

MEASURING OFFLINE EFFECTS OF ONLINE SOCIAL MEDIA

John Bollenbacher

Submitted to the faculty of the Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of Informatics
Indiana University
December 2023

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

Doctoral Committee:

Filippo Menczer, PhD
Informatics
Indiana University

Alessandro Flammini, PhD
Informatics
Indiana University

Yong Yeol Ahn, PhD
Informatics
Indiana University

Michael Trosset, PhD
Statistics
Indiana University

December 4th, 2023

ACKNOWLEDGMENTS

The body of work presented here was developed over the years from early 2020 to mid 2023, and it would not have been possible without support from friends, family, mentors, and collaborators.

I'd like to thank my advisors and mentors, Fil Menczer and John Bryden for their help. Fil's contributions to writing and refining the main case study of this work significantly improved it, and John's steady collaboration through 2020-2022 made possible each of the published papers that comprise the core of this dissertation. I'd also like to acknowledge Fil and the NaN research group for the foundational education in social media studies they provided in the early years of my graduate studies from 2016-2019, which provided the conceptual frameworks and skills for my projects presented here.

Thanks to my family, who have provided significant support through these years, especially my parents Michael Bollenbacher and Shonna Cole and my grandparents Les and Judy Cole. Their support made the long work of completing a PhD program easier and less troubled by the difficulties of a grad student's wages. I'd also like to acknowledge their significant efforts in my prior education and upbringing which made possible pursuing a PhD program in the first place, especially for the opportunities to attend Webb and Georgia Tech which laid excellent foundations for my later studies.

Thanks to Marissa Donofrio, whose companionship during the most difficult and uncertain years of our PhD programs made those years more fun.

Finally, I'd like to thank all the collaborators of the CoVaxxy team, especially Matthew DeVerna, Francesco Pierri, Bao Tran Truong, and John Bryden, whose timely and careful work in the early months of the pandemic collected the data that made this dissertation's main case study possible. This body of work could not have been completed without their efforts.

John Bollenbacher

MEASURING OFFLINE EFFECTS OF ONLINE SOCIAL MEDIA

Online social media is widely believed to have offline effects on society, including in politics and public health. However, to date it has been difficult to establish causal connections between online content and offline outcomes. In this work, I show how modern causal inference methods and data science tools can be applied to link online causes to offline effects, and measure the magnitude of these effects.

I note three significant points of the proposed method: 1. Online populations are identified with offline populations through linkage variables, such as the geolocation of online users. 2. A metric of social exposure to online content is defined, and is used to estimate the exposure of offline populations, 3. Observational causal inference methods including causal graphical modeling are used to estimate the average treatment effect of exposures on populations, leveraging existing domain knowledge of the offline phenomena to control for confounders.

I demonstrate the proposed methods through a public health case study in which we show that exposure to antivaccine content on Twitter reduced vaccine uptake during the COVID-19 pandemic, leading to increased cases and deaths. Additionally, I detail the necessary data collection and data preparation procedures, especially methods for identifying the relevant online content (e.g. antivaccine tweets) and linking online users to offline populations (e.g. via geolocation). Finally, I also demonstrate an individual-level mechanism for the population-scale effects observed in the case studies; in particular, I show that when an individual interacts with a single piece of online content, this changes their attention to topics and sentiments toward named entities (e.g. people, places, groups).

This body of work advances social media studies by establishing methods for linking online causes to offline effects, defining the limits of these methods, and demonstrating a plausible mechanism by which these effects may be produced.

TABLE OF CONTENTS

Acceptance Page	ii
Acknowledgments	iii
Abstract	iv
Chapter 1: Introduction	1
1.1 Motivation	1
1.2 Aims	2
1.3 Approach	2
Chapter 2: Literature and Concepts of Causation for Social Media Studies	5
2.1 Overview of Philosophy of Causation	5
2.2 Operationalizing Causal Philosophy for Social Media Studies	9
2.3 Key Causal Concepts	10
2.4 Overview of Popular Causal Inference Frameworks	14
2.5 Prior Use of Causal Methods in Social Media Studies	21
2.6 Conclusion	25
Chapter 3: A Method for Linking Online Causes to Offline Effects	28
3.1 Online and Offline Data	28

3.2	Linking Online Users to Offline Populations	29
3.3	Measuring Social Exposure to Online Content	30
3.4	Causal Modeling of Social Exposures and Offline Outcomes	32
3.4.1	Considerations for Choosing Causal Methods	32
3.4.2	Using Domain Knowledge to Control for Confounders	35
3.4.3	Additional Modeling Tricks	38
3.5	Summarizing the Method	39
3.6	Necessary Conditions and Limitations the Method	40
3.7	A Worked Example of the Method in Brief	42
3.8	Conclusion	45
Chapter 4: Data Collection for Antivaccine Tweets Case Study		46
4.1	Introduction	47
4.2	Dataset Curation	48
4.2.1	Identifying COVID-19 Vaccines Content	48
4.2.2	Content Coverage	49
4.3	CoVaxxy Infrastructure	51
4.3.1	Data Collection Architecture	51
4.3.2	Dashboard	52
4.4	Data Characterization	53
4.4.1	Volume	54
4.4.2	Hashtags	55
4.4.3	Sources	57

4.5	Discussion	58
Chapter 5: Case Study: Antivaccine Tweets and COVID-19 Vaccine Hesitancy		62
5.1	Introduction	63
5.2	Antivaccine Tweets Increase Vaccine Hesitancy	65
5.3	Antivaccine Tweets Prevent Vaccinations	67
5.4	Discussion	68
5.5	Methods	71
5.5.1	Data	71
5.5.2	Antivaccine Tweet Classifier	72
5.5.3	Exposure to Antivaccine Tweets	73
5.5.4	SIRVA Model	73
5.5.5	Estimating Parameters	75
5.5.6	Effect Estimations	78
5.5.7	Model Selection Criteria	85
Chapter 6: Case Study: Social Media and Legislative Agenda Setting		86
6.1	Introduction	87
6.2	Methods	90
6.2.1	Datasets	90
6.2.2	Measuring Changes in Text Content	92
6.2.3	Statistical Analyses	96
6.3	Results	97
6.3.1	Overall Content Flow Between Domains	97

6.3.2	Marginal Change with Social Media interactions	100
6.4	Discussion & Conclusions	102
Chapter 7: A Mechanism: Social Media’s Memetic Effects on Individuals		105
7.1	Approach	106
7.1.1	Causal Considerations	107
7.2	Methods	110
7.3	Results	112
7.4	Discussion	114
7.5	Future Work	115
7.6	Methods for Creating Event Sequences	116
7.6.1	Topic Extraction	117
7.6.2	Measuring Expressed Sentiment Toward Entities	118
7.6.3	Textual Similarity Measures	120
Chapter 8: Discussion and Conclusion		123
8.1	Main Contributions	123
8.2	Limitations and Revisions	123
8.3	Future Work	125
Appendices		127
Appendix A: Causal Graphical Modeling		127
A.1	Causal Graphs	127
A.2	Identifiability	128

A.2.1	Conditions for Identifiability	128
A.2.2	Definitions of Key Concepts in CGM Identifiability	129
A.2.3	Identifiability Cases	130
A.3	Effect Estimation	132
A.3.1	Do-Calculus Axioms	132
A.3.2	Average Treatment Effect (ATE) Estimates	133
A.4	CGM and Causal Loops	134
Appendix B: Transformer Language Models		136
Appendix C: Multivariate Hawkes Processes		139
C.1	Point Processes	139
C.2	Hawkes Processes	139
C.3	Multivariate Hawkes Processes	140
C.4	Inferring Excitation Functions	141
C.5	Causation and Confounders in Hawkes Processes	143
Appendix D: Granger Causality		144
Appendix E: Compartmental Epidemic Models		146
References		148
Curriculum Vitae		

CHAPTER 1

INTRODUCTION

Online social media is widely believed to influence society, especially in domains like politics and public health. For over a decade now, there has been intense academic study of social media's impacts, with significant success in analyzing the online social phenomena and some success in demonstrating offline effects. However, it remains difficult in general to prove and quantify offline effects of online social media activity. In this dissertation, I propose and demonstrate a methodology for linking online causes to offline effects on society and explore plausible mechanisms for these effects.

1.1 Motivation

Strong methods for linking online causes to offline effects may be of significant scientific utility in a number of domains. Social media content has been hypothesized to induce a variety of effects in populations including changes in exercise habits [1], increases eating disorders [2], worsened mental health [3, 4], changes in vaccination decisions [5], financial market effects [6], changes in citizens' voting behaviors [7], changes in politicians' legislative votes [8], increases in political protest activity [9, 10, 11], increases in hate crimes [12, 13], and even inciting genocides [14]. Enabling us to better understand and quantify these kinds of effects is the primary motive for this work. Some of these hypothesized impacts of social media are well proven and quantified, but most are not. This work aims to enable researchers to approach these hypotheses with stronger methods, to more rigorously establish causation and quantify the magnitude of these effects.

1.2 Aims

The primary aim in this body of work is develop a methodology for causally linking on-line social exposures to offline outcomes in populations, and quantifying the magnitude of these effects. This methodology will be grounded in prior literature on the philosophy of causation, causal inference methods, and prior causal studies of social media. I will clearly define the limits of the methodology, state the conditions under which it can be applied, and demonstrate its use in a major case study. This methodology is the central contribution of this work to the field of social media studies.

A secondary aim of this work is to propose a plausible set of mechanisms by which social media activity might produce the observed population-level effects through measurable effects on individual users who are exposed to social media content, and take steps toward demonstrating these mechanisms. This study of plausible mechanisms is a key part of making justifiable claims of causation in social media studies.

1.3 Approach

In this section, I'll outline in brief my approach to these aims in this work. I'll start by reviewing the literature on philosophy of causation, causal inference methods, and prior causal inference work in social media studies (Chapter 2). Through this review, I'll assess the applicability of various causal methods and definitions of causation for our purposes in social media studies. Throughout this assessment, I'll apply a pragmatic criterion for evaluating the definitions and methods: namely, does the definition allow us to address our goals of identifying and managing the offline effects of social media, and do the inference methods allow us to appropriately quantify and model these effects considering known confounders?

Next, I'll outline the methodology for linking online causes to offline effects (Chapter 3), consisting of three key points. First, online users are linked with offline population

groups via linkage variables such as geolocation and demographic features (Section 3.2). Second, I'll define a metric of social exposures to online content, and show how it can be used to estimate the social exposures of offline populations for causal inferences involving online social exposure and offline outcomes (Section 3.3). Finally, having obtained temporal data for both online exposures and offline outcomes in many populations, we can apply modern causal inference methods and domain knowledge about the offline phenomena to estimate the magnitude of effects of online social exposure on the offline population outcomes (Section 3.4). This chapter will also detail limits of the method (Section 3.6) and illustrate the method with a brief summary of one of the case studies, with special focus on the three main points of the method outlined above (Section 3.7).

The following chapters will be a series of case studies. The first case study exemplifies the typical data collection and processing methods required to obtain robust estimates of online social exposures in offline populations (Chapter 4). The second case study (Chapter 5) demonstrates the use of the method proposed in Chapter 3 in the public health domain, and establishes that exposure to online social media content can indeed have significant effects on public health outcomes. The final case study illustrates a more traditional non-causal analysis in social media studies, representing the kind of work I aim to move beyond and improve (Chapter 6).

Finally Chapter 7 is an exploratory analysis of proposed mechanisms by which these observed population-scale effects may be enacted through effects on individual people. I propose that social media content induces “memetic effects” in individuals, namely changes in their attention, sentiments, and beliefs. My analysis demonstrates that individual interactions with online content do induce such memetic effects, and that these are strongest in the initial minutes and hours after exposure but that effects persist into subsequent days. I hypothesize that these effects may accumulate and reinforce through constant, repeated exposures to similar content, leading to large, long term effects on behaviors, and that these resulting individual behavior changes constitute the population-scale effects I observe in

the case studies.

Together, this body of work provides a starting point for future work that assesses the offline effects of online social media.

CHAPTER 2

LITERATURE AND CONCEPTS OF CAUSATION FOR SOCIAL MEDIA STUDIES

I aim to demonstrate a method for linking online causes to offline effects. This is a special case of a broader class of problems called observational causal inference, in which we can only observe the system we're studying and cannot intervene on it as one would do in a classic experiment. In this chapter, I'll look at the foundational principles of observational causal inference and how they might be applied to perform causal inference in social media studies.

I'll start with a brief overview of philosophical theories of causation. Next, I'll review causal inference methodologies. Finally, I'll look at major prior works in social media studies which have used causal inference methods. Along the way, I'll define some key concepts in causal inference, and discuss how causation might be operationalized in social media studies.

2.1 Overview of Philosophy of Causation

The philosophy of causation has a long history from antiquity to the present, and contains difficult problems without generally accepted solutions. In this section, we'll restrict ourselves to a concise overview of contemporary positions about what may constitute causation: what does it mean for A to "cause" B (see Table 2.1)? Although abstract, having clear definitions and terminology for causation is critical to beginning a study of causal methods and applying them appropriately to new domains.

Regularity Theories. The regularity theories of causation originate with skeptical philosopher David Hume [15]. Hume posited that it was impossible to truly know about

School	Gist	Key Authors
Regularity Theories	A causes B if A always precedes B	David Hume
Counterfactual Theories	A causes B if B would not happen if A did not happen	David Lewis, J.L. Mackie, John Stuart Mill
Causal Mechanisms	A causes B if an identifiable mechanism links A to B	Peter Machamer, Lindley Darden, Carl Craver
Probabilistic Theories	A causes B if A increases the probability of B	Patrick Suppes, Ellery Eells
Manipulationist / Interventionalist Theories	A causes B if manipulating A changes B	James Woodward, Judea Pearl
Process Theories / Transmission Theories	A causes B if A transmits a certain conserved physical quantity to B	Wesley Salmon, Phil Dowe
Causal Powers and Dispositions	A causes B if A has the intrinsic power to bring about B	Nancy Cartwright, Stephen Mumford, Rani Lill Anjum

Table 2.1: Main contemporary schools of thought in philosophy of causation

causal relationships or even prove causation is a real principal in the universe. In practice, the best we can do is observe regularities of occurrences, or “constant conjunctions” as Hume puts it. For the regularists, what we really mean by “A causes B” is merely that we have always observed B to be preceded by A. By logical induction, we assume that A is related to B, but we can’t ever truly prove it. Regularist theories have the virtue of simplicity, but their skepticism is unnecessarily restrictive on our ability to learn causal facts about the world, which is why Humes initial position is no longer popular with contemporary philosophers of causation. Nevertheless, this skeptical attitude toward causation remains influential in some disciplines today, notably including many social sciences rooted in observational inferences, which often eschew causal claims.

Counterfactual Theories. The counterfactual theories of causation, as exemplified by David Lewis’s 1973 paper “Causation” [16], ask the question: if A had not happened,

would B still have happened? In this counterfactual perspective, we assert that A is a cause of B if and only if B only happens in hypothetical scenarios where A happened. A refined version of this idea was also developed by J.L. Mackie, who defines a cause as an “Insufficient, but Necessary part of an Unnecessary but Sufficient condition” (called the INUS condition) [17]. Pragmatically, this definition of causality has the virtue of being similar to what people tend to want to use causal reasoning for: to understand hypothetical counterfactual circumstances. However, the theory often is somewhat difficult to operationalize in the sciences, owing to the difficulty of fully modeling counterfactuals in complex systems or scenarios.

Causal Mechanisms Theories. Contemporary mechanistic theories of causation are exemplified by the 2000 paper “Thinking about Mechanisms” by Lindley Darden, and Carl Craver [18]. In the mechanistic theories of causation, the key criteria for causation is that we are able to identify a chain of simpler mechanisms that connect A and B. This theory is rooted in the principle of scientific reductionism, wherein we try to explain phenomena by deriving them from simpler known phenomena. This idea of causation is often considered the gold standard for causal thinking in physical sciences, where systems can be fully modeled in theory and controlled in experiments. However, it necessarily precludes making causal claims about processes that are too complex to model with complete fidelity, such as many processes in biology or social science. Scientists outside the physical science often accept mechanistic causal theories in principle, and yet still use orthogonal notions of causation to support their study of more complex systems.

Probabilistic Theories. Probabilistic theories of causation emerged in the mid 20th century in response to the difficulties of applying the existing theories of causation to scientific work outside the physical sciences. The probabilistic notion of causation says that A causes B if B is more probable conditional on A, considering all other relevant factors. Clive Granger and Patrick Suppes respectively introduced this idea to statistics in 1969 and

philosophy in 1970 [19, 20]. Recently, Ellery Eells has updated the probabilistic notion of causation to better account for the intricacies of causation in complex systems [21]. The probabilistic idea of causation became popular in some scientific disciplines because of its easy operationalization in a wide range of circumstances (e.g., via Granger causality, discussed below). The down side of this easy operationalization, however, is that it is often sloppily applied by failing to properly account for other influences on outcomes. This problem has led to accusations that the probabilistic causal theory underspecifies what it means for A to cause B, leading to ambiguities and false attributions.

Manipulationist / Interventionist Theories. In manipulation-based theories of causation, A is said to cause B if and only if manipulating A changes B. This theory is the straightforward interpretation of the causation present in Randomized Controlled Trials (RCTs), which are often considered the gold standard for causal inference outside the physical sciences. This theory was explained in detail by James Woodward in his 2003 book “Making Things Happen: A Theory of Causal Explanation” [22] and used as a foundation for Judea Pearl’s work on causal graphical modeling [23]. Interventionist theories of causation have a number of philosophical advantages: (i) They match what scientists often do in practice (such as in RCTs), (ii) they capture a pragmatic goal of causal understanding, namely enabling manipulation or control of a system, and (iii) they are generally compatible with other theories of causation, such as mechanistic or probabilistic theories.

Process Theories. In “process theories” of transmission, the key idea is that A causes B if and only if A transmits a conserved quantity (such as information or energy) to B. This idea was developed by Wesley Salmon [24] and Phil Dowe [25] to: (i) put causation on a physically realist footing, where it is an ontologically real process, rather than a mere epistemic convention, (ii) address a number of shortcomings in prior theories of causation (regularist, probabilistic, counterfactualist), and (iii) accommodate the processes of new physics (e.g., quantum physics) within the logic of causal philosophy. Although abstract,

process theories have a number of virtues, chiefly that they are largely compatible with other theories (e.g., mechanistic, probabilistic, and manipulationist) while offering greater specificity and explanatory power in some circumstances. In practice, however, process theories of causation are most often not directly operationalizable, and so process theorists must lean on other causal theories that are compatible with them to operationalize causation in science.

Causal Powers / Dispositionalist Theories. The causal powers theory of causation posits that objects possess intrinsic properties or powers that inherently determine their causal effects. Unlike theories that see causation as mere regularity or counterfactual dependence, the causal powers perspective grounds causation in the nature of the entities involved. A causal dispositionalist might say that A causes B if A has the intrinsic property of being able to bring about B. These theories are promoted by Nancy Cartwright [26] and Stephen Mumford & Rani Lill Anjum [27]. For these theorists, this approach offers some advantages: (i) it is realist, in the sense that it asserts that causation is an ontologically real process rather than a mere epistemic observation, (ii) it allows for context sensitivity, recognizing that the same object can have varying effects based on different conditions; (iii) it offers a more holistic view of causation by integrating the inherent tendencies or “dispositions” of objects into the causal narrative. However, causal powers theories offer little to the working scientist, as they are not clearly operationalizable.

2.2 Operationalizing Causal Philosophy for Social Media Studies

Rather than attempt to adjudicate the merits of these conflicting philosophies of causation,¹ the working scientist may instead take a pragmatic approach wherein we define causation based on the reason for making the causal claim. For instance, do we want to claim that A causes B because we would like to know how we can change B by changing A? Then

¹Although I do have strong opinions and a well-developed causal philosophy of my own.

an interventionist definition of causation may be appropriate. So when operationalizing the notion of causation in a new context, it's critical to ask what the purpose of our causal claims is, and how our audience may try to use these claims. So in this section, we'll briefly look at the common motivations for and interpretations of causal claims in the context of social media studies.

One of our primary goals for making causal claims is to inform policy *interventions* on social media to change the outcomes. Sometimes, we may also want to imagine *counterfactual* scenarios in which an intervention had happened in order to attribute particular known outcomes to specific preceding events (e.g., did some particular event determine an election outcome). We are often also interested in predicting what is more *probable* to happen as a result of observed events happening now. Finally, we may also want to establish that online and offline events are *mechanistically* linked for the purpose of better supporting causal claims to jurists, policy makers, and other scientists, who often employ mechanistic criteria in their understanding of causation and may ultimately be responsible for enacting our proposed interventions.

These motives for proving causation pragmatically suggest that the Interventionist, Counterfactual, Probabilistic, and Mechanistic theories of causation may be most relevant, so we'd like to employ methodologies that give us tools and opportunities to satisfy these philosophical criteria in a variety of circumstances. Fortunately, we are not the first to have these criteria for causal inference in the social sciences, and causal inference frameworks have been developed attempting to satisfy these criteria. In the following sections, we'll explore these causal inference methods and how people have applied them in the context of social media studies.

2.3 Key Causal Concepts

Before we can discuss causal inference methods in more detail, we need to define some key terms and ideas in causation. We'll start by introducing a simple diagrammatic language

for causation: causal graphs.

Causal Graphs. In causal graphs we represent the relationships between causes and effects as a directed graph, where nodes in the graph represent variables or factors or events, and directed edges represent a causal relationship pointing from cause to effect. For instance, in the relationship $X \rightarrow Y$ denotes that “X causes Y” (see Table 2.2). In causal graphs, we may also commonly refer to relationships between three variables, such as “mediators” ($X \rightarrow M \rightarrow Y$), “forks” ($X \leftarrow F \rightarrow Y$), and “colliders” ($X \rightarrow C \leftarrow Y$). In causal graphs, causation is transitive, such that if the causal relationship between X and Y is mediated by M , we may still say that X causes Y .


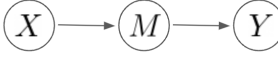
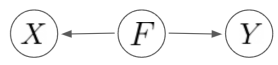
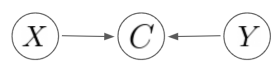
Concept	Definition	Diagram
Causal Influence	X causes Y	
Mediators	M mediates the causal influence of X on Y	
Forks	F causes both X and Y	
Colliders	X and Y both causally influence C	

Table 2.2: Key Concepts in Causal Graphs

Confounders. A common problem in causal analyses is the issue of confounding. For instance, we may have a relationship $X \rightarrow Y$ which we would like to study, but the system also contains the fork, $X \leftarrow W \rightarrow Y$ (see Figure 2.1a). In this circumstance, the fork relationship introduces additional influences on X and Y that make it harder to observe the

effect of X on Y clearly by introducing a systematic bias. In this circumstance, we might call W a “confounder” on the $X \rightarrow Y$ relationship and say that it creates a problem of “confounding” between X and Y , via W (Figure 2.1b). Sometimes, we may also have a more ideal case in which other variables like W are present in the system, but do not affect the relationship between X and Y (Figure 2.1d); when this happens, we call the $X \rightarrow Y$ relationship an “unconfounded” causal relationship.

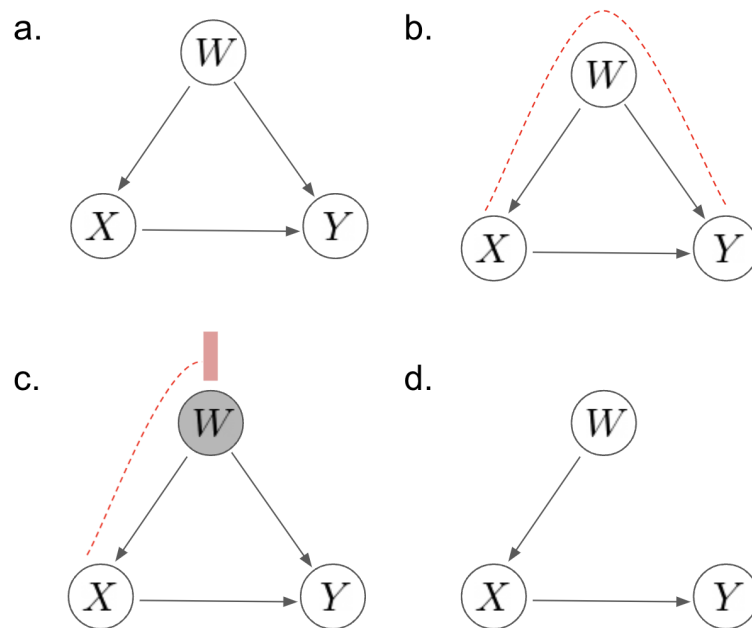


Figure 2.1: Causal graphs for the XYW example, where we would like to study the $X \rightarrow Y$ causal relationship. (a) A causal graph where W is a confounder on the $X \rightarrow Y$ relationship. (b) A diagram showing the confounding relationship in red. (c) A diagram showing that controlling for W blocks the confounding relationship, allowing us to study the $X \rightarrow Y$ relationship. (d) An unconfounded causal graph.

Controlling for Confounders. If we want to study the influence that X has on Y when the $X \rightarrow Y$ relationship is confounded, then we will have to “control for” this confounder by accounting for the influence of W on Y *before* we estimate the influence of X on Y (Figure 2.1c). “Controlling for” confounders looks different in different kinds of analyses. In regression-based estimations of effect sizes, controlling for a variable generally means including it as an additional predictor in the regression. In estimations of outcome prob-

abilities, controlling for a variable generally means conditioning on the variable in your estimations of the outcomes' probabilities. One can also "stratify" over a confounding variable by measuring the average treatment effect only within groups with similar values of the confounding variable, thereby eliminating the variation in the confounder and its impact on the outcome.

Controlling, Colliders, and Mediators. An additional problem that we can run into with causal studies is mistakenly controlling for the wrong variables. Consider the case where we are attempting to control for confounding forks F (of the form $X \leftarrow F \rightarrow Y$) and we mistakenly control for a collider C (of the form $X \rightarrow C \leftarrow Y$). This mistake biases our measurement of the $X \rightarrow Y$ relationship. Consider that the collider C is causally downstream of both X and Y and so initially does not affect the relationship $X \rightarrow Y$. However, if we mistakenly believe C to be a confounder and control for C in our analysis of the $X \rightarrow Y$ relationship, this introduces a spurious correlation in our measurements of the $X \rightarrow Y$ relationship. Similarly, if we mistakenly control for a mediator M (of the form $X \rightarrow M \rightarrow Y$), we can inadvertently destroy this causal path in our analysis, thus biasing our estimation of the effect size toward a null effect. We must therefore carefully distinguish between mediators, forks, and colliders, and avoid controlling for mediators or colliders in our causal analyses.

Instrumental Variables. Another a useful concept in observational causal inference studies is instrumental variables. An instrumental variable is a variable that: (i) is correlated with the treatment variable but, (ii) is independent of the unobserved confounders affecting the outcome, and (iii) does not affect the outcome except through its influence on the treatment or exposure. Identifying an instrumental variable can enable robust observational causal inferences by allowing us to properly control for unobserved confounding variables, via controlling for the instrumental variable that mediates their influence on the treatment variable. However, instrumental variables are relatively uncommon, and so cannot always

be relied upon as a general solution to observational causal inference.

Identification. In causal inference, “identification” refers to the ability to disentangle and isolate a particular causal effect from observed data.² In general, the process of identification involves coming up with a strategy that allows us to observe the specific impacts of treatments on outcomes while accounting for confounders in some way. Causal identification strategies always rest on both inference methods and assumptions, such as the assumption that our causal graph is correct and complete. These assumptions are often reasonable, which is what enables us to advance our understanding of causes in the special sciences, but sometimes they are difficult to justify, putting causal arguments on shaky ground. For this reason, it’s critical to bear in mind the assumptions embedded in causal inference methodologies, and to validate them as far as possible.

Estimation. In causal inference, “estimation” refers to the quantification of a identified effect’s magnitude. Estimation requires having first established an identification strategy. Causal estimations may be performed in a variety of ways, as we’ll see below, but notably include panel regressions, DiD methods, Hawkes process methods, and causal estimands derived from do-calculus.

2.4 Overview of Popular Causal Inference Frameworks

Now that we have introduced a conceptual vocabulary for causal scenarios, we’ll look at a selection of key ideas in causal inference that provide us methodological tools for causal identification and estimation (see Table 2.3). Until recently, causal inference has been a niche field with much contention on best practices outside a few special cases (like Randomized Controlled Trials), but contemporary methods provide us with robust general frameworks for thinking about causal inference in a variety of more complex scenarios. In

²Note that this is separate from “identifying confounders” which is a more colloquial use of identification, meaning simply to consider possible confounders and determine which ones are likely to be actual confounders.

this section, we'll highlight a few of the more popular and successful conceptual ideas for causal inference.

Note that regardless of which methods are selected for demonstrating the causal relationship and quantifying the effect size, the most critical step for causal inference is always identifying the confounders and choosing a method with enables the study to control for them appropriately. For that reason, it's good to begin any causal analysis by first constructing a causal graph of the phenomena being studied. This will be discussed in more detail in Chapter 3.

Laboratory Experiments. The earliest formal causal studies were laboratory experiments, in which all aspects of a system are tightly controlled and one aspect is varied in isolation to determine its effects on the system. Because of the tight control over other possible influences on the system, the changes observed in the system can be attributed specifically to the deliberately varied aspect of the system. This original method of causal inference is philosophically built on manipulationist and mechanistic understandings of causation, where we observe the outcome of interventions on particular physical mechanisms in the system. However virtually all contemporary philosophies of causation are compatible with laboratory experiments, because it is taken as an archetypical test case which philosophies of causation must account for. In the physical sciences, laboratory experiments and the epistemic standards they exemplify are the norm for causal studies and causal claims. However, in the social sciences true laboratory experiments are not possible for the simple reason that social systems and processes involve human subjects, whose diverse character and independent agency makes fully controlling the studied system both impossible and unethical.

Randomized Controlled Trials. Outside the physical sciences, Randomized Controlled Trials (RCTs) are the gold standard for causal studies. RCTs attempt to replicate the certainties of laboratory experiments in contexts where human subjects are involved. In RCTs,

a group of human subjects are randomly assigned to a treatment or control condition, and the difference in outcomes between the two groups is assumed to be the causal impact of the treatment. Unlike a true laboratory experiment, RCTs do not control all possible confounding influences on the outcomes. Instead, RCTs use each individual's unique identity as an instrumental variable to control for all of the other unique features of the individuals in the study. The central epistemological assumption is that by averaging over the randomly-assigned diverse set of individuals, we can measure an average treatment effect (ATE) that controls for all of the unique features of the individuals in the study. Like laboratory experiments, most philosophies of causation find RCTs a satisfactory means of causal inference, but the method itself is ultimately rooted in a manipulationist philosophy of causation.

A/B Testing. The term “A/B testing” is used to refer to RCT-type causal studies in which the treatment intervention is generally a change to a website, advertisement, product, or software, and the outcome is some easily observable behavior of the subjects. This kind of experiment is most often conducted for commercial rather than scientific purposes (e.g., to refine product design or marketing materials). In the case of social media, generally only the social media platforms themselves have the ability to perform systematic random interventions of this kind, and so studies using A/B testing is generally not an option for third-party researchers. Additionally, while A/B testing is often conducted with no informed consent in commercial contexts, in formal research context it generally requires IRB approval as human subjects research.

Granger Causality. The earliest methods for causal studies (laboratory experiments and RCTs) require some ability to directly manipulate or intervene on the system we're studying (such as random assignment of treatments), but often in the social sciences we cannot significantly intervene on the systems we're interested in studying, either for practical or ethical reasons. Granger causality was one of the early attempts to develop causal inference methods for systems which we can only observe, not manipulate. In Granger causality, we

use a probabilistic definition of causality, where: (i) Causes precede effects in time, (ii) Causes contain *unique* information helpful for predicting effects. The assumption is that if you can appropriately control for confounders, you will identify only the treatments which have unique predictive power, and thus presumed causal effect. The central problem with this reasoning is that it fails to systematically account for causal influences on the observed treatment variable itself, and so it is unable to fully disentangle the causal relationship between treatment and outcome. Furthermore, Granger Causality also fails to explicitly satisfy most other philosophical understandings of causation, such as the mechanistic and interventionist theories used in RCTs and laboratory experiments. For these reasons, Granger Causality is often not considered “real causality” today by many working scientists. Nevertheless, it is still a much stronger evidentiary criterion than mere correlation, because it establishes both the temporal precedence and predictive power of the “cause” over the “effect,” which in many cases is useful in its own right.

Propensity Score Matching. Partially to address these deficiencies of Grange causality, methods were developed to control for causal influences on observed natural treatments. In “propensity score matching” we estimate the probability that subjects would naturally receive a given treatment, based on other observed variables. The estimated likelihood of treatment is called a “propensity score.” We can then examine groups of subjects with comparable propensity scores but different treatment statuses, and examine the difference in their outcomes. This can allow us to estimate the causal impacts of the treatment within groups of similar propensity score. The propensity score method is ultimately attempting to approximate the form of RCTs in contexts where treatment cannot be properly randomized. As such, it is ultimately based on a interventionist philosophy of causation. The key assumption of this method is that the observed confounders contain enough information to provide an unbiased estimate of treatment probability, and that remaining observed variations in treatment status can be reasonably modeled as simple random noise. This assump-

tion is quite strong, and its verification depends on having a strong existing understanding of the causal structure of the system we're studying. In practice, however, propensity score methods are quite flexible, and are readily applied in a variety of observational sciences, including epidemiology, clinical studies, and economics. Furthermore, propensity score methods can be applied in conjunction with other causal inference methods, such as DiD (discussed below).

Difference in Differences. The difference in differences (DiD) framework for causal inference involves observing multiple similar subjects (e.g., people, populations, firms) as they experience different treatment levels over time, and using the observed differences in treatment levels and outcomes to estimate the treatment effect. In traditional DiD, there are only two groups of subjects (treatment and control), and the treatment is a discrete event which divides time into “before” and “after treatment periods, which are compared between the two groups to estimate the effect size. However, the causal logic of DiD can be generalized to use multiple time periods, multiple treatment levels, or even continuous treatments (such as when panel regressions are used in causal inference, see below). The critical limiting assumption of this DiD framework for causal inference is that, on average, the similar subjects we observe would've followed similar outcome trend lines if not for the treatment. This assumption is more easily justifiable when confounding factors are appropriately identified and controlled. Another perennial problem of DiD methods is that it can sometimes be difficult to control for different subjects' varying likelihood to be naturally treated; this can be remedied somewhat using propensity score methods. In spirit, the DiD method is another attempt to replicate the RCT paradigm of causal inference in a purely observational setting, and thus it is built implicitly on an interventionist causal philosophy. In the following two paragraphs, we'll briefly discuss two generalizations of the DiD causal logic to continuous treatment and discrete outcome events, respectively.

Causal Panel Regressions and Doubly-Robust Estimation. Another broad class of analytical tools used in causal inference is “panel regressions” in which time series data of continuous treatment, outcome, and confounder variables are analyzed using regression strategies to determine the relationships between the treatment and outcome variables. This inference method might be understood as a generalization of the logic of DiD for continuous treatment (rather than discrete events), and it is subject to the same assumptions: namely that similar subjects would have followed the same outcome trajectories if not for differences in treatments. Like DiD, this assumption can be supported by appropriate identification strategies, control of confounders, and/or use of propensity score methods. When panel regression methods are paired appropriately with propensity score methods and well-controlled confounders, they are sometimes called “doubly-robust” causal estimators (because they control other influences on both the treatment and outcome, hence “doubly”).

Causal Hawkes Processes. Another generalization on the logic of DiD is causal Hawkes processes, which analyses a set of discrete treatment and outcome events in time (see Appendix C). In a Hawkes process, events of one class (treatments) may induce additional events of another class (outcomes) for a period of time. By fitting a Hawkes process model to data about the events that subjects have experienced over time, we can estimate the degree to which different types of events affect each other and over what time periods. With appropriate identification strategies to handle confounding variables in the model, these estimates can even be causal. As with causal panel regressions, these are still subject to similar assumptions as in DiD, namely that but for the treatment events similar subjects should have similar rates of outcome events over time. As with other DiD-like methods, this assumption can be supported by controlling for confounders and treatment propensities appropriately. In this case, we might call it a “doubly-robust” Hawkes model for causal estimates, although this term is not often used as Hawkes models are underused in causal

analyses to date.

Potential Outcomes Framework. The potential outcomes (PO) framework is an abstract conceptual framework for approaching causal inference, and it is built on a counterfactual theory of causation. The framework evaluates the effects of an intervention by imagining two different hypothetical worlds: one where the intervention (or treatment) took place and another where it did not. For each subject in the study, there's a "potential outcome" under treatment and another under control, and the difference between them represents the effect of the treatment. However in any given study only one of these potential outcomes can be observed for each subject, rendering the other counterfactual outcome "missing data." Thus in the potential outcomes framework, causal inference is treated as a problem of imputing the missing data points (the counterfactual scenarios), and then comparing the outcomes between the treatment and control cases. Various methods of missing data imputation can be applied to this general conceptual framework. Indeed, some of the more concrete methods of causal inference methods like DiD can be viewed as specific operationalizations of PO. Partially for this reason, the PO framework is one of the more broadly applied frameworks for causal inference in the social sciences.

Causal Graphical Modeling. In Causal Graphical Modeling (CGM, see Appendix A), we begin by constructing a causal graph. CGM is a mathematical formalism which allows us to determine all possible identification strategies for a particular causal relationship in a given causal graph, and provides us with estimand expressions for each (see Figure 2.2), which may be estimated using a variety of strategies (e.g., DiD, panel regressions, or direct computation from model equations, as in Section 3.7). CGM offers two main advantages. First, it provides a clear visual language, which allows easier collaboration and validation of domain knowledge assumptions about causal relationships. Second, its formalism is the only systematic and universal approach to solving causal identification problems. Philosophically, CGM is premised on an interventionist concept of causation, and its mathe-

mathematical formalism provides mechanisms for analyzing interventions. In practice, the CGM framework has proven invaluable in a variety of fields, including epidemiology, economics, and social sciences, where it has been used to untangle complex causal relationships and guide empirical research.

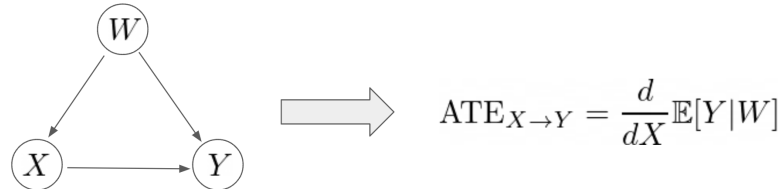


Figure 2.2: The Causal Graphical Modeling (CGM) formalism yields an estimand of the average treatment effect (ATE) of the $X \rightarrow Y$ relationship in the XYW example causal graph.

2.5 Prior Use of Causal Methods in Social Media Studies

In this section, I'll discuss how causal inference methods have previously been applied to study the offline effects of online social media content. Specifically, I'll focus on studies that use both online social media data and offline outcomes data. Perhaps surprisingly, this is a very small portion of all studies of social media's impacts on people and society, probably owing to the difficulty of obtaining good datasets and the lack of general methodology for linking specific online activity to offline outcomes that I am attempting to develop in this work.

There are three large categories of related social media research that I will not consider here: (1) surveillance of offline outcomes via online social media, a vast category of research especially within public health, (2) studies of how offline events affect online behaviors, and (3) studies which use surveys of individual social media users to assess the associations between self-reported online social media use and offline effects. These literatures have distinct methodologies and purposes that do not align with our goals in the present work.

Surveillance studies do not attempt to establish causal relationships, and studies of how offline events impact online behaviors are the inverse of our goal in this work; neither are relevant here. Survey-based studies, however, are much closer in motivation to our own, so their exclusion here is worth explaining in detail. Self-report methodologies are extremely limited in their ability to study the impacts of exposure to specific kinds of online content, because respondents are generally not able to recall with specificity the particular kinds and amount of content they have engaged with on social media. For this reason, survey questions are generally about overall usage rates (e.g., how often do you open instagram, how long do you scroll, etc.). This limits surveys' ability to study the offline effects of specific kinds of online content. The methods developed throughout this work aim for more specificity and statistical power than survey methods can offer. Additionally, survey studies employ entirely different methods than those discussed here, and so they offer little as methodological precedents to the present work. For these reasons, we'll exclude them in this review.

RCT Studies. Only a small handful of RCT experiments have been conducted with on-line treatments and offline outcomes. We'll look at three here. In the first [36], researchers at Facebook conducted an RCT to test if they could increase voter turnout by adding a banner to the top of the website's home page, encouraging users to vote. They found that this intervention did increase a user's likelihood to vote (as measured by self-reported voting behavior), and that the effect was stronger when the banner included mentions of specific Facebook friends who reported that they had voted. In a different (famously controversial) study, Facebook users were randomly shown either more positive or more negative posts to test if these users would experience "emotional contagion" and start expressing more positive or negative affects themselves [37]. The study found that emotional contagion via social media does happen, although the effect size is modest. In another study, a small number of participating US Facebook users' accounts were disabled for four weeks during

an election season in the US, and various outcome metrics were tracked, including subjective wellbeing and knowledge of current events [38]. This study found that the intervention improved subjective wellbeing, and reduced knowledge of current events. In two of these studies, the randomized intervention was only possible due to the researcher's control over the social media platform itself, namely Facebook, and in the third volunteers were recruited and asked to perform the relatively simple intervention (deactivating their account) themselves. While methodologically robust, in general RCT-style interventions on social media are infeasible for most researchers due to an inability to perform the necessary interventions on the platform. Furthermore, ethics concerns may often prevent responsible researchers from conducting such studies.

DiD Studies. Traditional DiD methods are better represented in social media studies owing to relatively wide applicability in the domain. In traditional DiD studies, the treatment event which affects some subjects of the study is a discrete event at a particular moment in time, and the outcome is some continuous variable that is tracked over time, before and after the treatment, in both the treated and control groups. This approach has been applied most often to look at the adoption or rollout of social media platforms by individuals and populations. For instance, when politicians adopt Twitter, studies find that this increases campaign contributions (especially if the politician is new to elected office) and increases politicians' alignment with constituency preferences in legislative votes [39, 8]. Other DiD studies have shown how the adoption of Facebook by whole populations have significant effects, such as the decline in self-reported mental wellbeing among college students and the increase in protest activity in countries following Facebook rollouts [4, 9]. Finally, some have also used DiD methods to study how gaining additional social connections in online social networks may change offline behaviors, such as demonstrating that a new social connection in a social fitness app increased physical activity, measured by steps taken each day. In general, DiD has been demonstrated as a reasonably effective and widely

applicable method for conducting causal studies with social media data. The key limitation that should be noted however is that it is confined to studies where the treatment is a discrete event that is either exogenous (as with Facebook rollouts to new colleges and countries) or that can be well predicted with propensity score methods.

Studies With Panel Data. Another larger class of causal studies of social media's offline outcomes are studies that use panel data. Panel regressions (and the closely-related doubly-robust estimations) have been used in studies of political engagement, violence rates, and economic outcomes. In one study, researchers showed that Twitter adoption rates in US counties caused a small reduction in Republican voter share in the 2016 and 2020 presidential elections [7]; they hypothesize that Twitter's more left-leaning user base may have influenced moderates to not vote Republican. In another study, Twitter sentiment toward specific companies was found to be a strong predictor of stock market prices for those companies [6]; however, the authors note that the Twitter sentiment may merely have been an early signal of influential news events that would impact the companies, which they suspect is the real underlying cause. Finally, two related studies by the same author have examined the effects of hate speech on the incidence of offline hate crimes. In the first, anti-refugee sentiment among Facebook users in specific German municipalities was found to cause increases in anti-refugee crimes in those municipalities [12]. In a later study, estimated exposure to anti-Muslim sentiment on Twitter was found to increase anti-Muslim hate crimes in US counties [13]. In general, these methods are reasonably strong for estimating effect magnitudes, but they vary in their reliability for drawing robust causal conclusions depending on the identification strategy used. When instrumental variables are used (as in [13] and [7]) or when careful consideration to confounding factors and treatment propensities is taken (as in [12]), then the causal conclusions are likely to be reasonably robust. These panel data based methods are in general likely to be the most flexible and widely applicable methods for estimating causal effects for social media studies.

Cross-sectional and Longitudinal Studies. Finally, a handful of studies have used other methods beside those discussed above. In one study, researchers used a cross-sectional OLS regression with an instrumental variable to show that usage rates of a Russian social media platform, VK, influenced protest activity in Russian cities [10]. In another, researchers used a longitudinal Granger causality analysis to try to show that protest-related hashtag usage increased subsequent protest activity during the Arab Spring in Egypt [11]. These two studies are good examples of how researchers sometimes attempt to work with less-than-optimal data in causal studies, and they show the causal reasoning pitfalls associated with these alternative strategies. While typically we might have preferred panel data for these kinds of causal analyses, which contain data for multiple subjects (cross-sectional) and data over time (longitudinal), sometimes we only have one or the other. In the study of protests in Russian cities, the authors had cross sectional but not longitudinal data, so they performed identification with an instrumental variable and estimated the magnitude by cross-sectional inference only; this has the weakness of failing to establish temporal precedent between cause and effect. In the other study of protests during the Arab Spring, they only had longitudinal data for a single group, namely Egyptian Twitter users and Egyptian protests; the lack of multiple groups to compare across destroys our ability to control for confounding variables that might've influenced the outcomes. In both the longitudinal and cross-sectional cases, we are unable to reach fully satisfactory causal conclusions. We generally need both, as in panel data, in order to establish temporal precedent of causes before effects and to control for confounders appropriately.

2.6 Conclusion

In this chapter, I introduced key ideas in philosophy of causation, several prominent quantitative causal inference frameworks, and finally reviewed their application in the field of social media studies. Along the way, I defined some important concepts in causal thinking, like confounders and instrumental variables, and briefly considered what definitions of

causation may be most appropriate in social media studies (a question I'll return to shortly). This conceptual background lays the groundwork for the methodological developments of Chapter 3.

Causal Inference Idea	Gist	Key Authors
Laboratory Experiments	Effects are studied by observing what happens when we change one aspect of a system while holding all others constant.	Francis Bacon [28], others
Randomized Controlled Trials (RCT) and A/B Testing	Effects on subjects are studied by randomly assigning test subjects to either receive a treatment or not, and the difference in outcomes between the two groups is assumed to be the effect.	Ronald A. Fisher [29], Ronny Kohavi [30]
Granger Causality	By assuming that “causes precede and uniquely predict effects,” we identify causes of an effect by finding reliable predictors of the effect.	Clive W.J. Granger [19]
Propensity Score Matching	When we cannot randomly assign treatment, causal relationships are studied by attempting to account for the likelihood of a subject being treated by using observed confounders.	Donald Rubin, Paul Rosenbaum [31]
Difference in Differences (DiD), and generalizations	By observing multiple similar subjects, some which received the treatment naturally and some which did not, we can conduct a natural quasi-experiment.	David Card [32, 33]
Potential Outcomes (PO)	Effects are quantified by comparing hypothetical treatment and control scenarios, where counterfactual scenario outcomes are estimated as a missing data imputation problem	Donald Rubin [34], Paul Holland [35]
Causal Graphical Models (CGM)	By constructing a “causal graph” which encodes the known causal relationships between variables, CGM’s mathematical formalism allows us estimate average treatment effects by controlling for known confounders.	Judea Pearl [23]

Table 2.3: Important contemporary ideas in causal inference

CHAPTER 3

A METHOD FOR LINKING ONLINE CAUSES TO OFFLINE EFFECTS

In the previous chapter I reviewed literature on causal philosophy and causal inference, introduced some key concepts and vocabulary in causal analysis, and looked at some examples of how causal methods have been applied to social media studies in the past. In this chapter, I'll leverage these concepts and methods, along with practical experience from our case studies (especially Chapter 5), to articulate a general methodology for causally linking online social exposures to offline population outcomes.

3.1 Online and Offline Data

I'll start by first defining the context in which these analyses might be performed by considering the kinds of data typically available. This will help describe the methodology more clearly in the following sections.

Online Social Media Data. Typically, our data about online social media activity will be records of specific interactions between users and social media posts. Namely, we can observe that a particular user saw or responded to a particular piece of content at a particular time. I also generally observe characteristics of the content, for instance any text in the social media post, or photos or links that it contains. In some cases, we may also get information about the user, such as their username, their self-description in their profile, and perhaps demographic or geographic metadata about the users.

Offline Outcomes Data. The offline effects of online social media that we're interested in measuring might be quite diverse, but generally they are population-level outcomes, such as public health statistics, crime statistics, market data, or surveys and polls. The data avail-

able are generally collected on a regular basis temporally (such as yearly or weekly), and are tabulated by specific geographic or demographic groups (such as geographic regions or population age groups). Sometimes, they may be reported as aggregate statistics like the number of COVID cases in a county in the past week, and sometimes they are reported as individual records of specific events like a particular criminal arrest or a hospital admission. This is the offline data that we'll use to estimate outcomes of online social exposures.

3.2 Linking Online Users to Offline Populations

The first requirement for being able to show a causal relationship between the online and offline phenomena is to be able to establish a clear connection between specific online populations and specific offline populations. That is, we need to be able to identify some demographic or geographic information about online users in order to show that they belong to the specific offline populations about which we have offline outcomes data. This identification of online users with offline population groups allows us to link the online data to offline data, and this the first key step in any analysis of offline effects of online social media. It also defines clear population groups that can be used for causal analyses across multiple populations, such as in causal panel regressions.

Geolocating Users. Often a convenient way to divide online users up into offline population groups is by geography. By geolocating users to specific geographic areas (for instance cities or counties), we can map our observed online populations onto offline populations about which we may have data. However, because of the inherent sensitivity of geolocation data, it is often difficult to obtain this data on many platforms. Nevertheless, it is sometimes possible to obtain geolocation data, for instance by examining self-reported current cities on Twitter or Facebook, which is a standard part of user profiles that many users opt to fill out. User locations may also sometimes be identified through explicit geolocation data collected by the social media platform; however this data collection is often opt-in and

therefore very sparse. In the case of Twitter, a tool called Carmen [40, 41] can analyze user profiles to determine their geolocations based on both self-reported profile cities and explicit geolocation data; the Carmen tool was used throughout this dissertation.

Identifying User Demographics. Similarly, we may also identify users with offline population groups about which we have data through their demographic features, such as age, gender, and ethnicity. As with geolocation, we may often identify these features based on user self-reports within their user profiles. Sometimes, these features may be inferred based on other features of the users’ profile and behaviors, as in these studies that estimated age [42], political affiliation [43], ethnicity [44], and gender [45]. In general, when categorizing users into demographic bins to link online data to offline data and perform inferences, we need to be able to split our users into many sufficiently distinct groups in order to make robust inferences. For instance, dividing only into two groups by gender is likely inadequate for most analyses. Therefore, it’s often necessary to obtain a significant degree of specificity across multiple demographic features, or augment demographic groupings with geographic groupings.

3.3 Measuring Social Exposure to Online Content

The next part of our methodology for linking online social media content to offline outcomes is to estimate each population’s social exposure to online content. There’s two key steps to this: (i) identify the content of interest (the “treatment content”), and (ii) define and apply a measure of social exposure to the content.

Identifying the Online Treatment Content. We also need to be able to identify instances of the online content that we want to use as the “treatment” in our study. For instance, imagine we would like to study the public health impacts of online antivaccine content. We will first need to be able to identify this content. In general, observing and identifying the online treatment amounts to a classification problem where we would like to determine

if a given text or image is a member of the treatment class of content (e.g., antivaccine content). Robust methods are now available for these text and image classification problems [46, 47], and we can apply them readily to identify a given type of treatment content. These classification methods always involve clearly defining the treatment content of interest and manually identifying instances of the treatment content to validate the classification model’s performance (and often also to train the models). An example of this workflow is described in greater depth in Chapter 5. Having clear definitions of the treatment content is important both for the pragmatic reason of needing clear criteria when labeling, but also because of the methodological consideration that the kind of treatment content identified ultimately affects the conclusions we can draw from our analyses.

A Measure of Exposure in Social Networks. Next, I’ll define a measure of exposure to specific content in a social network. I’ll make two assumptions based on the previous sections: (1) we can identify instances of the treatment content, and (2) we can categorize online users into offline population groups. Intuitively, our measure of exposure is defined such that if a population group is strongly socially connected to other population groups that produce a lot of treatment content, then its exposure to the treatment content is high. Formally, the per-capita treatment exposure rate in population group i at time t is defined as

$$E_{i,t} = \frac{1}{N_i} \frac{\sum_j W_{ij} T_{j,t}}{\sum_j W_{ij}} \quad (3.1)$$

where N_i is the number of people in of population group i , $T_{j,t}$ is the number of treatment posts made by population group j during time window t , and W_{ij} represents the strength of online social influence of population group j on population group i . The term W_{ij} might be defined in different ways, such as by looking at the follower or friendship relationships between users in these groups, or by tracking their interactions such as retweets or likes. For instance, W_{ij} might represent the number of times that topically-relevant posts authored by

users in population group j were viewed (or interacted with) by users in population group i during the observation period. This social exposure metric is quite general can be applied in a wide variety of circumstances, and it is generally easily computable from online social media data about individual users and their interactions.

3.4 Causal Modeling of Social Exposures and Offline Outcomes

In the previous sections, I discussed how to: (i) collect data on online interactions and offline outcomes, (ii) connect them via linkage variables (such as geolocation), and (iii) estimate exposures to specified treatment content. This results in a joint dataset of offline outcomes and online exposures for each of our observed offline populations over a particular observation period. In general, the resulting joint dataset can consist either of time series data or of records of discrete events in time (such as particular online posts or particular offline incidents) which we may treat as a point process.

In this section I'll examine how we can apply modern causal inference methods to observational data of this kind.

3.4.1 Considerations for Choosing Causal Methods

Unfortunately, there is no one-size-fits-all solution to observational causal inference. Instead, we need to consider a few key questions to shape our decisions about which methods to use: (i) What data are available? (ii) What is our goal with proving causation in this specific study? Let's consider these questions in turn and consider how they constrain our choices of methods.

Data Constraints. Regarding data constraints, we have a clear answer provided by the prior sections: our data is either time series or point process data of online social exposures and offline outcomes. In general, time series data over several populations will lend itself nicely to a DiD or causal panel regression analysis, augmented with CGM to determine

identification strategies and/or propensity score methods to help control for differences in observed treatment likelihoods. For point process data, such as records of individual online posts and offline events (e.g., crime incidents), we may apply Hawkes processes (as describe in Appendix C and applied in Chapter 7) to infer average treatment effects, and again augmented by CGM and PSM to determine and implement identification strategies. These data constraints narrow the range of applicable methods considerably, but still allow enough flexibility to properly consider and account for confounding variables, as well as known dynamical structure in the data (e.g., temporal or spatial autocorrelations in outcomes).

Goals and Definitions of Causation. One of the most important questions to ask in causal inference is what exactly do we mean by causation in this particular context (see Table 2.1 and Section 2.2)? What is our reason for using the word “cause,” and what might such a claim allow us to do?

In social media studies, the ultimate goal of proving causation is most often to enable us anticipate the effect of some imagined intervention on social media platforms, such as changing content moderation policies. For instance, if we say that “antivaccine tweets caused additional COVID-19 deaths,” the implicit conclusion is that if we moderate anti-vaccine content more strictly, we may prevent future deaths from infectious diseases. In short, our causal claim is targeted at enabling interventions, and so we ought to apply an interventionist definition of causality. This constrains our choice of methods to those that can most robustly satisfy this philosophical criterion for causation. In particular, RCT, CGM, and methods that can approximate the RCT formulation (such as DiD and PO, both with controlled confounders) may be appropriate methods, but methods like Granger Causality may not (see Table 2.3).

Occasionally, we may be interested in counterfactual claims such as “did online social media content sway the electorate and determine the outcome of an election?” In this case,

the process involves: (1) determining and identification strategy for the causal relationship, (2) estimating the magnitude of the effect with estimator methods, (3) imagining and initial-izing a plausible, consistent counterfactual scenario in which some alternative intervention had happened, (4) simulating the outcomes in that counterfactual scenario. This overall strategy allows us to estimate counterfactual scenarios either with explicitly counterfactualist causal methods like PO, or with strong interventionist methods (like using CGM for identification and doubly-robust panel regression for estimation). However, in practice it is often quite difficult to perform such counterfactual estimates, largely due to the difficulties of plausibly constructing and simulating the counterfactual scenarios (steps 3 and 4 of the described procedure) in complex systems such as a large social system affected by social media. For this reason, counterfactual claims are strictly stronger and harder than interventionist causal claims, and are often infeasible when the counterfactual scenario is complex or significantly different from the observed scenario. However, counterfactual analyses featuring smaller and simpler differences from the observed scenario may sometimes be plausible and worth analysing.

Finally, we may often be interested in merely predictive claims, such as forecasting consumer purchasing behaviors induced by observed social media content. In these cases, weaker forms of analysis such as Granger Causality may be applicable, as implemented in simple autoregressive time series analyses. However, when our goals are merely predictive, we probably want to avoid the word “cause” to prevent readers from erroneously applying an interventionist or counterfactualist understanding of our claims.

A General Recommendation. While there is no definitive one-size-fits-all solution to observational causal inference, I will still recommend a strategy that may serve as a good starting point based on the above considerations: try using causal graphical modeling (CGM, Appendix A) to determine the identification strategies, and panel regressions or DiD methods to estimate effect sizes. CGM as a causal reasoning framework has a number

of features to recommend it. First, it is philosophically compatible with the main goals for causal inference in social media studies, as explained above. Second, it lends itself nicely to studying the complex systems often encountered in social media studies because causal graphs can easily encode many complex relationships between many variables. Third, the visual nature of causal graphs allows modeling assumptions to be made explicit and easily interpretable, and this enables us to more easily involve subject matter experts who can help us appropriately map out the causal graph for our particular study. Finally, the CGM framework can be used in conjunction with other inference methods, such as panel regressions or causal Hawkes processes, to estimate the specific effect sizes after CGM identifies the confounders that must be controlled. For these reasons, CGM should be the initial go-to option for causal modeling in social media studies, often augmented with other estimator methods like DiD or doubly-robust panel regression to enable greater flexibility.

3.4.2 Using Domain Knowledge to Control for Confounders

Regardless of the causal inference methods used, the most critical part of a causal analysis is always the incorporation of existing domain knowledge, which allows us identify and account for confounders and known dynamical structure in the outcome phenomena. In this section, I'll look at two key aspects of this portion of the analysis: (1) constructing a causal graph in consultation with domain experts, and (2) applying existing quantitative models of the offline phenomena within our causal inferences.

Constructing the Causal Graph With Domain Knowledge. Regardless of whether one is using CGM, constructing a causal graph is still a good place to start when seeking to identify confounders in collaboration with domain experts. Once the graph is constructed, we can determine how to appropriately control for confounders (using tools like CGM). A good causal graph aims to identify all the significant causal influences on both the treatment and outcome variables, and the causal relationships between any phenomena that might be

causally “up stream” of both the treatment and the outcome variables. There are no hard and fast rules for constructing causal graphs, but ultimately the goal is to construct a causal graph that most domain experts will agree covers the important related phenomena. In general, the validity of inferences will depend on the completeness and correctness of the causal graph, so arriving at a causal graph that isn’t contested is key to arriving at causal conclusions that can be accepted. Once the graph is constructed, tools like CGM can instruct us on how to appropriately control for confounders. Importantly, in some domains even experts lack adequate knowledge of the causal relationships to construct a sufficiently-complete causal graph; in these cases, robust causal inference may not be possible.

Leveraging Existing Models of Offline Phenomena. Another critical aspect of controlling for confounding in causal inference is appropriately modeling the known dynamical structure of the outcome and treatment phenomena. After all, a causal graph does not tell the full story about how a system works. Simply knowing what causes what does not tell us how exactly changes in one part of the system will change other parts of the system. Therefore to perform robust quantitative causal inferences, we will generally need to leverage existing quantitative models of the offline outcome phenomena of interest. For instance, if we are interested in epidemic outcomes and their relationship to online antivaccine content, we should leverage existing quantitative models of epidemics that account for their known dynamical structure.

This raises the question: how exactly can we apply known dynamical models of the offline phenomena in causal inference? In general, the process goes like this: (1) Find or construct an existing model of offline outcome phenomena that predicts outcomes from offline data only, (2) Minimally modify this model to include an additional mechanism by which online social exposures may influence outcomes, (3) Fit the model to the observed data, and (4) Apply a causal inference framework to extract a causal conclusion from the fitted parameters, data, and model specification. This last point is the most variable, so let’s

discuss a few possible ways of doing it, based on the kind of predictive outcomes model used.

Autoregression Models. In this simplest case, the available predictive models of the outcome phenomena might be simple autoregressive time series models using data about confounding influences on the outcome variable. In this case, the model can be minimally modified to include an additional term for the online social exposures over time, and then the autoregressive model can be used in a DiD or causal panel regression analysis based on the time series data. In such inferences, the fitted model parameters generally directly relate to the average treatment effect.

Dynamical Systems Models. In more well-theorized domains, we might have a dynamics systems models based on differential equations relating several observed quantities, such as the compartmental epidemic models in a study of infectious diseases. In this case, the involvement of the model is somewhat more complex. Again, we'll start by minimally modify the dynamical system to include a mechanism by which social exposures may influence outcomes, and then fit the model to data. Then, we can use CGM to determine the form of the ATE we're aiming to measure. The resulting ATE estimand will be some expression involving the model variables, the data, and the fitted parameters. In general, we can simply plug in the model equations, the data, and the parameters, and arrive at the ATE (e.g., as in Figure 3.2). A more thorough illustration of this process, will be detailed below in Section 3.7.

Point Process Models. Another important class of models is point process models, which describe the rate at which discrete events happen over time, such as the incidence of offline political protests at a particular location or criminal incidents throughout a city at known times and locations. Specifically, we are interested in a subtype called Hawkes models in which events make future events more likely for a period of time (see Appendix C); this

“excitation” of future events is the causal mechanism in the model. Confounders may be accounted for as additional event types, or additional predictors of the baseline event rate. Similar to time series models, these kinds of self-excitatory point process models can be minimally modified to include an additional event type or baseline rate representing online social exposure. When a properly constructed model containing the relevant confounders is fitted to the data, the causal conclusions may generally be read directly from the fitted excitation functions (see Appendix C).

Micro-simulations and Agent Based Models. Finally, the last major class of models we’ll consider are blackbox computer simulations, which may leverage agent based or micro-simulation methods. These models are somewhat common in complex systems modeling, which is often used in domains in which we may want to predict outcomes, such as health behaviors, violence behaviors, etc. In this case, I know of no clear best practices for robust causal inference (although they may exist). However, one possible approach might be to treat the simulation’s predictions of outcomes as an additional predictive input variable to a simpler form of causal analyses, such as DiD, causal panel regressions, or causal Hawkes models. In this way, the predictions of the simulation are being treated as a kind of propensity score, accounting for the phenomena and confounders that the simulation models. This approach has the limitations that all propensity score methods have: notably, the the score is only as good as the model used to produce it, and failing to use the score properly in downstream causal modeling may render it useless.

3.4.3 Additional Modeling Tricks

In this last subsection, we’ll highlight two additional modeling tricks that can be important tools in these causal inferences.

Handling Causal Loops. First, often when constructing our causal graphs, we will encounter causal loops (e.g., $A \rightarrow B \rightarrow C \rightarrow A$). To handle causal loops in a formal causal

framework like CGM, we'll need to employ the trick of “unrolling” the causal graph over time (see Section A.4). In this procedure, we discretize time and make the variables in each timestep depend only on the variables in the past timestep; this “unrolls” the cyclic causal graph into a DAG, as required by CGM.

Handling Initial Differences in Populations. Second, when we are studying changes in offline behaviors, often we will have to account for the fact that each population may have a different initial propensity to engage in the behavior, prior to the online social exposures we're studying. In this case, it may sometimes be useful to treat these initial propensities as latent variables that are inferred in the process of fitting our model to the data. In this way, we can better study how online social exposures change these propensities over time, because we know we have fully accounted for initial difference in propensity. This has the additional advantage of accounting for many static, population-level confounding variables that may influence the initial propensity. An example of the use of this kind of latent variable is the initial vaccine hesitancy parameters α_i in the SIRVA model (subsection 5.5.4) in Chapter 5.

3.5 Summarizing the Method

Now let's summarize the proposed method for linking online causes to offline effects. We'll start by collecting the online social media and offline outcomes data and ensuring that we can link them together by categorizing online users into offline population groups for which we have offline outcomes data. Next, we'll identify our content of interest online, and determine its production and exposure rates in each population group over time. Then, we'll construct a causal graph of the whole system, including both the hypothesized causal pathway between online and offline phenomena and the known offline influences on outcomes. Then using CGM or similar analyses, we'll determine our identification strategy and which confounders must be controlled for. Finally, we'll fit a predictive model to the data (per-

haps leveraging existing models of offline phenomena), while being sure to account for confounders appropriately in the model. This will then provide us with the information we need to compute the ATE (which in the simplest cases may be merely a fitted parameter of the model).

3.6 Necessary Conditions and Limitations the Method

Although the methodology described above is reasonably general in its formulation, it has five key requirements that are in fact quite restrictive.

1. **Identifiable Online Treatment Content.** Instances of the online treatment content must be readily definable and identifiable, so that we may determine production rates and exposures in each population group, per Section 3.3.
2. **Observable Online Social Connections.** Relevant social connections between online users must be observable. For instance, their “followers,” their “friends,” or their interactions with other users. This is critical to determining the exposure as defined in Section 3.3.
3. **High Quality Online User Data.** Online users must be identifiable with offline populations about which we have offline outcomes data (e.g., via user geolocation or user demographics). Additionally, the users in our corpus of data must be a significant fraction of the offline populations (especially the set of users whose interactions or connections are used to generate the social network W_{ij} in Equation 3.1). This requirement is necessary to link online exposure data to offline outcomes data.
4. **High Quality Offline Outcomes Data.** The offline outcomes data must be quite high resolution, both temporally and also in terms of geographic or demographic specificity. These are essential requirements for robust causal inference analyses. Insufficient temporal resolution will likely conceal the effects of social exposures, which

are often short-lived and immediate, and insufficient geographic or demographic resolution will limit the power of inferences and the ability to appropriately control for confounders.

5. **Robust Domain Knowledge of Outcome Phenomena.** Strong existing domain knowledge about offline outcomes is critical to identify, measure, and control for possible confounding variables. Additionally, because our causal inference methods are quantitative, we also generally will need to have a strong quantitative, predictive model of the offline outcomes phenomena. It is strongly preferable that this quantitative model is well established and vetted by prior work, because the validity of our causal inferences will generally hinge on our ability to appropriately control for the known internal dynamics of offline phenomena.

The strictness of these requirements may be the reason for the relative lateness of the formulation of this methodology within social media studies. Simply put: it's not often that we can use it.

Limitations. Finally, it is critical to note two limitations of this method that apply even when the conditions above are met.

First, this method is necessarily subject to all of the normal pitfalls of observational causal inference methods. Most especially, sometimes not all significant confounders are known, observable, and quantifiable, and those confounders that have not been accounted for can threaten the validity of the causal inference. For this reason, the most important step of causal inference is the proper and complete construction of the causal graph, on which every inference will necessarily depend.

Second, the method proposed in this chapter does not demonstrate mechanisms for causation, and in many contexts causal claims are suspect unless one can articulate and demonstrate plausible mechanisms by which the effects might be produced. This limitation will be addressed in Chapter 7, in which we will articulate a plausible set of mechanisms

by which online social exposures may produce offline population-level effects, and offer some evidence that they occur.

3.7 A Worked Example of the Method in Brief

The general method above allows us to link online causes to offline outcomes, in some special cases. In this section, we'll walk through a high level summary of one of our case studies (Chapter 5) to demonstrate more clearly the application of this method.

Consider the following case: we observe that antivaccine content became widely shared on Twitter during the early period of the COVID-19 vaccine roll-out. We hypothesize that higher exposure to antivaccine content causes a measurable reduction in vaccine uptake by increasing vaccine hesitancy, and that this subsequently produces increased COVID cases and COVID-attributable deaths. We would like to measure this effect.

This case study satisfies the five requirements discussed in the previous section: (1) We can identify the treatment content (antivaccine tweets), (2) We can observe social connections between users (vaccine-related and COVID-related retweets), (3) We can identify online users with offline populations (via self-reported user geolocation), (4) We have high quality offline data with high geographic and temporal resolution (CDC records of vaccinations and infections), and (5) We have robust existing domain knowledge and models of the offline outcome phenomena, namely epidemic dynamics and vaccination behaviors. These features make this an ideal case study for the method.

First, we collect and prepare the data. The offline data consists of vaccinations and cases per capita each day during the initial COVID-19 vaccine rollout in the US, and the online data consists of a collection of all COVID and vaccine related tweets originating in the US. We prepare the online data by identifying all the tweets that are antivaccine and the geolocations of observed twitter users, and then computing the exposure rates over time for each US county. We then match these data with the county-level case and vaccination records, resulting in a time series dataset of cases, vaccinations, and exposures for

each population group (US counties, specifically). This is the joint data of treatments and outcomes that we will use.

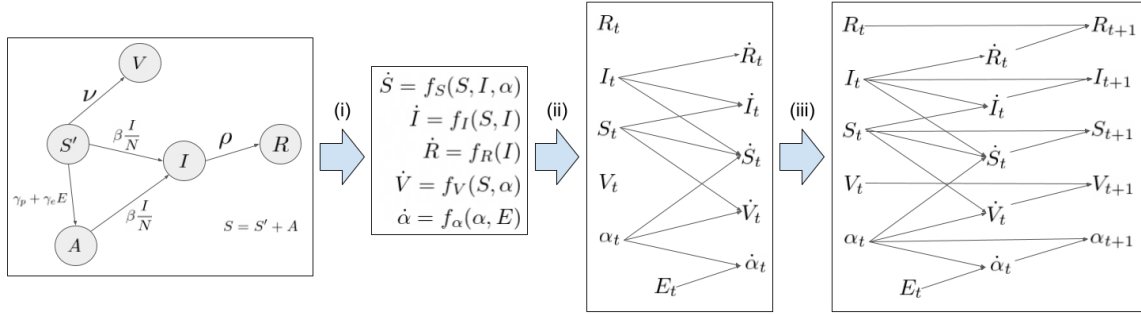


Figure 3.1: Steps to construct a causal graphical model from a dynamical equations model. Note that $\dot{X} = \frac{dX}{dt}$.

Next, we'll construct a predictive model of our offline phenomena and its relationship to the online content. The exact details of this model are not important here, but we'll offer a brief sketch to highlight some key points. We begin with an off-the-shelf model of the offline phenomena, namely a compartmental epidemic model that includes vaccination (see Appendix E). We then will minimally modify this well-known model to include the influence of the exposure to online content. Specifically, we'll add a term for vaccine hesitancy (which prevents people from becoming vaccinated), and a mechanism by which exposure to antivaccine content can increase or decrease vaccine hesitancy (the final compartmental model is shown in the first panel of Figure 3.1).

The dynamical equations that make up this model (given in abstract in the second panel of Figure 3.1) implicitly encode causal relationships, which we can represent as a causal graph (third and fourth panels of Figure 3.1). The process here is relatively straightforward: we look at the functions in the dynamical model, and each functional dependency is a causal relationship from function input to function output (in this case, from variables to their time derivatives). We also include the trivial observation that the future values of variables depend on their past values and time derivatives (fourth panel of Figure 3.1). In this way, a dynamical systems model can be translated into a causal graph (more details on this process available in subsection 5.5.6). A similar procedure is possible for any

dynamical systems model, and may be possible in some other predictive models with clear mechanisms (e.g., some microsimulation and agent based models).

$$\begin{aligned}
 \text{ATE}_{E \rightarrow \dot{V}} &= \frac{d}{dE_{t-1}} \mathbb{E}[\dot{V}_t] \\
 &\Downarrow \text{1. Plug in model equations} \\
 \text{ATE}_{E \rightarrow \dot{V}} &= -\mathbb{E}[\nu_{t-1} S_{t-1} \gamma_e (1 - \alpha_{t-1})] \\
 &\Downarrow \text{2. Evaluate with data and fitted parameters} \\
 \text{ATE}_{E \rightarrow \dot{V}} &= -757
 \end{aligned}$$

Figure 3.2: Process for evaluating an ATE estimand, based on the causal graph in Figure 3.1 and the dynamical equations in subsection 5.5.4. The expectation value is taken over the joint posterior distribution of the inferred parameters and the observed data points, as specified more explicitly in Equation 5.10.

Once we have constructed our causal graph, we can apply standard methods of causal graphical modeling. First, we consult with domain experts to determine if our graph is missing any known relationships, and add them to the graph. Then having confirmed the sufficiency of our causal graph, we can compute an estimand for average treatment effect (ATE) on the relationship of interest. In this case, we're interested in the relationship between Exposure (E) and change in vaccine uptake (\dot{V}).

Given the causal graph, we can compute the ATE estimand using CGM's do-calculus. The details of this process can be abstracted, as this computation can be performed automatically by modern causal inference software libraries that implement do-calculus, such as the Python package DoWhy [48] (see Figure A.1). The outputs of this computation are three things: (i) a determination of whether or not an ATE estimand can be computed (some causal graphs make it impossible); (ii) a generic form of the estimand, written in terms of partial derivatives of conditional expectation values of the observed variables in the graph; and (iii) a set of assumptions attached to the estimand, which must be checked manually

for the ATE estimand to be valid. In general, the resulting estimand expression can then be evaluated to yield a numerical result by simply plugging in the model equations and evaluating with the fitted parameters and data of the observed variables (see Figure 3.2).

To summarize, the general procedure used in this case study was: (1) Collect both online exposure data and offline outcomes data for each of several population groups, (2) Model the offline phenomena using a well-established quantitative, mechanistic model, (3) Modify the model minimally to include a linkage between online and offline phenomena, (4) Fit the model to the data, (5) Write down the causal graph of the model, and confirm it is sufficiently complete, (6) Compute the ATE estimand from the causal graph and evaluate it using the model equations, fitted parameters, and observed data.

3.8 Conclusion

In this chapter, I defined a method for linking online social exposures to offline outcomes in populations, and defined the conditions under which it is applicable. I presented a few key methodological innovations. First, I defined how online social media data and offline outcomes data can be linked by identifying online users' memberships in offline population groups. Second, I defined a metric for social exposure to online content through social networks. Third, I discussed how causal modeling may be applied to the resulting dataset of exposures and outcomes of offline population groups, with special emphases on how domain knowledge should be incorporated to account for confounders and known dynamical structure of offline phenomena. Finally, I defined the limits of applicability of this proposed method. Although some of these ideas have been applied before in social media studies, this chapters still represents a significant contribution for its systematization, review, and clarification of available tools and best practices.

In the following three chapters, I'll present two case studies: one in which this method was applicable, and one in which was not. In Chapter 7, I will delve further into the issue of possible mechanisms for the effects observed in these case studies.

CHAPTER 4

DATA COLLECTION FOR ANTIVACCINE TWEETS CASE STUDY

In this chapter, I explore the necessary but less glamorous side of social media studies: data collection. This chapter documents how the Observatory on Social Media team collected Twitter data on public sentiments toward the COVID-19 vaccinations during the initial 2021 vaccination campaign in the United States. This timely and careful data collection by our team facilitated the primary case study of this dissertation: the study of how antivaccine tweets impacted public health outcomes during the COVID-19 behavior (see Chapter 5).

This data collection had a number of key features that made it especially suitable for our later causal study. First, we were able to collect the geolocation of a large fraction of users in the dataset, allowing us to link the online data to offline outcomes- a key prerequisite for our causal method as defined in Chapter 3. Second, the data was collected over the key outcome-determinative period in which many people were initially becoming vaccinated, and so the impacts of these online conversations on offline vaccination behaviors were more easily measurable. Finally, the data collected was narrowly targeted to discussions of vaccination and COVID-19, and related antivaccine narratives. This specificity allowed us to observe the precise subset of the Twitter social network that communicated antivaccine sentiments to users, allowing us to precisely define our measure of exposure to antivaccine sentiments.

Any work aiming to conduct a causal study of social media's offline effects must be able to perform a data collection process similar this, in order to obtain the high quality and high specificity data required to use the methods outlined in Chapter 3.

4.1 Introduction

The COVID-19 pandemic has killed two million people and infected 93 million around the world as of mid-January, 2021 [49]. Vaccines were critical in our fight to end the COVID-19 pandemic [50]. It was estimated that around 60-70% of the population would need to be vaccinated against COVID-19 to achieve herd immunity [51]. However, surveys have found that only 40-60% of American adults reported that they would take a COVID-19 vaccine [52, 53]. With these levels of *vaccine hesitancy*, it is unlikely we will reach herd immunity; COVID-19 will remain endemic.

A possible driver for vaccine hesitancy is the anti-vaccination movement. This movement has been on the rise in the U.S. for two decades, beginning with unfounded fears over a Measles, Mumps and Rubella (MMR) vaccine [54]. The vocal online presence of the anti-vaccination movement has undermined confidence in vaccines. Worse, resistance to the COVID-19 vaccines is currently much more prevalent than resistance to the MMR vaccine. Since COVID-19 vaccine hesitancy and its drivers remain understudied, a goal of our project is to help address this gap.

There is a growing body of evidence linking social media and the anti-vaccination movement to vaccine hesitancy [55, 56, 57]. Studies show that vaccine hesitancy in one's peer group is associated with future hesitancy [58], and that misinformation spread on social networks is linked to poor compliance with public health guidance about COVID-19 [59]. Based on these findings, the core hypothesis behind this project is that the social spread of vaccine misinformation and vaccine hesitancy impacts public health outcomes such as vaccine uptake and COVID-19 mortality rate.

Here we present a collection of English-language posts related to the COVID-19 vaccines on Twitter. The collection is exempt from IRB review as it only includes tweet IDs of public messages. This allows us to comply with the Twitter Terms of Service while making the data available to both researchers and the general public. Although there has

been previous work presenting COVID-19 Twitter datasets [60, 61, 62], our work focuses specifically on discussion of COVID-19 vaccines and related public health outcomes.

The *CoVaxxy* dataset will enable researchers to study vaccine misinformation and hesitancy, and their relationship to public health outcomes. We will use established techniques to track vaccine misinformation within the data, along with misinformation superspreaders, coordinated campaigns, and automated accounts [63, 64, 65, 66, 67]. We will also relate this social media data to geographic public health data (such as COVID-19 mortality and vaccine uptake rates) by using geolocation data within the dataset.

In this chapter we describe the methods used to create the *CoVaxxy* dataset. Using one week of data, we provide a descriptive analysis and illustrate how our data could be used to answer various research questions. We also present the *CoVaxxy Dashboard*, a tool intended for the public to track key insights drawn from the data. Opportunities and limitations are discussed as we draw conclusions.

4.2 Dataset Curation

Our key data collection goal is to download a complete set of Twitter posts related to COVID-19 vaccines. In this section we describe our methodology for selecting appropriate keywords to achieve such a coverage. We then describe our architecture with server redundancy to maintain an unbroken stream of Twitter data containing these keywords.

4.2.1 Identifying COVID-19 Vaccines Content

To create as complete a set of Twitter posts related to COVID-19 vaccines as possible, we carefully select a list of keywords through a snowball sampling technique [68, 63]. We start with the two most relevant keywords, i.e., `covid` and `vaccine`, as our initial seeds. Keywords also match hashtags, URLs, and substrings. For example, `covid` matches “`cnn.com/covid`” and “`#covid.`” Next, we gather tweets utilizing the filtered stream

endpoint of the Twitter API¹ for three hours. From these gathered tweets, we then identify potential keywords that frequently co-occur with the seeds. These keywords are separately reviewed by two authors and added to the seed list if both agree that a keyword is related to our topic. This process was repeated six times between Dec. 15, 2020 and Jan. 2, 2021 with each iteration’s data collection taking place at different times of the day to capture tweets from different geographic areas and demographics. The seed list serves as our initial keyword list.

We further refine the keyword list by manually combining certain keywords into composites, leveraging the query syntax of Twitter’s filtered stream API. For example, using `covid19 pfizer` as a single composite matching phrase will capture tweets that contain *both* “covid19” *and* “pfizer.” On the other hand, including `covid19` and `pfizer` as separate keywords will capture tweets that contain “covid19” *or* “pfizer,” which we consider as too broad for our analysis. The final keyword list includes 76 (single or composite) keywords. Constructing various composites of relevant keywords in this way ensures the dataset is broad enough to include most relevant conversations while excluding tweets that are not related to the vaccine discussion.

4.2.2 Content Coverage

To demonstrate the effectiveness of the snowball sampling technique introduced above, we calculate the popularity of each keyword in the final list by the number of unique tweets and unique users associated with it.

Figure 4.1, where keywords are ranked by popularity, shows that additional keywords beyond the 60 most popular ones tend to capture very small numbers of users and tweets, relative to other keywords in the collection. This suggests that including more keywords in the seed list described above is not likely to alter the size and structure of the dataset significantly. In fact, the inclusion of additional keywords could be redundant, due to the

¹<https://developer.twitter.com/en/docs/twitter-api/v1/tweets/filter-realtime/overview>

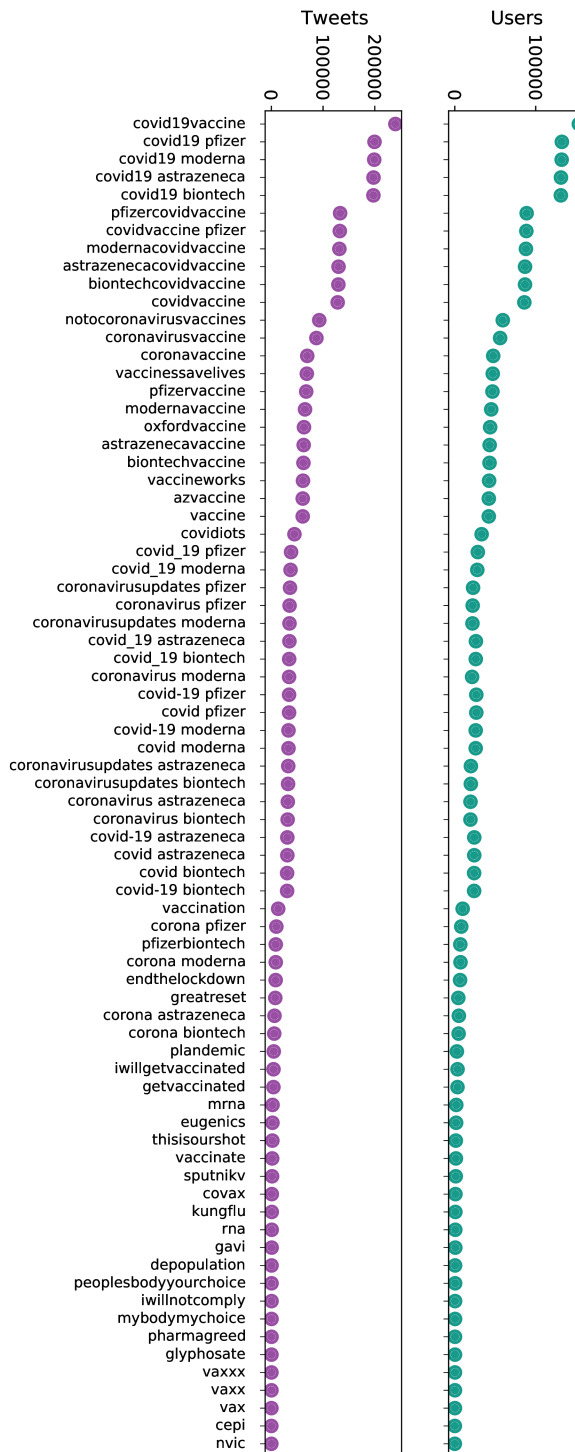


Figure 4.1: Number of tweets (purple, left) and users (green, right) captured by each key-word/phrase in the final list (ranked by popularity) between January 4–11, 2021.

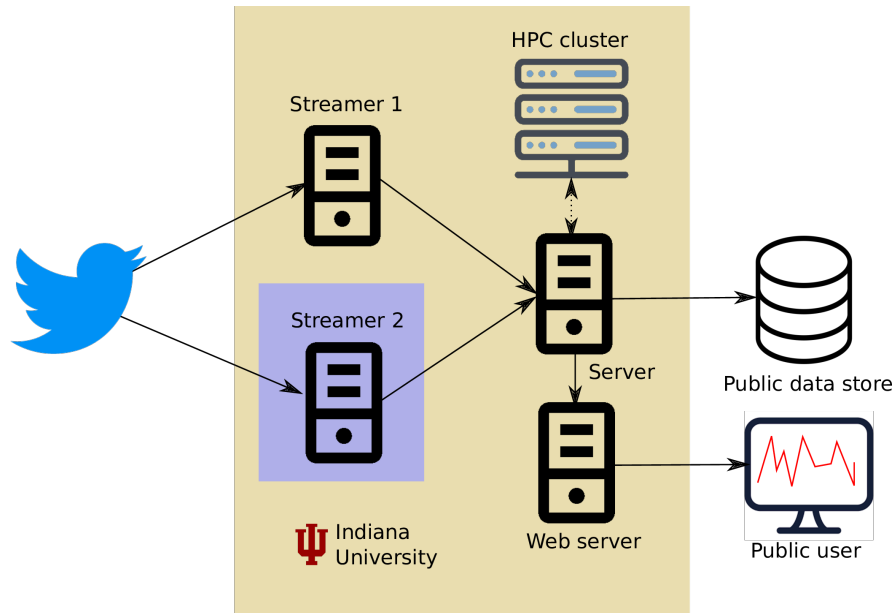


Figure 4.2: The VM server architecture for the *CoVaxxy* project. Data flows in the direction of the arrows. Machines in the larger yellow box are hosted by Indiana University. The VM “Streamer 2,” in the embedded blue box, is hosted by the Texas Advanced Computing Center.

co-occurrence of multiple keywords and hashtags in a single tweet, especially for the most popular terms. Thus, we believe that our set of keywords provides reasonable coverage and is representative of tweets communicating about COVID-19 vaccines.

As the collection of tweets is intended to persist over time, new relevant keywords may emerge. To ensure that the keyword list remains comprehensive throughout the entire data collection period, our team will continue to monitor the ongoing public discussion related to COVID-19 vaccinations and update the list with important emerging keywords, if necessary.

4.3 CoVaxxy Infrastructure

4.3.1 Data Collection Architecture

Our server architecture (Figure 4.2) is designed to collect and process large quantities of data. This infrastructure is hosted by Extreme Science and Engineering Discovery Environ-

ment (XSEDE) Jetstream virtual machines (VMs) [69, 70]. To maintain the integrity of our tweet streaming pipeline, we have incorporated redundancy. We maintain two *streamer* (stream collection) VMs in different U.S. states so that if one suffers a fault we can use data from the other. These servers connect to Twitter’s filtered stream API to collect tweets that match any of the keywords in real time. We use the language metadata to filter out non-English tweets.

The data from the two streamers is collated on a general purpose server VM where we run data analysis. The server VM is also linked to Indiana University’s high performance computing infrastructure for running advanced analyses.

We upload new data files to a public data repository [71] each day² and will continue to do so as long as the topic of COVID-19 vaccinations remains relevant in public discourse. This repository also includes our list of keywords. In compliance with Twitter’s Terms, we are only able to share tweet IDs with the public.

At the time of our original publication of this work, one could re-hydrate this dataset by querying the Twitter API or using tools like Hydrator³ or twarc⁴. However, today Twitter’s pricing and data access policies have changes, making this infeasible for nearly all researchers. The data is unfortunately no longer generally available.

Finally, a web server provides access to the data on the server VM through the interactive *CoVaxxy* dashboard, described next.

4.3.2 Dashboard

Existing COVID-19 visualization tools include those by Johns Hopkins University [49] and The Atlantic.⁵ These trackers address hospitalization and mortality. Another dashboard from the Fondazione Bruno Kessler reports on the proportions of misinformation

²<https://doi.org/10.5281/zenodo.4526494>

³<https://github.com/DocNow/hydrator>

⁴<https://github.com/DocNow/twarc>

⁵<https://covidtracking.com/>

Users	Tweets	Hashtags	URLs
1,847,067	4,768,204	39,857	983,158

Table 4.1: Breakdown of the data collected between January 4 and January 11, 2021 in terms of unique users, tweets, hashtags, and URLs.

and epidemic-related statistics (e.g., confirmed cases and deaths) per country.⁶ Finally, the Our World in Data COVID-19 vaccination dataset publishes vaccine uptake information by country.⁷

We are not aware of any tools that concurrently explore the relationships between COVID-19 vaccine conversations, vaccine uptake, and epidemic trends. Consequently, we have created a web-based dashboard to fill this void. The *CoVaxxy* dashboard will track and quantify credible information and misinformation narratives over time, as well as their sources and related popular keywords.⁸ Although we collect English tweets related to vaccines globally, the dashboard provides state-level statistics in the United States. Additionally, it shows global hashtag and domain sharing trends. It is updated daily. Figure 4.3 illustrates one example of an interactive visualization that lets users visualize the relationship between various misinformation-related and COVID-19 pandemic data. This data will be displayed alongside COVID-19 pandemic and vaccine trends. By highlighting the connection between misinformation and public health actions and outcomes, we hope to encourage the public to be more vigilant about the information they consume from their daily social media feeds in the fight against COVID-19.

4.4 Data Characterization

Our system started to gather tweets on Jan. 4, 2021. Table 4.1 provides a breakdown of the dataset (as of January 11) in terms of the number of unique users, number of tweets they shared, and numbers of unique hashtags and URLs contained in these tweets. Next let us

⁶<https://covid19obs.fbk.eu>

⁷<https://ourworldindata.org/covid-vaccinations>

⁸<https://osome.iu.edu/tools/covaxxy>

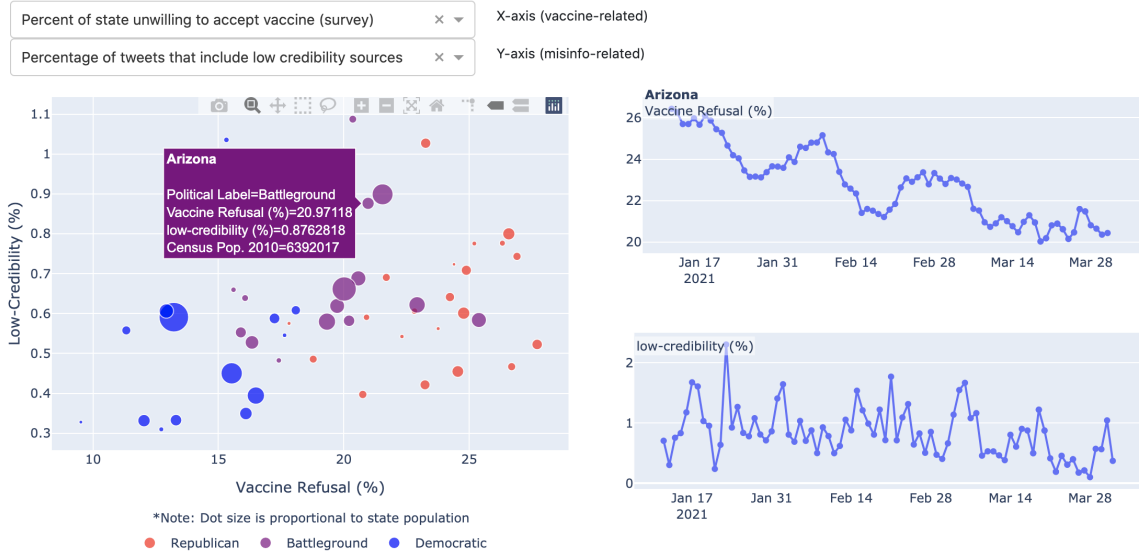


Figure 4.3: Example visualization from the *CoVaxxy* web dashboard. This visualization lets users plot relationships (at the state-level) between vaccine-related and misinformation-related data. The left figure’s axes are selected from the dropdowns, displaying the aggregate relationship. The two figures on the right illustrate the same relationship from a temporal perspective for an individual state. The user chooses which state to visualize in the figures on the right by hovering over a dot within the left figure.

analyze the data from that week to illustrate how our dataset might be used for different research projects.

4.4.1 Volume

We show in Figure 4.4 a time series for the number of tweets collected in our dataset, on an hourly basis. We can notice a decrease in the number of tweets after January 6, which might be driven by the increased media attention surrounding the storming of the U.S. Capitol.⁹ In fact, the mean daily number of tweets decreases from 900k tweets in the period of Jan 4–6 to 400k tweets in the period of Jan 7–11.

In Figure 4.5 we show the distribution of the tweets geo-located in the contiguous United States. We use a naive approach to match tweets to U.S. states: we first extract the user location from the profile (if present) and then match it against a dictionary of U.S. states. Finally, we compute the number of tweets for each state based on the activity of

⁹<https://www.nytimes.com/2021/01/06/us/politics/protesters-storm-capitol-hill-building.html>

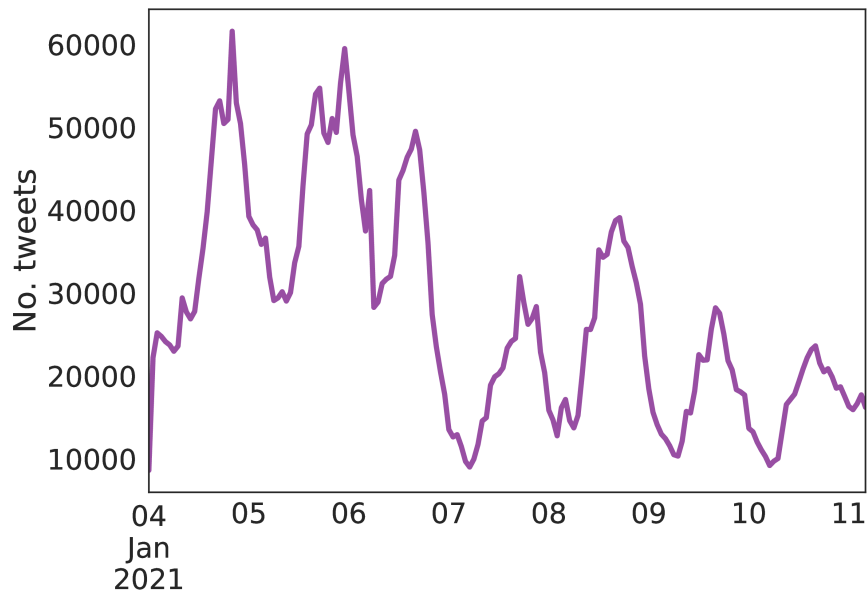


Figure 4.4: Number of collected tweets on an hourly basis since the beginning of the collection.

users geo-located in that state. Over 1M users in our dataset have location metadata in their profile; we were able to match approximately 40k users resulting in 600k geo-located tweets. Providing an accurate methodology to geo-locate users is outside the scope of this paper; the reader should consider these results only as an illustration of the insights that can be gained from the *CoVaxxy* data.

4.4.2 Hashtags

Figure 4.6 lists the most tweeted hashtags between January 4 and 11. We can see that they are largely related to the SARS-CoV-2 vaccine, with one (“#covidiot”) referring to COVID-19 deniers.

Many different conversations can occur concurrently on Twitter, using different hashtags for different topics. To cluster related hashtags, we have grouped them together using a network algorithm. We form a co-occurrence network with hashtags as nodes and edges weighted according to how often the linked hashtags co-occur within tweets. Nodes are clustered using the Louvain method [72]. Groups with hashtags that are used the most are

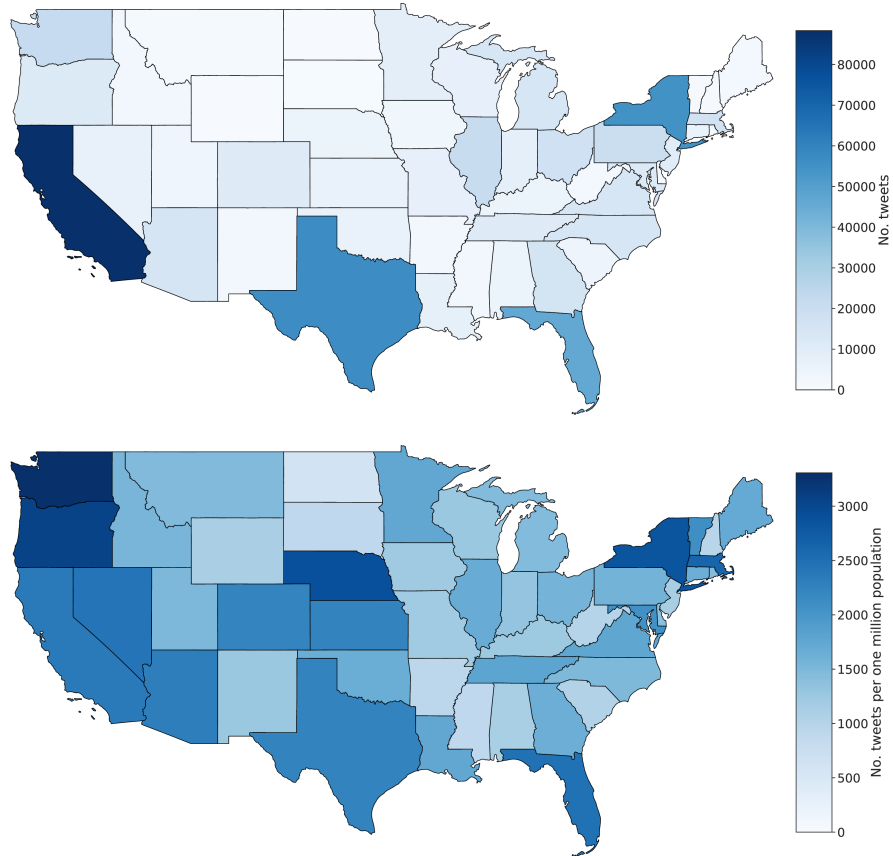


Figure 4.5: Distribution of the number of inferred geo-located tweets per U.S. state (excluding Alaska and Hawaii) by absolute numbers (top) and normalized by 2010 state population (bottom).

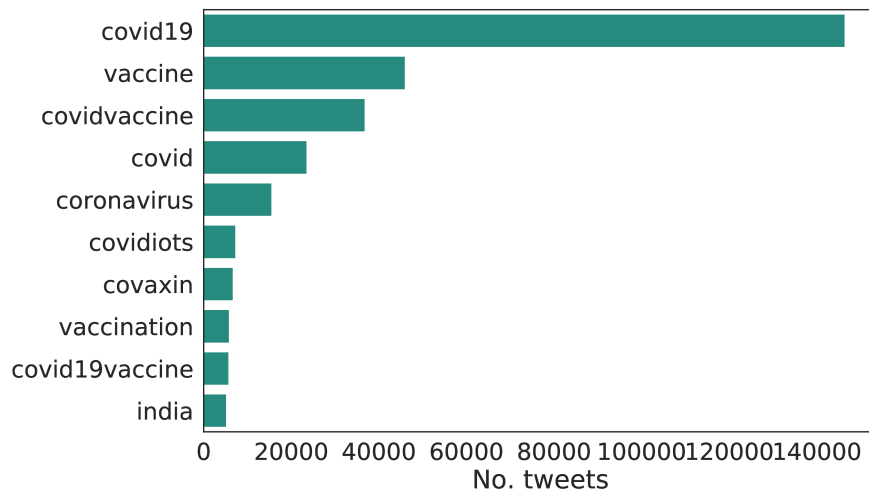


Figure 4.6: Top-10 shared hashtags.

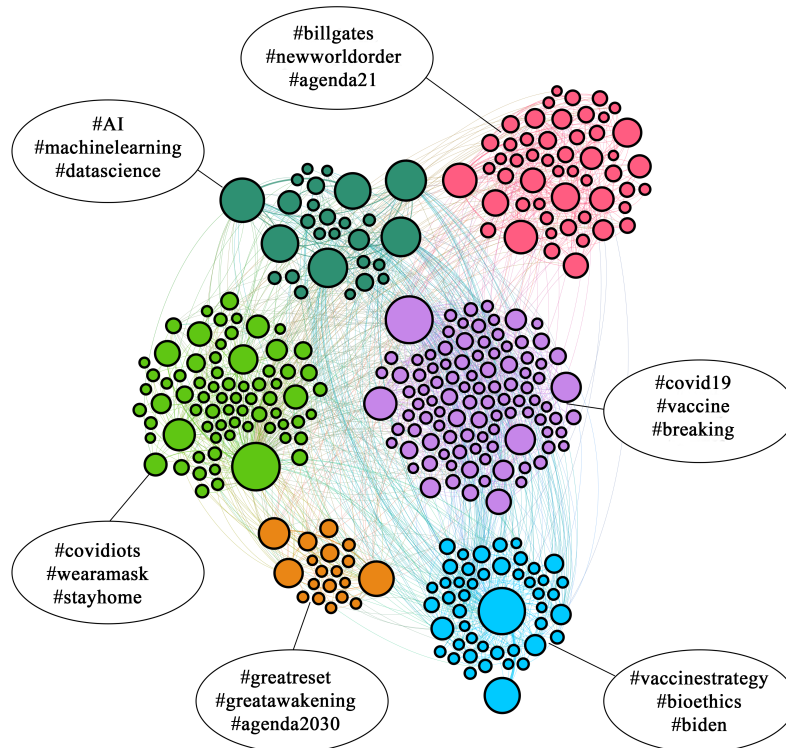


Figure 4.7: An overview of the prominent hashtags in the data, clustered into topic groups. A few hashtags characterizing each cluster are shown.

plotted in Figure 4.7. We observe groups of hashtags associated with vaccine conspiracy theories (“#greatreset,” “#billgates”) as well as positive messages (“#stayhome”).

4.4.3 Sources

In Figure 4.8 we show the top-10 most shared websites. We exclude “twitter.com,” which accounts for over 3M tweets. These sites are comprised mostly of high-credibility information sources. However, one low-credibility source — “zerohedge.com” — also makes this list (see below for details on the classification). We also observe a large number of links to YouTube, which suggests further investigation will be needed to assess the nature of this shared content.

Figure 4.9 provides time series data illustrating the prevalence of low- and high-credibility information. We follow an approach widely adopted in the literature [73, 74, 75, 76, 64]

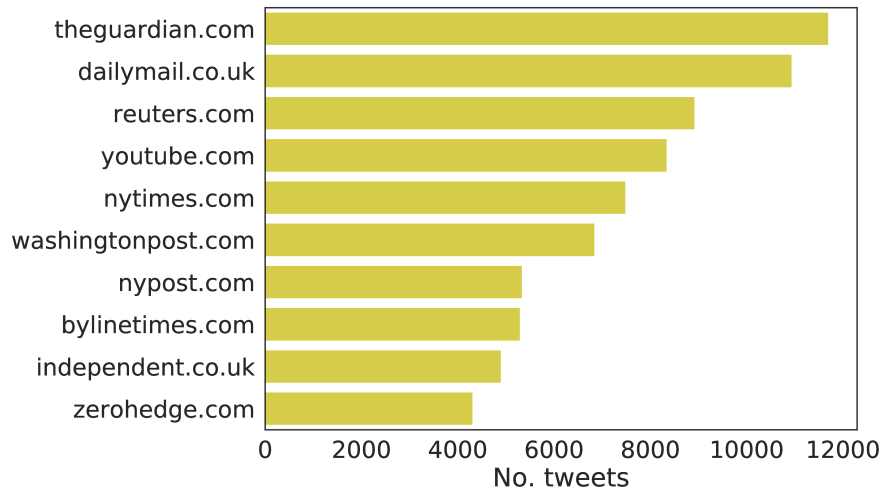


Figure 4.8: Top-10 sources shared in vaccine-related tweets.

to label links to news articles based on source reliability. In particular, we use a third-party list of 675 low-credibility sources¹⁰ and 26 hand-selected mainstream sources. The mainstream sources in this list are labeled by the Media Bias / Fact Check organization as having a factual reporting record as “very high”, “high”, “mostly factual” or “mixed.” We refer to them as “high-credibility” throughout the paper for simplicity. Overall, links to low-credibility sources account for 24,841 tweets compared to 72,680 tweets linking to our sample of mainstream sources. Readers should note that these numbers do not fully capture the news circulating on Twitter, as the lists we employ cannot be exhaustive.

We further list in Figure 4.10 the 20 most shared news sources in both classes. We notice several unreliable sources (cf. “zerohedge.com” and “bitchute.com”) that exhibit prevalence comparable to more reliable websites.

4.5 Discussion

In this chapter we present a dataset tracking discourse about COVID-19 vaccines on Twitter. We characterize the data in several ways, including prominent keywords, geographic distribution of tweets, and clusters of related hashtags. We also present a data dashboard

¹⁰<https://iffy.news/iffy-plus/> (accessed November 2020)

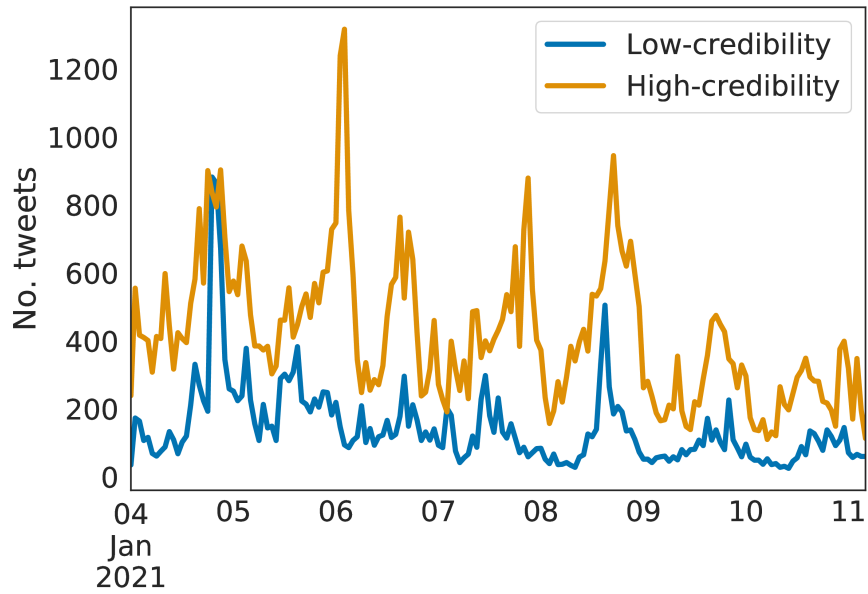


Figure 4.9: Number of hourly tweets containing links to low- (blue) and high-credibility (orange) sources.

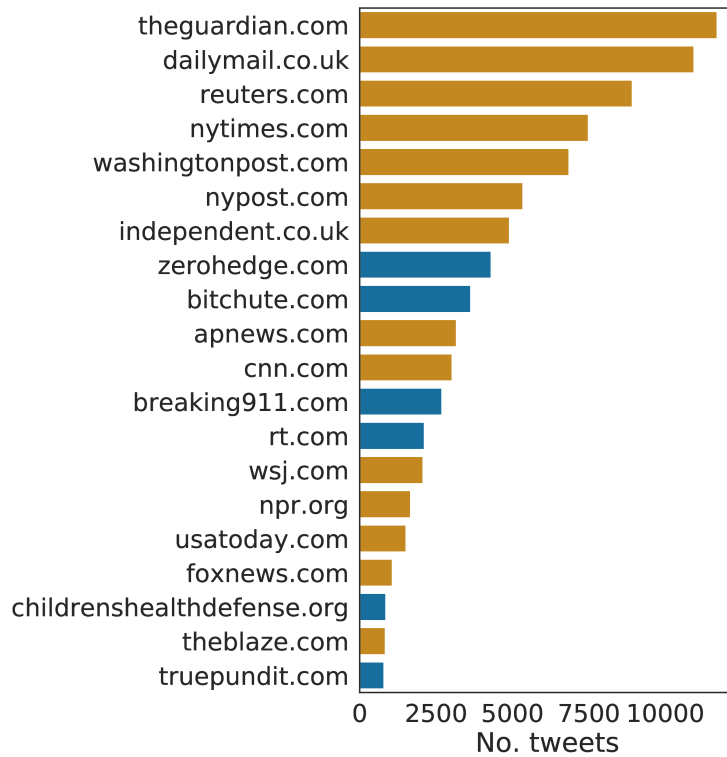


Figure 4.10: Top-20 shared low- (blue) and high-credibility (orange) sources.

that visualizes statistics and insights from this data.

The next chapter explores the relationship between online discussion of COVID-19 vaccines and public health outcomes, like COVID-19 mortality and vaccine uptake. We will also leverage existing social media analysis tools to track emerging narratives and suspicious accounts, such as bots, coordinated campaigns, and troll farms [63, 64, 65, 66, 67]. Finally, we plan to explore models to better understand how vaccine misinformation and anti-vaccine sentiment spreads on social media.

This dataset has some key limitations. Critically, Twitter users are not a representative sample of the population, nor are their posts a representative sample of public opinions [77]. Additionally, filtering our stream to include only English-language tweets comes at the price of occasionally excluding some variants of this language. This is because our stream gathers tweets that have been marked as containing English by Twitter’s automatic language identification system, which may not capture some tweets by minority dialect speakers and multilingual speakers [78].

The Twitter Filtered Stream API imposed a rate limitation of 1% of all public tweets. During the week described herein, we did not encounter this limitation.

Another potential source of bias is the keyword sampling procedure used to identify and collect COVID-19 vaccine related content, which involved evaluation of keywords to determine what was relevant. We are unable to fully exclude irrelevant content using only keyword-based filtering. However, further filtering is possible at a later stage. Other researchers may also refine the data to properly address their own topics of interest.

Given the large-scale, real-time nature of our data collection infrastructure, users do not have the ability to opt-out. This raises important ethical concerns related to anonymity. To address this concern, we note that (1) our dashboard only displays aggregate data, obfuscating the ability of users to identify those captured within our data; and (2) should a user delete a tweet or account, the related information will not be returned by Twitter during the re-hydration process.

The long-term aim of this project is to tackle the ambitious challenge of linking social media observations directly to public health, which we do in the next chapter.

CHAPTER 5
CASE STUDY: ANTIVACCINE TWEETS AND COVID-19 VACCINE
HESITANCY

The case study in this chapter constitutes the culminating demonstration of the causal methods outlined in Chapter 3. In this paper, we study how antivaccine content on Twitter affected vaccine hesitancy, vaccine uptake rates, and infection and death rates during the COVID-19 pandemic. We find that antivaccine tweets do cause increases in vaccine hesitancy, reductions in vaccine uptake, and increases in resulting infections and deaths. This novel result demonstrated that social media content (and therefore social media moderation policies) have significant public health implications.

Analyzing the impacts of antivaccine tweets during the COVID-19 pandemic is an especially good case study for our novel causal inference methodology in Chapter 3 because an abundance of high quality data was available and the underlying epidemic dynamics are well understood. In particular: (1) the offline public health data was highly granular in both time and space (daily records and county-level geographic resolution), (2) the online social media data was abundant, geolocatable, and available, (3) off-the-shelf quantitative models are available for understanding the interactions between epidemics and vaccination. These features of the data and the problem allowed us to link the online and offline data via geolocations, and to model the impacts of online activity on offline outcomes without having to develop and validate novel models of the offline phenomena. This is the ideal case for studies of offline effects of online social media.

This chapter's success demonstrates the viability of the general method outlined in Chapter 3. It is our hope that it may serve as a proof of concept for future proposed work studying the offline effects of online social media

Abstract Vaccines were critical in reducing hospitalizations and mortality during the COVID-19 pandemic [79, 80, 81]. Despite their wide availability in the United States, 62% of Americans chose not to be vaccinated during 2021 [82]. While online misinformation about COVID-19 is correlated to vaccine hesitancy [83, 84], prior work has not established a causal link between real-world exposure to antivaccine content and vaccine uptake. Here we present a compartmental epidemic model that includes vaccination, vaccine hesitancy, and exposure to antivaccine content. We fit the model to observational data and use causal graphical modeling to determine that a geographical pattern of exposure to online antivaccine content across US counties is responsible for a pattern of reduced vaccine uptake in the same counties. We find that exposure to antivaccine content on Twitter caused about 750,000 people to refuse vaccination between February and August 2021 in the US, resulting in at least 27,000 additional cases and 400 additional deaths. This work provides a methodology for linking online speech to offline epidemic outcomes. Our findings should inform social media moderation policy as well as public health interventions.

5.1 Introduction

Several socio-economic factors have been linked to vaccine hesitancy using correlational studies [85], but finding evidence of drivers remains a challenge. One potential driver is exposure to anti-vaccine content on social media. A laboratory study demonstrated that exposure to COVID-19 misinformation decreased willingness to be vaccinated [83]. Using social media data, higher rates of misinformation were found to precede increases in COVID-19 infections in countries during early 2020 [86], though this may have been due to a general wave of COVID-19 discussion in early 2020 [87]. A temporal correlation was also found between increased production of online vaccine misinformation and higher vaccine hesitancy, as well as lower uptake, across US states and counties [84].

These studies focus on associations and do not establish a causal link between real-world exposure to antivaccine content and vaccine uptake. Here we wish to measure the

extent to which changes in levels of exposure to antivaccine content on Twitter in US counties caused changes in vaccine uptake rates. To estimate the causal link between exposure to antivaccine tweets, vaccine hesitancy, and reduced vaccine uptake, we extend the SIR model with states that represent vaccinated and vaccine-hesitant people. This model fits empirical data about COVID cases, vaccinations, and exposure to antivaccine tweets better than simpler models that ignore vaccine hesitancy. A key parameter is the rate at which people exposed to antivaccine content become vaccine hesitant. Fitting the model to the data yields a positive value for this parameter, indicating that increased exposure increases vaccine hesitancy. As more people become vaccine hesitant, vaccine uptake rates decline, leading to increased infections and deaths.

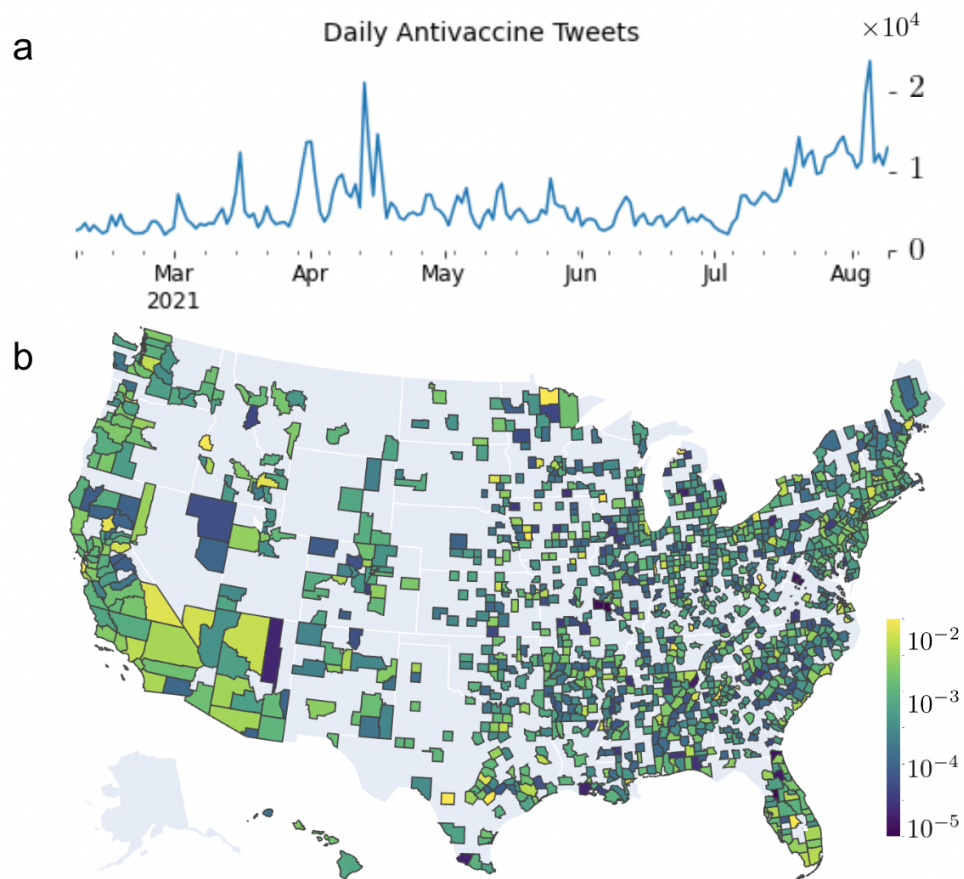


Figure 5.1: **Antivaccine tweets.** (a) Number of geolocated antivaccine tweets observed each day of the observation period. (b) Antivaccine tweets per capita per day, geolocated in each US county during the observation period. Grey coloring denotes counties where we have insufficient geolocated Twitter data.

We analyze the records of cases and vaccinations in US counties between February and August 2021 (see Methods). To connect these data to antivaccine content exposure, we identify and geolocate antivaccine tweets. We use a text classifier (see Methods) to identify COVID-related tweets as “Antivax” or “Other.” About 10% of tweets in our data set can be geolocated to specific US counties, and about 8% are identified as containing antivaccine content. This yields a dataset of 26 million geolocated tweets, 2.2 million of which are identified as containing antivaccine content. Figure 5.1 shows that antivaccine content increased around July 4th, 2021 and was broadly distributed throughout the US with limited geographic clustering; some counties produced orders of magnitude more antivaccine content per capita than others. To measure county-level exposure to antivaccine content, we combine the number of antivaccine tweets produced in each county with a network that captures the spread of this content from one county to another via retweets (see Methods).

We combine this antivaccine content exposure data with COVID case and vaccination data for each county. Next we describe the model and fit its parameters to this data, allowing us to infer the effect of antivaccine tweets on vaccine hesitancy, vaccinations, cases, and deaths.

5.2 Antivaccine Tweets Increase Vaccine Hesitancy

To infer the impact of antivaccine tweets on vaccine hesitancy, we model the COVID epidemic with an SIR-like compartmental model that we call SIRVA (see Methods). In addition to the standard Infected (I) and Recovered (R) compartments, the model has compartments for vaccinated people (V) and divides the Susceptible group (S) into those who are willing (S') and unwilling (A) to be vaccinated, see Figure 5.2a. The key epidemic model parameters are the infection rate β , the recovery rate ρ , the vaccination rate ν , and the rate γ at which people become unwilling to be vaccinated. The latter can be written as $\gamma = \gamma_e E + \gamma_p$, where γ_e is the rate at which people become vaccine-hesitant per unit of exposure to antivaccine content (E) and γ_p is the rate of conversion to vaccine hesitancy due

to other factors. Finally, we express the vaccine-hesitant population at time t as $A = \alpha_t S$, where the vaccine hesitancy ratio α_t changes by rate γ each day and its initial value is an additional parameter α_0 .

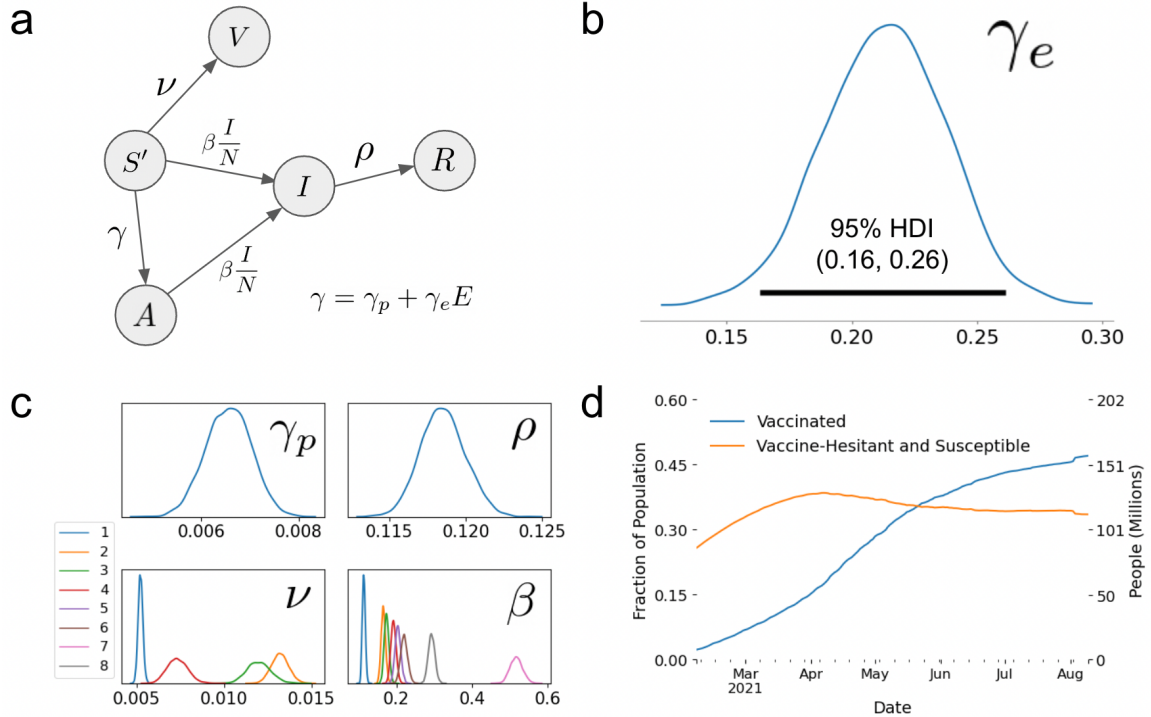


Figure 5.2: **SIRVA model.** (a) Compartmental model diagram (see Methods). Note that $A = \alpha S$ and $S' = (1 - \alpha)S$, where α is the vaccine hesitancy ratio, i.e., the fraction of susceptibles who are unwilling to be vaccinated at time t . E is the magnitude of exposure to antivaccine tweets and γ_e is the rate at which people become vaccine-hesitant due to this exposure. (b) Posterior distribution of γ_e , with 95% high-density interval. The maximum-likelihood value is $\gamma_e \approx 0.21$. (c) Posterior distributions of other global model parameters. Note that there are multiple curves for ν and β because we used different values of these parameters during different time periods to account for changing infectivity and national vaccine availability; these are ordered from earliest to latest. (d) Vaccinated population (V), and population who are both susceptible and vaccine-hesitant (A).

We apply this model to each county and use Bayesian Markov Chain Monte Carlo (MCMC) to infer the posterior distributions of the parameters from the data (see Methods). We are primarily interested in the parameter γ_e , which quantifies the impact of exposure on vaccine hesitancy. Inspecting the posterior distributions of the parameters (see Figure 5.2b,c), we find that γ_e is greater than zero ($p = 0.00022$) with an approximate

magnitude of 0.21 and a 95% credible interval between 0.16 and 0.26. This indicates that increases in exposure to antivaccine content predict future increases in vaccine hesitancy, and subsequent decreases in the vaccine uptake rate.

5.3 Antivaccine Tweets Prevent Vaccinations

The SIRVA model does not specify a direct relationship between exposure and vaccine uptake rates, however we can use causal graphical modeling [88] to assess this relationship (see Methods). We define the Average Treatment Effect (ATE) as the change in vaccinations per exposure to antivaccine tweets, where exposure is expressed in units of antivaccine tweets per capita. We derive an expression for the ATE and find its magnitude to be -757 (see Methods), with a 95% credible interval between -575 and -931 and high confidence that it is strictly less than zero ($p = 0.00022$). This result allows us to infer that the relationship is causal; we consider potential confounding factors in the Discussion. We also compare these results to a simpler linear model relating exposure to vaccine uptake, and find a similar (but stronger) negative relationship; our method improves over the simpler linear model by accounting for additional causal confounders (see Discussion).

Based on the ATE, we estimate that antivaccine tweets induced approximately 750,000 people to refuse COVID vaccinations nationwide (with a 95% credible interval of 572,000–926,000) during the period from February 6th to August 9th, 2021 in the United States (see Methods). This represents only a small fraction of the total number of Americans who were unwilling to be vaccinated and were susceptible to infection (A), which we estimate at approximately 113 million people in August 2021, up from 86 million in February (see Figure 5.2d and Methods).

We can use measurements of vaccine effectiveness to provide a lower bound for the number of COVID cases and deaths that may have resulted from this reduction in vaccination. We estimate that among the 750,000 people who remained unvaccinated as a result of antivaccine tweets, there were about 27,000 COVID cases and 400 COVID-attributable

Table 5.1: **Model comparison.** We compare models based on their expected out-of-sample model performance (ELPD) using leave-one-out (LOO) cross validation. ELPD-LOO closer to zero indicates better fit to the data. The standard errors give a natural scale for comparing the relative accuracy of the models. Metrics were computed on a sampled data set of 400 random counties and 24 evenly spaced dates spanning the observation period from February to August 2021.

Model	ELPD-LOO
SIRVA	$-(1698 \pm 6.7) \times 10^2$
SIRV	$-(1716 \pm 6.6) \times 10^2$

deaths during February-August 2021 that would have been prevented without this additional vaccine hesitancy (see Methods). This represents a lower bound on the total impact because there would have been secondary infections outside the vaccine-hesitant population. Additionally, there would have been more cases and deaths following August 2021, which are not counted here.

To test if accounting for vaccine hesitancy improves model accuracy, we compared the predictions of our SIRVA model against a simpler model that does not include vaccine hesitancy (called SIRV, see Methods) using leave-one-out cross-validation and a Bayesian model fit score (see Methods). The results are shown in Table 5.1. The SIRVA score is more than two standard errors better than that of SIRV, indicating that SIRVA is substantially more accurate at predicting unobserved data points.

5.4 Discussion

The proposed SIRVA model provides us with a novel approach to capture the role of vaccine hesitancy in epidemics. We use this model to measure the effect of antivaccine Twitter content on vaccine uptake during the COVID-19 pandemic in the United States. We find that exposure to antivaccine content on Twitter caused decreased vaccine uptake rates and increased cases and deaths.

Our causal analysis hinges on a few key points. First and most critically, by leveraging the retweet network of COVID-related tweets between US counties, our measure of

exposure to antivaccine content is specific to both the Twitter platform and the particular geographic distribution of Twitter users discussing COVID, allowing us to rule out many possible confounding factors that act through other social networks (e.g. Facebook) or with different geographic distributions. We test this geographic and platform specificity of our measure of exposure in two ways: (i) we tested the correlation of the COVID-related retweet network, W_{ij} , with other social networks (i.e. Meta’s Social Connectedness Index [89]), and found no correlation, and (ii) we tested whether shuffling exposure data by county would destroy the measured relationship between antivaccine tweets and increased vaccine hesitancy, and found that the relationship was null in this case.

In addition to this platform and geographic specificity, we also accounted for other possible causal confounders in our model, including preexisting vaccine hesitancy in each county, nationwide drift in vaccine hesitancy overtime, and possible differential antivaccine content exposure rates based on preexisting vaccine hesitancy. We accounted for preexisting antivaccine sentiments in each county by including an inferred free parameter for the initial antivaccine hesitancy ratio α_0 in each county. We accounted for nationwide drift in vaccine hesitancy over time through an additional parameter γ_p allowing for conversion to vaccine hesitancy without exposure to antivaccine tweets. Additionally, in our causal graphical model we accounted for the tendency of people with higher vaccine hesitancy to be more likely to be exposed to antivaccine content on social media by including an additional causal path from vaccine hesitancy to exposure, and this is reflected in our average treatment effect estimation. Finally, we also tested an alternative SIRVA model in which existing vaccine hesitancy may produce additional vaccine hesitancy within a county, for instance by word-of-mouth spread of antivaccine sentiment (see Methods); this more complex model produced qualitatively similar results as the SIRVA model, ruling out the possibility that word-of-mouth spread can explain the relationships we find.

Although our causal inference analysis accounts for confounding factors that affect the vaccine uptake rate by inducing vaccine hesitancy, we have not accounted for confounders

that might act via vaccine availability. We assume that the processes governing Twitter's social dynamics and those determining vaccine availability are largely independent. Based on this assumption, we believe there is no significant common cause of both vaccine unavailability and exposure to antivaccine tweets. Because effects on vaccine uptake must act either by vaccine hesitancy or vaccine availability, we believe we have accounted for the possible confounders. We therefore conclude that the observed relationship between exposure to antivaccine tweets and reduced vaccine uptake rate is causal.

The accuracy of our method may be affected by data limitations. First, the official CDC data on COVID vaccinations, cases, and deaths contain imputed values and reporting lags. Second, our Twitter user geolocation data has limited coverage of the user population. Third, our observations may not capture the full spectrum of all antivaccine content on Twitter. These limitations may have created sample biases in the measured antivaccine content exposure and vaccine uptake rates.

With these caveats, our analysis estimates that 750,000 people became vaccine hesitant as a result of antivaccine content on Twitter. These are only a small fraction of our model's estimate of 27 million Americans who became vaccine hesitant between February and August 2021. This larger increase is likely due to other sources of antivaccine messaging outside Twitter, including other social media, traditional media, and word-of-mouth interactions. Further work could analyze data from larger platforms, such as Facebook, Instagram, and TikTok, which also carry antivaccination content [90].

This work constitutes a significant contribution to both public health research and social media studies because it establishes a causal link between online content and offline public health outcomes. These conclusions should inform future social media policy and epidemic modeling efforts.

5.5 Methods

5.5.1 Data

We use three main data sources in this work: the CoVaxxy Tweets database, which contains Tweets related to COVID vaccines; CDC records of COVID cases and vaccinations in US counties; and Mønsted and Lehmann’s antivaccine tweets dataset, which supplements our own labeled data for training a tweet classifier. Details are provided below.

The primary data source for this work is the CoVaxxy project [91] from Indiana University’s Observatory on Social Media. CoVaxxy collects tweets related to COVID vaccines and vaccine hesitancy, and geolocates the tweets to US counties when possible. Using this dataset, we can track the online discourse surrounding COVID vaccination in individual US counties. In this study we used data from February 6th to August 9th, 2021, and used only tweets geolocated to US counties. The beginning of this window coincides with vaccines becoming widely available in the US, and the end of the window marks approximately the time when the vaccine uptake rate began to slow substantially. The primary features of this data are the text content, timestamps, and geolocations of the tweets.

The US Center for Disease Control (CDC) collects and publishes data on COVID health outcomes and vaccination in US counties, including cumulative cases, deaths, and vaccinations [92] for each county, for each day. This is the source of the public health metrics in this work. Due to differences in local reporting systems and reporting schedules, some entries in the cumulative counts are imputed by CDC as the last known value until they are updated with new reports from the counties or states. Counties within the state of Texas are excluded from our dataset because the official CDC data on COVID vaccinations does not contain information about Texas counties until after October 22, 2021.

To train our antivaccine tweet classifier, we require a dataset of tweets labeled as “antivaccine” or “other.” We supplement our own labeled data with a similar dataset created by Mønsted and Lehmann [93]. Although this dataset is not specific to COVID antivaccine

sentiment, we found its content to be similar enough to improve the performance of our classifier. We used the Mønsted and Lehmann dataset for training our model but not as part of test data used to evaluate our classifier’s performance.

5.5.2 Antivaccine Tweet Classifier

To track the volume of antivaccine content produced in each county, we built a text classifier that determines if a tweet is expressing antivaccine sentiment. The classifier takes the text of a tweet as input and returns a label, either “antivaccine” or “other.” The classifier is a neural network based on the RoBERTa language model [94], and is trained on a set of manually labeled tweets.

Our labeled training and test data were coded by two human annotators and examined for agreement. Cases where annotators disagreed were discarded. To be labeled as antivaccine, a tweet must express the belief that safe, effective COVID vaccines are bad, ineffective, not actually a vaccine, or harmful (without specific evidence). Most tweets expressing the belief that COVID vaccines are harmful made one of a few common claims, so these common claims were manually checked using reputable fact-checkers (e.g., PolitiFact, FactCheck.org) and CDC publications; in general, the common claims of harm were found to be false and labeled as antivaccine.

The classifier was trained on 4,200 labeled tweets from the CoVaxxy dataset. To increase training data volume and variety, we also added to this training set 2,000 tweets randomly selected from the Mønsted and Lehmann dataset [93], which were labeled by three human annotators; cases where annotators disagreed were discarded. The model was evaluated on a hold-out set of 900 labeled tweets from the CoVaxxy dataset, labeled by the same method and annotators as the training data. We did not use cross-validation because (i) the training data includes Mønsted and Lehmann’s tweets, which do not match 2021 COVID-related tweets, and (ii) our sample of training tweets from the CoVaxxy dataset deliberately included a greater proportion of antivaccine tweets to help with classifier learning. Instead

we produced the test dataset by pure random sampling from the CoVaxxy data to ensure that classification metrics like F_1 are unbiased.

The classifier was evaluated using common classification metrics. It has an accuracy of 0.94 ± 0.02 , an F_1 of 0.74 ± 0.07 , and a Matthews Correlation Coefficient [95] of 0.72 ± 0.08 , where 95% confidence intervals are computed by the bootstrap procedure [96]. This classifier was used to determine the number of antivaccine tweets geolocated in each county on each date. While some individual tweets may be misclassified, we believe the performance of the classifier is adequate to determine population-level trends and relative magnitudes in the prevalence of antivaccine tweets.

5.5.3 Exposure to Antivaccine Tweets

To measure the impact of antivaccine tweets on people’s propensity to get vaccinated, we measure the amount of antivaccine Twitter content to which people are exposed at the county level. Intuitively, if a population is strongly connected to other populations that produce a lot of antivaccine content, then its exposure is high. The per-capita antivaccine exposure rate in county i at time t is defined as

$$E_{i,t} = \frac{1}{N_i} \frac{\sum_j W_{ij} T_{j,t}}{\sum_j W_{ij}} \quad (5.1)$$

where N_i is the population of county i , $T_{j,t}$ is the number of antivaccine tweets in county j during time window t , and W_{ij} is the number of times that COVID-related tweets posted by users in county j were retweeted by users in county i during our observation period.

5.5.4 SIRVA Model

In the SIRVA epidemic model (see Figure 5.2), we assume that vaccinated people (V) cannot become infected and that vaccine-hesitant people ($A = \alpha S$) cannot become vaccinated.

The dynamic equations for each county can be written as:

$$\frac{dS}{dt} = -\beta(I/N)S - \nu S(1 - \alpha) \quad (5.2)$$

$$\frac{dI}{dt} = \beta(I/N)S - \rho I \quad (5.3)$$

$$\frac{dR}{dt} = \rho I \quad (5.4)$$

$$\frac{dV}{dt} = \nu S(1 - \alpha) \quad (5.5)$$

$$\frac{d\alpha}{dt} = \gamma(1 - \alpha) \quad (5.6)$$

where N is the population of the county, S is the number of people who are susceptible to infection, I is the number of people who are currently infected, R is the number of people who have either recovered or died from the infection, V is the number of people who are vaccinated, α is the vaccine hesitancy ratio, γ is the conversion rate to vaccine-hesitancy, and β , ρ , and ν are infection, recovery, and vaccination rate parameters, respectively. To keep the notation simple, we omit the explicit time dependency of various parameters, such as γ .

To infer the latent variable α from the data, we assume that there is an initial vaccine hesitancy ratio, α_0 , and people convert to vaccine hesitancy from the non-vaccine-hesitant portion of the population $(1 - \alpha)$ at rate γ_t . We thus compute α_t as:

$$\alpha_t = \alpha_0 + \int_{t' < t} \frac{d\alpha_{t'}}{dt'} dt' = \alpha_0 + \int_{t' < t} \gamma(1 - \alpha_{t'}) dt'. \quad (5.7)$$

Finally, we break the conversion rate γ into components due to exposure to antivaccine tweets (γ_e) and other factors (γ_p): $\gamma = \gamma_p + \gamma_e E_t$, where E_t is the antivaccine exposure rate in the county at time t (specifically, the exposure over the previous eight days).

The parameter α_0 is inferred for each county. All the other parameters (β , ρ , ν , γ_e , and γ_p) are inferred from the data across all counties. We use Bayesian Markov Chain Monte Carlo (MCMC) sampling to infer their posterior distributions given the data (see details in

following section).

We also define a simpler comparison model, SIRV, which is a special case of SIRVA where $\alpha_t = 0$ for all t . This allows us to test how including vaccine hesitancy and its dynamics impacts the model’s predictive performance.

Another possible model would include a feedback term, $\gamma_a\alpha$, in γ , accounting for the spread of vaccine hesitancy within a county as a social contagion. We also tested a model with this dynamic effect and found similar results to the SIRVA model described here. We dropped this feature of the model for simplicity.

5.5.5 Estimating Parameters

Our goal in this section is to infer the parameters of the SIRVA model from the data. In particular, we want to infer γ_e to understand whether exposure to antivaccine tweets leads additional people to become unwilling to be vaccinated. To estimate the likely range of the model parameters, we use Bayesian inference with MCMC sampling [97]. The goal of Bayesian inference is to find the probability distribution (which we call the posterior distribution) that describes the likely values of a parameter given the data. From the posterior distributions, we can get a mean estimate, the 95% high probability interval (“HDI”), and a p -value for each parameter.

At a high level, Bayesian MCMC inference samples random parameter values proportionally to their likelihood given the data. The MCMC algorithm begins by sampling from a prior distribution, which is defined by specifying plausible ranges of the parameters, and gradually converges to the posterior distribution. We use the standard NUTS algorithm [98] implemented in the NumPyro package [99].

To specify a likelihood function for the SIRVA model given the data, we will define probability distributions for the daily changes in each of our observed variables: cumulative cases ($C = I + R$), vaccinations (V), susceptible individuals ($S = N - (V + C)$), and recovered individuals (R). We assume $R_t \approx C_{t-8}$, based on a typical time from ini-

tial symptoms to non-infectious status of 10 days [100], and a lag time between initial symptoms and a positive test of about 2 days.

We want to define the probability of observing the daily changes in cases, vaccinations, recoveries, and susceptibles (i.e., $x_j = \Delta C, \Delta V, \Delta R, -\Delta S$) given estimates of these data (μ_j) from the SIRVA model equations using the sampled parameters. We define the probability of observing data x_j given estimate μ_j with the negative binomial distribution, which accounts for noise in the data through a concentration parameter ϕ_j :

$$P_{\text{NegBin}}(x_j | \mu_j, \phi_j) = \binom{x_j + \phi_j - 1}{x_j} \left(\frac{\mu_j}{\mu_j + \phi_j} \right)^{x_j} \left(\frac{\phi_j}{\mu_j + \phi_j} \right)^{\phi_j}. \quad (5.8)$$

Each observed variable is distributed according to the negative binomial, for example, $\Delta C \sim P_{\text{NegBin}}(x_C | \mu_C, \phi_C)$. The ϕ parameters are inferred by the MCMC sampler in the same way as other model parameters.

We define the likelihood function of our parameters given the data as the product of these four negative binomial distributions:

$$L = \sum_j \sum_i \log(P_{\text{NegBin}}(x_{ij} | \mu_j, \phi_j)) \quad (5.9)$$

where the index i represents a data point for a particular date and county.

The MCMC sampler also requires us to specify prior distributions for each model parameter we want to infer. For the basic epidemic parameters, we use normal distributions as priors:

$$\rho \sim \text{Norm}(\mu = 0.1, \sigma = 0.3)$$

$$\beta \sim \text{Norm}(\mu = 0.2, \sigma = 0.3)$$

$$\nu \sim \text{Norm}(\mu = 0.0025, \sigma = 0.1).$$

These distributions have high variance relative to the plausible parameter ranges so that

the priors do not strongly influence the final inferred parameter values. For the vaccine-hesitancy parameters, γ_e and γ_p , we choose priors centered at zero. This way, if the posterior is found to be non-zero, we know the prior did not bias that conclusion. For α_0 , we choose a weak prior based on general estimates of vaccine hesitancy in the population [101]. We use these priors:

$$\gamma_e \sim \text{Norm}(\mu = 0, \sigma = 1)$$

$$\gamma_p \sim \text{Norm}(\mu = 0, \sigma = 0.5)$$

$$\alpha_0 \sim \text{Norm}(\mu = 0.2, \sigma = 0.5).$$

Finally, the concentration parameters defined above are given weak Gamma priors, which flexibly capture variance in the data:

$$\phi_j \sim \text{Gamma}(\phi_j \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \phi_j^{\alpha-1}$$

with parameters $\alpha = 1$ and $\beta = 6$.

We set weak upper and lower bounds for all the priors to prevent runtime errors associated with negative or very large parameter values. In particular, the parameters that must be positive (ρ , β , ν , A_0) are restricted to be positive definite with a lower bound of 10^{-15} . Upper bounds are set well above realistic ranges, e.g., 1.0 for A_0 , the initial fraction of the population unwilling to be vaccinated.

In our model, the parameters ν and β can vary over time, following prior work [102, 103]. The ν parameter is allowed to change once every six weeks to account for changing nationwide vaccine availability. The β parameter is allowed to change once every three weeks to account for changing mean reproduction numbers associated with the emergence of new variants (e.g., the Delta variant late in our observation window) and changing public health policies (e.g., the imposition or lifting of lockdowns and mask mandates). The multiple values of these parameters are inferred by the MCMC sampler just like the other

parameters.

Our dataset comprises 1,319 counties and 188 dates for which we have sufficient Twitter and public health data. We use a subsample of this data to perform our inferences. We sample by date to minimize the effect of temporal autocorrelations in the data; specifically, we use every 8th day in the time series data for each county. This interval is chosen for two reasons: (i) it's slightly more than a week, so it smooths over data lags associated with weekly reporting cycles, and (ii) it is the approximate recovery time used to compute R . We also randomly sample by county to limit the number of model parameters (each county i introduces an additional parameter $\alpha_{0,i}$). Our Bayesian MCMC inference tools struggle with numerical stability when the numbers of parameters and data points get too large. We therefore use 400 randomly selected counties and 24 evenly-spaced dates from our data. Although this sampling ultimately reduces the precision of our inferred parameter values, we do not believe it introduces any systematic biases; results are consistent across different random samples of the counties.

5.5.6 Effect Estimations

In this section we first summarize our methods for the average treatment effect (ATE) of exposure E on the vaccine uptake rate $\dot{V} = \frac{dV}{dt}$, and then offer further details on the derivations of the key equation.

We use a causal graphical model (see subsection 5.5.6) of the SIRVA dynamical system (Equation 5.2–Equation 5.7) to derive an expression for the average treatment effect of exposure E on vaccine uptake rate \dot{V} :

$$\text{ATE} = \frac{d}{dE_{t-1}} \mathbb{E}[\dot{V}_t \mid \alpha_{t-1}] \approx -\mathbb{E}[\nu_{t-1} S_{t-1,i} \gamma_e (1 - \alpha_{t-1,i})] \quad (5.10)$$

where the expectation value is taken over times t , counties i , and the posteriors of ν , γ_e , α .

See

Given the ATE from Equation 5.10, we wish to calculate how many vaccinations were prevented in each county and nationwide. Assuming the total change in vaccine uptake rate is relatively small, we can estimate the number of vaccinations prevented each day in each county as $\Delta V_{t,i} \approx (\text{ATE})E_{t-1,i}$, and then use this estimation to calculate the total number of vaccines prevented nationwide as $\Delta V = N \Delta t \mathbb{E}_{i,t}[\frac{\Delta V_{t,i}}{N_i}]$, where N is the total population of the US and Δt is the length of the observation period (see details in following sections).

After calculating the number of people across the whole US who remained unvaccinated as a result of exposure to antivaccine tweets (750,000), we wish to estimate how many cases and deaths may have been prevented among these people if they had been vaccinated instead. Assuming their infection and death rates were typical for unvaccinated people during the February-August 2021 time period (3,870 cases and 57 deaths per 100,000 people, see details in following sections), we estimate that about 29,000 cases and 430 deaths occurred in this population. Of these cases and deaths, approximately 93% of cases and 94% of deaths may have been prevented by vaccination [104, 105], yielding the numbers of COVID cases and deaths attributable to antivaccine tweets, as reported in our results.

In the following subsections, we detail the construction of the causal graphical model (CGM), the derivation of the ATE expression from the CGM, and finally the estimation of case and death rates among the unvaccinated population, which were used in the above computation of total estimated deaths resulting from antivaccine tweets. cccccbgnhrvb

Constructing the Causal Graphical Model

We construct a Causal Graphical Model (CGM) corresponding to the SIRVA model in a few steps, illustrated in Figure 5.3. (i) We consider the derivatives of the key dynamic variables $(\dot{S}, \dot{I}, \dot{R}, \dot{V}, \dot{\alpha})$, and identify the variables on which they functionally depend in the model equations. (ii) We construct a simple bipartite graph from the variables to their

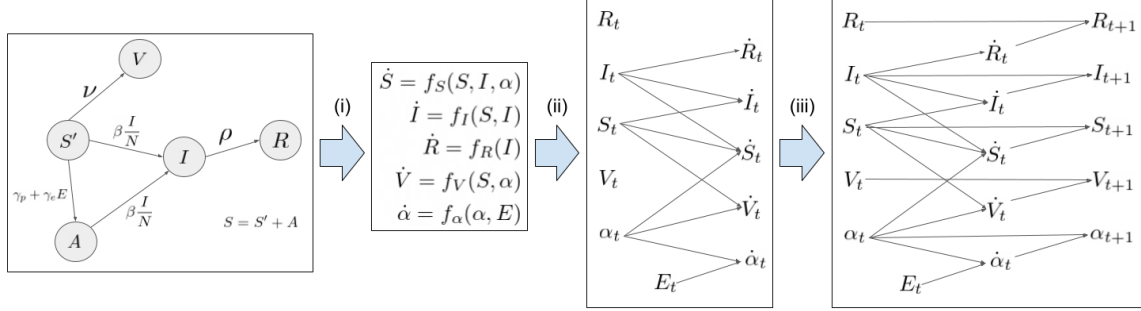


Figure 5.3: Steps to construct the causal graphical model from the SIRVA compartmental model.

derivatives, where each arrow represents a dependency. (iii) We complete a full time step in the CGM by creating additional links from the variables and their derivatives at time t to the variables at time $t + 1$.

Figure 5.4 illustrates the final CGM, which includes additional dotted lines representing a confounding relationship from α to E not explicitly captured by the SIRVA model equations. This potential confounding relationship is included because vaccine-hesitant people may be more likely to engage with antivaccine content online, increasing their exposure. The chain of time steps is also extended backward in time to include the variables at times $t - 1, \dots, 0$, back to the initial conditions (e.g., α_0).

Deriving the ATE Estimand

We leverage the causal graphical model constructed in the previous methods section to find the average treatment effect (ATE) of exposure E on the vaccine uptake rate \dot{V} . The CGM of Figure 5.4 shows a path $E_{t-1} \rightarrow \dot{\alpha}_{t-1} \rightarrow \alpha_t \rightarrow \dot{V}_t$, indicating that there is a causal chain of effects from E to \dot{V} . In addition, there is a causal chain from α_{t-1} to both \dot{V} and E , indicating that α_{t-1} is a confounding variable that must be considered. To account for these relationships, we use do-calculus [88], as implemented in the DoWhy python package, to

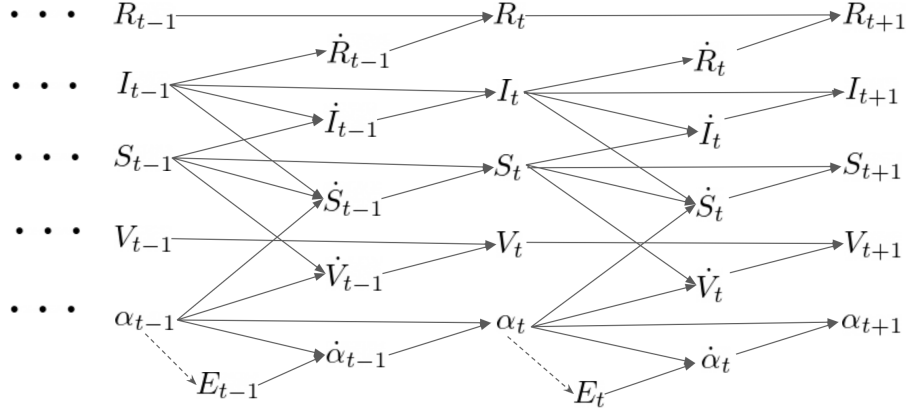


Figure 5.4: Causal graphical model (CGM) corresponding to the SIRVA model. The diagram shows the causal antecedents of the model variables at time t . Here the notation \dot{X} denotes the quantity $\frac{dX}{dt}$, and each arrow represents a causal relationship, where the source variable (on the left) causes the target variable (on the right). The dotted arrows denote an additional possible confounding relationship from α to E in the model; this relationship is assumed to exist in the CGM, and our Average Treatment Effect accounts for its presence. Note that the causal graph is a lattice that can be extended backward in time to an initial state (e.g., α_0).

find the generic form of the ATE:

$$\frac{d}{dE_{t-1}} \left(\mathbb{E}[\dot{V}_t \mid \alpha_{t-1}] \right)$$

where $\mathbb{E}[x]$ denotes the expectation value of x over the data and the posterior samples.

To write this expression in terms of our model variables and parameters, we plug in the expression for \dot{V}_t in terms of E_{t-1} according to our model equations. We can look at the causal path through the CGM ($E_{t-1} \rightarrow \dot{\alpha}_{t-1} \rightarrow \alpha_{t-1} \rightarrow \dot{V}_t$) to find the relevant model equations:

$$\dot{\alpha}_{t-1} = (\gamma_p + \gamma_e E_{t-1})(1 - \alpha_{t-1})$$

$$\alpha_{t-1} \approx \dot{\alpha}_{t-1} \Delta t + \alpha_{t-1}$$

$$\dot{V}_t = \nu S_{t-1}(1 - \alpha_{t-1}).$$

Plugging in these expressions and $\Delta t = 1$ for a 1 day change, we get:

$$\dot{V}_t \approx \nu S_{t-1} (1 - ((\gamma_p + \gamma_e E_{t-1})(1 - \alpha_{t-1}) + \alpha_{t-1})).$$

Next, we need to take the expectation value over our data set and the joint posterior distribution of our parameters. Specifically, we compute an expectation value over our counties i , times t , and posterior samples s , and weight the expectation by county population, N_i .

So the expression for $\mathbb{E}_{t,i,s}[\dot{V}_t]$ expands out as:

$$\begin{aligned} \mathbb{E}_{t,i,s}[\dot{V}_t] &\approx \frac{1}{(n_t - 1)n_s(\sum_i N_i)} \sum_{t=1}^{n_t} \sum_i \sum_s \\ &N_i \nu_s S_{t-1,i} (1 - ((\gamma_{p,s} + \gamma_{e,s} E_{t-1,i})(1 - \alpha_{t-1,i,s}) + \alpha_{t-1,i,s})) \end{aligned}$$

where n_t and n_s are the number of dates and posterior samples, respectively. Note that this quantity depends explicitly on α_{t-1} , so, $\mathbb{E}[\dot{V}_t | \alpha_{t-1}] = \mathbb{E}_{t,i,s}[\dot{V}_t]$. Taking the derivative with respect to E_{t-1} and collapsing the expectation value to a more concise notation, we get our estimand:

$$\begin{aligned} \frac{d}{dE_{t-1}} \left(\mathbb{E}[\dot{V}_t | \alpha_{t-1}] \right) &\approx - \frac{1}{(n_t - 1)n_s(\sum_i N_i)} \sum_{t=1}^{n_t} \sum_i \sum_s N_i \nu_s S_{t-1,i} \gamma_{e,s} (1 - \alpha_{t-1,i,s}) \\ &= - \mathbb{E}_{t,i,s} \left[\nu_s S_{t-1,i} \gamma_{e,s} (1 - \alpha_{t-1,i,s}) \right] \\ &= - \mathbb{E} \left[\nu S_{t-1} \gamma_e (1 - \alpha_{t-1}) \right]. \end{aligned}$$

Assuming the total change in vaccination is small (which is valid in this case), we can estimate the change in vaccine uptake in each county on each date using a linear approximation of the total effect as

$$\Delta V_{t,i} \approx E_{t-1,i} \left(\frac{d}{dE_{t-1}} \left(\mathbb{E}[\dot{V}_t | \alpha_{t-1}] \right) \right)$$

and the total change in vaccine uptake over the whole time period and nation-wide population as

$$\Delta V = (N\Delta t)(\mathbb{E}_{t,i}[\Delta V_{t,i}]) = N\Delta t\left(\frac{1}{n_t(\sum_i N_i)} \sum_{i,t} N_i \Delta V_{t,i}\right)$$

where N is the total population (about 337 million people in the US) and Δt is the length of the observation period in days (192 days from February to August 2021). This results in a figure of about 750,000 vaccines prevented between February and August 2021 in the US.

We also computed this quantity by simply simulating the SIRVA model with the inferred parameters in a counterfactual scenario where the exposure on all counties was set to zero, and found a comparable estimate of 980,000 people; this method was not used for our final results because the ATE-based method is better able to explicitly account for potentially confounding factors.

Estimating Case and Death Rates among the Unvaccinated Population

We want to know the probability that a person who was unvaccinated would have become infected with COVID during the observation period from February to August 2021 in the United States. We can estimate this probability from the case and vaccination data, with the help of an estimate of vaccine effectiveness. Consider the probability that any person would have been infected, $P(C)$. We can break this probability down into two parts: the cases of vaccinated people $P(C, V)$ and the cases of unvaccinated people $P(C, \bar{V})$:

$$P(C) = P(C, V) + P(C, \bar{V}) = P(C | V)P(V) + P(C | \bar{V})P(\bar{V}).$$

We can relate these two parts using the effectiveness of vaccinations at preventing cases, defined as $\lambda_C = 1 - \frac{P(C|V)}{P(C|\bar{V})}$:

$$P(C) = (P(C | \bar{V})(1 - \lambda_C))P(V) + P(C | \bar{V})P(\bar{V}).$$

We can solve for $P(C | \bar{V})$, the probability that an unvaccinated person would become infected:

$$P(C | \bar{V}) = P(C)[P(V)(1 - \lambda_C) + P(\bar{V})]^{-1}.$$

We can get a real value of this quantity by plugging in mean estimates of these probabilities for the cases that occurred over a particular time period (e.g. the previous day) and the number of vaccinations at that point in time:

$$P_t(C | \bar{V}) = \frac{\Delta C_t}{N - C} \left[\frac{V_t}{N - C} (1 - \lambda_C) + \frac{N - C - V_t}{N - C} \right]^{-1}.$$

Summing over time, we find the total risk is:

$$P(C | \bar{V}) = \sum_t \frac{\Delta C_t}{N - C} \left[\frac{V_t}{N - C} (1 - \lambda_C) + \frac{N - C - V_t}{N - C} \right]^{-1}.$$

We can similarly derive the risk of an unvaccinated person dying using the effectiveness of vaccines at preventing deaths, λ_D :

$$P(D | \bar{V}) = \sum_t \frac{\Delta D_t}{N - D} \left[\frac{V_t}{N - D} (1 - \lambda_D) + \frac{N - D - V_t}{N - D} \right]^{-1}.$$

Computing these risks for an unvaccinated person in the United States over the period February-August 2021 using CDC data for cases, deaths, and vaccinations and values of $\lambda_C = 0.93$ and $\lambda_D = 0.94$ [104, 105], we find $P(C | \bar{V}) = 0.0387$ and $P(D | \bar{V}) = 0.00057$. This equates to 3,870 cases and 57 deaths per 100,000 unvaccinated people,

which we use in the main text to estimate the numbers of cases and deaths resulting from exposure to antivaccine content on Twitter.

5.5.7 Model Selection Criteria

To compare the performance of the SIRVA model to simpler models, we use leave-one-out cross validation (LOO), approximated with Pareto-smoothed importance sampling (PSIS). The PSIS-LOO criterion [106] is designed to estimate out-of-sample predictive performance by approximating the expected log pointwise predictive density for a new dataset from the observed dataset without refitting the model to the data. The criterion is robust and efficient and represents the current state of the art for Bayesian model comparison.

CHAPTER 6

CASE STUDY: SOCIAL MEDIA AND LEGISLATIVE AGENDA SETTING

In this chapter, I'll look at a case study in which we could not apply the causal inference method described in Chapter 3. In this context of the larger dissertation, this case represents the more typical, non-causal research that is generally found in social media studies; it is what this dissertation aims to advance beyond. The case attempts to study the effects of Twitter and Facebook content on the legislative discourse in the UK House of Commons. We measure how new topics appear first on social media and then later appear in Commons, suggesting that social media may lead the attention of legislative bodies like the UK House of Commons. However, this study is only able to establish temporal precedence and association (i.e., "Granger causation," see Appendix D), but is not able to deliver more complete causal conclusions.

In retrospect the method described in Chapter 3 may have been applicable to this case. It may have been possible to quantify the exposure of each Member of Parliament (MP) to new topics based on their own personal interaction networks on Twitter, and then assess the resulting changes in the speeches of their respective parties in Commons. However, this would only have been possible for Twitter data, where we had information about the UK Parliamentarians' Twitter accounts (and therefore their exposure levels), and not for Facebook where we had no such data. Additionally, the method described in Chapter 3 was designed for much larger population groups (not just a handful of individuals), so applying it to this problem may have resulted in excessively noisy results due to the limited volume of data associated with each individual MP. Regardless, the method of Chapter 3 had not been fully developed when this project began, and so we were unable to attempt to apply it.

6.1 Introduction

The relationship between the public, the news media, and legislative discourse has been transformed by widespread adoption of social media in the last 15 years. Many people now consume news articles on social media provided by algorithms which choose which articles to display partly based on their popularity. Furthermore, many politicians use social media themselves to get their news and to interact with the public. However, it's unclear to what extent public reactions to news articles guide or predict the attention of legislatures to the content of those articles.

In this chapter, we aim to measure how changes in the discourse of the UK Parliament may be related to public social media activity. In particular, we measure how statements by MPs change after the sharing of a news article on Facebook. To do this, we focus on analyzing the transcripts of special sessions of the UK Parliament in which politicians ask questions of the government to hold the government to account on policy issues. The transcripts of these question sessions offer us a way to track the attention of Parliament as a whole, and allow us to examine the relationship between parliamentary attention and social media posts and news articles.

Understanding how parliament and social media discourse interact may be considered in the context of the broader literature on agenda setting. Most prior works on social media and agenda setting look primarily at intermedia agenda setting, i.e. the extent to which different kinds of media (e.g. Twitter, print news media, TV news) influence one-another [107, 108]. There is evidence that politicians' use of Twitter can influence news media [109] and other online Twitter users [110]. Conversely, the impact that the public may have on politicians via social media is still a relatively under-studied area [111]. Work has found petitions shared online may influence parliament [112] and the public may respond to politicians' online Twitter behavior [110]. This chapter aims to contribute to this under-studied area by assessing the degree to which public online social platforms may act as agenda-setters for

official legislative discourse.

Another large class of related literature studies how and why politicians use social media platforms like Twitter. In general these studies find a pattern of most politicians using Twitter as a means to reach journalists first and foremost, and the public only secondarily [113, 114]. Many politicians also use social media to publicly signal loyalty to the most visible party standard-bearers who align with their policy agendas [115], as this is an effective shorthand for communicating their political stance.

A prior work that is particularly relevant to this study is Barberá *et al.* [110]. This work looked at agenda-setting between the public and members of the US Congress in their interactions on Twitter, and like our work it investigates the direction and magnitude of information flow between these two groups. However, many studies looking at how politicians use Twitter have found that they tend to use it to speak to other politicians and political journalists, rather than interacting with the general public [113, 114]. While the prior work did find that the public may lead politicians' Twitter statements in some cases, it is important to extend this work to look beyond Twitter. In particular, by looking at the official legislative speech of politicians, we may get a better understanding of how the public impacts legislative behavior. Moreover, tracking public attention to content via Facebook data may be more representative than using Twitter data, because Facebook has a larger and more politically diverse user base.

Given this background, we identified four Research Questions (RQs) to study in more detail:

RQ2 Do metrics of public engagement with articles on Facebook predict parliamentary mentions of the content of those articles?

RQ3 Does the content of MP Twitter accounts follow or precede content publicly posted on Facebook? In this respect, do their tweets differ from their Commons speeches?

RQ4 Regarding the previous three research questions, are there any differences between

Labour and Conservative MPs?

To study these research questions we looked for methods that investigate social information transfer between different domains (e.g. between Facebook and Parliament) or groups of people (e.g. news media writers and Members of Parliament). The prior literature contains three broad classes of methods. The first method looks at directional interaction events (e.g. retweets) in an online social network, and to use these as a proxy for information flow between users and groups [116]. The second class of methods tracks the occurrences of defined topics over time, which are treated as time series and analysed using Granger Causality [117, 110] or Transfer Entropy [118, 119] to infer directional transfer between the domains. A third approach [120], which we dub “Content Flow Analysis,” is more general and does not use topics. It simply looks to see how words are transferred between domains over time.

In this study, we are unable to observe direct responses to content from one domain to another (e.g., an MP reading a Facebook post), which we could use to track flow of information. Moreover, the use of discrete topics was too narrow for our more general research question. We therefore choose to use the content flow analysis method, which works around these limitations. In particular, we examine how words used by UK MPs become more or less similar to content of news articles after the articles are shared on Facebook. We interpret this change in similarity to articles posted online as a change in attention toward (or away from) the content discussed in the articles. As well as tracking content flow between Facebook and Parliament, we are also able to track content flow between MPs Twitter posts to and from both Facebook and Parliament. We are able to measure public interactions with the Facebook and Twitter posts. Interactions are actions someone has performed on Facebook or Twitter, including retweets, likes, clicks, views, and comments, among others. We are then able to examine the extent to which changes in word use are related to public interactions to articles.

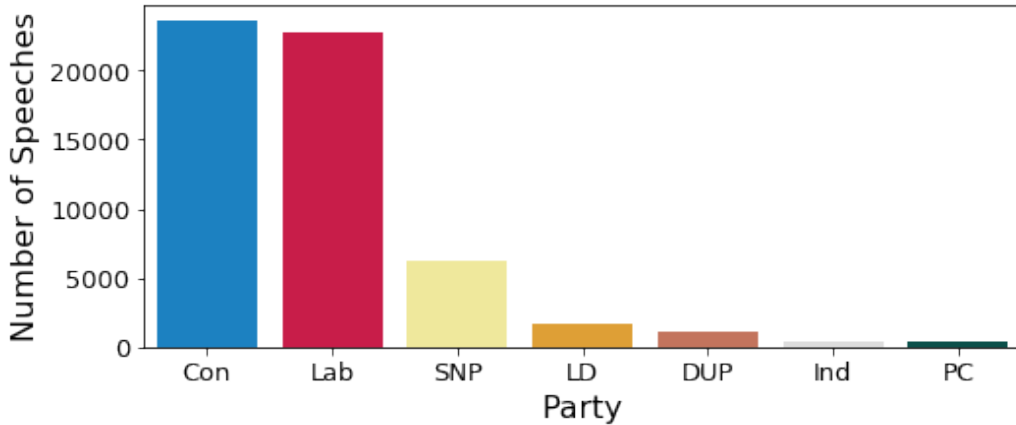


Figure 6.1: Top 7 parties by number of Commons Speeches in our dataset. Abbreviations denote: Con=Conservative, Lab=Labor, SNP=Scottish National Party, LD=Liberal Democrats, DUP=Democratic Unionist Party, Ind=Independent, PC=Plaid Cymru.

6.2 Methods

6.2.1 Datasets

We use three main datasets in this work: (1) records of statements in the UK House of Commons, (2) tweets of UK Members of Parliament, and (3) a dataset about news links shared on Facebook and the interactions with those links (e.g. number of shares, likes, etc).

Hansard data. Hansard is the official report of all parliamentary debates. Since 1909, speeches in the UK parliament have been transcribed and archived. We downloaded the online archive for 2016-2020 from hansard.parliament.uk during January 2021. These were comprised of html files which were parsed to extract speeches. We show a summary of the number of speeches from each party in our dataset in Figure 6.1. In this work, we focus on Questions Sessions in which MPs ask the government questions about current policy issues.

tweets from parliamentarians. We gratefully received a dataset of Tweet IDs sent by UK members of parliament from the Twitter parliamentarian database [121]. The IDs cover tweets between May 21, 2017 until December 24, 2020. The tweets were rehydrated

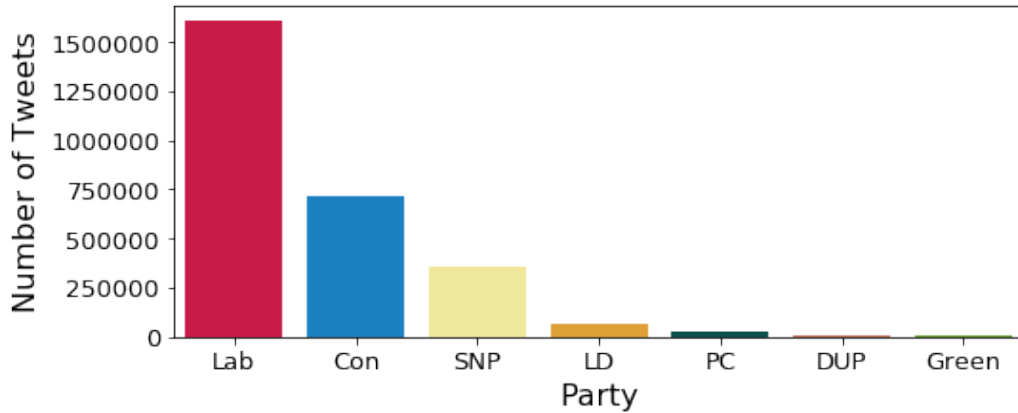


Figure 6.2: Top 7 parties by number of tweets in our dataset. Note that Labour is overrepresented in online discourse, and Conservatives are underrepresented. Abbreviations denote: Con=Conservative, Lab=Labour, SNP=Scottish National Party, LD=Liberal Democrats, DUP=Democratic Unionist Party, PC=Plaid Cymru.

using the twarc2 library during July 2021. We show a summary of the number of tweets from each party in our dataset in Figure 6.2.

URL shares on Facebook. Access to the Condor dataset [122] was coordinated by the Social Science One organization [123] and provided by Facebook. The data we used from the Condor dataset consists of 284,861 URLs that were publicly shared by at least 100 users on the Facebook platform between 1 January, 2017 and 1 August, 2019, and were tagged by Facebook as having been most-shared in the UK. The numbers of shares, clicks, likes and other interactions are aggregated for each URL on a monthly basis, with Gaussian noise added to protect privacy.

To characterize the content of this dataset, we look at which website domains were commonly shared on Facebook during our period of interest. The top 10 are plotted in Figure 6.3. We find that these sources account for the vast majority of URLs shared in our data. These sources are all mainstream UK-based news media. The quantity of other websites, including those that are known to share misinformation, was negligible in our data set.

Finally, a key feature to note about this dataset is that large amounts of Gaussian noise

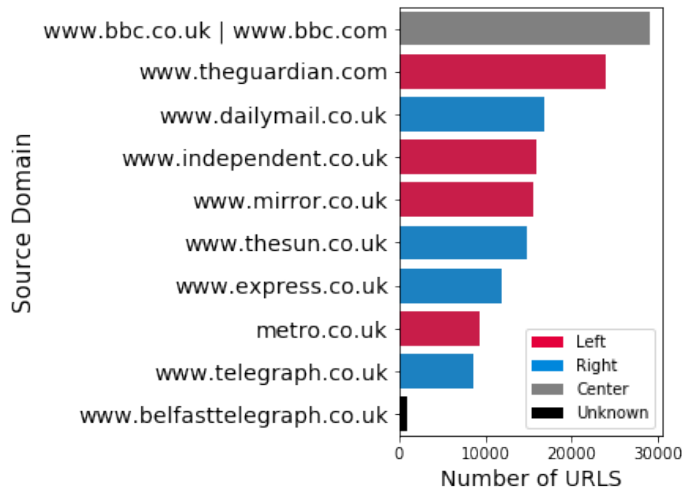


Figure 6.3: Top 10 domains shared on Facebook in our dataset. Note that these domains are overwhelmingly news media from mainstream UK media sources

have been added to the reaction counts for the purposes of differential privacy [122]. This means that the results in our chapter that stem from the reaction counts are necessarily less precise and log-transformed reaction counts display heteroscedasticity. This reduces the statistical power of our regression results, and we believe that many otherwise-significant results are rendered insignificant because of this loss of statistical power.

6.2.2 Measuring Changes in Text Content

The main goal of our analysis is to quantify the extent to which discourse in one domain (e.g. commons speeches) changes after a new statement (the “reference text”) is made in a different domain (e.g. articles on Facebook). To do this, we take the approach of measuring changes in word usage frequency, excluding stop words and Twitter handles. For instance, if the reference text is a news article shared on social media, our method will indicate the extent to which words used by the article are more or less likely to be used in Commons after the news article is published.

The method is illustrated in Figure 6.4. We focus on reference texts that are news articles published on Facebook, tweets from MPs, or statements made by MPs in Commons. For each reference text in a *stimulus* domain, we measure changes in other texts in *respond-*

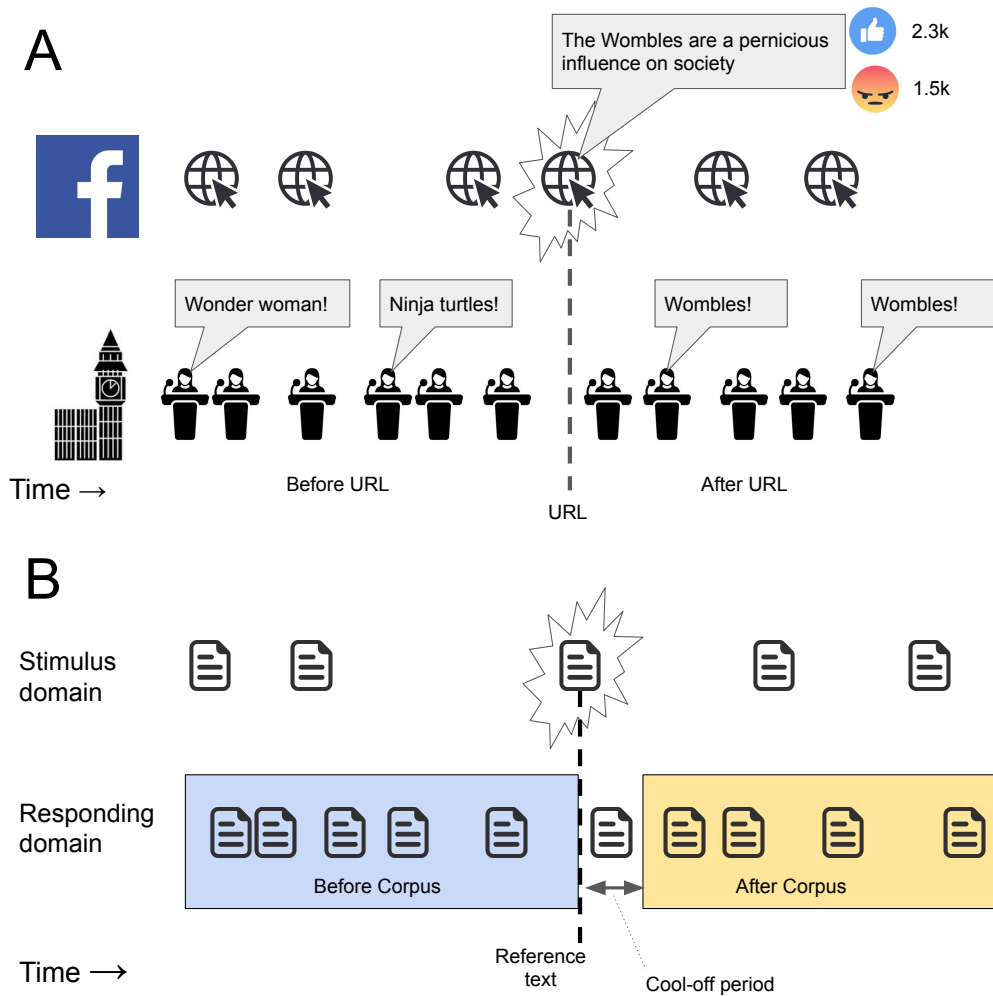


Figure 6.4: Diagram showing our method with a specific example (panel A) of how an Article shared on Facebook might influence Commons Speeches. Note that the news article contains the word Wombles, and then Wombles appears in the commons speeches after the article is shared on Facebook. Panel B shows a more general method that is used to measure content flow between domains.

ing domains. To do this, we select texts in the responding domain to create a *before corpus* and an *after corpus*, split around the publication time of the reference text. In simple terms, our method seeks to quantify the extent to which texts in the after corpus are more similar to the reference texts, compared to texts in the before corpus.

For every text in the responding domain, we compute a similarity measure to the reference text based on the number of words they have in common. The main measure of similarity between texts that we use is the Bray Curtis Similarity [124]. We also applied cosine similarity and found the results are consistent with the Bray Curtis analysis. We ultimately chose Bray-Curtis because it yields more easily interpretable magnitudes. The Bray-Curtis similarity is defined as

$$S_{ij} = \frac{2C_{ij}}{n_i + n_j}$$

where n_i is the number of words in text i ; C_{ij} is the number of words shared in both texts, i.e. $C_{ij} = \sum_k \min(w_{k,i}, w_{k,j})$; and $w_{k,i}$ is the number of times word k occurs in text i .

From these similarity measures on each responding domain text, we then compute a mean *before similarity* from the before corpus and a mean *after similarity* from the after corpus. We call the difference between the before and after similarity values the *content flow* with respect to the reference text. For reference text i and sets of texts A and B , where B is the before corpus and A is the after corpus (see Figure 6.4), we define the content flow from the reference text i as the difference in mean before similarity and after similarity:

$$R(i) = \left(\frac{1}{|A|} \sum_{j \in A} S_{ij} \right) - \left(\frac{1}{|B|} \sum_{j \in B} S_{ij} \right)$$

where $|A|$ denotes the number of texts in corpus A.

Finally, we compute the mean *content flow* over the set of all possible reference texts in the stimulus domain. We define the overall content flow of texts in domain D_1 with respect

to reference texts in domain D_2 as the mean of content flows for each text:

$$R(D_1, D_2) = \frac{1}{|D_1|} \sum_{i \in D_1} R(i) \quad (6.1)$$

where $A \subset D_2$ and $B \subset D_2$. This measures the content flow from the stimulus domain to the response domain.

Cool-off Period When creating the before corpus and after corpus, we only include texts from the two weeks before and two weeks after the split (i.e., the time of the reference text’s publication). We also exclude texts from the first 24 hours after the split. We introduce this “cool-off period” so that the social media interactions (e.g. likes, shares) to the reference text will already have accumulated before we measure their impact on subsequent discourse. For Facebook, it’s well known in industry that the vast majority of interactions take place within the first several hours after posting (variously quoted as “75% within 3hrs”, “90% within 12 hrs”, “a lifespan of 6hrs”, and “80% of total engagement within 24hrs” on industry blogs [125, 126, 127, 128]). For Twitter the timescale is known to be even shorter. We also compute our results using a 0hr and 48hr cool-off period and find they are substantially similar, suggesting that our results are robust to changes in choice of cool-off period.

Resampling Texts Due to differences in lengths of texts, the content flow from reference texts to corpuses, $R(i)$, may not be comparable to one another and are difficult to interpret. For more easily interpretable Bray-Curtis similarity measurements, and to correct for differences in text lengths between domains, we resampled each text to a standard length of 1,000 words. That is, we resampled 1,000 words from each text with replacement, and used these resampled lists of words to compute the Bray-Curtis similarities. A sensitivity analysis of the number of words found similar results with any sufficiently large value (e.g. greater than 1000). This resampling method allows us to interpret a Bray-Curtis similarity of 10^{-4} as “the texts share 1 word in 10,000,” and therefore we can interpret a

content flow of 10^{-4} as “1 in 10,000 words in the responding domain were adopted from the reference text.”

6.2.3 Statistical Analyses

To estimate the overall flow of words between domains, we compute the mean content flow over the set of texts for each category of reference text and response text (see Equation 6.1). We use the bootstrap method to generate error bars and corresponding p -values [96] for these overall mean content flows.

To measure the marginal change of content flow associated with social media interaction metrics, we perform a linear regression from log-transformed reaction counts to the content flows we measured. The reaction counts must be log-transformed before performing the linear regression because their distributions are heavy-tailed (e.g. power-law distributions and log-normals), and the log-transformed distributions are appropriate for standard linear regression. The linear regression is defined in the standard way:

$$\vec{R} = X\vec{\beta} + \vec{\epsilon} \quad (6.2)$$

where \vec{R} is a vector containing the measured content flows for each text, X is the data matrix of the log-transformed reaction counts associated with each text (and a constant column), $\vec{\beta}$ are the regression coefficients associated with each reaction (which we call the *marginal change* associated with the reaction), and $\vec{\epsilon}$ is Gaussian noise associated with each text. We show an illustrative example of these regressions in Figure 6.5.

For the statistically-significant interactions, we present regression coefficients, 95% confidence windows, and p -values. We set a conservative Bonferroni-corrected p -value threshold at $0.05/n$ where n is the number of distinct regressions we performed on the data. In this case $n = 12$ and the threshold is 4.2×10^{-3} . There are 12 separate regressions because we did one for each combination of the two stimulus domains (articles on

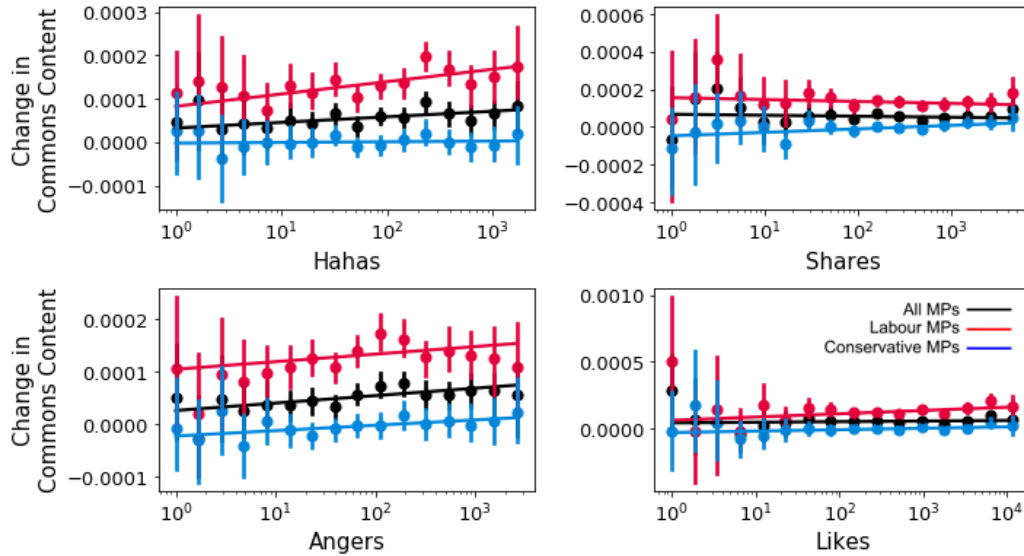


Figure 6.5: Word flow from Facebook to Commons speeches significantly increases with levels of Hahas. We show regression analyses of the marginal changes in content flow of Commons Speeches, associated with Facebook interactions. y-axis is content flow and x-axis is reaction count on a log scale. See Table 6.2 for p-values and slopes of significant effects.

Facebook, MP tweets), response domains (MP tweets, Commons Speeches), and parties (Labour, Conservative, All Parties). These stimulus domains are chosen because they are the only ones that have social media interactions to measure the marginal change, and these response domains were chosen because they are those in which only MPs communicate.

6.3 Results

6.3.1 Overall Content Flow Between Domains

We use content flow (Equation 6.1) to measure word flow between domains. A positive measurement indicates that words from the reference texts are promoted, becoming more common in the other domain. A negative measurement indicates that words from the reference text are suppressed, becoming less common in the other domain. We find evidence of both promotion and suppression of words between the three domains (see Figure 6.6 and Table 6.1).

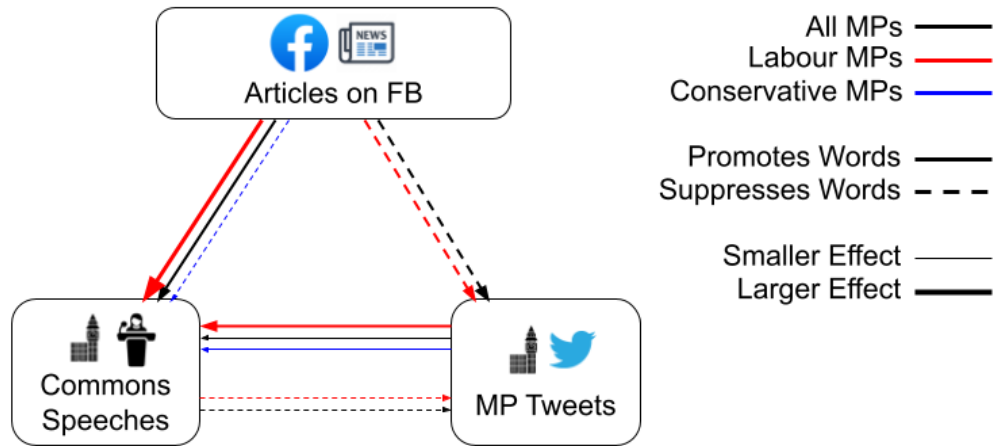


Figure 6.6: Overall word transfer between domains, by party. Lines indicate statistically significant measurements. Colors indicate the party of the MPs for whom we’re measuring a response. Note that lines from Commons to Articles and lines from tweets to Articles are omitted because Articles do not have clear party affiliations, and so these values are undefined. Line thickness is proportional to the log of the measured effect. See Table 6.1 for precise magnitudes, p-values, and 95% confidence intervals.

Party	Stimulus	Response	Mean	+/-	p
All	News Article on FB	Commons Q&A	3.96e-05	4e-06	0.0001
		MP Tweet	-2.48e-05	2e-06	0.0001
	Commons Q&A	Commons Q&A	-1.43e-06	2e-06	0.094
		MP Tweet	-2.4e-06	5e-07	0.0001
	MP Tweet	Commons Q&A	4.97e-06	7e-07	0.0001
		MP Tweet	-2.29e-06	2e-07	0.0001
Lab	News Article on FB	Commons Q&A	0.000111	6e-06	0.0001
		MP Tweet	-3.06e-05	3e-06	0.0001
	Commons Q&A	Commons Q&A	1.04e-05	3e-06	0.0001
		MP Tweet	-2.77e-06	6e-07	0.0001
	MP Tweet	Commons Q&A	1.16e-05	1e-06	0.0001
		MP Tweet	-2.69e-06	3e-07	0.0001
Con	News Article on FB	Commons Q&A	-6.2e-06	4e-06	0.0036
		MP Tweet	-2.92e-07	3e-06	0.43
	Commons Q&A	Commons Q&A	1.4e-05	3e-06	0.0001
		MP Tweet	-1.81e-08	8e-07	0.48
	MP Tweet	Commons Q&A	5.68e-06	8e-07	0.0001
		MP Tweet	-1.51e-06	3e-07	0.0001

Table 6.1: Overall mean content flows between domains. Error bars (at 95% confidence) and p-values were obtained by bootstrapping with 10^4 resamples. FB stands for Facebook.

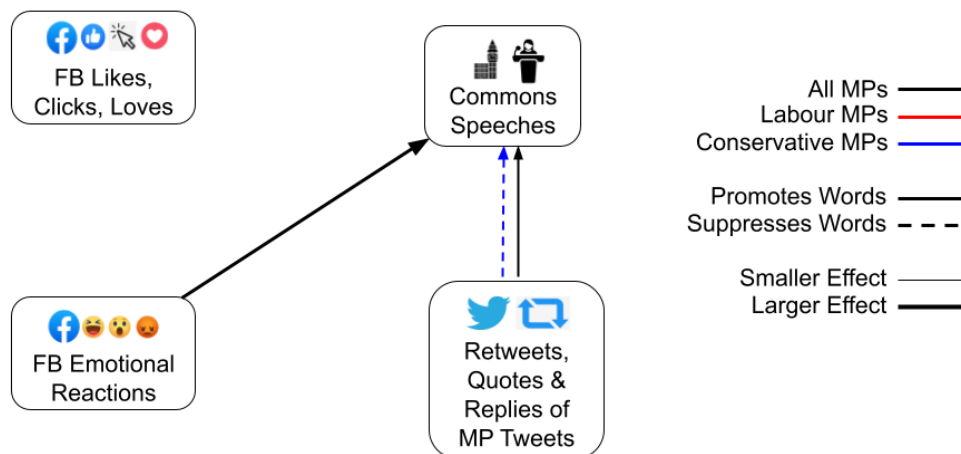


Figure 6.7: Marginal change in word transfer to Commons Speeches associated with different kinds of interactions. Lines indicate statistically significant measurements. Colors indicate the party of the MPs for whom we’re measuring a response. Line thickness is proportional to the log of the measured effect. See Table 6.2 for precise magnitudes, p-values, an 95% confidence intervals.

We find evidence for word flow between domains. Specifically, we find Commons Speeches tend to use more words which previously appeared in Articles on Facebook (see RQ1). However, there are significant differences by Party (RQ4). In general Labour MPs use many more words from Articles on Facebook, while Conservative MPs avoid words used previously in Articles on Facebook. We also find that Commons Speeches use words previously used in MP tweets, and this phenomenon is stronger for Labour. Finally, we find that MP tweets tend to avoid language previously used in the Commons and Articles domains (RQ3); however this effect is not observed for Conservative MPs’ tweets (RQ4).

We also look at which specific words flow from Facebook articles to Commons speeches. To do this, we rank articles by their content flow to Commons speeches, $R(i)$. Selecting the top 10% of these articles, we look at the words in those articles that were use more often relative to other articles. We found these words were generally about the important topics of the period studied: Brexit, the European Union, Theresa May, customs union, etc.

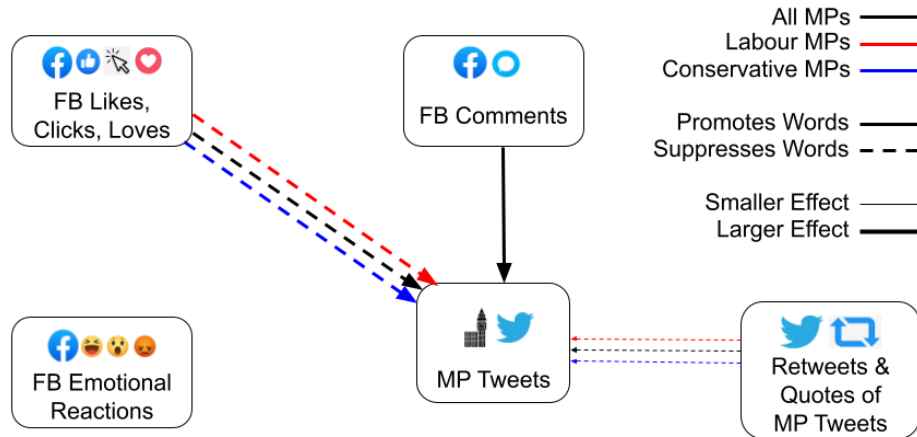


Figure 6.8: Marginal change in word transfer to MP tweets associated with different kinds of interactions. Lines indicate statistically significant measurements. Colors indicate the party of the MPs for whom we’re measuring a response. Solid lines indicate a positive magnitude; dotted lines indicate a negative magnitude. Line thickness is proportional to the log of the measured effect. See Table 6.2 for precise magnitudes, p-values, an 95% confidence intervals.

6.3.2 Marginal Change with Social Media interactions

To study how public engagement with social media posts may have some impact (see RQ2), we looked at the marginal effects of social media interactions on the word flow from a stimulus domain to a responding domain. Specifically, we performed multiple linear regression from the reaction counts in a stimulus domain to the content flows in the responding domain (see Equation 6.2). We present the regression coefficients as a measure of the marginal change to interactions. In general we find associations between levels of social media interactions and changes of the sizes of content flows (see Figure 6.7 for significant changes).

Considering public engagement on Facebook, we find increased Facebook reaction counts are generally associated with increased word flow (RQ2). A multi-variate regression including all interactions finds only the laugh reactions (aka the “Haha” reaction) on Facebook are significant in light of the other interactions (see Table 6.2), after including a conservative Bonferroni correction to the p-value threshold. We estimate that an additional 1.4 words in 100,000 are adopted for a 10-fold increase in Hahas.

Party	Stimulus	Response	Interaction	Marginal Change	p
All	MP Tweet	Commons Q&A	replies	1.5e-06	0.002
		MP Tweet	retweets	-4.6e-07	5.2e-19
	News Article on FB	Commons Q&A	hahas	1.4e-05	0.0034
		MP Tweet	clicks	-2.4e-05	2.3e-07
		MP Tweet	likes	-3.1e-05	3.9e-07
Lab	MP Tweet	MP Tweet	comments	2.2e-05	0.00014
		MP Tweet	retweets	-5.6e-07	2.3e-19
	News Article on FB	MP Tweet	quotes	6.1e-07	0.0019
		MP Tweet	hahas	1.2e-05	0.00025
		MP Tweet	loves	-1.4e-05	0.00029
		MP Tweet	clicks	-1.7e-05	0.0027
		MP Tweet	quotes	-2.4e-06	0.00014
Con	MP Tweet	Commons Q&A	quotes	-2.4e-06	0.00014
		MP Tweet	retweets	-2.9e-07	0.0003
	News Article on FB	MP Tweet	clicks	-3.6e-05	5e-06
		MP Tweet	hahas	-1.4e-05	0.0017

Table 6.2: Marginal change in word transfer between domains associated with logged counts of different kinds of interactions. Linear regression coefficients are the magnitude of the marginal change coefficient. Only statistically-significant interactions are included here.

Considering public engagement with MP tweets, we looked at content flow from Twitter to Commons Speeches (see Figure 6.7). We find in general that increased word flow is associated with the counts of replies to MP tweets. We also found that the number of retweets, replies or quotes of a tweet reduced word flow from those tweets to Conservative MPs Commons Speeches (RQ4).

We also consider MPs tweets, and consider how public engagement metrics on Facebook or Twitter posts impact content flow to the MPs tweets (see Figure 6.8, and RQ2). Using a multi-variate regression to correct for all types of reaction, we find that increased counts of comments on Facebook are associated with increased word transfer, and that increased counts of retweets on Twitter are associated with decreased word transfer to future MP tweets. We find that, in general, MPs tend to Tweet using word that occurred previously in posts that received many comments on Facebook, and avoid words from tweets that had already received many retweets. We also find that Like, Click, and Love interactions are

associated with decreased word flow to Twitter, regardless of whether or not we correct for the effects of other kinds of interactions.

6.4 Discussion & Conclusions

Through applying our content flow analysis method, we have found significant relationships between parliamentary discourse, news media content, and social media engagement metrics. Specifically we find that about 4 words in 100,000 (per news article) transfer from Facebook articles to Commons Speeches. We also find that 10-fold increases in Facebook reaction counts (Likes, Hahas, etc) can increase this word transfer by up to about 35% (or 1.4 words in 100,000), depending on the reaction type. Although 4 words in 100,000 per shared news article may seem small, we believe this is actually a rather large transfer rate because there are approximately 8,000 UK news articles shared to Facebook every two weeks (the length of the observation window used in these measurements).

We also see clear party differences in overall content flow and to a lesser extent in marginal effects from public engagement with social media posts (Figure 6.6, Figure 6.7, Figure 6.8). In general, statements by Labour MPs use more words from news media content and have larger marginal changes in word use associated with additional social media interactions. We note that Labour MPs and their constituents are more active on social media (see Figure 6.2) and a slight bias toward left-leaning articles (shown in Figure 6.3). We hypothesize that this greater activity may explain the party differences we measure.

One might expect Parliament to continue discussing content from one day in subsequent days. This was not detected by our measurements of the Commons (Stimulus) to Commons (Response) in Table 6.1. We do however find that individual parties (Conservative and Labour) maintain specific content over time more consistently. These results seem counter intuitive. We speculate that they may be explained by parties talking past each other and not taking up each others' topics, e.g. by diversion, distraction, or changing the terms of discussion.

We also find that the statements by legislators on Twitter do not have the same relationships to social media as their statements in official legislative proceedings. This finding contextualizes prior work that examined the statements of politicians on Twitter [110]. In particular, it indicates that studies of the dynamics of legislators' tweets may not have direct implications for their official legislative behavior. This distinction between the two political arenas, parliamentary question sessions, and Twitter, warrants further study.

We found that MPs statements on Twitter tend to avoid words previously used in the other domains (see Figure 6.6). We hypothesize that this is due to Twitter's faster response rate. Twitter discussions take place over minutes and hours, while discourse in Commons and news articles takes place over days or weeks. Because of this, Twitter discussions may lead the other two domains in content, and avoid content that has already been discussed previously because interest has already waned.

While we have found clear evidence of directional word flows between domains, it is unclear from our study if this relationship is causal. It is possible that social media engagement metrics are acting as a measurement device which gauges public interest and controversy, and it is the underlying public interest and controversy which entice MPs to pay attention to that content. We are unable to measure the underlying public interest and controversy independently of social media metrics, so we cannot control for these potential confounding variables and establish a clear causal link between social media and parliamentary discourse. Future studies may try to control for other confounding factors, such as other media sources or social processes.

The second major limitation of this work concerns the Facebook dataset. A large amount of Gaussian noise has been added to the reaction counts (e.g. number of likes, shares, etc.) by Facebook for the purpose of differential privacy. We believe this increases the p -values associated with our analysis of marginal changes with interactions, and therefore eliminates many otherwise-significant results. This effect is especially visible in Figure 6.7, where measurements for all MPs are found to be significant (due to larger

data volumes) but measures of individual parties are not found to be significant in some cases. In general, we suspect that the networks in Figure 6.7 and Figure 6.8 are significantly sparser than would be the case without the added noise. Another factor is the very conservative Bonferroni correction that we use, which also contributes to this sparsity of significant results.

Our results are part of a growing body of evidence indicating that official legislative discourse in parliament follows public discussion of news on social media, and may be vulnerable to manipulations of social media. We have seen in recent years a number of high profile cases in which social media is believed to have played a large role in major popular political upheavals, such as the 2016 U.S. Presidential election [129, 130] and the 2016 UK Brexit referendum [131], and there is some evidence that deliberate information campaigns using bots and other inauthentic activity contributed significantly to the social media activity related to these political events [132]. While it is clear that public discourse plays a role in legislation has democratic value, vulnerabilities of social media are a concern. This highlights an ongoing need to protect legislative processes from social media manipulation.

CHAPTER 7

A MECHANISM: SOCIAL MEDIA'S MEMETIC EFFECTS ON INDIVIDUALS

The preceding case study chapters attempt to measure the effects of social media content on public health and on political discourse. I tacitly assume in these case studies that the mechanism for these population-level effects must be that individual people who interact with social media content are affected by it and change their behaviors as a result. In this chapter, I propose some plausible mechanisms by which these effects may be produced, and take the first steps toward demonstrating them.

I hypothesize that when an individual see or interacts with a piece of online content (e.g. a text or image in a social media post), this affects the individual's attention, sentiments, and beliefs (which collectively I call "memetic effects"), and that these memetic effects in turn produce behavioral changes in the individual. We then observe these individual behavior changes as population-scale changes in the propensity to behave in certain ways (e.g. get vaccinated, or speak politically on a given topic). In particular, we suggest that induced changes in attention may produce the behavioral changes seen in Chapter 6 (namely, engaging with specific topics in legislative discourse) and induced changes in sentiment are likely relevant for producing the effects observed in Chapter 5 (namely, refusing vaccination).

This chapter aims to operationalize and demonstrate this hypothesis. Specifically, we will attempt to show that interacting with social media content on Twitter changes people's attention to topics and the sentiments they express toward entities (e.g politicians, places, products, etc.). For instance, interacting with a piece of content about cats may induce you to produce and view more content about cats, and interacting with content that expresses positive sentiment toward cats may induce your expressed sentiments towards cats to become more positive (see Figure 7.1).

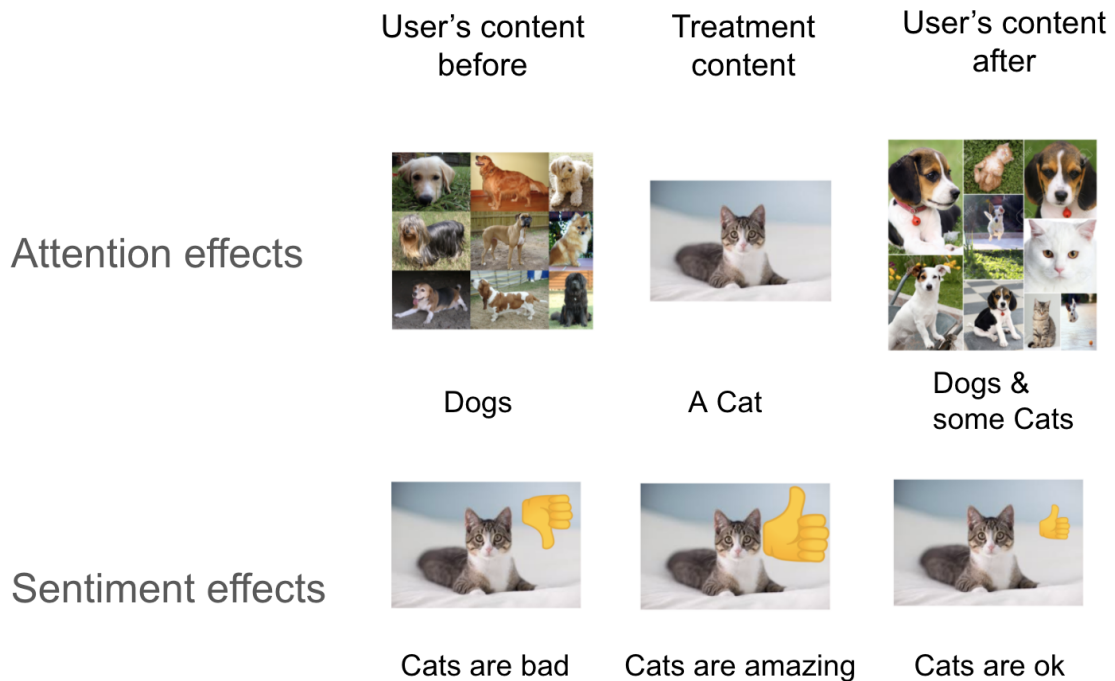


Figure 7.1: An example of attention and sentiment effects on a user after an interaction event with the treatment content.

7.1 Approach

Intuitively, we understand that individual interactions with social media content cannot generally have large, lasting effects on individual users. Nevertheless, it seems obvious that there is a cumulative effect of many interactions, and this must result from the accumulation of many very small interaction effects. In this chapter, we aim to measure these small interaction effects on both attention and sentiment.

In the case of attention, we specifically want to measure if a person becomes more likely to discuss a topic after interacting with a piece of content about that topic. For sentiments, we want to measure if a person changes their sentiments toward entities (e.g politicians, places, products, etc.) after interacting with content about those entities. To do this, we'll track a user's own posts over time, and study how they change in topic and sentiment after the user interacts with another user's post. In this way, we'll use a user's posting behaviors

as an observable proxy for studying the persons’ changes in attention and sentiments.

7.1.1 Causal Considerations

In order to measure the effects of individual social media posts on individual users, we need to first describe the causal graph that we’ll use in our analyses. We want to observe how a user’s online interactions change their online posting behaviors, which is our observable proxy for their actual attention and sentiments. To begin, let’s assume (implausibly) that a user’s online posts depend only on their prior online behaviors and exposures; we’ll return to this assumption shortly. Given this assumption, the causal graph can be written as in Figure 7.2.

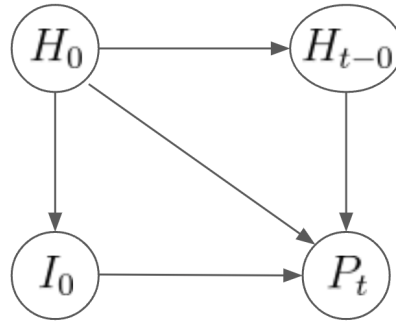


Figure 7.2: A simple causal graph for online influences on a user’s social media posts. Consider the case where a user interacts with a treatment post at time 0, and we would like to estimate the probability that the user posts about the topics or sentiments in the treatment post at time t . Let I_0 be a user’s an interaction event with the treatment post at time $t = 0$, let H_0 be the user’s post and interaction history before time 0, let H_{t-0} denote the user’s post and interaction history after time 0 and before time t , and let P_t be the the user’s probability to post about the topics or sentiments in the treatment post at time t . In this graph, we are interested in the average treatment effect of the relationship $I_0 \rightarrow P_t$.

In this causal graph, we suppose that a users probability to post about a particular topic or with a particular sentiment at time t (P_t) depends on their recent interaction event (I_0) with a post containing that topic or sentiment. Applying do-calculus to this causal graph (see Appendix A for details on do-calculus), we can obtain an average treatment effect (ATE) estimand for the $I_0 \rightarrow P_t$ relationship: $\frac{d}{dI_0} \mathbb{E}[P_t|H_0]$. We would like to estimate the ATE as a function of time t because the strength of the relationship likely diminishes over

time. Fortunately for us, a mathematical model exists for exactly this kind of estimation: Hawkes processes.

Hawkes processes are a type of point process in which prior events may induce greater likelihood of future events (see Appendix C, which explains Hawkes processes in detail). In this case, we believe that prior interaction events (such as viewing or liking a social media post) may induce future post events. By fitting a multivariate Hawkes process to our user’s interaction and post data, we can estimate the increase in likelihood of post events due to an interaction event, as a function of time from the interaction event. In the language of Hawkes processes, this is the “excitation function,” but it also happens to correspond precisely to the ATE estimand in this case. Thus if the causal graph in Figure 7.2 is correct, fitting a Hawkes process to our data yields the ATE as its excitation function.

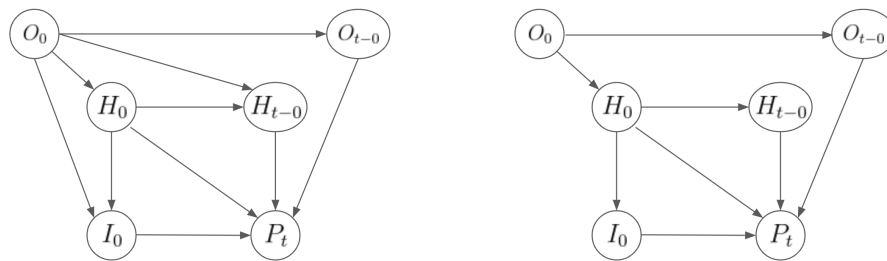


Figure 7.3: A confounded causal graph for online influences on a user’s social media posts. Elements defined as in Figure 7.2, but with additional confounders, O_0 and O_{t-0} , representing the unobserved offline influences on the user’s online behaviors prior to and after the interaction event, respectively. Left: fully confounded causal graph. Right: an unconfounded causal graph resulting from simplifying assumptions.

However, as noted above, the causal graph given in Figure 7.2 is not quite right. It ignores a known unobserved confounder: the offline influences on a user’s post and interaction behaviors. A more complete causal graph is given in the left panel of Figure 7.3. The ATE estimand for the $I_0 \rightarrow P_0$ resulting from this graph requires us to condition on the unobservable variable, O_0 , which represents the offline influences on the user’s online behaviors before time 0. Obviously, we cannot condition on this unobservable variable, and so we cannot make a fully robust causal analysis of this problem.

Instead, let's make a simplifying assumption which may be approximately (but not entirely) true: the user's prior online behaviors (H_0) will contain the information about a user's prior offline influences on online behavior (O_0). Intuitively, this assumption makes some sense. If a user's offline circumstances are influencing their online behavior, surely these influences will be visible, at least in part, in past online behaviors. This assumption amounts to treating H_0 as an instrumental variable, and this allows us to rewrite the causal graph to look like the right panel of Figure 7.3, in which O_{t-0} is absent, and H_0 mediates the influence of O_0 on the other variables. This revised causal graph has the same ATE estimand for the $I_0 \rightarrow P_t$ relationship as the original unconfounded graph, namely $\frac{d}{dI_0} \mathbb{E}[P_t | H_0]$.

For users with sufficient online history, this assumption should approximately hold. To satisfy this condition and use this causal graph in our analyses, we'll restrict our analyses to users with sufficient post and interaction history. Under these conditions, our estimation of the Hawkes process excitation function for the $I_0 \rightarrow P_t$ relationship should *approximately* correspond to the ATE estimand for the $I_0 \rightarrow P_t$ relationship.

Our problem of estimating the effects of online interactions on future online posting behavior is therefore reduced to the problem of estimating the excitation functions of a carefully defined Hawkes process. These online post behaviors can in turn give us insight into the user's changes in topical attention and entity-sentiments, based on the topics they post about and the sentiments they express in posts. Thus we can attempt to estimate the memetic effects of social media content on individual users by fitting a multivariate Hawkes process to our data.

Note, however, that the instrumental variable assumption we've adopted here may not be fully robust, and therefore that this analysis cannot yield ironclad causal conclusions.

7.2 Methods

In the previous section, we outlined a general strategy and causal argument for how to attack this problem of inferring the impact of an online interaction event on a user’s attentions and sentiments. In this section, we’ll define the Hawkes process formally, and describe how we transformed our raw Twitter data of a user’s post and interaction events into a set of event sequences that we can use to fit the Hawkes process.

Twitter Dataset. The primary dataset data we use in this work is the Observatory on Social Media’s archive of the Twitter Decahose [133]. This data is a 10% sample of all public Tweet events and all the Like events attached to those Tweets. We use data from the 30-day period from December 23, 2020 to January 21, 2021 for a random sample of 100,000 active users; more active users are not more likely to be selected in this sampling. This time window was selected because it was the most recent month of data when the project began; we also tested other shorter windows in August 2020 and February 2021 and found similar results. For each user, we collected all available Tweet events and Like events. Tweet events consist of Twitter posts, including features such as the userID, the tweetID, the text content of the Tweet, and the time the Tweet was posted. Like events are records of when a user “favorites” a Tweet, and each record consists of a userID, a TweetID, and the time of the Like event. We also collected all of the “Liked Tweets,” which are the Tweets that the users Liked during the observation period. We processes the users’ Likes, Tweets, and Liked Tweets for into the event sequences to which we will fit the Hawkes process model. In the subsequent sections, we’ll call “Tweet” events the users’ “posts,” and we’ll call Likes and their associated Liked Tweets the users’ “interactions.”

Event Sequences. To produce this collection of event sequences, we randomly draw a sample of users and a single treatment event I_0 from their interaction history to study for each user, and then we produce one event sequence for this user relative to this treatment

event. To obtain the events sequence, we review the users’ post and interactions histories and classify each observed event into one of a several types with respect to the content of the treatment post (see Table 7.1). This results in an event sequence with times and event types. In Section 7.6, we describe the precise process for classifying the users’ post and interaction events into these different event types using natural language processing (NLP) methods.

#	Event Type
1	Liked a Tweet (any tweet)
2	Authored a Tweet (any tweet)
3	Liked a Tweet with a treatment topic
4	Authored a Tweet with a treatment topic
5	Liked a Tweet with a treatment sentiment toward an entity
6	Authored a Tweet with a treatment sentiment toward an entity

Table 7.1: Event types in the multivariate Hawkes analysis. These event types are defined relative to the treatment post that the user interacted with in event I_0 . “Treatment topics” are topics contained in the treatment post, and “treatment sentiments toward an entity” are entity-sentiments expressed in the treatment post.

We obtain one such sequence for each user and their randomly-selected treatment event. This collection of event sequences, with unique random users and random topics and sentiments, is the collection event sequences to which we’ll fit a multivariate Hawkes process. Our inferred excitation functions will be average treatment effects, averaged over users as well as over topics and entity-sentiments. It is therefore very general.

Model. The Hawkes process that we will fit to this data is given as:

$$\lambda_a(t) = \mu_a + \sum_{b \in B} \sum_{i : t_{b,i} < t} \phi_{a,b}(t - t_{b,i}) \quad (7.1)$$

where $\lambda_a(t)$ is the excitation function for events of type a , μ_a is the baseline rate for events of type a , and $\phi_{a,b}$ is the excitation function that determines how events of type b precipitate subsequent events of type a . Following Xu [134], we define our kernel functions ϕ as a very

flexible sum of gaussians with fixed positions and variances, but fitted amplitudes, A :

$$\phi_{a,b}(t) = \sum_j A_{a,b,j} e^{-(t-t_j)^2/(2\sigma^2)} \quad (7.2)$$

where j is an index over the different gaussians in the model, and t_j and σ are fixed parameters, such that the t_j s are evenly spaced over the temporal axis (i.e. $t_j = t_{j-1} + c$ for some constant c) and σ is defined to allow optimal smooth interpolation between the gaussians' centers. This flexible form allows us to smoothly approximate arbitrary shapes of our unknown excitation functions.

The multivariate Hawkes process model is then fit to the data using the method described in Appendix Section C.4 and Xu [134]. This fitting processes results in 36 excitation functions, one for the effects of each of the 6 event types on each other and itself. In the following section we'll detail our results from this analysis. We define $\phi_{a,b}$ such that if $\phi_{a,b} > 0$ then events of type b cause events of type a , where a and b are number indices referring to the event types in Table 7.1.

7.3 Results

In this analysis, we're primarily interested in two of the 36 excitation functions: namely $\phi_{4,3}$ (describing the effects of treatment events on user attention) and $\phi_{6,5}$ (describing the effects of treatment events on user sentiment).

Immediate Effects. In Figure 7.4, we can see that interactions with Tweets do induce a change in attention to the topics in those tweets over a short timer period of minutes to hours. Similarly in Figure 7.5, we can see that interactions with Tweets do induce users to express sentiments more similar to those expressed in the tweet over a short period of minutes to hours.

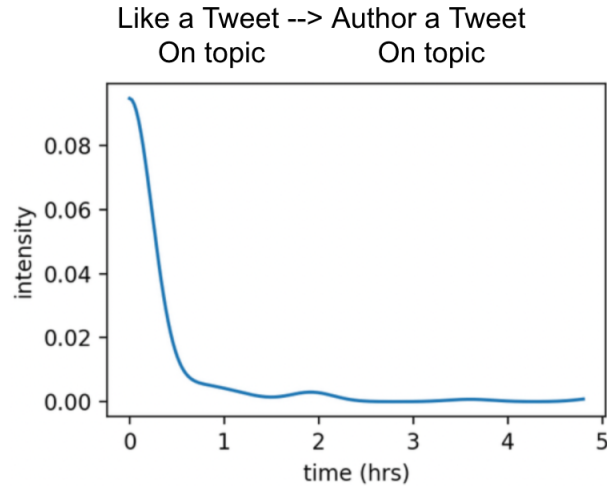


Figure 7.4: Excitation function $\phi_{4,3}$, showing that liking a tweet about a topic makes the user about 10% more likely to author a tweet on that topic in the next 20 minutes, with decreasing influence thereafter.

Multiday Effects. In Figure 7.6, we can see that memetic effects persist with echos over subsequent days at approximately the same time of day as the treatment interaction. This daily temporal cycle is due to the fact that social media users tend to use social media at about the same time of day each day. This figure shows the $\phi_{3,3}$ excitation function, which represents the increased likelihood for a user to like a post containing a topic that was present in the treatment post (which they previously liked). This excitation gave the clearest signal of this recurrence over a multi-day time period, so it is used as an example here to introduce the phenomena of these next day echos.

In Figure 7.7, we can see the same 2.5-day timescale for the $\phi_{4,3}$ and $\phi_{6,5}$ excitation functions. These are considerably noisier and lower resolution than the prior $\phi_{3,3}$ example, due to a much sparser dataset of events in these cases. The $\phi_{4,3}$ excitation function (which describes attention effects) clearly shows next day echos for days 2 and 3, though they are coarsely resolved. The $\phi_{6,5}$ excitation function (which describes sentiment effects) suggests some longer term effects, but it is too noisy to be sure without much more data. Note that in both cases, the next-day echo effects are much smaller than the immediate response at $t < 1\text{hr}$.

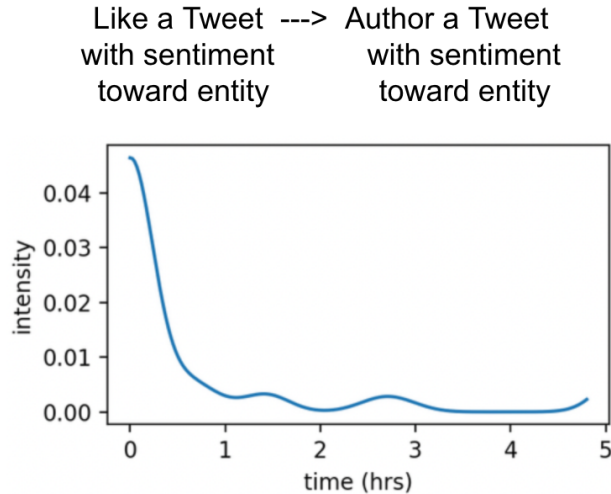


Figure 7.5: Excitation function $\phi_{6,5}$, showing that liking a tweet containing a given sentiment toward a given entity makes the user about 5% more likely to author a tweet containing that entity-sentiment in the next 20 minutes, with decreasing influence thereafter.

7.4 Discussion

In this chapter I have shown that interacting with a single social media post does induce individual people to change their expressed topical attention and sentiments toward entities. I posit that these micro-scale “memetic effects” constitute the key mechanism for the macro-scale behavioral effects of social media on society that we have observed in the case study chapters. We further showed that such memetic effects can persist for at least two days subsequent to the initial interaction event, and may persist longer. With daily social media usage, the observed multi-day effects may self-reinforce over time and persist indefinitely, even if the effects of individual interactions eventually diminish over time to zero. We believe that the cumulative effect of constant engagement with social media content builds up over time, producing durable and significant changes in people’s attention and sentiment. We speculate similar processes may also act on people’s beliefs, and that this could be demonstrated with similar methods (see Future Work below).

This analysis has a number of key limitations. First and most critically, our Twitter dataset was a convenience sample chosen because it was the most-recent month of data

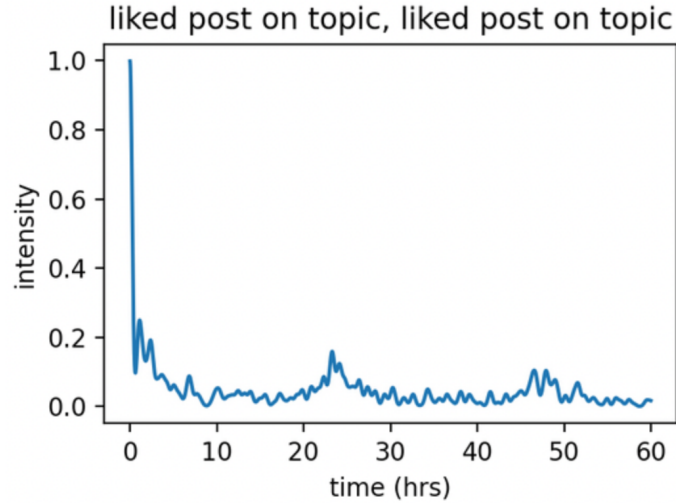


Figure 7.6: Excitation function $\phi_{3,3}$, showing that liking a treatment tweet about a topic makes the user more likely to like another tweet about the same topic for a period of at least two days, with notable "echos" of the initial effect at 24 and 48 hours after the initial treatment event.

when we began the project. Our intention had been to replace this with a larger dataset spanning a wider time period, but the data source (Twitter) revoked all researchers data before the analysis could be fully concluded. The results presented here are therefore not what we had hoped, and in particular Figure 7.7 failed to fully resolve the effect we hoped to show due to inadequate data. Second, our causal analysis hinges on a key assumption that is only approximately true (see subsection 7.1.1). This analysis may be seen as strong evidence for a causal relationship between interaction with social media content and changes in sentiment and attention, but it may not be fully definitive due to our inability to fully validate this assumption.

7.5 Future Work

Further work is warranted in this area. First, I believe the observed effects persist much longer than the 60 hour time window we were able to study here, albeit with diminishing force. Future work should attempt to measure how long these effects persist, and whether it can be shown that effects of individual interactions persist indefinitely (e.g., by showing

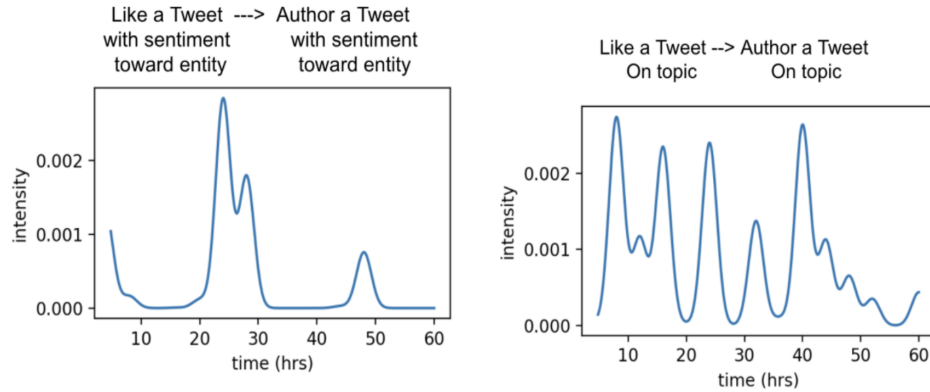


Figure 7.7: Excitation functions $\phi_{4,3}$ (left) and $\phi_{6,5}$ (right) excitation functions over a 2.5 day timescale. The plot has been truncated on the left side, dropping the $t < 1$ hr domain (which has much higher maximum values of ϕ) to allow clearer resolution of the smaller next-day effects.

that they approach an positive-definite asymptote over a few weeks).

Second, I speculate that similar effects operate on people’s beliefs (not just sentiments and attention), and that these belief effects could be measured by observing changes in users stated beliefs using similar methods to those applied to sentiment here. Although this analysis was not possible when I began this work, new methods using large language models (LLMs) now make it possible to extract stated beliefs from social media posts (similar to how we extracted keywords and entities) and to define similarity scores between stated beliefs. Using these tools, researchers could potentially demonstrate that social media posts directly and persistently influence beliefs of individuals, and quantify the magnitude and durability of these effects.

7.6 Methods for Creating Event Sequences

This section documents how we produced the event sequences described in Section 7.2.

At a high level, we examine each post in the user’s history (both the ones they authored and the ones they interacted with), and extract from each post a set of topical keywords and a set of entity-sentiments expressed by the posts author. We then compare each post in the users’ history to the treatment post and determine if they share some topics or sentiments.

This allows us to classify each event in a user’s history into one (or more) of the 6 categories in Table 7.1, resulting in an event sequence for the user, relative to the selected treatment tweet.

To do this, we’ll define a few intermediate tools. First, we need a way to extract topical keywords from very short texts like Tweets (see subsection 7.6.1). Second we need a way to measure an author’s expressed sentiment toward an entity in a text (see subsection 7.6.2). Third, we need to define topical and sentiment similarity scores which can tell us how similar two posts are based on their extracted topical keywords and entity-sentiments. Each of these methods will be explained below.

7.6.1 Topic Extraction

To measure topical attention, we extract topic keywords from each post. This section describes our method for this task.

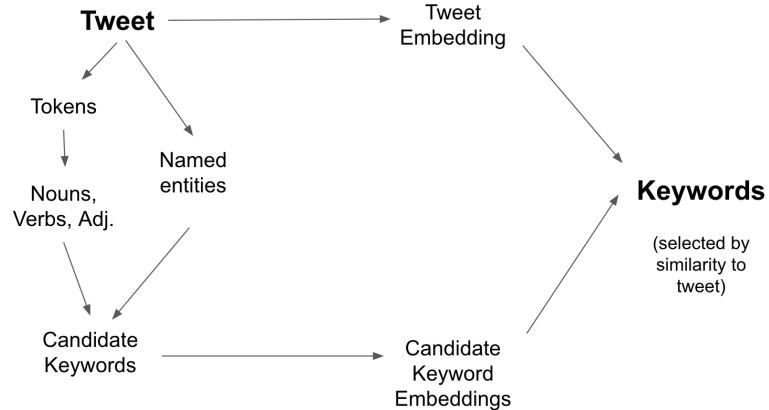


Figure 7.8: Diagram of how keywords are extracted from a Tweet. First candidate keywords are produced by using a part of speech tagger and a named entity recognition model. Next we produce semantic text embeddings of the candidate keywords and the Tweet. Finally, we choose candidate keyword based on semantic similarity to the Tweet.

To extract topical keywords from Tweets, we use a part of speech tagger to identify the nouns, verbs, adjectives, and multi-word named entities (e.g. places, people) in the Tweet. These words are treated as candidate keywords for the Tweet. Next, we select

between 0 and 4 keywords among the candidates by scoring each candidate for its topical and semantic relevance to the overall text of the Tweet. To obtain these relevance scores, we produce text embeddings of the Tweet and the candidate keywords using the pre-trained SentenceBERT language model [135] (see Appendix B), and compute cosine similarities between the embeddings. The keyword candidates that are found to be most similar to the overall text, and more similar than a set threshold, are selected as keywords.

Note that it is possible for the set of keywords of a Tweet to be the empty set if no identifiable keywords are present in the text. This can happen if the text contains only stop-words, or if the text is wholly uninterpretable to the text embedding model (e.g. because it's non-English or because it's gibberish bot-generated content, such as a list of Twitter usernames with no other text).

By manual inspection, we find that our keyword extraction method generally identifies relevant keywords (see Table 7.2). However, it struggles somewhat with slang, multi-word entities (e.g. "united states"), and especially the inclusion of non-English text. The diversity, informality, and brevity of the texts makes keyword extraction especially difficult for Tweets, and this results in some degree of noise.

7.6.2 Measuring Expressed Sentiment Toward Entities

To measure a post's sentiment toward mentioned entities, we first find the subset of topical keywords that are nouns, and then determine the post's sentiment toward those nouns. To do this, we use targeted sentiment analysis, also known as aspect based sentiment analysis or ABSA.

ABSA is an increasingly common task in natural language processing. In this task, a text and a target word are given to the model, and a sentiment score is produced. The score should reflect if the text expresses a positive, negative, or neutral sentiment about the target word. Generally the model expresses this as a number between -1 and +1 or as a three class classifier output (-1, 0, or 1).

text	keywords
a map of the united states https://t.co/lgbP8zOn4E	united,map,states
Joe Biden has a plan to install 500,000 electric vehicle charging cords by 2030, roughly a five-fold increase in the nation's EV infrastructure that could cost more than \$5 billion https://t.co/hSSHULGqP1	charging,plan,biden,cords
@str1ing ANDROID GANG WITH THE WIN!!! 🙌🏽	win,android,gang
first home alone 2 actor to be impeached twice i'm guessing	impeached,guessing,home,actor
roommate: ay bro ur gift is making noise\nthe gift: https://t.co/PJoMXuQQf8	noise,bro,gift,roommate
interesting	interesting
wang yibo's rap for the opening of day day up!!! 🎤\n#WangYibo #王一博 #왕이보 https://t.co/jtekhXlhCh	rap,왕이보,yibo,王一博
gerard way sending the uk government coronavirus information to put the country in yet another lockdown so he doesn't have to play a gazillion shows in milton keynes next year https://t.co/F5LqleD6AY	gerard,coronavirus,lockdown,keynes
@FantasyFBStoner Lol imagine not having the wherewithal to start David Johnson in a smash spot on Championship week. 🤔	smash,david,championship,johnson
capricorn: be real with yourself. is it really worth bringing with you into this year - if it makes you question your own feelings this much? the emotional rollercoaster either ends on your terms or on its own. the changes that come aren't as scary as they seem	rollercoaster,feelings,scary,changes

Table 7.2: Examples of keywords extracted from random Tweets in our data

We employ a BERT-based language model as a text-pair regression model (see Figure 7.9, and Appendix B). The neural network model takes in a pair of short texts, in this case a Tweet and a target entity, and produces a number between -1 and 1 (inclusive). We train this model on an aggregated collection of Tweets (described below) and targets which have been manually assigned sentiment scores (-1, 0, or +1), and test the performance of the regression model on a new set of Tweets that we labeled manually to confirm it generalizes well to our specific data set.

For a given Tweet, we extract a set of entity-sentiments, where the entities are all the extracted keywords that are nouns and the sentiments are determined by our sentiment model. These entity-sentiment pairs are used to characterize the sentiment of the Tweet, just as the keywords are used to characterize the topical content of the Tweet.

Note that it is possible for the set of entity-sentiments of a Tweet to be the empty set if no identifiable entities are present in the text.

We find that our sentiment analysis model produced relatively unbiased estimates of sentiment (see Figure 7.10). However, these estimates were very imprecise. Fortunately,

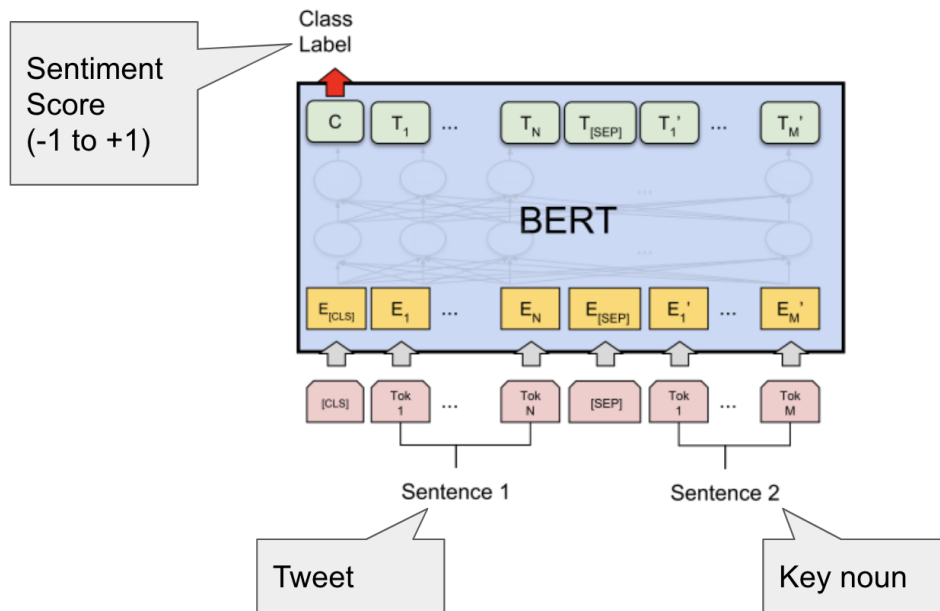


Figure 7.9: Diagram of a BERT language model being used for Sentiment regression. The neural network model takes in two texts and produces a number between -1 and +1.

with sufficiently large volumes of data, we can still use these unbiased estimates to measure an average effect precisely.

Targeted Sentiment Analysis Training Data. Targeted sentiment analysis on tweets is an established task and a number of public datasets are available for it. The datasets consist of three fields: (1) Tweet text, (2) the target word, (3) a sentiment label (-1, 0, or +1). We aggregated multiple public datasets [136, 133, 137] and used the combined English-language Tweets data to train our targeted sentiment regression model. We tested our sentiment model on data from our Twitter Decahose corpus that we labeled ourselves.

7.6.3 Textual Similarity Measures

To define which two posts are sharing the same topics or sentiments, we first compute a topical and sentiment similarity score between the two, and then apply a threshold to determine if they should be classed as expressing the same topics or sentiments. For topics

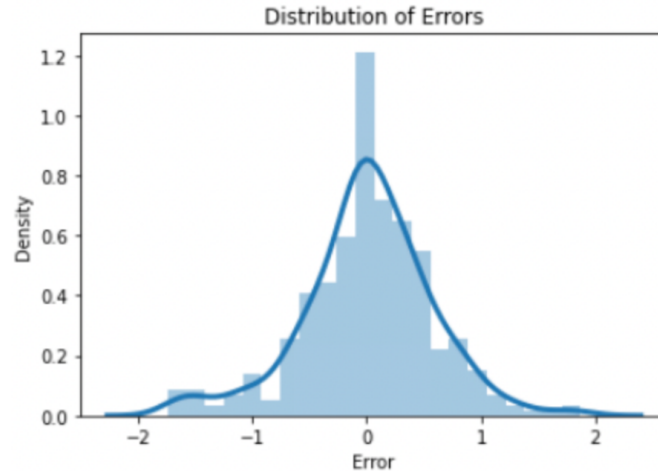


Figure 7.10: Distribution of errors in targeted sentiment regression. Mean = 0.0078; Skewness = -0.28.

we set the threshold at simply being non-zero (i.e. containing at least one topic in common). For sentiment, we set the threshold such that the two posts share a similar sentiment for a majority of the entities mentioned in both posts.

This intermediate step of first computing similarity scores and then thresholding was required because entity-sentiments consist of a word (the entity) and a continuous sentiment value on $[-1, 1]$, so a simple check for common entities would not have been applicable. A distance or similarity measure between the continuous sentiment scores was also required. We applied the same procedure of computing a similarity score and then thresholding to topics primarily so that both sentiment and topic would have the same algorithmic structure for finding matched posts. This has the additional benefit of giving us general similarity measures, which can be used for other non-discretized analyses of these phenomena.

Topical Similarity. To compute a "topical similarity" score between two Tweets, we'll use the Jaccard index between the two sets of Tweet keywords. For Tweet A and Tweet B ,

the topical similarity is given as:

$$S_t(A, B) = \begin{cases} \frac{|W_A \cap W_B|}{|W_A \cup W_B|} & |W_A \cup W_B| > 0 \\ 0 & |W_A \cup W_B| = 0 \end{cases} \quad (7.3)$$

where S_t is the topical similarity, and W_A is the set of keywords for Tweet A . Note that S_t is defined on the $[0, 1]$ interval.

Sentiment Similarity. To compute a "sentiment similarity" score between two Tweets, we'll use the entity-sentiments for each Tweet. For Tweet A and Tweet B , we first find the set of entities (i.e. keywords which are nouns) that they have in common, $E = E_A \cap E_B$. We can then define a sentiment similarity score in terms of the mean L1 distance between entity-similarities:

$$S_s(A, B) = \begin{cases} 1 - \frac{1}{2|E|} \sum_{i \in E} |v_{A,i} - v_{B,i}| & |E| > 0 \\ \text{undefined} & |E| = 0 \end{cases} \quad (7.4)$$

where S_s is the sentiment similarity, $v_{A,i}$ is the entity-sentiment score for Tweet A and entity i , and 2 is the maximum possible sentiment distance for sentiment scores defined on the $[-1, 1]$ interval. Note that S_s is defined on the $[0, 1]$ interval, and that it may frequently be undefined for a pair of tweets because they often do not share any common entities.

Thresholds. For S_t , the similarity threshold required to say that two tweets discuss the same topic is simply that $S_t > 0$, which indicates that they have at least one topic in common. For S_s , the threshold is that $S_s > 0.8$, which indicates that they share very similar sentiments on a majority of the entities mentioned in both tweets. The S_s threshold was subjected to a sensitivity analysis, and we found that results do not vary significantly based on small changes to S_s , but that the results were most clear for values near 0.8.

CHAPTER 8

DISCUSSION AND CONCLUSION

8.1 Main Contributions

This work has established a methodology for measuring online effects of offline social media, defined its limits, demonstrated it through case studies, and proposed and demonstrated mechanisms by which the observed effects may be produced.

The causal method involves three key points: (1) linking online and offline populations via linkage variables such as user geolocation, (2) defining and using a measure of exposure to quantify how offline populations are exposed to online content, and (3) applying causal inference methods such as causal graphical modeling and doubly-robust panel regression to infer the relationship between online exposures and offline outcomes.

The proposed mechanism for the observed population-level effects is individual “memetic effects,” in which individual users’ attention, sentiments, and beliefs are altered by exposure to online content. We demonstrate that online content does affect individual user’s expressed attention and sentiments, and propose how similar methods may be used to show that beliefs are also affected.

8.2 Limitations and Revisions

I’d like to note a significant limitation to the proposed method, and two possible revisions to the larger body of work presented here.

The proposed causal inference method for linking online causes to offline effects hinges ultimately on the methodological assumption that in the absence of treatment differences, similar population groups will have similar outcomes, after controlling for confounders. In essence this amounts to asserting that there are no significant unobserved confounders

and that confounders are the only possible outcome-determinative difference between subjects other than treatment. These are quite strong assumptions, so I feel I need to offer some defense of them here. In general, the assumptions of inference methods, whether in statistics or causal inference, are only ever approximately true. In practice we find that the approximate truth of these assumptions is often adequate to yield useful conclusions, where “useful” here means that if we assume the conclusions are correct and act on them, things go as we expected. This is the pragmatic philosophy behind working science, and it is one I adopt here. That is my only and hopefully adequate defense of adopting this central assumption of all observational causal inference.

Next, I’d like to note two things I would have done differently in this body of work, if time had permitted. First, the antivaccine tweets case study was concluded prior to fully articulating the general causal method in Chapter 3. The proposed method was conceived as a generalization of the lessons learned through that case study, not the other way around. However, with the more thorough development of the method now in hand, I believe that this case study could have used a more systematic process to account for confounders and used simpler effect estimation methods than were ultimately employed. Despite these deficiencies of organization and parsimony, the case study does nevertheless appropriately control for confounders and estimate effect magnitudes reasonably, and I stand by its results.

Second, the study of proposed mechanisms was left in a state of partial completion, due in part to Twitter’s data use agreement changing which limited access to new data. I would like to have completed this work, in particular by demonstrating longer term cumulative effects and effects on beliefs. Nevertheless, the body of evidence for these proposed mechanisms remains compelling because it does demonstrate measurable changes to sentiment and attention after only a single interaction event. The precision reflected in this measurement of such small individual effects on social media users is, to my knowledge, unmatched in prior work, and so I count it as a significant contribution, despite its relative

incompleteness.

8.3 Future Work

I would like to indicate three future directions of work that would follow naturally after this body of work.

First, I would like to explore additional use cases of the proposed causal method, particularly to improve and refine prior works studying the influence of online hate speech on offline hate crimes by employing my novel measure of online content exposure and improved control models for baseline predictions of hate crime events.

Second, I suggest in Chapter 3 that the proposed causal method may be applicable in cases where the existing understanding of offline phenomena is encapsulated in arbitrary blackbox simulations of the offline phenomena, rather than in explicit statistical relationships between observable variables. This idea is so far underdeveloped and imprecise, and I believe future work could offer more specificity, opening up this method to a larger class of use cases.

Third and most ambitiously, I believe the notion of “memetic effects” that I have articulated here warrants further study. For starters, these effects should be better quantified in their magnitude, duration, and durability, and measurements of changes in stated beliefs should also be conducted. Additionally, I believe this mode of measuring memetic objects like attention, sentiments, and belief of individuals could be applied to studies of other related memetic phenomena, such as how specific kinds of ideas and sentiments interact and evolve in a memetic ecosystem such as online social media. Beginning to model and understand the internal dynamics of such phenomena could be a new avenue of research with wide applications for designing interventions in domains such as public health communications and extremism prevention. However, the course of such a research program would likely need to be decades long and span a broader research community.

In sum, there’s much more to do in the domain of measuring offline effects of online

social media. I hope that the work and methods presented here may offer an organized starting point to future studies in this field.

APPENDIX A

CAUSAL GRAPHICAL MODELING

Causal Graphical Models (CGMs) are used in causal studies to depict causal relationships among variables. These models use directed acyclic graphs (DAGs) where nodes represent variables and edges indicate direct causal relationships. By providing a graphical representation of the underlying causal structure, these models aid in understanding the relationships between different variables, predicting the outcomes of interventions, and guiding data analysis strategies.

A.1 Causal Graphs

In causal graphs we represent the relationships between causes and effects as a directed graph, where nodes in the graph represent variables or factors or events, and directed edges represent a causal relationship pointing from cause to effect. For instance, in the relationship $X \rightarrow Y$ denotes that "X causes Y" (see Table A.1). In causal graphs, we may also commonly refer to proximate relationships between three variables (known as "trails"). The three kinds of trails are "mediators" ($X \rightarrow M \rightarrow Y$), "forks" ($X \leftarrow F \rightarrow Y$), and "colliders" ($X \rightarrow C \leftarrow Y$). In causal graphs, causation is transitive, such that if the causal relationship between X and Y is mediated by M , we may still say that X causes Y .

In general, a causal graph for a particular causal inference must contain all the significant influences on both the treatment and outcome variables. Constructing causal graphs is, at present, a primarily manual task which involves conferring with subject matter experts and arriving at a consensus. The validity of inferences will depend on the validity of the assumed causal graph. However, because this graph is clearly documented and easily interpretable, the causal assumptions are easily checked in peer review, rendering causal graphical modeling more easily verifiable than some other causal inference frameworks.


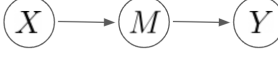
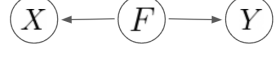
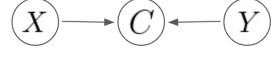
Concept	Definition	Diagram
Causal Influence	X causes Y	
Mediators	M mediates the causal influence of X on Y	
Forks	F causes both X and Y	
Colliders	X and Y both causally influence C	

Table A.1: Key Concepts in Causal Graphs

A.2 Identifiability

A causal relationship in a give causal graph is said to be “identifiable” in CGM if it is possible to estimate its average treatment effect solely from the structure of the causal graph and the observed joint distribution of the variables in the graph. In short, identifiability is what enables purely observational causal studies in CGM.

A.2.1 Conditions for Identifiability

A causal relationship is identifiable under the following conditions:

1. **Absence of Confounding (aka Exchangeability):** There must be no confounding causal paths between the treatment and outcome in the causal graph, or all confounding paths must be blocked by conditioning on observed variables in the graph. A confounding path is defined as a backdoor path (see next section for precise technical definition).

2. **Consistency:** The observed outcome under any observed treatment level must be equivalent to the potential outcome at that treatment level, mathematically represented as: $P(Y(x)) = P(Y|X = x)$, where $Y(x)$ denotes the potential outcome when treatment X is set to level x .
3. **Positivity:** Every level of the treatment has a non-zero probability of occurring, ensuring that there is sufficient data to estimate the causal effect. Formally, $P(X = x|Z = z) > 0$, for every level x of the treatment variable X and every level z of the covariate Z .

In the following subsections, we'll define specific three cases in which these conditions can be satisfied. First, we need to define a couple formal concepts in CGM: d-separation and backdoor paths.

A.2.2 Definitions of Key Concepts in CGM Identifiability

d-Separation. d-Separation (d-Sep) has to do with the conditional independence of sets of variables. Two sets of variables \mathbf{X} and \mathbf{Y} are said to be “d-separated” if they are conditionally independent when conditioned on a third set of variables \mathbf{Z} . Formally, two sets of nodes \mathbf{X} and \mathbf{Y} in a DAG are said to be d-separated by a set of nodes \mathbf{Z} if there is no active trail between any node in \mathbf{X} and any node in \mathbf{Y} given \mathbf{Z} . A trail is active given \mathbf{Z} if: 1. Every non-collider on the trail is not in \mathbf{Z} , and 2. Every collider on the trail has a descendant in \mathbf{Z} . In general, if the treatment and outcome nodes are d-separated given a set of observed variables, the causal relationship will be identifiable.

Backdoor Path. Intuitively, a backdoor path is a generalization of the concept of a fork. It is an *undirected* path through the causal graph that contains a node which is a causal ancestor of both the treatment and outcome variables. It is called a “backdoor” path because it does not lie along any of the directed paths from treatment X to outcome Y , but rather comes in from the “backdoor” upstream of X . Consider the following formal definition.

Let $G = V, E$ be a DAG with nodes V and edges E . A directed path is a sequence of nodes $v_i \in V$ such that for each subsequent node in the sequence there exists an edge $e_{i,i+1} \in E$. A directed path p in G is a backdoor path from X to Y if: 1. the first directed edge on path p is of the form $u \rightarrow X$, and 2. Y is a descendant of the last vertex in the path p , i.e., there exists a directed path from the last vertex of p to Y .

Blocking Backdoor Paths. To obtain unconfounded estimates of causal effects, it is often necessary to "block" the influence of backdoor paths on a causal relationship. Blocking a backdoor path means conditioning on a set of variables such that the treatment and outcome are conditionally independent, and no confounding bias is transmitted through the path. A path is blocked by conditioning on: 1. Any non-collider variable on the path, or 2. The descendant of any collider variable on the path.

A.2.3 Identifiability Cases

Backdoor Criterion. The Backdoor Criterion is the simplest of the cases in which a causal relationship is identifiable. A causal effect is identifiable by the Backdoor Criterion if there exists a set of observed variables Z such that:

1. No variable in Z is a descendant of the treatment variable.
2. The set Z d-separates the treatment and outcome variables, blocking all backdoor paths between them and rendering them conditionally independent given Z .

Adjusting for or conditioning on the variables in set Z allows for the estimation of causal effects from observational data by isolating the association between treatment and outcome from confounding influences.

Front-door Criterion. The Front-door Criterion provides an alternative approach for identifying causal effects when the Backdoor Criterion cannot be satisfied, especially when

there is an unobserved confounder between the treatment and outcome. It also utilizes d-separation to determine the conditions under which causal effects are identifiable. A causal effect is identifiable via the Front-door Criterion if there exists a mediator M of the effect such that

1. M intercepts all directed paths from the treatment to the outcome
2. All backdoor paths from the treatment to M are blocked, implying that there is no confounder between the treatment and M .
3. All backdoor paths from the outcome to M are blocked, given the treatment.

If these conditions are satisfied, the Front-door Criterion allows for the decomposition of the causal effect into a sequence of identifiable components, enabling the estimation of the causal effect even in the presence of unobserved confounding between the treatment and outcome variables.

Instrumental Variable Criterion. The Instrumental Variable (IV) Criterion is another common approach used in CGM. It is particularly effective when treatment and outcome are confounded by unobserved variables. An instrumental variable is one that is associated with the treatment variable but is conditionally independent of the outcome variable given the unobserved confounders. Mathematically, a variable Z is an instrumental variable if:

1. Z is associated with the treatment variable X , i.e., $P(X|Z) \neq P(X)$.
2. Z is independent of the unobserved confounders U , i.e., $P(U|Z) = P(U)$.
3. Z affects the outcome Y only through its effect on X .

When a valid instrumental variable is identified, condition on it isolate the causal effect of the treatment on the outcome by eliminating the confounding due to unobserved variables.

Other Identifiability Cases. The cases described above are the most common identifiability cases in CGM. However, other cases exist which satisfy the identifiable conditions, such as the Swigler Criterion. These will not be detailed in this concise overview of CGM.

A.3 Effect Estimation

Once a causal effect is deemed identifiable, the next step is to estimate its magnitude from the data. The estimation process often relies on the structure of the causal graph, the observed data distribution, and a mathematical formalism called do-calculus.

A.3.1 Do-Calculus Axioms

Do-calculus is a mathematical framework to compute the causal effects of interventions from a known causal graph and a joint distribution over its variables [88]. The essential idea is to introduce a “do-operator,” to represent interventions in a probabilistic causal system. For instance, with $\text{do}(X = x)$ (or simply $\text{do}(x)$), we imagine that we set the random variable X to the value x by some exogenous intervention, thereby cutting all the causal paths flowing into X in the causal graph. This has cutting of the causal paths has predictable effects on the joint distribution of the set of variables in the causal graph, and the axioms of do-calculus articulate these effects. By defining the do-operator through these axioms, we can create a way to imagine hypothetical interventions in causal graphs. For causal graph G and sets of variables X , Y , W , and Z in the graph, the axioms of do-calculus are given as follows [138]:

1. **Insertion/deletion of observations:** If Y and Z are d-separated by $X \cup W$ in G^* (the graph obtained from G by removing all arrows pointing into variables in X), then:

$$P(Y|\text{do}(X), Z, W) = P(Y|\text{do}(X), W)$$

2. **Action/observation exchange:** If Y and Z are d-separated by $X \cup W$ in G^\dagger (the

graph obtained from G by removing all arrows pointing into variables in X and all arrows pointing out of variables in Z), then:

$$P(Y|\text{do}(X), \text{do}(Z), W) = P(Y|\text{do}(X), W)$$

3. **Insertion/deletion of actions:** If Y and Z -subscript are d-separated by $X \cup W$ in G^* (the graph obtained from G by removing all arrows pointing into variables in X), then:

$$P(Y|\text{do}(X), \text{do}(Z), W) = P(Y|\text{do}(X), W)$$

These rules allow us to transform expressions involving interventions into expressions that can be computed directly from observational data about the variables. This formalism is the crux of CGM's power for determining identification criteria and estimating causal relationships from observational data, because it allows us to determine which variables must be observed and conditioned on to measure the causal magnitude of a particular relationship in a given causal graph.

A.3.2 Average Treatment Effect (ATE) Estimates

The Average Treatment Effect (ATE) quantifies the expected difference in outcomes due to a treatment or intervention across the entire population. For instance, for a discrete treatment with no confounding variables the ATE is given as

$$\text{ATE} = \mathbb{E}[Y|\text{do}(X = 1)] - \mathbb{E}[Y|\text{do}(X = 0)]$$

where $Y|\text{do}(X = 1)$ represents the outcome under treatment conditions, and $Y|\text{do}(X = 0)$ represents the outcome under control conditions. When the ATE is identifiable, per the identifiability conditions described above, the do-calculus axioms can always be used to

reduce the do-notation expression to a set of simple conditional probabilities expressions in terms of the observed data.

This computation of ATE expressions can be very complicated in complex graphs and difficult identifiability situations. However in practice these expressions can be obtained using modern causal inference libraries, such as the python package DoWhy [48], which implement do-calculus. DoWhy can be used to determine identifiability and to compute the ATE expressions. Additionally, DoWhy lists the key assumptions associated with each ATE expression (such as no unobserved confounding in the case of the Backdoor Criterion). See Figure A.1.

```

from dowhy import CausalModel
import networkx as nx

CGM = nx.DiGraph()
CGM.add_edges_from([('X', 'Y'), ('W', 'Y'), ('W', 'X')])

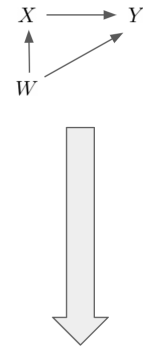
data = pd.DataFrame(columns=['X', 'Y', 'W'])

model = CausalModel(
    data = data,
    treatment='X',
    outcome='Y',
    #convert graph to a GML string, and pass to model
    graph=''.join(list(nx.generate_gml(CGM))) |
)

identified_estimand = model.identify_effect()
print(identified_estimand)

### Estimand : 1
Estimand name: backdoor
Estimand expression:
d
——(Expectation(Y|W))
d[X]
Estimand assumption 1, Unconfoundedness: If U→{X} and U→Y then P(Y|X,W,U) = P(Y|X,W)

```



$$ATE_{X \rightarrow Y} = \frac{d}{dX} \mathbb{E}[Y|W]$$

Figure A.1: A demonstration of using DoWhy to compute average treatment effect (ATE) estimands from causal graphs.

A.4 CGM and Causal Loops

Causal graphical modeling makes a key assumption that somewhat complicates its applicability to complex systems which evolve over time: it assumes that the causal graph is a DAG, i.e., that there are no feedback loops in the system. Fortunately, a relatively simple approach exists to address this limitation and apply CGM to systems with feedback loops. Consider a simple causal graph with a feedback loop. We can simply “unroll” the

graph over time, so that the variables of the prior timestep causally influence the same set of variables at the subsequent timestep, in this way producing a DAG from a non-DAG graph (see Figure A.2). This results in a repeating lattice over time, where each discrete time step is identical, and the “weights” of the causal graph are shared across timesteps. In particular, if we measure the ATE of edge $X_t \rightarrow Y_t$, it is definitionally identical to the ATE of $X_{t-dt} \rightarrow Y_{t-dt}$. When one has time series data (as is generally required for these causal inference methods), this form of the causal graph is easy to utilize.

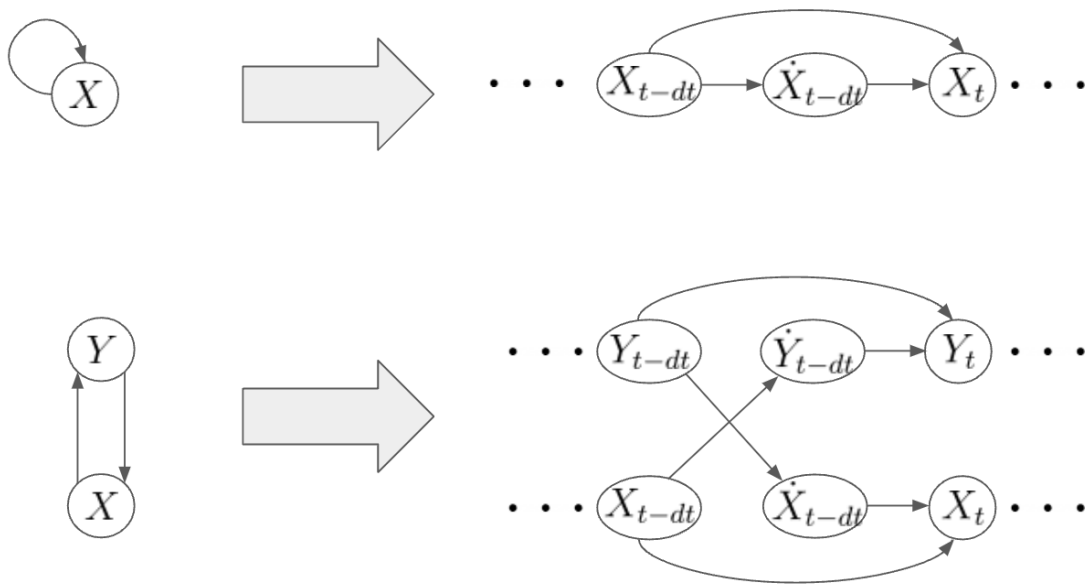


Figure A.2: A diagram of how to “unroll” causal graphs over discretized time, to produce a causal DAG out of a non-DAG causal graph. Note that $\dot{X} = \frac{dX}{dt}$, and dt is a small amount of time.

APPENDIX B

TRANSFORMER LANGUAGE MODELS

In December 2017, Google researchers introduced a class of neural network models called “transformers” [139], which take in an ordered sequence of numbers (generally encoded words) and produces an ordered sequence of numbers (again, often decoded as words). Transformer models leverage a neural network architecture called “cross attention” in which the neural network can adjust its own weights and biases for a given element in the sequence based on the context surrounding the element in the rest of the sequence. In language applications, this allows the network to learn complex grammars and use context across the entire length of the input text.

In May 2019, these researchers debuted BERT [140], a transformer model pretrained on large volumes of English text, which is capable of many different text-related tasks. A followup paper by Facebook produced RoBERTa[94], a version of BERT which performs better on web text such as social media. These language models, especially RoBERTa, enable the work in Chapter 5 and Chapter 7, which would not previously have been possible. In this appendix, I’ll briefly explain how I apply transformer models to do text classification, text regression, and text embeddings.

Transformers for Classification and Regression. BERT-based models can be specialized to do supervise machine learning for text classification and regression. For classification, the model takes in a text and produces class labels as its first outputs in the sequence (see Figure B.1). The model can be trained on manually-labeled text examples. Regression works similarly, with numerical outputs replacing class labels. In 2022, transformer-based models still produce state-of-the-art results for text classification and regression (e.g. tweet classification and targeted sentiment analysis).

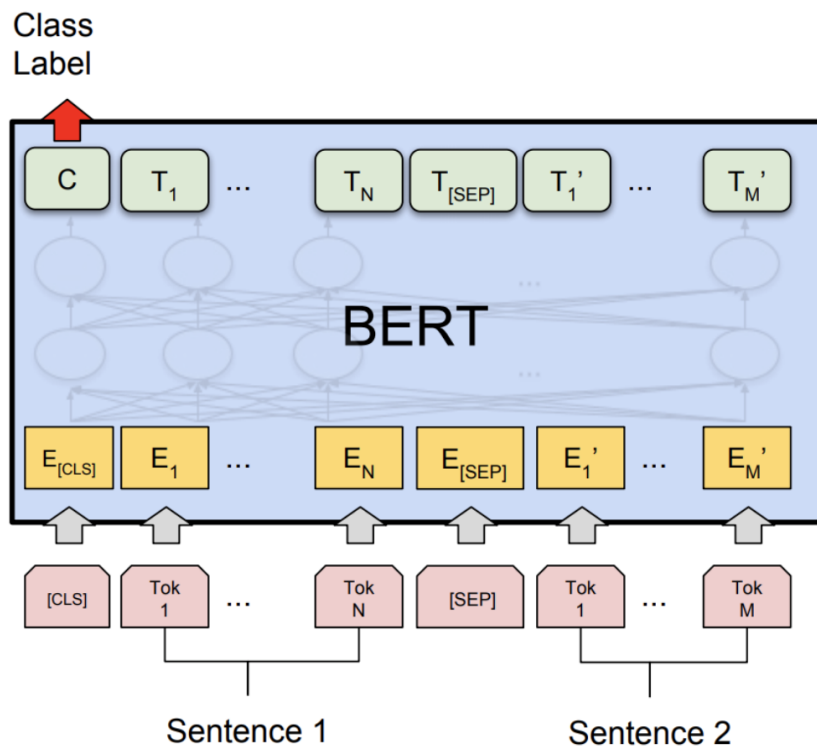


Figure B.1: An illustration of the BERT language model. It takes in a pair of short texts, encodes them as a sequence of vectors, and produces a numerical output, often interpreted as a class label. Illustration reproduced from [140]

Transformers for Semantic Text Embeddings. BERT-based models have been trained to produce semantic text embeddings [135], i.e. vector representations of texts which are defined such that similar vectors represent semantically-similar short texts (generally sentences or short paragraphs). In applications, we pass this pretrained model a short text and it produces a vector embedding of each token in the text, as well as an overall embedding of the whole text. These embeddings are useful in a variety of applications, and in Chapter 7 we apply them to perform keyword extraction by measuring the semantic relevance of different candidate keywords to a text (see subsection 7.6.1).

Large Language Models. In late 2022, the state of the art for natural language processing (NLP) began to shift dramatically with the release of ChatGPT and the mainstreaming of Large Language Models (LLMs). LLMs now present a novel way to perform many of

the NLP tasks mentioned in this work, including text classification, keyword extraction, target sentiment analysis, and stated belief extraction. While traditional methods like those used in this work remain preferable in many cases due to their greater scalability, LLMs offer greater power and flexibility enabling social media researchers to perform more complex assessments of text data with less labor hours. Of particular interest in this work is stated belief extraction, which proved infeasible with traditional methods and given time constraints for this work. Small tests conducted after the conclusion of the analyses in Chapter 7 showed that LLMs like GPT-4 and Mistral-7b can perform stated belief extraction adequately well with relatively simple prompting. However, in late 2023, we find that using LLMs for this kind of data processing is infeasible as the volume of data approaches millions of posts, as were used in these analyses, so we opted not to update and extend our work at this time.

APPENDIX C

MULTIVARIATE HAWKES PROCESSES

In Chapter 7, we treat sequences of actions by individual Twitter users as a point process. Specifically, we let their Like events and Tweet events be different types of events in a Multivariate Hawkes Process, and we use this formalism to infer the causal interactions between events. In this appendix, we briefly introduce the Multivariate Hawkes Process and an inference method for determining how events in the process are causally related.

C.1 Point Processes

Xu [134] defines a temporal point process as a random process whose realization consists of a set of discrete events in time t_i with $t_i \in [0, T]$. The simplest example is a Poisson process, in which events randomly occur at a constant mean rate, μ . For a Poisson process, the expected number of events that occur in the time interval $[0, T]$ is given as:

$$\mathbb{E}[N(T)] = \int_0^T \lambda(t) dt = \int_0^T \mu dt = \mu T \quad (\text{C.1})$$

where N is the number of events, \mathbb{E} indicates an expectation value, and λ is called the "intensity function" and it indicates the expected rate at which events will occur at time t . For Poisson processes, the event rate is constant over time, so $\lambda(t) = \mu$. We can define more complex point processes using this idea of an intensity function.

C.2 Hawkes Processes

Another useful temporal point process is called a Hawkes Process, in which each event increases the likelihood of subsequent events for some time (see Figure C.1). Hawkes

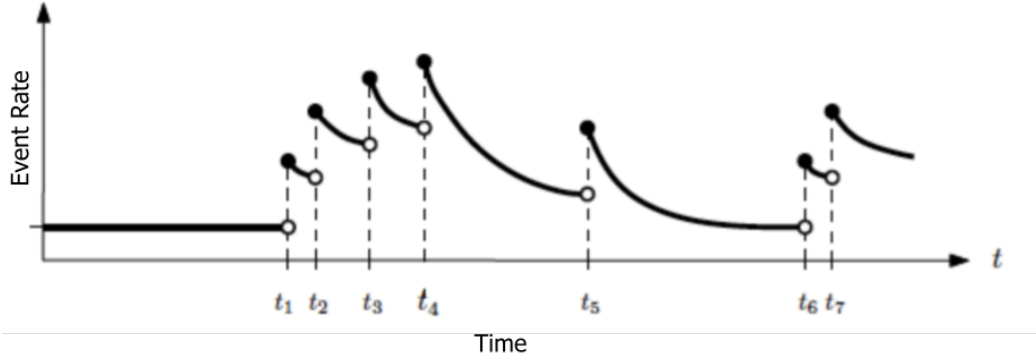


Figure C.1: An illustration of a Hawkes process intensity function. Illustration reproduced from [141]

Processes are defined by the form of their intensity function, which is given as:

$$\lambda(t) = \mu + \sum_{i: t_i < t} \phi(t - t_i) \quad (\text{C.2})$$

where $\phi(x)$ is the "excitation function." Excitation functions are always nonnegative and are usually defined so that they decay exponentially as $x \rightarrow \infty$. This fast decay ensures that the intensity function $\lambda(t)$ remains finite even as the point process accumulates more events over time (i.e. as $t \rightarrow \infty$ and $|\{t_i : t_i < t\}| \rightarrow \infty$).

Note that the excitation function ϕ characterizes how events can cause future events. This is what makes the Hawkes process formalism useful for us.

C.3 Multivariate Hawkes Processes

Hawkes process of the form Equation C.2 are designed to model temporal point processes where all events are identical. For temporal point processes with multiple different kinds of events, such as Tweet events and Like events, we can generalize Hawkes processes to multivariate Hawkes processes (see Figure C.2).

To define multivariate Hawkes processes, let B be the set of types of events involved in the multivariate Hawkes process (e.g. $B = \{\text{Likes}, \text{Tweets}\}$); let $a \in B$ and $b \in B$ be indexes for specific event types (e.g. $a = \text{Likes}$, $b = \text{Tweets}$); and let $t_{b,i}$ be specific events

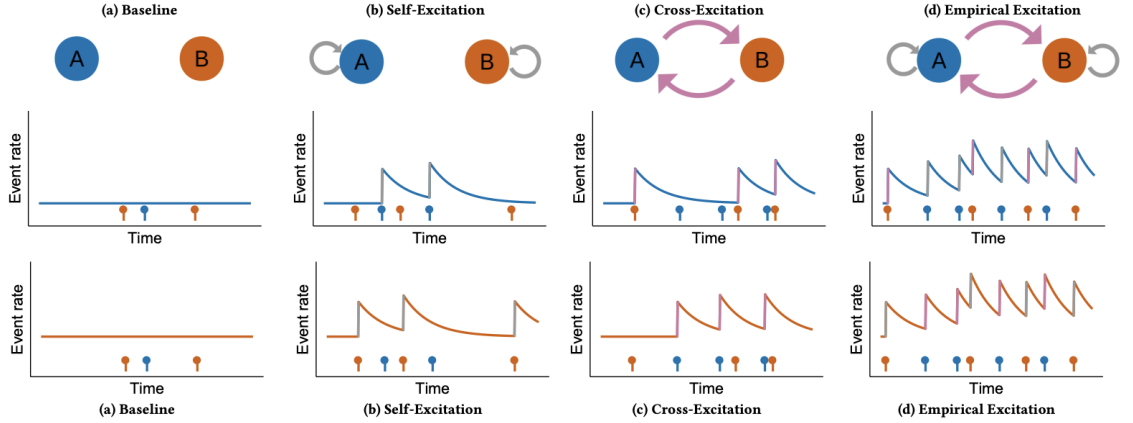


Figure C.2: An illustration of intensity functions of a multivariate Hawkes process. Illustration reproduced from [142]

of type b (e.g. $t_{b,i}$ is the time a specific Tweet occurred). We can then define multivariate Hawkes processes as temporal point processes with intensity functions of the form:

$$\lambda_a(t) = \mu_a + \sum_{b \in B} \sum_{i : t_{b,i} < t} \phi_{a,b}(t - t_{b,i}) \quad (\text{C.3})$$

where $\lambda_a(t)$ is the intensity function for events of type a , μ_a is the baseline rate for events of type a , and $\phi_{a,b}$ is the excitation function which determines how events of type b precipitate subsequent events of type a . Note that $a = b$ is the self-excitation case where events of type a precipitate more events of type a . Also notice that for the case $|B| = 1$ (and therefore $b = a \forall b \in B$), this reduces to the univariate Hawkes process defined in Equation C.2.

C.4 Inferring Excitation Functions

To use multivariate Hawkes models to understand real observed event streams, such as the Like and Tweet events performed by a Twitter user, we are primarily concerned with using the observed event sequences $t_{b,i}$ to infer the excitation functions $\phi_{a,b}$. The excitation functions allow us to see to what degree events of type b produce additional events of type a , and over what time period. This is of particular importance for understanding the causal relationships between different types of events in the point process.

A number of methods have been developed to perform this inference. We tried several and found the method described by Xu [134] to be most robust. To apply this method in our work, we used the published reference implementation [143]. I'll offer an overview of the method here, but for a full treatment I refer the reader to the original paper.

This inference method strives to achieve a flexible nonparametric fit of $\phi_{a,b}$ and μ_a for all $a \in B$ and $b \in B$. In practice, it approximates the function $\phi_{a,b}$ as a sum of Gaussian curves with fixed means (at regular intervals on the time axis) and fixed variances (typically set to be the interval width divided by π). The method works by then parameterizing the function $\phi_{a,b}$ by the amplitudes of the gaussians, $A_{a,b,j}$.

$$\phi_{a,b} = \sum_j A_{a,b,j} e^{-(x-c_j)^2/(2\sigma^2)} \quad (\text{C.4})$$

These amplitudes $A_{a,b,j}$ and baseline rates μ_a are inferred by an expectation maximization (EM) algorithm, which is a convex optimization that guarantees convergence on the global optimum (after sufficient iterations).

The major innovation of this method in particular is that it also applies a regularization scheme in which near-zero amplitudes are suppressed by an L1 regularization, and whole intensity functions $\phi_{a,b}$ which are near-zero are suppressed by a group lasso regularization (i.e. a single L1 regularization shared across the multiple amplitude parameters which specify $\phi_{a,b}$). This regularization scheme ensures that false inferences of nonzero excitation functions are prevented, and therefore that an inferred nonzero excitation can be taken as evidence of Granger causality (see Appendix D). This causal inference property is the key reason we needed to use this method for Chapter 7.

For appropriate choices of σ and c_j , this inference method can reliably yield an empirical approximation of $\phi_{a,b}$ for the observed event sequences. However in practice its efficacy is limited by data density. For instance, if $c_j - c_{j-1}$ is too small, the amplitude $A_{a,b,j}$ can be underconstrained, and this may yield noise or a false zero value. Therefore large volumes

of data are required for high temporal resolution of $\phi_{a,b}$.

C.5 Causation and Confounders in Hawkes Processes

Hawkes processes can lend themselves nicely to causal analyses, because they inherently establish both precedence and predictive relationships between events and they can be complemented by tools like CGM (Appendix A) to identify and then account for confounders, as illustrated in Chapter 7. In Hawkes processes, confounders may be controlled for in two main ways. First, we may include them as additional event types (as in Chapter 7). Second, we may make the baseline event rates, μ_a , a fitted function of confounding variables, either as static features or as time series features. When using methods like these, we may sometimes refer to "causal Hawkes processes," by which we mean the use of Hawkes processes in conjunction with CGM or other causal reasoning frameworks to perform causal inferences.

APPENDIX D

GRANGER CAUSALITY

Granger causality is simple heuristic used for causal inference. Granger causality operates on two philosophical principals: 1. Causes precede effects in time, 2. Causes contain unique information helpful for predicting effects. If event X has "Granger-caused" event Y , we define the general form of Granger causality as [144]:

$$P[Y(t)|I(t-1)] \neq P[Y(t)|I(t-1), X(t-c)] \quad (\text{D.1})$$

where P is probability, X is a possibly causal event, c is some temporal offset (often 1), and I is the all the information which may be relevant to the evolution of the system, excluding the information in the causal event X . We can apply this general idea to time series and to point processes.

Assumptions: Completeness of I . The reliability of the Granger-causal method hinges on whether or not your set of confounding variables I are complete; if you exclude relevant information, the analysis is flawed. In practice, the causal inference is trusted if researchers agree that the confounding variables which were included were reasonably comprehensive.

Time Series. In Chapter 5, we treat public health outcomes (e.g. COVID deaths per capita) and social media post volumes (i.e. number of antivax Tweets per capita) as time series. Typically when analyzing time series, Granger causation is tested using a regression model where lagged independent variables X (e.g. antivax Tweet volumes) and confounding variables I (e.g. socioeconomic confounders) are used to predict dependent variables Y [145]. If it is found that the lagged independent variables X significantly predict the dependent variables Y after controlling for confounders I , we can say that X Granger-cause

Y .

Point Processes. In multivariate point processes, if an event of type X tends to precipitate events of type Y (after considering the effects of other relevant events I), we can say that X Granger-causes Y . This inference process is somewhat less common than Granger inference from time series, so I'll explain it here in some detail.

In Chapter 7, we treat user behaviors on social media as a point process, and try to infer the causal relationships between different behaviors. For instance, causal events of type X might be the user "liking a tweet about a topic," and response events of type Y might be the user "authoring a tweet about that topic." If we find that, after including the effects of other user behaviors, events of type X make events of type Y more likely in the future, we can say that X Granger-causes Y . The formalism of Multivariate Hawkes processes (see Appendix C) lend themselves nicely to this formulation.

In Multivariate Hawkes processes (see Appendix C), we quantify these causal influences by using "excitation functions," $\phi_{Y,X}$, which indicate the increased rate at which events of type Y happen after an event of type X occurs (accounting for the effects of all other events). If the excitation function $\phi_{Y,X}$ is nonzero, we can say that X Granger-causes Y . The task of inferring Granger causality in Hawkes processes is then reduced to a task of inferring the excitation functions with appropriate regularization; the regularization ensures that excitation functions which are actually zero are inferred as zero. Fortunately, methods for performing inference of excitation functions have been thoroughly developed by other researchers [134, 143], and we apply their method in Chapter 7. This method is briefly described in Section C.4.

APPENDIX E

COMPARTMENTAL EPIDEMIC MODELS

Compartmental epidemic models [146] are a type of mathematical model used to understand the dynamics of infectious diseases. They divide the population into "compartments," each representing a different stage of the disease, and describe the transition of individuals between these compartments using differential equations. In this appendix, we provide a brief overview of the Susceptible-Infectious-Recovered (SIR) and Susceptible-Infectious-Recovered-Vaccinated (SIRV) models.

SIR. The SIR model [147, 146] is one of the simplest compartmental epidemic models, dividing the population into three compartments: Susceptible (S), Infectious (I), and Recovered (R). Susceptible individuals can become infected when they come into contact with infectious individuals, and infectious individuals can recover and gain immunity. The model is usually represented by the following set of ordinary differential equations:

$$\begin{aligned}\frac{dS}{dt} &= -\beta SI \\ \frac{dI}{dt} &= \beta S \frac{I}{N} - \rho I \\ \frac{dR}{dt} &= \gamma I\end{aligned}$$

where β is the transmission rate and ρ is the recovery rate, and N is the total population, $N = S + I + R$. The basic reproduction number, R_0 , is given by $\frac{\beta}{\rho}$ and represents the average number of secondary infections produced by one infected individual in a completely susceptible population.

SIRV. The SIRV model extends the SIR model by including a Vaccinated (V) compartment. This compartment represents individuals who have been vaccinated and have acquired immunity to the disease. The vaccination is modeled as a rate at which susceptible individuals move to the vaccinated compartment, avoiding the infectious compartment. The model can be represented by the following set of ordinary differential equations:

$$\begin{aligned}\frac{dS}{dt} &= -\beta SI - \nu S \\ \frac{dI}{dt} &= \beta S \frac{I}{N} - \rho I \\ \frac{dR}{dt} &= \rho I \\ \frac{dV}{dt} &= \nu S\end{aligned}$$

where ν is the vaccination rate, β is the transmission rate, and ρ is the recovery rate. Individuals in the Vaccinated compartment are assumed to be immune to infection and do not contribute to disease transmission.

REFERENCES

- [1] T. Althoff, P. Jindal, and J. Leskovec, “Online Actions with Offline Impact: How Online Social Networks Influence Online and Offline User Behavior,” in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, Cambridge United Kingdom: ACM, Feb. 2017, pp. 537–546, ISBN: 978-1-4503-4675-7.
- [2] P. G. Turner and C. E. Lefevre, “Instagram use is linked to increased symptoms of orthorexia nervosa,” *Eating and Weight Disorders - Studies on Anorexia, Bulimia and Obesity*, vol. 22, no. 2, pp. 277–284, Jun. 2017.
- [3] H. B. Shakya and N. A. Christakis, “Association of Facebook Use With Compromised Well-Being: A Longitudinal Study,” *American Journal of Epidemiology*, amjepid, kww189v2, Jan. 2017.
- [4] L. Braghieri, R. Levy, and A. Makarin, “Social Media and Mental Health,” *American Economic Review*, vol. 112, no. 11, pp. 3660–3693, Nov. 2022.
- [5] M. R. DeVerna *et al.*, “CoVaxxy: A global collection of English-language Twitter posts about COVID-19 vaccines,” *arXiv:2101.07694 [cs]*, Feb. 2021. arXiv: 2101.07694 [cs].
- [6] N. Tabari, P. Biswas, B. Praneeth, A. Seyeditabari, M. Hadzikadic, and W. Zadrozny, “Causality Analysis of Twitter Sentiments and Stock Market Returns,” in *Proceedings of the First Workshop on Economics and Natural Language Processing*, Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 11–19.
- [7] T. Fujiwara, K. Müller, and C. Schwarz, “THE EFFECT OF SOCIAL MEDIA ON ELECTIONS: EVIDENCE FROM THE UNITED STATES,” *Journal of the European Economic Association*, 2023.
- [8] R. Mousavi and B. Gu, “The Impact of Twitter Adoption on Lawmakers’ Voting Orientations,” *Information Systems Research*, vol. 30, no. 1, pp. 133–153, Mar. 2019.
- [9] L. Fergusson and C. Molina, “Facebook Causes Protests,” *SSRN Electronic Journal*, 2019.
- [10] R. Enikolopov, A. Makarin, and M. Petrova, “Social Media and Protest Participation: Evidence From Russia,” *Econometrica*, vol. 88, no. 4, pp. 1479–1514, 2020.

- [11] D. Acemoglu, T. A. Hassan, and A. Tahoun, “The Power of the Street: Evidence from Egypt’s Arab Spring,” *The Review of Financial Studies*, vol. 31, no. 1, pp. 1–42, Jan. 2018.
- [12] K. Muller and C. Schwarz, “Fanning the Flames of Hate: Social Media and Hate Crime,” *SSRN Electronic Journal*, Jun. 2020.
- [13] K. Müller and C. Schwarz, “From Hashtag to Hate Crime: Twitter and Antiminority Sentiment,” *American Economic Journal: Applied Economics*, vol. 15, no. 3, pp. 270–312, Jul. 2023.
- [14] P. Mozur, “A Genocide Incited on Facebook, With Posts From Myanmar’s Military,” *The New York Times*, Oct. 2018.
- [15] D. Hume, Ed., *Enquiry Concerning Human Understanding*. Clarendon Press, 1904.
- [16] D. Lewis, “Causation,” *Journal of Philosophy*, vol. 70, no. 17, pp. 556–567, 1973.
- [17] J. L. Mackie, “Causes and conditions,” *American Philosophical Quarterly*, vol. 2, no. 4, pp. 245–264, 1965.
- [18] P. Machamer, L. Darden, and C. F. Craver, “Thinking about mechanisms,” *Philosophy of Science*, vol. 67, no. 1, pp. 1–25, 2000.
- [19] C. W. J. Granger, “Investigating causal relations by econometric models and cross-spectral methods,” *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.
- [20] P. Suppes, *A Probabilistic Theory of Causality*. Amsterdam: North-Holland Pub. Co., 1970.
- [21] E. Eells, *Probabilistic Causality*, ser. Cambridge Studies in Probability, Induction and Decision Theory. Cambridge University Press, 1991, ISBN: 9780521392440.
- [22] J. F. Woodward, *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press, 2003.
- [23] J. Pearl, *Causality: Models, Reasoning and Inference*, 2nd. USA: Cambridge University Press, 2009, ISBN: 052189560X.
- [24] W. C. Salmon, *Scientific Explanation and the Causal Structure of the World*. Princeton University Press, 1984, ISBN: 9780691101705.
- [25] P. Dowe, *Physical Causation*, ser. Cambridge Studies in Probability, Induction and Decision Theory. Cambridge University Press, 2000.

- [26] N. Cartwright, *Nature's Capacities and Their Measurement*. New York: Oxford University Press, 1989.
- [27] S. Mumford and R. L. Anjum, *Getting Causes From Powers*, R. L. Anjum, Ed. Oxford, GB: Oxford University Press, 2011.
- [28] F. Bacon, *Bacon's Novum Organum*. T. and G. Shrimpton, 1854.
- [29] R. A. Fisher, *The Design of Experiments*. Edinburgh: Oliver and Boyd, 1935.
- [30] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne, "Controlled experiments on the web: Survey and practical guide," *Data Mining and Knowledge Discovery*, vol. 18, no. 1, pp. 140–181, Jul. 2008.
- [31] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.
- [32] O. Ashenfelter and D. Card, "Using the longitudinal structure of earnings to estimate the effect of training programs," *The Review of Economics and Statistics*, vol. 67, no. 4, pp. 648–60, 1985.
- [33] D. Card and A. Krueger, "Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania," *American Economic Review*, vol. 84, no. 4, pp. 772–93, 1994.
- [34] D. Rubin, "Estimating causal effects of treatments in experimental and observational studies," *ETS Research Bulletin Series*, vol. 1972, no. 2, pp. i–31, 1972.
- [35] P. W. Holland, "Statistics and causal inference," *Journal of the American Statistical Association*, vol. 81, no. 396, pp. 945–960, 1986.
- [36] R. M. Bond *et al.*, "A 61-million-person experiment in social influence and political mobilization," *Nature*, vol. 489, no. 7415, pp. 295–298, Sep. 2012.
- [37] A. D. I. Kramer, J. E. Guillory, and J. T. Hancock, "Experimental evidence of massive-scale emotional contagion through social networks," *Proceedings of the National Academy of Sciences*, vol. 111, no. 24, pp. 8788–8790, Jun. 2014.
- [38] H. Allcott, L. Braghieri, S. Eichmeyer, and M. Gentzkow, "The Welfare Effects of Social Media," *American Economic Review*, vol. 110, no. 3, pp. 629–676, Mar. 2020.
- [39] M. Petrova, A. Sen, and P. Yildirim, "Social Media and Political Contributions: The Impact of New Technology on Political Competition," *Management Science*, vol. 67, no. 5, pp. 2997–3021, May 2021.

- [40] M. Dredze, M. J. Paul, S. Bergsma, and H. Tran, “Carmen: A twitter geolocation system with applications to public health,” Association for the Advancement of Artificial Intelligence, 2013.
- [41] J. Zhang, A. DeLucia, and M. Dredze, “Changes in tweet geolocation over time: A study with carmen 2.0,” in *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, Gyeongju, Republic of Korea: Association for Computational Linguistics, Oct. 2022, pp. 1–14.
- [42] A. E. Kim *et al.*, “Estimated ages of juul twitter followers,” *JAMA Pediatr*, vol. 173, no. 7, pp. 690–692, 2019, Erratum in: *JAMA Pediatr*. 2019 Jul 1;173(7):704.
- [43] M. D. Conover, B. Goncalves, J. Ratkiewicz, A. Flammini, and F. Menczer, “Predicting the political alignment of twitter users,” in *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, 2011, pp. 192–199.
- [44] S. Golder, R. Stevens, K. O’Connor, R. James, and G. Gonzalez-Hernandez, “Methods to establish race or ethnicity of twitter users: Scoping review,” *Journal of Medical Internet Research*, vol. 24, no. 4, e35788, Apr. 2022.
- [45] Y.-C. Yang, M. A. Al-Garadi, J. S. Love, J. Perrone, and A. Sarker, “Automatic gender detection in twitter profiles for health-related cohort studies,” *JAMIA Open*, vol. 4, no. 2, Apr. 2021.
- [46] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, “Deep learning-based text classification,” *ACM Computing Surveys*, vol. 54, no. 3, pp. 1–40, Apr. 2021.
- [47] L. Chen, S. Li, Q. Bai, J. Yang, S. Jiang, and Y. Miao, “Review of image classification algorithms based on convolutional neural networks,” *Remote Sensing*, vol. 13, no. 22, p. 4712, Nov. 2021.
- [48] A. Sharma and E. Kiciman, *Dowhy: An end-to-end library for causal inference*, 2020. arXiv: 2011.04216 [stat.ME].
- [49] E. Dong, H. Du, and L. Gardner, “An interactive web-based dashboard to track covid-19 in real time,” *The Lancet Infectious Diseases*, vol. 20, no. 5, pp. 533–534, May 2020.
- [50] W. A. Orenstein and R. Ahmed, “Simply put: Vaccination saves lives,” *Proceedings of the National Academy of Sciences*, vol. 114, no. 16, pp. 4031–4033, 2017. eprint: <https://www.pnas.org/content/114/16/4031.full.pdf>.

- [51] R. Aguas, R. M. Corder, J. G. King, G. Gonçalves, M. U. Ferreira, and M. G. M. Gomes, “Herd immunity thresholds for sars-cov-2 estimated from unfolding epidemics,” *medRxiv*, 2020. eprint: <https://www.medrxiv.org/content/early/2020/11/16/2020.07.23.20160762.full.pdf>.
- [52] C. Funk and A. Tyson, *Intent to get a covid-19 vaccine rises to 60% as confidence in research and development process increases*, Pew Research Center, <https://www.pewresearch.org/science/2020/12/03/intent-to-get-a-covid-19-vaccine-rises-to-60-as-confidence-in-research-and-development-process-increases/> (accessed January, 2021), 2020.
- [53] L. Hamel, A. Kirzinger, and M. Brodie, *Kff covid-19 vaccine monitor: December 2020*, KFF Health Tracking Poll, <https://www.kff.org/coronavirus-covid-19/report/kff-covid-19-vaccine-monitor-december-2020> (accessed January, 2021), 2020.
- [54] A. Hussain, S. Ali, M. Ahmed, and S. Hussain, “The Anti-vaccination Movement: A Regression in Modern Medicine,” *Cureus*, vol. 10, no. 7, 2018. pmid: 30186724.
- [55] D. A. Broniatowski *et al.*, “Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate,” *American Journal of Public Health*, vol. 108, no. 10, pp. 1378–1384, 2018.
- [56] T. Burki, “Vaccine misinformation and social media,” *The Lancet Digital Health*, vol. 1, no. 6, e258–e259, 2019.
- [57] N. F. Johnson *et al.*, “The online competition between pro- and anti-vaccination views,” *Nature*, vol. 582, no. 7811, pp. 230–233, 7811 Jun. 2020.
- [58] E. K. Brunson, “The Impact of Social Networks on Parents’ Vaccination Decisions,” *Pediatrics*, vol. 131, no. 5, e1397–e1404, 2013.
- [59] J. Roozenbeek *et al.*, “Susceptibility to misinformation about COVID-19 around the world,” *Royal Society Open Science*, vol. 7, no. 10, p. 201199, Oct. 2020.
- [60] E. Chen, K. Lerman, and E. Ferrara, “Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set,” *JMIR Public Health and Surveillance*, vol. 6, no. 2, e19273, 2020.
- [61] X. Huang, A. Jamison, D. Broniatowski, S. Quinn, and M. Dredze, *Coronavirus twitter data: A collection of covid-19 tweets with automated annotations*, Zenodo, Mar. 2020.
- [62] R. Lamsal, *Coronavirus (covid-19) tweets dataset*, IEEE Dataport, 2020.

- [63] K.-C. Yang, P.-M. Hui, and F. Menczer, “Bot electioneering volume: Visualizing social bot activity during elections,” in *Companion Proceedings of The 2019 World Wide Web Conference*, 2019, pp. 214–217.
- [64] K.-C. Yang *et al.*, “The covid-19 infodemic: Twitter versus facebook,” *Big Data & Society*, 2021, In press; preprint arXiv:2012.09353.
- [65] F. Pierri, C. Piccardi, and S. Ceri, “A multi-layer approach to disinformation detection in us and italian news spreading on twitter,” *EPJ Data Science*, vol. 9, no. 35, 2020.
- [66] —, “Topology comparison of Twitter diffusion networks effectively reveals misleading news,” *Scientific Reports*, vol. 10, p. 1372, 2020.
- [67] D. Pacheco, P.-M. Hui, C. Torres-Lugo, B. T. Truong, A. Flammini, and F. Menczer, “Uncovering coordinated networks on social media,” in *Proc. AAAI Intl. Conf. on Web and Social Media (ICWSM)*, In press; preprint arXiv:2001.05658, 2021.
- [68] M. D. Conover, B. Gonçalves, A. Flammini, and F. Menczer, “Partisan asymmetries in online political activity,” *EPJ Data Science*, vol. 1, no. 6, 2012.
- [69] J. Towns *et al.*, “Xsede: Accelerating scientific discovery,” *Computing in Science & Engineering*, vol. 16, no. 5, pp. 62–74, Oct. 2014.
- [70] C. A. Stewart *et al.*, “Jetstream: A self-provisioned, scalable science and engineering cloud environment,” in *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*, ser. XSEDE ’15, New York, NY, USA: Association for Computing Machinery, Jul. 26, 2015, pp. 1–8.
- [71] M. DeVerna *et al.*, *Covaxxy tweet ids dataset*, Zenodo, version 1, Feb. 2021.
- [72] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, P10008, 2008.
- [73] D. Lazer, M. Baum, Y. Benkler, A. Berinsky, K. Greenhill, *et al.*, “The science of fake news,” *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.
- [74] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer, “The spread of low-credibility content by social bots,” *Nature Communications*, vol. 9, p. 4787, 2018.
- [75] A. Bovet and H. A. Makse, “Influence of fake news in Twitter during the 2016 US presidential election,” *Nature Communications*, vol. 10, no. 1, p. 7, 2019.

- [76] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer, “Fake news on Twitter during the 2016 U.S. presidential election,” *Science*, vol. 363, no. 6425, pp. 374–378, 2019.
- [77] S. Wojick and A. Hughes, *Sizing up twitter users*, Pew Research Center, <https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/> (accessed January, 2021), 2020.
- [78] D. Jurgens, Y. Tsvetkov, and D. Jurafsky, “Incorporating dialectal variability for socially equitable language identification,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 51–57.
- [79] A. B. Suthar, J. Wang, V. Seffren, R. E. Wiegand, S. Griffing, and E. Zell, “Public health impact of covid-19 vaccines in the us: Observational study,” *BMJ*, vol. 377, 2022. eprint: <https://www.bmj.com/content/377/bmj-2021-069317.full.pdf>.
- [80] S. Gupta, J. Cantor, K. I. Simon, A. I. Bento, C. Wing, and C. M. Whaley, “Vaccinations against covid-19 may have averted up to 140,000 deaths in the united states,” *Health Affairs*, vol. 40, no. 9, pp. 1465–1472, 2021.
- [81] E. C. Schneider, A. Shah, P. Sah, S. M. Moghadas, T. Vilches, and A. Galvani, *The u.s. covid-19 vaccination program at one year: How many deaths and hospitalizations were averted?* 2021.
- [82] E. Mathieu *et al.*, “A global database of covid-19 vaccinations,” *Nature human behaviour*, vol. 5, no. 7, pp. 947–953, 2021.
- [83] S. Loomba, A. de Figueiredo, S. J. Piatek, K. de Graaf, and H. J. Larson, “Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa,” *Nature human behaviour*, vol. 5, no. 3, pp. 337–348, 2021.
- [84] F. Pierri *et al.*, “Online misinformation is linked to early covid-19 vaccination hesitancy and refusal,” *Scientific Reports*, vol. 12, no. 1, pp. 1–7, 2022.
- [85] F. Yasmin *et al.*, “COVID-19 Vaccine Hesitancy in the United States: A Systematic Review,” *Frontiers in Public Health*, vol. 9, p. 770 985, Nov. 2021.
- [86] R. Gallotti, F. Valle, N. Castaldo, P. Sacco, and M. De Domenico, “Assessing the risks of ‘infodemics’ in response to COVID-19 epidemics,” *Nature Human Behaviour*, vol. 4, no. 12, pp. 1285–1293, 2020.
- [87] K.-C. Yang *et al.*, “The covid-19 infodemic: Twitter versus facebook,” *Big Data & Society*, vol. 8, no. 1, p. 20 539 517 211 013 861, 2021.

- [88] J. Pearl, “The Do-Calculus Revisited,” in *Proc. Twenty-Eighth Conference on Uncertainty in Artificial Intelligence (UAI)*, Jul. 2012, pp. 3–11.
- [89] M. Bailey, R. Cao, T. Kuchler, J. Stroebel, and A. Wong, “Social connectedness: Measurement, determinants, and effects,” *Journal of Economic Perspectives*, vol. 32, no. 3, pp. 259–80, 2018.
- [90] E. Ortiz-Sánchez *et al.*, “Analysis of the Anti-Vaccine Movement in Social Networks: A Systematic Review,” *International Journal of Environmental Research and Public Health*, vol. 17, no. 15, p. 5394, Jul. 2020.
- [91] M. DeVerna *et al.*, “CoVaxxy: A Collection of English-Language Twitter Posts About COVID-19 Vaccines,” in *Proc. Intl. AAAI Conf. on Web and Social Media (ICWSM)*, vol. 15, 2021, pp. 992–999.
- [92] Centers for Disease Control and Prevention (CDC), *COVID-19 Vaccinations in the United States, County*, <https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh>, 2021.
- [93] B. Mønsted and S. Lehmann, “Algorithmic Detection and Analysis of Vaccine-Denialist Sentiment Clusters in Social Networks,” *arXiv preprint arXiv:1905.12908*, May 2019.
- [94] Y. Liu *et al.*, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” *arXiv:1907.11692 [cs]*, Jul. 2019, arXiv: 1907.11692.
- [95] D. Chicco, N. Tötsch, and G. Jurman, “The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation,” *BioData Mining*, vol. 14, no. 1, p. 13, Dec. 2021.
- [96] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Boston, MA: Springer US, 1993.
- [97] G. Hamra, R. MacLehose, and D. Richardson, “Markov Chain Monte Carlo: An introduction for epidemiologists,” *International Journal of Epidemiology*, vol. 42, no. 2, pp. 627–634, Apr. 2013.
- [98] M. D. Hoffman and A. Gelman, “The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo,” *Journal of Machine Learning Research*, p. 31, 2014.
- [99] D. Phan, N. Pradhan, and M. Jankowiak, “Composable effects for flexible and accelerated probabilistic programming in numpyro,” *arXiv preprint arXiv:1912.11554*, 2019.

- [100] Centers for Disease Control and Prevention, *Ending isolation and precautions for people with covid-19: Interim guidance*, 2023.
- [101] ASPE, *Vaccine hesitancy for covid-19: State, county, and local estimates*, 2021.
- [102] J. Dehning *et al.*, “Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions,” *Science*, vol. 369, no. 6500, eabb9789, Jul. 2020.
- [103] T. T. Marinov and R. S. Marinova, “Adaptive SIR model with vaccination: Simultaneous identification of rates and functions illustrated with COVID-19,” *Scientific Reports*, vol. 12, no. 1, p. 15 688, Sep. 2022.
- [104] M. W. Tenforde, W. H. Self, M. Gaglani, and et al., “Effectiveness of mrna vaccination in preventing covid-19–associated invasive mechanical ventilation and death — united states, march 2021–january 2022,” *Morbidity and Mortality Weekly Report*, vol. 71, no. 12, pp. 459–465, 2022.
- [105] J. Lopez Bernal *et al.*, “Effectiveness of covid-19 vaccines against the b.1.617.2 (delta) variant,” *New England Journal of Medicine*, vol. 385, no. 7, pp. 585–594, 2021, PMID: 34289274. eprint: <https://doi.org/10.1056/NEJMoa2108891>.
- [106] A. Vehtari, A. Gelman, and J. Gabry, “Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC,” *Statistics and Computing*, vol. 27, no. 5, pp. 1413–1432, Sep. 2017.
- [107] R. A. Harder, J. Sevenans, and P. Van Aelst, “Intermedia Agenda Setting in the Social Media Age: How Traditional Players Dominate the News Agenda in Election Times,” *The International Journal of Press/Politics*, vol. 22, no. 3, pp. 275–293, 2017.
- [108] Y. Su and P. Borah, “Who is the agenda setter? Examining the intermedia agenda-setting effect between Twitter and newspapers,” *Journal of Information Technology & Politics*, vol. 16, no. 3, pp. 236–249, 2019.
- [109] M. A. Shapiro and L. Hemphill, “Politicians and the Policy Agenda: Does Use of Twitter by the U.S. Congress Direct *New York Times* Content?: Politicians and the Policy Agenda,” *Policy & Internet*, vol. 9, no. 1, pp. 109–132, 2017.
- [110] P. Barberá *et al.*, “Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data,” *American Political Science Review*, vol. 113, no. 4, pp. 883–901, 2019.
- [111] A. Jungherr, G. Rivero, and D. Gayo-Avello, *Retooling Politics: How Digital Media Are Shaping Democracy*, 1st ed. Cambridge University Press, 2020.

- [112] C. Leston-Bandeira, “Parliamentary petitions and public engagement: An empirical analysis of the role of e-petitions,” *Policy & Politics*, vol. 47, no. 3, pp. 415–436, 2019.
- [113] U. Bernhard and M. Dohle, “Local Politics Online: The Influence of Presumed Influence on Local Politicians’ Online Communication Activities in Germany,” *Local Government Studies*, vol. 41, no. 5, pp. 755–773, 2015.
- [114] D. Kreiss, “Seizing the moment: The presidential campaigns’ use of Twitter during the 2012 electoral cycle,” *New Media & Society*, vol. 18, no. 8, pp. 1473–1490, 2016.
- [115] W. T. Daniel and L. Obholzer, “Reaching out to the voter? Campaigning on Twitter during the 2019 European elections,” *Research & Politics*, vol. 7, no. 2, p. 205 316 802 091 725, 2020.
- [116] A. Bovet and H. A. Makse, “Influence of fake news in Twitter during the 2016 US presidential election,” *Nature Communications*, vol. 10, no. 1, p. 7, Jan. 2019.
- [117] G. C. Edwards and B. D. Wood, “Who Influences Whom? The President, Congress, and the Media,” *American Political Science Review*, vol. 93, no. 2, pp. 327–344, Jun. 1999.
- [118] J. Borge-Holthoefer *et al.*, “The dynamics of information-driven coordination phenomena: A transfer entropy analysis,” *Science Advances*, vol. 2, no. 4, e1501158, Apr. 2016.
- [119] G. Ver Steeg and A. Galstyan, “Information Transfer in Social Media,” in *Proceedings of the 21st International Conference on World Wide Web*, ser. WWW ’12, New York, NY, USA: ACM, 2012, pp. 509–518, ISBN: 978-1-4503-1229-5.
- [120] J. Bryden, S. P. Wright, and V. A. A. Jansen, “How humans transmit language: Horizontal transmission matches word frequencies among peers on Twitter,” *Journal of The Royal Society Interface*, vol. 15, no. 139, 2018.
- [121] L. van Vliet, P. Törnberg, and J. Uitermark, “The Twitter parliamentarian database: Analyzing Twitter politics across 26 countries,” *PLOS ONE*, vol. 15, no. 9, S. Lozano, Ed., e0237073, Sep. 2020.
- [122] S. Messing *et al.*, *Facebook Privacy-Protected Full URLs Data Set*, Oct. 2021.
- [123] G. King and N. Persily, “A New Model for Industry–Academic Partnerships,” *Political Science & Politics*, vol. 53, no. 4, pp. 703–709, Oct. 2020.

- [124] J. R. Bray and J. T. Curtis, “An Ordination of the Upland Forest Communities of Southern Wisconsin,” *Ecological Monographs*, vol. 27, no. 4, pp. 325–349, Oct. 1957.
- [125] OptimalSocial, *75% of facebook engagement is in the first 180 minutes, says facebook competition*, <https://venturebeat.com/2013/03/28/75-of-facebook-engagement-is-in-the-first-180-minutes-says-facebook-competition-winning-tool>, 2013.
- [126] S. Ayres, *Shocking New Data about the Lifespan of Your Facebook Posts*, <https://www.postplanner.com/lifespan-of-facebook-posts/>, 2013.
- [127] J. Symonds, *Lifespan of Social Media Posts in 2021: How Long Do They Last?* <https://the-refinery.io/blog/how-long-does-a-social-media-post-last>, 2021.
- [128] J. Corless, *Measuring Your Social Media Engagement*, <https://universitymarketing.osu.edu/blog/measuring-your-social-engagement.html>, 2020.
- [129] H. Allcott and M. Gentzkow, “Social Media and Fake News in the 2016 Election,” *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–236, May 2017.
- [130] P. N. Howard, S. Woolley, and R. Calo, “Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration,” *Journal of Information Technology & Politics*, vol. 15, no. 2, pp. 81–93, Apr. 2018.
- [131] M. Hänska-Ahy and S. Bauchowitz, “Tweeting for Brexit: How social media influenced the referendum,” in *Brexit, Trump and the Media*, Bury St Edmunds: Abramis academic publishing, 2017, ISBN: 978-1-84549-709-5.
- [132] Y. Gorodnichenko, T. Pham, and O. Talavera, “Social media, sentiment and public opinions: Evidence from #Brexit and #USElection,” *European Economic Review*, vol. 136, p. 103 772, Jul. 2021.
- [133] I. Twitter, *Twitter Decahose*, 2022.
- [134] H. Xu, M. Farajtabar, and H. Zha, “Learning Granger Causality for Hawkes Processes,” *arXiv:1602.04511 [cs, stat]*, Jun. 2016, arXiv: 1602.04511.
- [135] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” *arXiv:1908.10084 [cs]*, Aug. 2019, arXiv: 1908.10084.
- [136] H. Yang, B. Zeng, J. Yang, Y. Song, and R. Xu, “A Multi-task Learning Model for Chinese-oriented Aspect Polarity Classification and Aspect Term Extraction,” *arXiv:1912.07976 [cs]*, Feb. 2020, arXiv: 1912.07976.

- [137] B. Wang, M. Liakata, A. Zubiaga, and R. Procter, *Multi-target UK election Twitter sentiment corpus*, 2016.
- [138] C. Hitchcock, “Causal Models,” in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta and U. Nodelman, Eds., Spring 2023, Metaphysics Research Lab, Stanford University, 2023.
- [139] A. Vaswani *et al.*, “Attention Is All You Need,” *arXiv:1706.03762 [cs]*, Dec. 2017, arXiv: 1706.03762.
- [140] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *arXiv:1810.04805 [cs]*, May 2019, arXiv: 1810.04805.
- [141] A. Swishchuk, *Hawkes Processes and their Applications in Finance and Insurance*, University of Calgary, 2017.
- [142] T. Santos, S. Walk, R. Kern, M. Strohmaier, and D. Helic, “Self- and Cross-Excitation in Stack Exchange Question & Answer Communities,” in *The World Wide Web Conference*, San Francisco CA USA: ACM, May 2019, pp. 1634–1645, ISBN: 978-1-4503-6674-8.
- [143] H. Xu, *Hawkes-Process-Toolkit*, 2018.
- [144] C. W. Granger, “Testing for causality: A personal viewpoint,” *Journal of Economic Dynamics and control*, vol. 2, pp. 329–352, 1980.
- [145] M. Eichler, “Causal Inference in Time Series Analysis,” in *Wiley Series in Probability and Statistics*, C. Berzuini, P. Dawid, and L. Bernardinelli, Eds., Chichester, UK: John Wiley & Sons, Ltd, Jun. 2012, pp. 327–354.
- [146] F. Brauer, “Compartmental models in epidemiology,” in *Mathematical Epidemiology*, ser. Lecture notes in mathematics, Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 19–79.
- [147] W. O. Kermack, A. G. McKendrick, and G. T. Walker, “A contribution to the mathematical theory of epidemics,” *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 115, no. 772, pp. 700–721, 1927. eprint: <https://royalsocietypublishing.org/doi/pdf/10.1098/rspa.1927.0118>.

John Bollenbacher
jbollenbacher.github.io

EDUCATION

Indiana University, Bloomington, IN
Doctor of Philosophy, Informatics, May 2024
Track: Complex Networks and Systems
Minor: Statistics

Indiana University, Bloomington, IN
Masters of Science, Informatics, May 2019

Georgia Tech, Atlanta, GA
Bachelor of Science, Physics, May 2016
Concentration: Astrophysics
Minor: Computer Science

PROFESSIONAL EXPERIENCE

Research Data Scientist
September 2022 - Present
Supervisor: Gayle Bieler

RTI International
Research Triangle, NC

Applying data science methods to conduct quantitative research in collaboration with domain experts in subject areas including public health, environmental science, and media studies. Specializing in NLP, causal inference, and task automation with generative AI.

Research Assistant
August 2021 - May 2022
Supervisor: John Bryden

Observatory on Social Media
Bloomington, IN

Analyzed social media data, news articles, and UK Parliamentary discourse to measure the influence of social media on the language of UK MPs. Wrote and published a paper.

Assistant Instructor
August 2019 - May 2022
Supervisors: Alexis Peirce Caudell, Nina Onesti

Indiana University Informatics Department
Remote & Bloomington, IN

Helped teach an Ethics of Technology course. Helped design course curricula including the syllabus, lectures, assignments, and tests. Occasionally lectured. Graded assignments.

Data Scientist
May 2021 - September 2021
Supervisor: Neerja Bharti

GeniusMesh
Remote

Performed data science, NLP, and feature engineering tasks to analyze career paths of Executive MBAs and make career decision recommendations. Worked with a development team to deploy a client-facing data dashboard to production.

Research Assistant
August 2017 - August 2019

Indiana University Informatics Department
Bloomington, IN

Supervisor: Filippo Menczer

Analyzed large social media datasets, and modeled and forecast social systems. Worked effectively with a large data science team. Wrote regular reports and presentations. Wrote and published academic papers.

Assistant Instructor

August 2016 - May 2017

Supervisors: Filippo Raddichi, Alexander J. Gates

Indiana University Informatics Department

Bloomington, IN

Helped teach a discrete mathematics course. Lectured and proctored in-class work weekly. Graded assignments.

Radiological Surveyer, CAD Technician

May 2014 - August 2014, August 2015

Supervisors: Vince Barlock, Beth Jensen

USA Environment LP

Golden, CO

Surveyed land for radiological contamination. Designed and managed instrument packages for radiological survey work; trained others in their use. Designed excavation plans for remediation of contaminated land in Civil3D CAD software.

Student Researcher

August 2013 - May 2014

Advisor: David Ballantyne

Georgia Tech Physics Department

Atlanta, GA

Analyzed spectral emissions from black hole accretion disks, following close instructions of David Ballantyne.

GRANTS

Google Cloud Research Credits Program

2021

PI: John Bollenbacher

Contribution: Wrote and submitted grant application

cloud.google.com/edu/researchers

Observatory on Social Media

Bloomington, IN

Social Media and Democracy Grant

2021

PI: John Bryden

Contribution: Data science work; wrote a paper as a grant deliverable

ssrc.org/grantees/how-does-facebook-influence-parliament

Observatory on Social Media

Bloomington, IN

DARPA SocialSim Grant

Fall 2017 - Spring 2019

PIs: Fil Menczer, Emilio Ferrara

Contribution: Added to grant proposal; produced grant deliverables

darpa.mil/program/computational-simulation-of-online-social-behavior

Indiana University

Bloomington, IN

PUBLICATIONS

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

Bollenbacher, J., Loynes, N., Bryden, J. Does United Kingdom parliamentary attention follow social media posts? EPJ Data Sci. 11, 51 (2022).

Bollenbacher, J. et al. On the challenges of predicting microscopic dynamics of online conversations. *Appl Netw Sci* 6, 12 (2021).

DeVerna, Matthew et al. CoVaxxy Tweet IDs data set. in Proceedings of the International AAAI Conference on Web and Social Media (2021).

Blythe, J. et al. Massive multi-agent data-driven simulations of the GitHub ecosystem. in International conference on practical applications of agents and multi-agent systems 3–15 (2019).

Blythe, J. et al. The DARPA SocialSim challenge: massive multi-agent simulations of the github ecosystem. in Proceedings of the 18th international conference on autonomous agents and MultiAgent systems 1835–1837 (2019).

Ballantyne, D. R. et al. NuSTAR reveals the Comptonizing corona of the broad-line radio galaxy 3C 382. *The Astrophysical Journal* 794, 62 (2014).

CONFERENCES PRESENTATIONS

John Bollenbacher, Fillipo Menczer, “Forecasting Vaccine Refusal Rates by Modeling Social Influence,” *NetSci*, Full Talk, September 2020.

John Bollenbacher, Diogo Pacheco, Pik-Mai Hui, Yong-Yeol Ahn, Alessandro Flammini, Filippo Menczer, “The empirical limits of prediction of microscopic dynamics of online conversation,” *NetSci*, e-Poster Session, September 2020.

John Bollenbacher, Fillipo Menczer, “Forecasting Vaccine Refusal Rates by Modeling Social Influence,” *IC2S2*, e-Poster Session, June 2020.

John Bollenbacher, Diogo Pacheco, Pik-Mai Hui, Yong-Yeol Ahn, Alessandro Flammini, Filippo Menczer, “The empirical limits of prediction of microscopic dynamics of online conversation,” *IC2S2*, e-Poster Session, June 2020.

OTHER PROJECTS

Comparison of spectrogram generation methods for audio classification using CNN classifiers, 2019

Used image classifiers (e.g. ResNet) to classify audio spectrograms
Generated and compared audio spectrograms

Forecasting Yellowstone Visitor Traffic, 2019

Used time series methods to forecast Yellowstone National Park visitor traffic
Performed ablation study on rich feature set

Predicting Academic Citations with Collaborative Filtering, 2018.

Used collaborative filtering methods to forecast academic citations

Radon Decay Chain, 2015.

Modeled concentrations of radioactive decay products over time for risk assessments

Modeling Gravitational Wave Detection, 2014.

Used MCMC methods to fit model parameters for gravitational wave detection events

Observations of the Faraday Instability 2014.

Used particle image velocimetry to measure internal fluid flows of Faraday waves
Wrote related software tools to enable data processing

INSTITUTIONAL SERVICE AND MEMBERSHIPS

Networks & agents Network (NaN)
Fall 2021 - Spring 2022
Position: Lab Manager

Indiana University
Bloomington, IN

Networks & agents Network (NaN)
Fall 2018 - Spring 2020
Position: Website Admin

Indiana University
Bloomington, IN

NetSci 2017 Conference
June 2017
Position: Volunteer

The Network Science Society
Indianapolis, IN

Society of Physics Students
Spring 2014 - Fall 2016
Position: Member

Georgia Institute of Technology
Atlanta, GA

ProQuest Number: 31237315

INFORMATION TO ALL USERS

The quality and completeness of this reproduction is dependent on the quality and completeness of the copy made available to ProQuest.



Distributed by ProQuest LLC (2024).

Copyright of the Dissertation is held by the Author unless otherwise noted.

This work may be used in accordance with the terms of the Creative Commons license or other rights statement, as indicated in the copyright statement or in the metadata associated with this work. Unless otherwise specified in the copyright statement or the metadata, all rights are reserved by the copyright holder.

This work is protected against unauthorized copying under Title 17, United States Code and other applicable copyright laws.

Microform Edition where available © ProQuest LLC. No reproduction or digitization of the Microform Edition is authorized without permission of ProQuest LLC.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346 USA