *Ethnic and Racial Studies*, 41(11):1934–1951.

Updegrove, A. H., Cooper, M. N., Orrick, E. A., and Piquero, A. R. (2020). Red states and black lives: Applying the racial threat hypothesis to the Black Lives Matter movement. *Justice Quarterly*, 37(1):85–108.

Urbatsch, R. (2014). Nominal partisanship: Names as political identity signals. *PS: Political Science & Politics*, 47(2):463–467.

Vach, W. and Gerke, O. (2023). Gwet's AC1 is not a substitute for Cohen's kappa – A comparison of basic properties. *MethodsX*, 10:102212.

van der Does, T., Galesic, M., Dunivin, Z. O., and Smaldino, P. E. (2022). Strategic identity signaling in heterogeneous networks. *Proceedings of the National Academy of Sciences*, 119(10):e2117898119.

Van Dyke Parunak, H. (2005). A survey of environments and mechanisms for human-human stigmergy. In *International Workshop on Environments for Multi-agent Systems*, pages 163–186. Springer.

Van Laer, J. and Van Aelst, P. (2010). Internet and social movement action repertoires: Opportunities and limitations. *Information, Communication & Society*, 13(8):1146–1171.

Van Valen, L. (1973). A new evolutionary law. *Evolutionary Theory*, 1:1–30.

Varela, F. G., Maturana, H. R., and Uribe, R. (1974). Autopoiesis: The organization of living systems, its characterization and a model. *Biosystems*, 5(4):187–196.

Vasi, I. B., Walker, E. T., Johnson, J. S., and Tan, H. F. (2015). "No fracking way!" Documentary film, discursive opportunity, and local opposition against hydraulic fracturing in the united states, 2010 to 2013. *American Sociological Review*, 80(5):934–959.

Vaughan, D. (1998). Rational choice, situated action, and the social control of organizations. *Law & Society Review*, 32(1):23–61.

Wasow, O. (2020). Agenda seeding: How 1960s black protests moved elites, public opinion and voting. *American Political Science Review*, 114(3):638–659.

Watson, D. R. and Weinberg, T. S. (1982). Interviews and the interactional construction of accounts of homosexual identity. *Social Analysis*, 11:56–78.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

Weaver, A. A. (2008). Does protest behavior mediate the effects of public opinion on national environmental policies? A simple question and a complex answer. *International Journal of Sociology*, 38(3):108–125.

Weaver, W. (1948). Science and complexity. *American Scientist*, 36(4):536–544.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Wilkins, D. J., Livingstone, A. G., and Levine, M. (2019). Whose tweets? The rhetorical functions of social media use in developing the Black Lives Matter movement. *British Journal of Social Psychology*, 58(4):786–805.

Williamson, V., Trump, K.-S., and Einstein, K. L. (2018). Black lives matter: Evidence that police-caused deaths predict protest activity. *Perspectives on Politics*, 16(2):400–415.

Wimmer, A. (2008). The making and unmaking of ethnic boundaries: A multilevel process theory. *American Journal of Sociology*, 113(4):970–1022.

Wolkewitz, M., Allignol, A., Graves, N., and Barnett, A. G. (2011). Is 27 really a dangerous age for famous musicians? Retrospective cohort study. *BMJ*, 343.

Wood, M. L., Stoltz, D. S., Van Ness, J., and Taylor, M. A. (2018). Schemas and frames. *Sociological Theory*, 36(3):244–261.

Xiao, Z., Yuan, X., Liao, Q. V., Abdelghani, R., and Oudeyer, P.-Y. (2023). Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*.

Xu, Z., Shi, S., Hu, B., Yu, J., Li, D., Zhang, M., and Wu, Y. (2023). Towards reasoning in large language models via multi-agent peer review collaboration. *arXiv preprint arXiv:2311.08152*.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. (2022). ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Yuan, W., Pang, R. Y., Cho, K., Sukhbaatar, S., Xu, J., and Weston, J. (2024). Self-rewarding language models. *arXiv preprint arXiv:2401.10020*.

Zhou, K., Jurafsky, D., and Hashimoto, T. (2023). Navigating the grey area: Expressions of overconfidence and uncertainty in language models. *arXiv preprint arXiv:2302.13439*.

APPENDIX A

CHAPTER 3

Table A.1: Google search terms by thematic category

BLM	**Victims**	**Historical**	**Historical Figure**
Black Lives Matter	Eric Garner	Abolition	Angela Davis
Police Shootings	Michael Brown	Black Panthers	Malcolm X
	Philando Castile	Black Power	Martin Luther King
	Tamir Rice	Civil Rights	Stokely Carmichael
	Trayvon Martin	Jim Crow	
		Racism	
		Segregation	
		Slavery	
Slogan	**Identity Politics**	**Policy**	**White Supremacy**
Hands Up Don't Shoot	Antiracist	Decriminalization	White Nationalism
I Can't Breath	Identity Politics	Defund the Police	White Power
Say Her Name	Inequality	Mass Incarceration	White Supremacy
	Institutional Racism	Prison Abolition	
	Social Justice	Prison Reform	
	White Privilege	Redlining	
		The New Jim Crow	
		War on Drugs	

Table A.2: Contagious protest events as modeled in Figure 3

Era	Event(s)	Start	End	Duration	Criterion
2014–2015	Eric Garner	2014-07-17	2014-07-25	9	80%
2014–2015	Michael Brown	2014-08-12	2014-08-21	8	80%
2014–2015	Hearings, Tamir Rice	2014-11-24	2014-12-05	12	80%
2014–2015	Freddie Gray	2015-04-20	2015-04-29	10	8 month max, 80%
2016	Philando Castile	2016-07-05	2016-07-16	12	8 month max
2016	National Anthem	2016-09-18	2016-09-28	11	8 month max
2017	Unite the Right	2017-08-11	2017-08-21	11	8 month max
2020	George Floyd	2020-05-25	2020-06-05	12	8 month max

Note: Each event belongs to a particular protest "era" that is represented by a single binary variable in the general additive models of search volume spikes.

Table A.3: Generalized additive model (GAM) of co-activation as nonparametric functions of time as visualized as the trend line for Figure 3

	Coefficient
Intercept	-0.14***
Approximate significance of smooth terms:	
Calendar Week[a]	8.94***
Time[a]	7.41***
N	2857
Adj. R^2	0.24

Notes: * P<.05; ** P<.01; *** P<.001

[a] Estimated Degrees of Freedom

Table A.4: Bootstrapped estimates for GAMs of Google search spikes for all thematic term categories as visualized in Figure 4

	All Terms	**BLM**	**Victim**	**Historical**	**Historical Figure**	**Slogan**	**Identity Politics**	**Policy**	**White Supremacy**
Intercept (no protest)	−0.03**	−0.08*	−0.03	−0.02**	−0.03	−0.07	−0.04	−0.01	−0.01
2014-2015	0.38***	1.39***	1.55***	0.15*	0.38**	1.07*	0.45	−0.17	0.24
Castile	0.97***	2.72***	1.41***	0.57***	0.52**	1.11	1.00**	1.05***	0.12
Unite the Right	0.75***	1.46***	−0.25	0.77***	0.34	1.24	1.06*	−0.25	3.09***
George Floyd	2.02***	4.02***	2.52***	1.18***	1.62***	1.91*	1.75***	1.81***	1.47**
y-lag	−0.34***	−0.10	−0.15***	−0.02	−0.01	−0.46***	−0.36***	−0.36***	−0.37***
Approximate significance of smooth terms:									
Calendar Week[a]	8.12***	1.86	1.00	8.59***	8.02***	1.00	4.07***	7.81***	3.47
Time[a]	5.12***	2.17	1.73	3.67	1.50	1.00	1.00	1.00	1.00
Adj. R^2	0.15	0.33	0.13	0.21	0.08	0.23	0.15	0.16	0.19

Notes: * *P*<.05; ** *P*<.01; *** *P*<.001

[a] Estimated Degrees of Freedom

GAM results in Table A.4 give the coefficients and goodness of fit for models of search spikes. The primary coefficients of interest, whether there is a spike in Google Searches during a particular protest are visualized in Figure 4 with 95% confidence intervals. As in Figure 4 each thematic category of search terms (see Supporting Table A.1) is fit to a separate model, with an additional model for all 41 terms. We make some observations here about these models in general and then focus on a handful of issues.

First, search data is relatively noisy in the sense that goodness of fit statistics indicate that modeling search behavior on seasonal patterns yields low to moderate adjusted r-squared estimates, even when including terms for protest waves. Goodness of fit statistics range from $R^2 = .08$ to $R^2 = .33$. Second, as noted in the main text, different clusters of terms are associated with different protest waves. The cluster of terms least explained by protest and seasonal variation are historical figures ($R^2 = .08$), which is an intuitive finding, as each figure has spikes on their birthdays, which are not accounted for in the model. The cluster that is most associated is protest is the group of terms denoting BLM and police violence ($R^2 = .33$). Third, the protest that has the largest and most consistent effect is the George Floyd wave. In contrast, the earliest BLM protest wave in 2014/2015 had inconsistent effects across different clusters of terms.

Table A.5: OLS models of daily tweet volume ($\log_{10}$) for four hashtag time series

	Post-Unite the Right[a]	**Pre- & Post-George Floyd[b]**
Intercept	3.5240***	3.2809***
Time[c]	0.0001	−0.0008***
#AllLivesMatter	−1.1230***	−0.9095**
#BlueLivesMatter	−0.7225***	−0.4892**
#WhiteLivesMatter	−1.7830***	−2.4131***
Time×#AllLivesMatter	−0.0001	0.0005**
Time×#BlueLivesMatter	0.0002	0.0005*
Time×#WhiteLivesMatter	−0.0020***	0.0005**
Post-George Floyd		1.6010***
Time×Post-GF		−0.0055***
#AllLivesMatter×Post-GF		−0.3373*
#BlueLivesMatter×Post-GF		−0.8242***
#WhiteLivesMatter×Post-GF		0.5901***
Time×#AllLivesMatter×Post-GF		−0.0015*
Time×#BlueLivesMatter×Post-GF		−0.0012
Time×#WhiteLivesMatter×Post-GF		−0.0024***
Adjusted R^2	0.8687	0.9324

Notes: * *P*<.05; ** *P*<.01; *** *P*<.001

[a] The first model estimates the decline of hashtag volume in the year following the Unite the Right rally and BLM counter-protests.

[b] The second model includes a structural break to compare the pre-George Floyd baseline (2019) to a post-George Floyd baseline (August-December 2020); all 8 lines from the second model (pre- and post-George Floyd for each of the 4 hashtags) are visualized in the bottom left panel of Figure A.1.

[c] Time is measured from the beginning of each period (2017-08-11, 2019-01-01, or 2020-08-01,) such that the intercept represents expected number of tweets on Day 0.

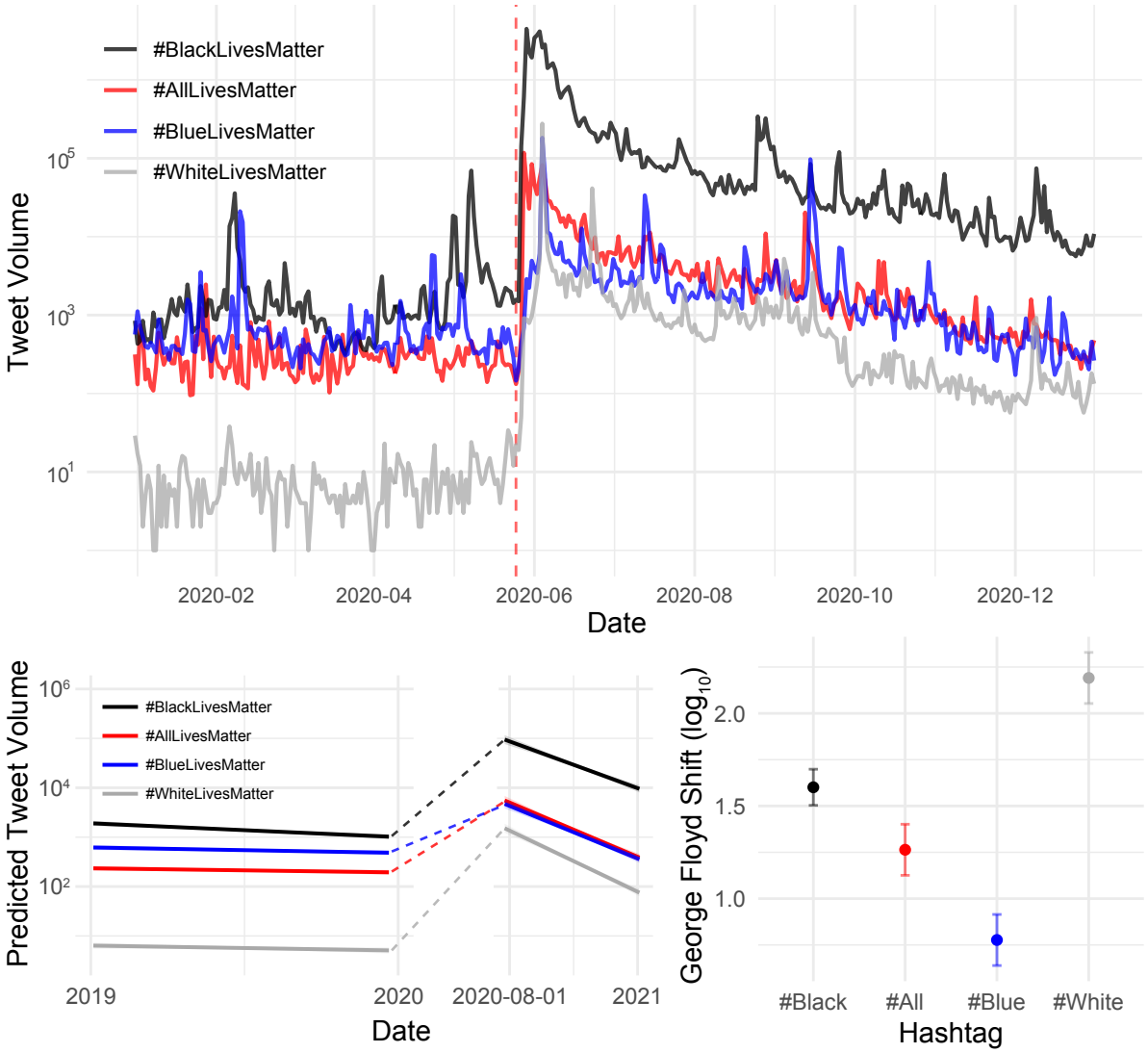

Figure A.1: **Weekly hashtag volume for #BlackLivesMatter and three opposing terms before and after George Floyd protests** Top: Daily hashtag volume for 2021 (George Floyd protests begin at the red dashed line.) Bottom left: Regression lines for pre- and post-George Floyd "baselines" for four hashtags corresponding to Table A.5. Bottom right: Expected shift in tweet volume from 2019 to new "baseline" beginning August 2020 with 95% confidence intervals, corresponding to the dashed lines in the the left panel. All pairwise comparisons are significantly different at P<0.0001, except #BlackLivesMatter–#WhiteLivesMatter (P=0.0014).

Table A.6: Baseline shift in term frequency for post- vs. pre-George Floyd protests

Term	Search	News		Twitter		Wikipedia		Wikipedia Page
	Ratio[a]	Ratio[a]	Volume[b]	Ratio[a]	Volume[b]	Ratio[a]	Volume[b]	
Abolition	0.92	2.20	9	2.57	4352	1.05	1500	Abolitionism
Angela Davis	1.72	3.32	2	1.76	1087	2.06	3817	Angela Davis
Antiracist	28.75	13.85	2	3.97	989	2.43	696	Anti-racism
Black Lives Matter	12.23	22.93	171	32.35	52 622	9.65	19 019	Black Lives Matter
Black Panthers	1.28	2.04	2	1.27	612	1.33	4704	Black Panther Party
Black Power	1.39	3.19	3	2.09	7212	1.12	580	Black Power
Civil Rights	1.11	1.89	134	2.02	21 890	1.16	3138	Civil rights movement
Decriminalization	1.38	1.28	3	1.22	531	0.66	231	Decriminalization
Defund the Police	inf	inf	24	3699.34	13 368			
Eric Garner	0.46	1.05	3	0.14	241	0.80	1562	Killing of Eric Garner
Hands Up Don't Shoot	nan	nan	0	6.77	187			
I Can't Breath	1.87	nan	0	1.62	1505			
Identity Politics	0.91	1.40	7	1.03	2904	1.53	1751	Identity politics
Inequality	1.19	1.61	76	1.19	14 753	1.05	1199	Wealth inequality
Institutional Racism	1.83	4.25	6	1.65	996	2.75	2927	Institutional racism
Jim Crow	1.29	2.21	16	2.72	3733	2.23	12 752	Jim Crow laws
Malcolm X	0.88	3.22	5	1.81	2695	1.02	9747	Malcolm X
Martin Luther King	0.76	1.94	27	1.37	4553	1.08	12 519	Martin Luther King Jr.
Mass Incarceration	1.28	1.45	9	1.29	1639			
Michael Brown	1.02	1.84	5	2.17	1340	1.38	3510	Shooting of Michael Brown
Philando Castile	inf	3.17	1	5.81	331	1.67	1814	Shooting of Philando Castile
Police Brutality	1.71	5.08	77	5.01	29 359	1.15	942	Police brutality
Police Shootings	2.06	2.15	6	2.31	11 485			
Prison Abolition	inf	inf	0	2.36	298	1.64	304	Prison abolition movement
Prison Reform	1.17	1.38	3	1.37	826	0.88	221	Prison reform
Racism	1.68	2.82	232	1.58	121 855	1.02	3678	Racism
Redlining	2.17	2.98	4	2.22	480	2.02	2193	Redlining
Say Her Name	2.59	inf	2	2.93	2928	5.20	298	SayHerName
Segregation	1.08	1.72	21	1.74	4554	1.11	1802	Racial segregation
Slavery	1.04	1.74	46	1.46	24 815	1.03	3057	Slavery
Social Justice	1.25	3.21	69	2.14	13 845	1.19	1952	Social justice
Stokely Carmichael	1.07	1.62	0	2.94	50	1.90	2040	Stokely Carmichael
Systemic Racism	9.61	7.73	7	9.44	10 313	5.60	150	Systemic racism
Tamir Rice	2.31	3.50	1	2.01	375	1.97	1986	Shooting of Tamir Rice
The New Jim Crow	1.29	30.48	0	2.02	133	1.57	553	The New Jim Crow
Trayvon Martin	7.68	2.65	4	5.98	1081	1.48	1967	Trayvon Martin
War on Drugs	1.16	1.25	6	1.17	1626	1.09	1893	War on drugs
White Nationalism	0.43	0.54	4	0.34	1039	0.81	1062	White nationalism
White Power	1.26	1.95	3	1.50	8253			
White Privilege	1.41	2.78	9	1.43	8829	1.32	1722	White privilege
White Supremacy	1.30	1.76	31	1.32	18 229	1.19	3630	White supremacy

Notes: In addition to the raw data for Figure 5 (Wikipedia page visits), we show the same data for Google Search, news media, and Twitter.

[a] The ratio of expected daily volume for August-December 2020 over the same period in 2019.

[b] The expected daily volume for the post-George Floyd period.

VITA

Zackary Okun Dunivin https://zduniv.in

Education

Indiana University
Ph.D. in Informatics, Complex Networks and Systems, Aug. 2024
Ph.D. in Sociology, Dec. 2024

M.A., Sociology, Oct. 2020

Reed College
B.A., Biology, Dec. 2015
Concentration in Philosophy and Computation

Research Interests

cultural sociology & cultural evolution
organizational topology & decision-making
text mining & natural language processing
techno-bureaucracy & information systems
complex networks & systems

BA Thesis

"Quantifying the Spatiotemporal Patterning of *atoh7* Expression in the Developing Zebrafish Retina."

Affiliations

University of California, Davis
Post-doctoral Scholar, 2024–, Davis, CA
Department of Communication
Department of Computer Science

Mimbres School
Center for Humanistic Exploration of Complex Problems
Teaching Fellow, 2021–2023, Mimbres Valley, NM

Bloomington Cooperative Living
70 member affordable housing collective
Board President, 2019–2022, Bloomington, IN
Board Member, 2017–2019