A Model of Social Norm Dynamics

Kristopher Overbo

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2025

Reading Committee:
Michael Hechter, Chair
Caitlin Ainsley
Cynthia Levine

Program Authorized to Offer Degree:
Individual Ph.D. Program

University of Washington

Abstract

A Model of Social Norm Dynamics

Kristopher Overbo

Chair of the Supervisory Committee:

Michael Hechter

Department of Sociology

This paper introduces a deterministic model of social norm dynamics, with foundations in rational choice and methodological individualism. The model complements traditional game-theoretic approaches by addressing how individual decisions aggregate to form societal norms. While game theory provides important insights into coordination and cooperation, insofar as it is formalized, it often focuses on static outcomes and small-scale situations. In contrast, the proposed model scales effectively and features temporal dynamics of norm development and stabilization. This utility-theory approach incorporates three primary forces to explain agent-level behavior: native preference, social influence, and habit formation. Native preference represents intrinsic and heterogeneous motivations, which ensure some behavioral variety, even in environments characterized by high conformity. Social influence reflects the pressure to change that individuals feel from observing the behaviors of others, which drives conformity. Habit formation stabilizes behavior over time, encouraging actions consistent with past decisions. These forces interact to explain how individuals embedded in a social environment contribute to the emergence of macrosocial patterns. A key feature of this model is the concept of "support," which captures the degree to which chosen behaviors align with a given social norm. Importantly, support is not just a measure of a

single behavior; it reflects the interrelation between various behaviors with respect to the norm. A norm is formally defined in the context of this model, and simulated environments are presented that demonstrate how changes in network structure affect the overall level of conformity to a norm within a population. While the work presented here is theoretical, the model provides a foundation for future empirical exploration and contributes to ongoing discussions on social norms.

Contents

List of Figures **v**

1 Introduction **1**

1.1 Outline 2

2 Existing Literature **4**

2.1 Norm Taxonomy and Terminology 4

2.1.1 Norm Targets and Beneficiaries 5

2.1.2 Conjoint and Disjoint Norms 6

2.1.3 Focal Action 6

2.1.4 Conventional Norms 7

2.1.5 Proscriptive, Prescriptive, and Bipolar Norms 7

2.1.6 Sanctions and Internalization 7

2.1.7 Descriptive and Injunctive Norms 8

2.2 Defining Social Norms 9

2.2.1 Coleman's Definition 10

2.2.2 Bicchieri's Definition 11

2.2.3 Other Definitions 12

2.3 Norms as Game Theory 13

2.3.1 Mechanics of Norm Emergence 15

2.3.2 Advantages and Limitations of Game Theoretic Modeling 17

2.4 Evolutionary Game Theory 19

2.5 Threshold Models 21

2.6 Other Relevant Research 22

2.6.1 Pluralistic Ignorance 22

2.6.2 Reasoned Action Approach and Behavioral Intention 23

2.6.3 DeGroot Learning 24

2.6.4 Habit Formation 25
2.7 Conclusion 26

3 The Model **27**
3.1 Notes on Notation 28
3.2 Specification 29
3.3 Structure of the Model 30
3.4 Support 32
3.5 Native Preference 35
3.6 Habit Formation 37
3.7 Influence 38
3.7.1 Exposure 39
3.7.2 Social Leverage 40

4 Macro-Properties of the Model **42**
4.1 Mean Support 42
4.2 Mean Social Pressure 43
4.3 Target Support 43
4.4 Standard Social Deviance and Standard Native Deviance 44
4.5 Norm strength 44
4.6 Defining a Social Norm 45

5 Model Dynamics **47**
5.1 The Baseline Simulation 47
5.2 Degenerate Forms 49
5.3 A Sparse Network 51
5.4 A Bridged Network 51
5.5 A Single Influential Individual 52
5.6 Disconnection from the Network 53

5.7 A Shock to Native Preferences 54
5.8 Continuous Shocks to $\mathbf{N^t}$ and $\dot{s}^t$ in a Larger Community 55

6 Relationship to Rational Choice **55**

6.1 Assumptions 56
6.2 Derivation of the Model from Utility 57
6.3 The Role of Information 61
6.4 Connecting Support and Behavior 61
6.4.1 The Single Behavior Case 62
6.4.2 The Two Behavior Case 63
6.4.3 The n Behavior Case 65

7 Comparison of the Model to Existing Work **66**

7.1 Traditional Game Theoretic Approaches 67
7.1.1 Comparison to Coleman 68
7.1.2 Comparison to Bicchieri 69
7.2 Evolutionary Game Theory 70
7.2.1 Comparison to Axelrod 71
7.3 Threshold Models 72
7.3.1 Comparison to Centola 73
7.4 Reasoned Action Approach 74
7.5 DeGroot Learning 75

8 Empirical Considerations **75**

8.1 Key Propositions 76
8.1.1 Proposition 1: Substitution Under Behavioral Restriction 77
8.1.2 Proposition 2: Long-Term Effects of Policy 78
8.1.3 Proposition 3: The Role of Influencers 80
8.1.4 Proposition 4: Endowment and Social Influence 82

8.1.5 Proposition 5: The Role of Deviation Costs in Norm Strength 83
8.2 Insights from the Model 85
8.2.1 Celebrity Influence 85
8.2.2 The Emergence of a Voting Norm 87
8.2.3 Variations in Voter Participation 89
8.2.4 Harmful Behaviors 90
8.2.5 Pluralistic Ignorance 91
8.2.6 Persistence of Behaviors in Changing Social Contexts 92
8.3 Managing Large Parameter Sets 93
8.4 Measurement of Parameters 94
8.4.1 Native Support $\dot{s}_i^t$ 95
8.4.2 Habit Formation Factor h_i 95
8.4.3 Support Weights $\vec{W}$ 95
8.4.4 Target Support τ^t 96

9 Conclusion **97**

References **99**

A Appendix: Voting as a Theoretical Case Study **107**
A.1 Voting and Welfare 107
A.2 The Paradox of Voting 109
A.3 Expressive Voting 111
A.4 Social Norms and Voting 114

B Appendix: Optimal Support with Forward-Looking Utility **116**

C Appendix: Proof of Convergence of Support **118**

D Appendix: Simulation Code **120**

List of Figures

1 A Basic Coordination Game 15

2 Three Player Common Project Game 16

3 Schematic of the Reasoned Action Approach 23

4 Schematic of the Model. 31

5 The Baseline Simulation 48

6 Degenerate Forms of the Model 50

7 Sparse Network 51

8 A Bridged Network 52

9 A Particularly Influential Individual 53

10 Removal from the Network 53

11 Shock to Native Preferences 54

12 Shocks Every Period to $\mathbf{N^t}$ and $\dot{s}^t$ in a Large Population 55

13 Voting Model Categories and Associated Reasoning 113

Acknowledgments

I would like to extend my deepest gratitude to Michael Hechter; I could not have hoped for a more knowledgeable mentor in the field of social norms. In addition to accepting the responsibility of chairing my dissertation committee, he has provided invaluable guidance. I would also like to thank my other committee members for their interest in this interdisciplinary work: Caitlin Ainsley for her insights into logic and the philosophy of science, Cynthia Levine for her lively seminars and knowledge of agent-level perspectives, and Xu Tan for her teachings and helpful critiques with respect to network theory and rational choice.

This work is the culmination of many lengthy discussions with colleagues and others who are close to me. Particular thanks of this kind are owed to Matthew Daniels for his cleverness, Promise Kamanga for his inspiration, Thor Morris for his wisdom, Peter Sand for his attention to detail, Joseph Skelley for teaching me how to think, and Mark Zachry for supporting this effort from the beginning. A debt is also owed to the late Yoram Barzel. There is no doubt that without his encouragement, insight, and support this work would not exist. I miss his company greatly.

1 Introduction

Social norms are foundational elements of human societies, guiding behavior and expectations in nearly all aspects of life. They are informal institutions that exert significant influence on actions, often without the need for legal structures or other explicit forms of enforcement. From economic transactions and political participation to everyday practices such as hygiene and social etiquette, norms shape behaviors that are crucial to the functioning of society. Even modest improvements in our understanding of the forces that shape social norms have value, given the broad implications such knowledge carries.

This paper introduces a deterministic model that attempts to describe the temporal dynamics of social norm development. Such dynamics are underserved by existing theory. This formal model was built with methodological individualism in mind, a doctrine that asserts that macrosocial outcomes should be explained through individual-level interactions (Weber, 1922). One of the model's strengths is its versatility; it can be applied to a wide variety of norms and allows for limited comparisons to be made among them. It also reliably generates long-run equilibria under time-invariant parameters, a feature that is not typically present in established approaches.

This effort attempts to address a key limitation of traditional game-theoretic approaches to social norm dynamics. While game theory has been instrumental in understanding coordination and cooperation at an individual level, it often fails to describe dynamic, large-scale societal phenomena. Specifically, traditional models tend to focus on static equilibria or small-scale interactions, which limits their empirical applicability to real-world, evolving macro-social contexts. In contrast, this model is designed to scale effectively and capture the temporal processes that characterize the development and stabilization of social norms.

This model incorporates three determinants of individual behavior that are salient to social norms: native preference, social influence, and personal habit. Native preferences are intrinsic, heterogeneous motivations that ensure behavioral variety, even in environments

with high conformity. Social influence represents the pressure individuals experience from observing others. Habit formation reinforces and stabilizes behaviors based on an individual's previous decisions. These forces interact iteratively, contributing to the emergence, persistence, or fading of collective social coordination.

A novel feature of this model is the concept of support, which measures the degree to which a set of behaviors aligns with a given social norm. Unlike simple measures of adherence to a single behavior, support reflects the interrelation of multiple behaviors relevant to a norm, offering a more nuanced understanding of how collective patterns emerge from individual actions.

Computational simulations of this model demonstrate how changes in network structure and social connectivity influence the overall level of conformity to a norm within a population. These findings highlight the role of both individual choices and social contexts in shaping societal norms.

While the focus of this work is theoretical, this research lays the groundwork for future empirical explorations. By addressing gaps in traditional approaches, this model offers an alternative tool for analyzing how norms evolve and how social systems might adapt to changes in behavior, network structure, or external shocks. The model also generates a number of propositions that serve as testable implications, guiding empirical investigations and refining its theoretical foundations.

1.1 Outline

The literature review in Section 2 covers the various formalized attempts at model social norms, drawing from classical game theory, evolutionary game theory, and threshold models. As this is an interdisciplinary work, other related research from social psychology and network theory is included due to its relevance to the proposed model. This section provides essential background for understanding the theoretical contributions of this paper, situating the proposed model within the broader context of social norm research.

In Section 3, the model is introduced in detail; the mathematical structure, notation, and key variables are described.

Section 4 explores the macro-properties of the model, demonstrating how individual behaviors aggregate to produce measurable societal-level outcomes. This section provides mathematical definitions of concepts like target support, social pressure, and norm strength as well as proposing a formal definition of a social norm. These macro-level definitions ground discussion of the properties of norms and illustrate how they are endogenized in the model. Such measures highlight the broader social implications of the model and allows for limited cross-norm comparisons.

Section 5 presents macro-dynamics using simulated networks to demonstrate how the model behaves under hypothetical social scenarios. Simulations of sparse networks, bridged networks, and the presence of influential individuals provide insight into how social norms develop and persist in various contexts. This section demonstrates the empirical potential of the model by showing how it can be used to predict and explain real-world social phenomena.

Section 6 discusses the relationship between the model and rational choice theory, grounding this work in a broader framework of utility maximization and rational decision-making. This section examines the assumptions of the model and compares them with those of traditional rational choice models. An argument is made that this model is rational. This section also explores the role of information and the link between support and behavior.

Section 7 provides a comparison of the model to existing work, highlighting the ways in which this effort incorporates other research and extends and improves upon traditional game-theoretic, evolutionary, and threshold models. This section demonstrates how the proposed model generates social norm dynamics in ways that other methods do not fully capture.

Section 8 makes a case for the model's empirical potential by offering examples of real-world applications. It lays out several propositions that emerge from the model. It also presents scenarios such as celebrity influence, voter participation, and persistent behaviors

under changing societal conditions to demonstrate how the model provides explanations for outcomes that are difficult to account for using other frameworks. Additionally, the section suggests how key parameters can be measured empirically.

2 Existing Literature

The objective of this literature review is to examine existing formalized models of social norms as well as other research areas related to the model developed in this paper. The review covers attempts to describe social norms grounded in classical game theory, evolutionary game theory, and threshold models. It also touches on relevant research in social psychology and network theory. Though some of these later research areas discussed are not models of norms per se, they are included for their relevance to the model presented in Section 3.

The review begins with a brief outline of terminology and definitions that are used in the discussion of norms across various theoretical frameworks. These definitions serve as a foundation for the subsequent discussion and highlight the diversity of treatments that norms have received. The review then evaluates game-theoretic explanations of norm emergence and stability, typically by considering them as instances of public goods problems, coordination games, and second-order free-rider issues. In addition, the strengths and weaknesses of these applications are covered.

By drawing insights from multiple disciplinary perspectives, this review aims to frame and inform the model of social norms developed in later sections. Readers already familiar with the relevant literature may choose to proceed to Section 2.7 for a concluding discussion, referring back to specific topics as needed.

2.1 Norm Taxonomy and Terminology

What follows is a breakdown of some common classification schemes and terminology related to norms. It is collected in one section for three reasons. First, it acts as a reference section to

quickly find definitions, should the reader need clarification on a term. Second, preemptively establishing such information avoids breaking the flow of later discussion with definitional sidebars. Third, to highlight the breadth of the academic discourse on this subject. This list of terms is by no means complete, and where definitional conflicts emerge, the dominant usage is favored.[1] When consensus is unclear, we defer to Coleman (1990).[2]

2.1.1 Norm Targets and Beneficiaries

A norm *target* refers to an individual or group whose behavior is directly influenced or constrained by a specific social, cultural, or legal norm. This influence regulates the target's actions, ensuring they align with certain accepted standards of behavior. Norm targets are subject to the expectations that the norm imposes, whether through formal mechanisms, such as laws, or informal social pressures.

A norm *beneficiary*, on the other hand, is the individual or group who gains from the norm's successful regulation of the target's behavior. The benefits can be material, social, or psychological, and they arise because the target's modified behavior creates a favorable or more equitable outcome for the beneficiary. These beneficiaries may include individuals directly impacted by the target's behavior or broader societal groups that experience improved conditions due to compliance with the norm.

By way of example, consider the social norm of holding a door open for someone whose hands are full. In this scenario, the beneficiary of the norm is the person carrying the groceries, while the targets of the norm are those nearby who are unencumbered.

Distinguishing between targets and beneficiaries is typically done when discussing mechanisms for norm enforcement. Beneficiaries have an incentive to modify the behavior of targets and are therefore willing to invest in encouraging the norm.

[1] e.g., Ullmann-Margalit (1977) uses different terminology, Cialdini et al. (1990) deviates from Coleman (1990) on what constitutes a "prescriptive" norm, Elster (1989) calls internalized norms "private" norms, etc.

[2] More on the centrality of Coleman can be found in Section 2.3

2.1.2 Conjoint and Disjoint Norms

Coleman (1990) describes norms as being either primarily *conjoint* or *disjoint*. He borrows these terms from set theory to represent extreme examples of the various possible ways norm target and beneficiary groups might overlap. Under a conjoint norm, the beneficiaries and targets are the same set of agents. If participating in political elections is a norm, it can be thought of as conjoint because those subjected to the social pressure to participate are presumed to benefit from the election outcomes as well. Disjoint norms, on the other hand, are those where the beneficiaries and targets are distinct groups. A norm requiring that dog owners pick up after their pet in public spaces could be considered an example, as dog owners might be burdened by the norm while the rest of the public benefits (pp. 247-248).

Conjoint and disjoint can be thought of as Weberian ideal types that rarely exist in their pure forms. Even in the examples given, one may be tempted to argue that people who are not eligible to vote may benefit from voting and dog owners also bear the cost of unclean public spaces.

Though there are various reasons one might wish to identify the actors on whom the costs and benefits fall, the primary purpose of this distinction is to group norms by the methods used to examine them. Purely conjoint norms are well-described as games of coordination and prisoner's dilemmas. Disjoint norms present a bit more difficulty. For his part, Coleman argues that the Coase (1960) theorem applies in purely disjoint cases. Rather than compensating the rights-holder in the traditional sense, the offended party simply bears the cost of exercising their socially defined right over the offender (Coleman, 1990, pp. 261, 292-295).

2.1.3 Focal Action

A *focal action* is a behavior that is the focus of a norm, rule, or social expectation. It is the specific action that individuals are expected to perform or refrain from performing in a given situation. This action is often seen as critical for maintaining order, achieving an objective,

or fulfilling a social obligation. In the norm of recycling, the focal action would be sorting and placing one's recyclables in the appropriate bins.

Behavioral coordination around a focal action is the most apparent evidence of the existence of a social norm but, as we will later see, it does not necessarily imply that a norm is at work.

2.1.4 Conventional Norms

A *conventional* norm refers to a social standard where the benefit does not stem from the specific focal action chosen, but rather from collective agreement on any focal action. A typical example is the convention that cars drive on the same side of the road. It does not much matter to anyone which side of the road people drive on, so long as they all choose to drive on the same side. In practice, this is a legal norm, but it is clear that even if the relevant law were repealed, social forces would most likely maintain the rule (Coleman, 1990).

2.1.5 Proscriptive, Prescriptive, and Bipolar Norms

This distinction concerns the nature of the behavior rule of the norm. *Proscriptive* norms discourage a particular action. For example, in many cultures it is a norm not to have loud conversations in a library. *Prescriptive* norms encourage actions, such as thanking someone who has helped you. Less commonly, a *bipolar* norm refers to a case in which the target action is to be followed in certain circumstances and not in others (Coleman, 1990; Jasso & Opp, 1997; Horne & Mollborn, 2020; Hechter & Opp, 2001).

2.1.6 Sanctions and Internalization

External enforcement of norms or *sanctioning* refers to the mechanisms through which others impose consequences (praise, punishment, or social pressure) to ensure compliance. In broad usage, this could include formal punishments, such as fines, but when referring to social

norms it is associated with informal pressure such as social disapproval or ostracism for norm violations.

Internal enforcement of norms, or *internalized* norms involve self-regulation of behavior by individuals according to beliefs, values, and feelings which come from within, such as guilt, pride, or conscience. Such feelings and beliefs engender a personal sense of reward or discomfort based on the agent's adherence to or violation of a norm.

For example, when a person drives their vehicle under the speed limit, they may be doing so because of the external risk of getting a traffic ticket, or they may have internalized the understanding that following this law is the responsible thing to do. Coleman treats internalization as the process by which disjoint norms become conjoint norms. Norm beneficiaries deliberately spend effort to socialize the desired behaviors into norm targets, which reduces long-term enforcement costs (Coleman, 1990, pp. 293-394).

Irrespective of the cause, internalization ensures that norms persist in the absence of sanctions.

2.1.7 Descriptive and Injunctive Norms

Cialdini et al. (1990) make a distinction between norms whose force comes through regularity of behavior and those that are powered by the perceived appropriateness of a behavior. The former is called *descriptive* and the latter *injunctive*. Descriptive norms are maintained by people using the behavior of others as a heuristic, while injunctive norms are powered by sanctions.

Descriptive norms are closely related to conventional norms, but they are distinguished by their basis in observed behavior rather than agreed-upon rules. Descriptive norms reflect what people commonly do, while conventional norms depend on shared expectations and social agreements about appropriate behavior in specific contexts.

An example of a descriptive norm might be a fad exercise program. As people notice more individuals participating in these exercises, whether through social media, fitness

classes, or seeing friends and acquaintances engaging in them, they may infer that these workouts must be effective or provide significant benefits, even if they do not fully understand the underlying science. The visibility of others performing these exercises signals that they are somehow good for fitness or well-being, leading more people to try them out. This creates a feedback loop, where the perceived effectiveness of the exercise boosts its adoption, and this increased popularity reinforces the assumption that the behavior is beneficial. In this case, there is no enforcement or sanction involved.

An example of an injunctive norm could be the norm of dressing in a particular manner for certain social occasions. For instance, in many cultures, wearing formal attire to a wedding is expected. If someone arrived in casual clothes, they would likely experience disapproval or subtle social pressure from their peers.

With respect to game theoretical literature, the novelty in noting this division is that identifying a norm as descriptive introduces a new motivation for adhering to a norm. Instead of threats of potentially costly sanctions or internalization, the focal behavior is followed because its popularity signals that it is a productive or sensible action for the individual, rather than simply an externally demanded action. This terminology remains popular in psychology and has also made its way into other fields (Bicchieri, 2005; Hechter, 2008; Compernolle, 2017; Horne & Mollborn, 2020; Legros & Cislaghi, 2020).

2.2 Defining Social Norms

In the study of any topic, it is important to have a shared definition of what is being discussed. This is especially desirable if one wishes to apply formal modeling. Unfortunately, standardizing a definition of a "social norm" has proven challenging for academics:

> "Truth to tell, there is no standard definition of the term" (Hechter & Opp, 2001, p. 402).

Similarly, Horne notes the lack of agreement among social scientists:

> "Whereas the concept is one of the most widely used in the social sciences, there is little consensus about what norms are and how they emerge" (Horne, 2009, pp. 2-3).

Villatoro et al. echo these sentiments:

> "There is still not a clear definition of what social norms are and the types of problems they solve" (Villatoro et al., 2010, p. 1).

In the absence of agreement on this subject, we will focus our attention on two definitions that are particularly relevant for our purposes, those of Coleman (1990) and Bicchieri (2005). The first for the popularity of its source material and the second for its precision. For each, we will briefly discuss the limitations that prevent them from serving as universal definitions, illustrating the challenges in establishing a single, agreed-upon standard.

2.2.1 Coleman's Definition

One influential definition comes from Coleman (1990), who offers up this as his "explicit definition of a norm":

> "A norm concerning a specific action exists when the socially defined right to control the action is held not by the actor, but by others" (p. 243).

By this criterion, each norm is tied to a specific focal action (pp. 246-250). It emphasizes that norms represent a transfer of control of one's own actions to others in society.

There are a few semantic nuances and limitations to Coleman's definition. First, rather than directly defining a norm, Coleman describes the conditions under which a norm exists. Describing norms by their manifestations rather than their underlying properties potentially obscures the more fundamental nature of norms, much in the manner describing an illness exclusively by its symptoms might. The reader may suppose that Coleman means that a norm is not just evidenced by a transferred right, a norm *is* a transferred right. If that is

true, the definition could be both simplified and clarified by stating so. There is also some ambiguity around what is meant by "socially defined right." He does elaborate that these rights are informal—neither legal nor contractual—but socially recognized and backed by sanctions, although he does not expand further on this (pp. 243, 266).

Another point of critique is that Coleman's definition does not address the role of regularity in norms. According to his framework, if an individual is socially pressured into an unusual action that others do not typically perform, it would still qualify as a norm. For example, if a student is bullied into eating paste by their classmates, Coleman's definition would categorize this behavior as a norm, despite its lack of regularity. Despite these minor issues, Coleman's definition is a useful attempt at generalization. It is simple and broad, with intuitive appeal.

2.2.2 Bicchieri's Definition

Bicchieri (2005) offers perhaps the most ambitious and specific definition of social norms. It is notable for applying some mathematical rigor to the task:

> Let R be a *behavioral rule* for situations of type S, where S can be represented as a mixed-motive game. We say R is a social norm in a population P if there exists a sufficiently large subset of $P_{cf} \subseteq P$ such that, for each individual $i \in P_{cf}$:
>
> **Contingency**: i knows that rule R exists and applies to situations of type S.
>
> **Conditional preference**: i prefers to conform to R in situations of type S, under the condition that:
>
> (a) **Empirical expectations**: i believes that a sufficiently large subset of P conforms to R in situations of type S.
>
> (b) **Normative expectations**: i believes that a sufficiently large subset of P expects i to conform to R in situations of type S;

or

(b') **Normative expectations with sanctions**: i believes that a sufficiently large subset of P expects i to conform to R in situations of type S, prefers i to conform and may sanction behavior.

(Bicchieri, 2005, p. 11)

Bicchieri argues that a rule qualifies as a norm if enough people follow it, if individuals believe others are following it, and if they believe they are either expected to comply or face sanctioning for non-compliance.

Despite its admirable nuance and forked logic, this definition does not account for all types of phenomena commonly considered normative. For instance, she excludes descriptive norms, providing fashion trends as an example. She declares that such norms are not social norms, but a distinct phenomenon. Interestingly, she declares that such norms carry no expectation from others that an individual conforms (Bicchieri, 2005, pp. 29-30).

2.2.3 Other Definitions

Consistent with the lack of consensus, various other definitions have been put forth by researchers. Indeed, this paper will contribute to that effort in Section 4.6. Many authors provide brief, one-sentence definitions of social norms, leaving clarification to be inferred from their broader work. The variety in definitions is understandable, as each researcher tailors the concept of norms to their specific study as appropriate. Still, this lack of consensus is unfortunate, given the importance of norms in social science research. A common, standardized definition would reduce confusion, make it easier to integrate contributions from diverse studies, and advance the field more cohesively.

Recognizing this difficulty, Opp (2001) provides a partial solution. He presents a framework for classifying definitions by identifying three common elements used in the description of norms:

Oughtness: A shared sense within the population as to what behaviors are expected.

Behavioral Regularity: Coordination of behavior across a population.

Sanctioning: A system that provides incentives for certain behaviors and disincentives for others.

Some combination of these components serve as key elements in most definitions of social norms. This paper will make use of these terms occasionally. Rather than further catalog the variety of existing definitions, the reader is referred to Opp's paper on the subject.

2.3 Norms as Game Theory

In the formal modeling of social norms, game-theoretic approaches have emerged as the dominant framework. Early contributions of this sort came largely from economics. Schelling (1960) introduced the idea that social norms could be understood as coordination problems. Sugden (1986) extended this work by applying it to micro-scale situations. Among the foundational works in this area, though, Coleman's *Foundations of Social Theory* (1990) stands out as arguably the most influential.[3] Coleman's insights have had a lasting impact, with both social psychology and economics frequently drawing on his ideas (Frank, 1992; Keefer & Knack, 2008; Anderson & Dunning, 2014). In sociology, the field to which Coleman himself belongs, much of his approach and terminology remain prevalent in the recent literature. Therefore, it is natural to center our discussion on Coleman's work.

According to Coleman, there are two requirements for the formation of a norm. There must be demand for the norm and there must be a practical way for that demand to be satisfied. His description of demand is straightforward. If a particular action taken by someone results in an externality, or in other words, if others can benefit from controlling that behavior, we say that such demand exists. For many actions, this condition is easily met. Indeed, it is difficult to imagine a choice that one could make that would not have some impact on others. Simply existing consumes resources that could be consumed by another.

[3]Building on earlier contributions by Ullmann-Margalit (1977) and Schelling (1978).

Presumably, it is for this reason that a large majority of game theoretical work concentrates on the conditions for the satisfaction of existing demand (pp. 266-299).

Coleman's exploration of demand satisfaction heavily features norms as a solution to the public goods problem. He models this as a prisoner's dilemma. In the two-person game, the optimal strategy of defect-defect (where defect equates to deviating from the focal action), depends on the assumption that players cannot communicate or negotiate a different outcome. Dropping this assumption allows players to find a solution to the free rider problem. He goes on to claim that, in practice, the free rider problem is solved by coordination through social relationships and the presence of zealous sanctioners, resulting in a socially beneficial outcome (pp. 252, 256, 273-274). This explanation for the emergence of norms will be discussed in more detail in section 2.3.1.

Some have made note that Coleman's treatment only covers norms that are collectively beneficial, leaving little room for norms that are not born of necessity, happenstance, or deliberate human effort (Elster, 2003; Opp, 2018). Notably, Coleman's definition of a norm does not explicitly exclude neutral or harmful norms. It is therefore unclear whether this is an omission in the model or that it is assumed that all norms must provide benefit to at least some subgroup. In either case, the issue is left unaddressed. Elster (2003) is particularly critical of Coleman, "I believe his treatment is deeply unsatisfactory. It is a piece of crypto-functionalism, in spite of his official rejection of that method and his professed methodological individualism."

Despite these criticisms, Coleman's terminology and methodology are widespread today, with most recent research on the microfoundations of norms making use of, and/or extending his approach. A variety of game theoretic descriptions have emerged along with the requisite terminology for classifying and labeling them. Depending on the environment and incentive structure, norms have been variously described as prisoner's dilemmas, pure coordination, Stag-hunt, and other games of varying levels of complexity (Coleman, 1990; Young, 2015; Fallucchi & Nosenzo, 2021; Bicchieri, 2005; Jindani & Young, 2020).

		A_2	
		a	b
A_1	a	$(3,3)$	$(0,0)$
	b	$(0,0)$	$(3,3)$

Figure 1: A Basic Coordination Game

The contribution of Bicchieri (2005) is noteworthy for its more social approach. Rather than treat game-theoretic payoffs as exogenous, she extends the microfoundations of those payoffs by emphasizing that individuals are motivated by social expectations. She proposes that an individual will modify their behavior to follow a norm if they believe others are following the norm (what she calls empirical expectations) and that others expect them to follow the norm (so-called normative expectations). This contrasts with most previous work, as it suggests that utility functions are shaped by a desire to align behavior with the expectations of others. In her view, social context is highlighted as central to decision-making as opposed to more conventional models which emphasize internally focused self-interest.

2.3.1 Mechanics of Norm Emergence

Next, we will discuss how game theory explains norm emergence. By examining the strategic interactions that give rise to behavioral regularities, we are afforded insight into the processes that lead to their establishment and persistence.

Explaining the formation of conventional norms is rather simple. They are modeled as coordination games where the payout is dependent upon each player choosing the same strategy (see Figure 1). There are two principal methods proposed to solve this sort of game in practice. First, the agents can communicate and agree on which option to choose. If they can discuss their decisions before they make them, coordination becomes trivial. Barring communication, focal equilibria might be identified the manner illustrated by Schelling

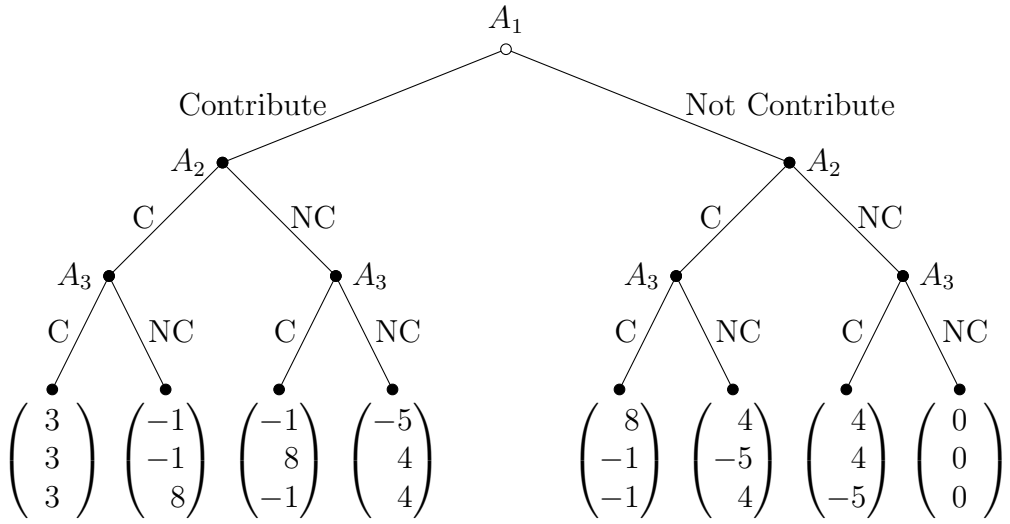


Figure 2: Three Player Common Project Game

(1960): In the absence of communication, certain solutions stand out because they are naturally more salient or obvious to all parties involved. The standard illustration of this would be a meeting in Paris: Two Americans are visiting Paris. They are in different locations and are unable to communicate with each other. They wish to meet, but they have not discussed a meeting location beforehand and the city is large with countless possible meeting spots. When presented with this problem, and knowing that the other is faced with the same problem, each is likely to go to the Eiffel Tower, which serves as a focal point owing to its iconic status, or salience with respect to Paris. Likewise, if any particular salience is assigned to a behavior in the choice set of a coordination game, agents will gravitate to that strategy, dictating the equilibrium outcome.

Non-conventional norms are most commonly illustrated as prisoner's dilemmas. In this game the first-order problem is that for any given player, only sub-optimal decisions from other players can improve her lot. Because of this some method is needed to ensure that each player cooperates despite their incentive within game the to defect. That nobody has an incentive to contribute to a system of controls to override this issue is the second-order free rider problem. The largest scale models formally expressed with respect to norms take the form of a three-player game. Under the appropriate payouts, such as those in Figure

2, each pair of players, knowing that the third would defect, can collude to sanction the third player. It happens that each of those games themselves is a prisoner's dilemma. In principle, these could be solved by a functioning market or pairwise contract enforcement, enabling an outcome of full cooperation. Under this reasoning, contracts, whether formal or informal, could explain norms. However, the plausibility of such a proposition is stretched to an extreme when we consider the scales on which real world norms exist. A norm across a population of just 100 individuals would likely require hundreds of pairwise contracts of nontrivial complexity to reach critical mass. Without a consensus in the literature on how equilibria are achieved, this theoretical avenue remains incomplete. Other researchers have made similar observations (Yamagishi, 1986; Yee, 1997).

Since societies often manage to solve public goods problems in the real world, it is tempting at times to take as an assumption that the issue of free ridership is solved exogenously. However, this assumption would beg the question, as solving the free rider problem is itself a public goods problem.

2.3.2 Advantages and Limitations of Game Theoretic Modeling

Game theory offers a powerful framework for understanding social norms, particularly due to its expositional clarity and its ability to provide simple, formal models of human interaction. By reducing complex social behaviors into strategic games, game theory allows researchers to simplify and clarify the exposition of norm formation and adherence, despite the apparent complexity of the process.

Another major strength of applying game theory to norms lies in the rich research ecosystem that has developed around it. Since its introduction, game theory has been applied and extended across disciplines such as economics, sociology, psychology, and political science. This vast body of prior work allows scholars to draw on a wealth of existing models, techniques, and empirical findings when applying game theory to social norms. This ecosystem offers a solid foundation from which researchers can build, test, and refine new theories

about norm dynamics.

Additionally, game theory adds rational foundations to the study of social norms by framing them as outcomes of strategic decision-making. Unlike other approaches that may rely on less formal or intuitive explanations of norm adherence, game theory anchors the analysis of norms in rational choice theory. This grounding provides improved theoretical depth to explanations for how norms emerge, persist, and change over time. By assuming that individuals act in their own self-interest, game theory provides a logical basis for explaining why people conform to norms, even when doing so may require costly sacrifices. The concept of equilibrium, central to game theory, shows how stable patterns of behavior can arise from individual decision-making processes, when such equilibria can be found.

While game theory has proven valuable for understanding strategic interactions, it is not without limitations when applied to social norms. Coleman's work, for example, frames norm internalization as a mechanism of cost reduction imposed by others, but he does not describe the process. He also does not offer an explanation that can account for why individuals would willfully accept internalization of norms that are costly to them or contradict their private beliefs. Additionally, his treatment of disjoint norms, where the beneficiaries and targets are separate groups, focuses heavily on coercive enforcement, overlooking more subtle forms of social pressure and influence.

More generally, traditional game-theoretic approaches, such as coordination games or prisoner's dilemmas, focus on static equilibria and are often restricted to small-scale interactions. Real-world norms, however, involve complex, multi-agent dynamics that evolve over time and across large populations. Game theory's narrow focus on payoff maximization and fixed strategies also fails to capture the fluid, adaptive nature of norms.

Another weakness is that game-theoretic models often overlook the emergence of harmful or neutral norms. As Coleman and others note, game theory assumes that norms must provide some benefit to the group, ignoring the fact that many norms appear to be suboptimal or even damaging to collective welfare. These harmful norms challenge the func-

tionalist assumptions that exist throughout the literature, revealing a gap in how we conceptualize norm emergence and sustainability. Substantial effort has been made to address this issue, but such solutions generally result in further exception handling rather than incorporation into a unified theory (Hechter & Opp, 2001; Opp, 2001; Bicchieri, 2005; Horne, 2009).

Furthermore, game theory struggles with the diversity of norms across different contexts. The highly specialized taxonomy of norms in existing models, non-conventional vs. conventional, conjoint vs. disjoint, prescriptive vs. proscriptive, etc., often results from the constraints of fitting norms into predefined game structures. These categories can fragment our understanding rather than unify it. They also put strain on the empiricist to apply or hybridize these models for their unique research task.

While game-theoretic explanations provide valuable insights into the emergence and equilibrium behavior of social norms, they often fall short in addressing the dynamic and macro-level effects that shape norm persistence and evolution over time. These limitations underscore the need for models that can better generalize across different types of norms and incorporate ongoing dynamics. In the next section, we shift focus to evolutionary game theory, which partly addresses these issues by modeling how norm adherence might evolve in populations over long periods of time.

2.4 Evolutionary Game Theory

In this section, we explore the application of evolutionary game theory (EGT) to the study of social norms. Unlike traditional game theory, which focuses on static equilibria and individual decision-making, EGT emphasizes population-level dynamics and intergenerational change. EGT models capture the adaptation of strategies and behaviors within a population. Through this lens, one can study broader macro-level forces that shape social norms.

Strictly speaking, EGT also qualifies as game theory, and it is again being applied to social norms. However, this approach is distinct enough to warrant separate treatment.

Though it does introduce macrosocial scale predictions, its microfoundations are quite different. Under EGT, a population of agents is assigned fixed strategies that they then employ in simulated interactions with each other. In the case of social norms, those strategies are generally whether or not to follow a norm with respect to a focal behavior. The payout each agent receives after the conclusion of a round of games determines their ability to replicate in the next generation. One simple and intuitive replication function might be that, in the subsequent generation, strategies are assigned to a new population in proportion to the total payouts received by that strategy in the previous generation. This process of interaction and replication is repeated, and the change in the distribution of strategies over time is observed (Axelrod, 1986; Bowles & Gintis, 2011; Jindani & Young, 2020).

This method introduces some interesting features, notably intertemporal dynamics. This evolution of behaviors in a population over time is difficult to replicate using classic game theory. The stochastic approach of EGT also avoids some of the difficulty in finding equilibria. The trial-and-error loop often, but not always, results in an evolutionarily stable distribution of strategies. Examining the parameter values under which various dominant strategies emerge can tell us something about why certain equilibria are favored over others.

There are important issues that do not make this approach well suited for some purposes, though. Importantly, the mechanics of EGT are far removed from our intuitions on how norms and individuals actually work. In this sense, it is not microfounded. Furthermore, a critical input to these models is the replicator function, which dictates the reproductive process. Where norms are concerned, this has no clear empirical analog.

While EGT does show population dynamics, this evolution does not include agent dynamics. Rather, agents are treated as having a fixed strategy. While there are good arguments that can be made for bounded rationality or imperfect information, this treatment discards all sense of rational calculation in the individual. Instead, it offers an evolutionary explanation for purely behavioral choices. It is for this reason, presumably, that EGT has found the most success in ecology.

2.5 Threshold Models

The threshold models are perhaps the most underdeveloped area of literature we will address. This class of computational simulations makes frequent use of social network structures. It is rooted in, and named for, Granovetter's threshold model (1978). Granovetter possibly drew inspiration from Conway's Game of Life and other cellular automata which were of growing interest at the time (Codd, 1968). Such models posit that agents make binary decisions, such as whether to follow a norm, based on observing the same behavior in their network neighbors. Individual thresholds, which vary across agents, play a critical role. These thresholds represent the proportion of neighbors who must adopt a norm before the observing agent also does. Heterogeneity is central in determining the tipping point at which norm adoption cascades through a population. Once enough individuals adopt a norm, a local "critical mass" is reached, triggering rapid diffusion across the network.

This diffusion is, of course, sensitive to network topology. Building on this topic, Centola (2005; 2018) examines how different network structures, such as clustered versus random networks, influence the spread of norms. Centola's concept of "complex contagions" adds nuance by showing that norms requiring reinforcement from multiple contacts (i.e., strong ties) spread more effectively in tightly connected clusters. This contrasts with Granovetter's original emphasis on weak ties, which are more effective in spreading information, or simple contagions. Centola's work asserts that different types of social interactions (weak versus strong ties) play distinct roles depending on the complexity of the behavior being adopted.

Further extending these models, Mäs & Opp (2016) incorporates flexible thresholds, allowing individuals' decisions to change dynamically based on evolving social and personal contexts. They explore how revealing or hiding norm violations can rapidly alter social behaviors within networks.

Broadly speaking, this area of research focuses on the importance of network topology combined with heterogeneous thresholds for norm participation to explain the diffusion and evolution of social norms.

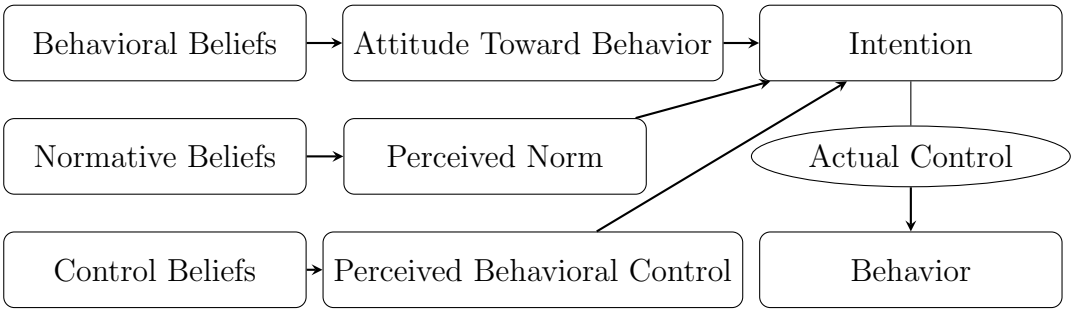
Threshold models, while valuable for exploring the dynamics of norm diffusion in social networks, exhibit several notable limitations. Their reliance on binary decision-making oversimplifies the complexity of real-world behaviors, which often involve degrees of adherence rather than all-or-nothing choices. Additionally, these models typically assume static thresholds for agents, failing to capture how individual behaviors and social pressures evolve over time. Furthermore, the lack of empirical alignment in many threshold models makes it challenging to validate their assumptions or apply their findings directly to real-world scenarios. These limitations highlight the need for more integrated microfoundations that detail the behavioral or rationalized processes that may be at work.

2.6 Other Relevant Research

2.6.1 Pluralistic Ignorance

As we have discussed, harmful social norms are particularly challenging to explain in a game theoretical framework. Perkins & Berkowitz (1986) noted that people often misperceive the behaviors and attitudes of those around them with respect to social norms. They found that students' reports about the popularity of drinking alcohol on college campuses was much higher than actual measures of alcohol consumption. Observations of this sort have been used to justify a policy intervention called the *social norms approach,* which is intended to undermine certain norms, such as those concerning tobacco or alcohol consumption. The approach focuses on providing accurate information to reduce the undesirable behavior. The underlying theory, first proposed by Katz et al. (1931), assumes that people have *pluralistic ignorance*, systematic biases which lead them to misperceive what typical behavior is. That, in turn, encourages them to act in accordance with an illusory standard, turning the imagined behavior into a real norm (Berkowitz, 2005; Bicchieri, 2005). Kuran (1995) attempts to add game theoretical grounding for this concept. He claims that agents strategically falsify their preferences. This leads to widespread pluralistic ignorance, which he calls "private knowledge distortion."

2.6.2 Reasoned Action Approach and Behavioral Intention



Adapted from Fishbein & Ajzen (2010).

Figure 3: Schematic of the Reasoned Action Approach

While neither a formal model, nor a model of social norms per se, the *Reasoned Action Approach* (RAA) from social psychology provides a well-studied theoretical basis for how individuals use norms to make decisions about their focal actions at the micro level. RAA is the modern incarnation of the earlier Theory of Planned Behavior and Theory of Reasoned Action (Fishbein & Ajzen, 2010). According to this view, behavior with respect to a specific focal action is determined primarily by what is called *behavioral intention* which itself is a function of three inputs:

1. Attitude toward personally performing the behavior.
2. Perceived behavioral control.
3. Perceived norm.

These determinants themselves are shaped by beliefs and perceptions of the individual in question. *Attitude toward behavior* represents a personal desire to perform the action based on one's isolated interests. This can be thought of as a sort of naive intrinsic utility which does not consider external social costs and forces[4]. *Perceived behavioral control* represents the influence of one's estimates of ability and circumstance. The *perceived norm* represents the result of the calculus of extracting social expectations (both injunctive and descriptive

[4]It is, however, worth noting that RAA does not assume rationality.

norms) from those around them. Once the strength of one's intention is established, shocks, errors, or bias in behavioral control, that is *actual control*, mediates how likely the individual is to take the focal action. In the sense that it takes the norm as exogenous this can be considered a well-developed extension of the model presented by Akerlof (1980).

2.6.3 DeGroot Learning

Also not a model of social norms, but mathematically similar to the model presented in this paper, is one by statistician Morris DeGroot (DeGroot, 1974). This model describes how individuals in a social network update their opinions or beliefs about what is true through interpersonal contact. The model captures the process by which a group of agents iteratively adjust their beliefs based on a weighted average of the opinions of their neighbors and their own prior beliefs. Though this process superficially resembles Bayesian updating, it differs in that agents do not incorporate likelihoods or new evidence. Instead, they simply update based on fixed interpersonal weights. Eventually, agents converge on persistent consensus or disagreement, depending on the network structure and the influence of endogenously set weights. The procedure can be described mathematically as follows.[5]

$$s_i^t = \sum_{j=1}^{n} w_{ij} s_j^{t-1} \tag{1}$$

where

$$\sum_{j=1}^{n} w_{ij} = 1, \ \forall i \in [1, 2, ..., n]. \tag{2}$$

The w_{ij} values represent the weight agent i places on the opinion of agent j. s_i^t is the expressed opinion held by agent i in period t. n is the number of agents in the network. Each period, opinions are updated until equilibrium is achieved.

This is not generally considered a rational model because there is no deliberate optimization done by the agents to arrive at the weights they assign to each opinion. Despite

[5]I have changed the notation from the source material to more closely match that of this paper. The purpose of doing so is to clarify the functional similarity.

this, this model is remarkably simple and intuitive. There is also a modest body of empirical evidence emerging that individuals do learn this way (Jadbabaie et al., 2012; Chandrasekhar et al., 2020).

The model presented in this paper is most structurally similar to that of Friedkin & Johnsen (1990), a variant of the DeGroot Learning Model which adds static preferences, though it differs in application, elements of time variance, and theoretical explanation of its components. See Section 7.5 for further discussion.

2.6.4 Habit Formation

The topic of habit formation is included in this discussion because it is a crucial element of the model developed in this paper. Habit will play a central role in stabilizing individual behaviors over time, contributing to the persistence of social norms even in the face of changing preferences or external conditions. By anchoring actions in past behaviors, habits reduce the volatility of individual responses, creating continuity in norm adherence.

There is a diverse body of research providing theoretical justification for habit formation. Becker & Murphy (1988), outlines rational foundations. In this abstract view of habit, it can be understood as a rational response to previous behaviors, where the utility derived from certain actions increases with its prior consumption. Under this framework, past behaviors synergize with matching current behaviors resulting in increased utility over time.

Alternatively, Cognitive Dissonance Theory and Balance Theory offer psychological explanations for habit formation. These frameworks suggest that individuals experience discomfort when their actions are inconsistent with their past behaviors. To resolve this psychic dissonance, they internalize justification for the behavior which then persists (Festinger, 1957; Heider, 1958).

From a behavioral perspective, habit formation is also related to the concept of automaticity, where actions become automatic responses to environmental cues over time. When

behaviors are repeated regularly, they require less cognitive effort and become ingrained as default choices. This phenomenon reduces the mental costs associated with decision-making, as agents no longer need to consciously evaluate their choices in each period. Instead, they rely on past behaviors as heuristic shortcuts, allowing them to respond efficiently to recurring situations without the need for deliberate thought (Verplanken & Aarts, 1999).

There is substantial empirical support for the habit-forming nature of social behaviors which are related to norms such as voting, handwashing, and others (Coppock & Green, 2016; Fujiwara et al., 2016; Hussam et al., 2016; Lally et al., 2010).

2.7 Conclusion

In this literature review, we explored various theoretical approaches to understanding social norms. Game theory has provided foundational insights, particularly in explaining coordination problems and collective action. However, traditional game-theoretic models have limitations, such as in their weak treatment of harmful norms and their static nature, which often fails to explain the dynamic processes that characterize the evolution of norms over time. Evolutionary game theory offers an alternative perspective by introducing population-level dynamics, but its lack of adaptive behavior at the individual level constrains its applicability to social norms. This is because in practice social norms routinely appear and disappear within sub-evolutionary timescales. Threshold models, particularly those incorporating network structures, show promise in explaining norm diffusion, but they disregard the nuance of real-world social interactions.

The limitations identified in the existing literature signal the need for a more comprehensive and adaptable model of social norm dynamics. Coleman advocates for methodological individualism, but he notes his theory is not complete in this regard. In his words,

> "No assumption is made that the explanation of systemic behavior consists of nothing more than individual actions and orientations, taken in aggregate. Furthermore, there is no implication that for a given purpose an explanation must

be taken all the way to the individual level to be satisfactory" (Coleman, 1990, p. 5).

While he is not incorrect, there is clearly benefit in linking individual actions to large scale phenomena in a more comprehensive theory.

Utility theory provides a foundational framework for modeling rational decision-making, while classical game theory elaborates on strategic interactions under specific conditions, often focusing on static equilibria. EGT extends this by incorporating population-level dynamics and long-term adaptation in an effort to find equilibria. Threshold models complement these by highlighting the network-dependent diffusion of behaviors, accounting for heterogeneous thresholds of adoption. However, disconnections between these frameworks limit their ability to combine under methodological individualism empirically. Their isolated approaches often fail to integrate the interplay between temporal dynamics, network structures, and agent-level adaptation, leaving a fragmented theoretical landscape that struggles to connect micro-level behaviors with emergent macro-level patterns.

The next section of this dissertation introduces a model which aims to address this issue by integrating elements of existing theories into a flexible framework capable of explaining both the micro-level interactions and macro-level social trends that govern norm formation, persistence, and change.

3 The Model

This section lays out both the mathematical description and the conceptual framework of the model this paper proposes. This model integrates several key ideas from existing theories: psychological inputs from the Reasoned Action Approach and Bicchieri, the mathematical structure of DeGroot learning, network incorporation from threshold models, and habit formation as a stabilizing force.

This model is unique for its emphasis on three main determinants of individual behav-

ior: native preference, social influence, and habit formation. Native preferences are intrinsic motivations that introduce behavioral diversity, even in highly conforming environments. Social influence accounts for the pressure individuals feel when observing the actions of others, while habit formation helps to stabilize behaviors based on previous decisions. The interaction of these forces generates dynamics that describe how societal norms emerge, persist, or fade.

What follows is a detailed explanation of the model. It begins by briefly establishing notation used throughout the rest of this paper. The discussion then opens with the formal specification of the model, describing how its central components interact at the agent level. After a brief overview of the mechanics at work, each major component of the model is discussed in turn.

3.1 Notes on Notation

Unless otherwise stated, the following conventions are used for notation in this work:

Matrices and Vectors:

Capital letters are used for both matrices and vectors.

Matrices are denoted in **bold**, e.g., $\mathbf{X}$.

Vectors are denoted with an arrow, e.g., $\vec{X}$.

Scalars: The scalar components of matrices and vectors are represented by the corresponding lowercase letters. For example, the scalar components of a matrix $\mathbf{X}$ would be x_{ij}, where i and j index the rows and columns, respectively.

Macrosocial Measures are denoted with Greek letters (e.g., α, β, γ). These variables represent aggregate social dynamics or macro-level indicators.

Unweighted Means are denoted with a bar, e.g., $\bar{x}$. These are simple population-wide means of corresponding values.

Choice variables are denoted with a tilde, e.g., $\tilde{x}$. These represent variables over which agents have control in optimization problems.

Exogenous variables: A dot notation is used to distinguish between different kinds of exogenous variables. Internal agent properties are represented as $\dot{x}$, while observable neighbor measures are represented as $\ddot{x}$.

3.2 Specification

Let $\mathbf{N^t}$ be a p by p matrix whose elements, $n_{i,j}^t$ correspond to the weights on the edges of a directed network at discrete time t. The *support* s_i^t of agent i at time t shall be recursively defined:

$$s_i^t = n_{i,i}^t\big((1-h_i)\dot{s}_i^t + h_i s_i^{t-1}\big) + \sum_{\substack{j=1,\\ j\neq i}}^{p} n_{i,j}^t s_j^{t-1}, \tag{3}$$

with the following conditions true for all $i \in \{1, 2, 3, ..., p\}$:

$$n_{i,j}^t = e_{i,j}^t l_{i,j}, \forall j \neq i, \tag{4}$$

$$\sum_{j=1}^{p} n_{i,j}^t = 1, \tag{5}$$

$$h_i, e_{i,j}^t, l_{i,j} \in [0, 1], \tag{6}$$

$$s_i^0 = \dot{s}_i^t. \tag{7}$$

The following variables are endogenous:

$n_{i,j}^t$: weight of influence of agent j on agent i. The weight or impact that one agent's past behavior has on another agent's current behavior.

s_i^t: support of agent i at time t. A scalar measure of the alignment of an agent's behaviors with respect to a social standard.

The following are exogenous parameters:

p: population size.

h_i: weight of habit formation for agent i. The extent to which an agent's self-influence is dictated by their own past behavior.

$\dot{s}_i^t$: native support of agent i at time t. The agent's intrinsic level of alignment with a social norm, independent of social influences or past behaviors.

$e_{i,j}^t$: exposure of agent i to agent j at time t. The level of an agent's attention directed toward another agent in the social network.

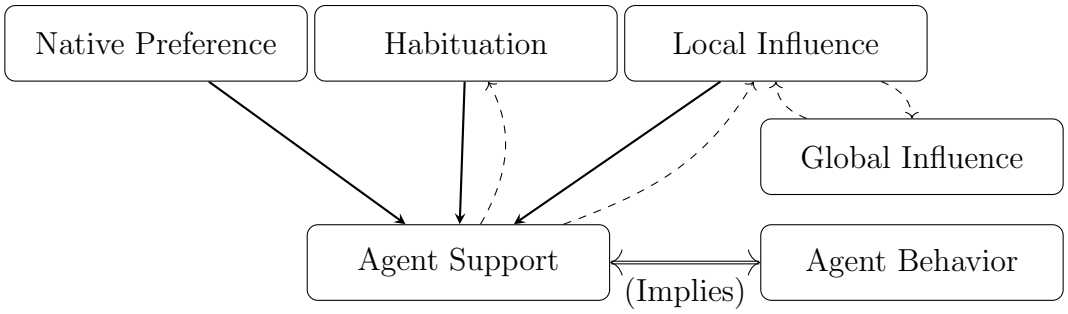
$l_{i,j}$: social leverage of agent j over agent i. The strength of influence that one agent holds over another, reflecting the degree to which the actions or opinions of the influencing agent can affect the support of another.

The diagonal elements (n_{ij}^t where $i = j$) represent the proportion of the agent's influence attributable to non-social factors in period t. These values are directly implied by conditions (5) and (4). Conditions (5) and (6) normalize weights to ensure convergence under stable parameters. Condition (7) initializes the model.

Next, we will discuss the general structure of this model.

3.3 Structure of the Model

When an individual makes a behavioral decision under this model, they go through a process shaped by three key forces: native preference, habit, and social influence. First, the agent considers their native preference. This represents the individual's internal motivations and desires. It is what they would naturally prefer to do if they existed in social isolation. Next,



Solid arrows indicate intra-period effects, dashed arrows represent inter-period effects.

Figure 4: Schematic of the Model.

the agent reflects on their past behavior. Choices that have been made in the past are valued for their familiarity and inertia. Finally, the agent evaluates social influence, which comes from observing the behavior of others. They scan their social environment, noting the support levels and relative importance of each individual they observe. Actions performed by admired individuals or friends carry more weight in this regard.

These three forces factor into the agent's decision-making process. The final behavioral choice emerges as the product of this balancing act. Over time, this interplay, repeated across agents and time, can regulate behaviors into the society-wide patterns we think of as social norms.

The three terms on the right side of Equation 3 correspond to the three factors determining support. Here we have distributed the $n_{i,i}$ factor below to clarify the distinction.

$$s_i^t = \overbrace{(1-h_i)n_{i,i}^t \dot{s}_i^t}^{\substack{\text{Native} \\ \text{Preference}}} + \overbrace{h_i n_{i,i}^t s_i^{t-1}}^{\text{Habit}} + \overbrace{\sum_{\substack{j=1, \\ j\neq i}}^{p} n_{i,j}^t s_j^{t-1}}^{\substack{\text{Social} \\ \text{Influence}}}. \tag{8}$$

After the $t = 0$ initial state, the support chosen by agent i in each period is a weighted average of three other support values that are observable to i: native preference, habit, and social influence. The $n_{i,j}^t$ weights indicate the extent to which each agent in the network, including themselves, influences future support, while the h_i weights further divide the non-social influences into the other two component parts. The support level generated by this

function feeds into the $t+1$ period behaviors through the habit term for the same agent and the social influence term for other agents connected directly via an influence network $\mathbf{N^{t+1}}$. Through this social influence channel, the time t support of a given agent affects neighbors at distance d at time $t+d$ and as well as providing some amount of self-feedback in alternating periods. This speed-of-light effect provides additional dynamics and links individual behavior to macrosocial effects. See Figure 4.

This model is generally concerned with dynamics, consistent with an expectation that large social systems frequently experience shocks. Still, it is worth noting that the model itself is not inherently divergent. The model reaches equilibrium in the absence of shocks to N^t and $\dot{s}^t$ as $t \to \infty$. This is apparent in the demonstrations in Section 5. See Appendix C for a proof of this convergence.

In the following sections we discuss each major term mentioned in greater detail, beginning with support.

3.4 Support

The concept of *support* is not discussed in existing literature on norms; it is introduced with this model. Support serves as a scalar value representing the degree to which an individual's actions reflect adherence to an ideal. In support, we are attempting to capture the nuanced nature of conformity. This section discusses the theoretical underpinnings of support, its relationship to observable behaviors, and its significance in shaping norm dynamics within the social system.

We begin the discussion with an example. Suppose we are interested in studying the social norm one might initially describe as "clapping at the end of a good performance." It may be claimed that, with respect to this norm, either a person claps in a particular instance, or they do not, and thus the individual's behavior is a binary variable for observational purposes. This idea aligns cleanly with the idea of a focal behavior as discussed in existing literature on norms in Section 2. Seldom, however, is this type of social behavior so simply

exhibited or interpreted by others. Some clap easily, some enthusiastically, some clap more or less frequently, or pause before beginning to clap. When a researcher considers such variations important, and they seldom do, they may deem it necessary to measure focal behavior as a numeric (scalar) value. To further complicate matters though, there are very close substitutes for clapping. Some may add vocal cheers or whistles. Some might boo, which could be considered a sort of anti-clapping. This range of behaviors is observed by others and serves as a key indicator of an individual's attitude toward the relevant norm.

When applying conventional game theoretic modeling to norms, capturing such nuance is particularly challenging. In principle, it is possible to construct a discrete game which divides the set of strategies available to an individual into "strong clapping", "weak clapping", "booing" etc., but introducing any reasonable amount of strategic variety into such a framework invites intractability for both the theorist and the empiricist even at the micro-scale. Still, the ability to accommodate more nuanced strategies would have an advantage in this regard over one that does not.

It is with this in mind that the concept of support is introduced. It will first be defined informally as the amount of enthusiasm with which one's collection of behaviors aligns to the principles of a norm. Behavior with respect to a particular focal action is typically a component of support, but not the entirety of it. It is quite possible to demonstrate meaningful support for a prescriptive norm while not engaging in the focal action it is built upon (e.g., "I would love dearly to shake your hand, but I have a flu").[6]

Support, as a concept, has the advantage of capturing meaningful relationships among behaviors beyond a single focal behavior. The disadvantage is that by adding this layer of abstraction, the methods of measurement become less obvious. Researchers can, and often do, simply survey a population about their local and subjective estimations of how behaviors relate to social expectations (Jasso & Opp, 1997; Gerber et al., 2008; Fishbein & Ajzen, 2010). In practice, a survey could also be constructed which captures perceptions

[6]Some may notice that support replaces what is sometimes called "conditionality" in other literature. Rather than having focal actions be conditional, in this view they are substitutable.

about how well a particular collection of behaviors align to the expectations of a norm.

Still, it is desirable to make the relationship between support and behavior explicit. Doing so clarifies the concept and provides an alternate method for measuring support when surveys are unavailable or unconvincing. To that end, we next define support formally in terms of observables.

Let $\vec{B}_i^t \in \mathbb{R}^m$ be a vector of m measurable observations of agent i's behaviors in some discrete time period t. Let *support weights* $\vec{W} \in \mathbb{R}^n$ be a vector of weights corresponding to those observations. Let i's support $s_i \in \mathbb{R}$ be the dot product of $\vec{W}$ and $\vec{B}_i^t$.

$$s_i^t = \vec{W} \cdot \vec{B}_i^t. \tag{9}$$

In s_i^t, we are reducing an individual's measurable behaviors down to a scalar value via a weighted sums of behaviors. In other words, we are measuring each behavior related to the norm and applying a factor which corresponds to how salient that behavior is to the norm.[7] Then, these salience-adjusted values are added together to generate a support value. This support measure describes an intensity of the normative behavior. As such, $\vec{W}$ uniquely identifies the behaviors targeted by the norm we are interested in, as well as their relative importance in corresponding to the norm. The model assumes $\vec{W}$ is stable and known to all. Conceptually, it may be convenient to think of s_i as the units of effort an agent i spends in contributing to an ideal specified by $\vec{W}$, with $s_i = 0$ indicating perfectly neutral behavior. $\vec{W}$ can be thought of as a perceived exchange rate between behaviors in their contribution to that ideal.

One way for agents to coordinate on support is to coordinate on behavior. If one behaves exactly the way another does, applying $\vec{W}$ results in identical support values. Such behavioral mimicry is consistent with the work of Cialdini et al. (1990) on descriptive norms.

[7]Note that s_i degenerates to a traditional focal behavior when $\vec{W}_i$ is set to a standard unit vector. In such cases, the above equation simplifies to $s_i^t = b_i^t$. This fact may be useful if one wishes to compare the performance of this model to previous theory, which is predominantly concerned with focal behaviors. b_i^t is still scalar though, so for comparison to binary output models further steps much be taken. Either the value must be interpreted probabilistically, or some threshold value must be established to convert b_i^t to binary.

Kuran (1995) uses the argument of bounded rationality to explain such behavior coordination. It has been found that teenagers often mimic each other's behavior (Robalino & Macy, 2018; Paluck et al., 2016). Repacholi et al. (2014) finds evidence that 15-month-old infants are able to interpret the social interactions of others to effectively mimic acceptable behaviors, suggesting that the phenomenon develops quite early in life.

A utilitarian argument for such mimicry can be expressed as follows: People have a preference for others whose interests appear to be aligned with their own. Agents, knowing this will, to varying degrees, modify their behavior to reflect the interest of people whose favor they seek; a child will mimic the behavior of their parent, an employee will mimic the behaviors of the boss, married couples will seek to mimic each other, etc. Benefits of mimicry can also extend beyond the direct relationship with the agent being mimicked. If agent A aligns their behavior with a successful or popular individual B, agent C may see A's behavior as a signal that A shares qualities with B. In this way, A's reputation may improve with C.

As mentioned, one could mimic support by copying the behaviors of another perfectly. However, there is another possibility. Through $\vec{W}$, the agent is afforded the option of substituting one behavior for another to coordinate on a concept instead of a specific behavior. A roommate who wishes to honor a norm of cooperation among housemates can prepare dinner while another washes the dishes. The model in this paper implies that such conceptual trade-offs are important to understanding complex social norms.

3.5 Native Preference

Intuitively, native preference is the agent's most preferred set of behaviors if they were to exist in a vacuum where they are ignorant of, or indifferent to, the past and the behavior of others. Since certain behavior choices typically have some intrinsic value to the individual, they are preferred on instrumental grounds. These preferred behaviors incidentally correspond to a particular support value when measured and $\vec{W}$ is applied. There will be a natural tendency

for the agent to gravitate toward that support level, all else equal.

The purpose of including this term in the model is to broadly account for the aggregate effects of none-the-less meaningful factors outside the scope of model, without specifying precisely what they are. If, for example, one of the behaviors which contribute to support is "wearing designer clothing," that choice comes with a very real monetary consideration which may discourage or even strictly prevent that focal action from being taken. Similarly, an activity like dancing might be socially valuable, but for an elderly person, dancing could pose physical risks, whereas for a youthful athlete, dancing might be easily embraced as an opportunity to display coordination and fitness at low cost. An introvert attending social events or someone with financial limitations trying to meet material display norms may face stress or resource strain, and so on.

It is true that the concept of support can partially compensate for such heterogeneities in preference. The agent can adjust somewhat for cost factors via the trade-off implied by $\vec{W}$. To continue an example above, a person for whom dancing is dangerous might watch and support others in their efforts in an attempt to satisfy normative requirements. Unfortunately, such trade-offs are not always available in real social situations. It is also worth recognizing that such trade-offs, even when they are available, are not always recorded by the empiricist. More fundamentally though, we must assume that any deviation from the set of behaviors one would natively choose comes with some cost, and sometimes that cost will be significant. The cost to a criminal in supporting high social stigma against ex-convicts, for example, might be particularly high.

Native preferences are an important inclusion in the model as they represent the behaviors individuals naturally prefer based on intrinsic motivations, capabilities, and values. When agents are required to deviate from these preferences to conform to social norms, they incur costs, which may manifest in a variety of ways including lost material wealth, harmed health, or lost time. The more an individual is forced to deviate from their native preferences, the higher the cumulative cost. This compromise between personal satisfaction and

social conformity shapes both individual behavior and broader social dynamics, explaining why certain norms are easier for some to follow and why the persistence of norms can depend on how well they align with individuals' natural inclinations.

3.6 Habit Formation

Habit formation is central to the model's ability to account for the persistence and stability of social norms over time. It is incorporated as an agent-level tendency for individuals to maintain consistent support levels over time.

Habit is expressed in the model as a single period look-back term. As simple as this implementation is, the recursive nature of the model ensures that when $h_i, n_{i,i} \neq 0$, current support levels are influenced by all past support choices, with the relative impact of a given period diminishing as t increases. This means that, once established, support has persistence, even when external conditions or preferences change. This framework could be adapted to more sophisticated or longer period habit, however this simple implementation suitably enables the desired dynamics without over-complicating the model.

This conceptualization of habit draws inspiration from the literature discussed in section 2.6.3, particularly the rational foundation of habit formation discussed by Becker & Murphy (1988). From a practical perspective, the empirical studies on voting, hand washing, and other socially normative behaviors previously discussed support the idea that habits play a crucial role with respect to maintaining social norms.

This inclusion of habit in the model provides a simple and well-supported mechanism for modeling the persistence of norms at the macrosocial level and the internalization of norms at the individual level. Since the model is recursive, this implementation has a meaningful dynamics as we will see in Section 5.

3.7 Influence

Influence, in this model, refers to the mechanism by which individuals are motivated to adjust their support based on what they observe in others. In addition to native preferences and habit formation, social influence represents the final, and perhaps most critical term in the model. It plays an important role in modifying individual behavior and, consequently, the dynamics of the larger social norm. Unlike habit formation and native preference, which are grounded in internal states of the agent, influence captures the external pressures and cues that are received from others.

The social influence factor $n_{i,j}^t \in [0, 1]$ is a measure of how strongly the support level of i changes in response to the support level of j at time t. Influence can be, and is often, asymmetric. A television personality, for example, transmits some information about their support level to viewers, but gets no such feedback from audience members. Cialdini & Trost (1998) point out that even in direct exchanges, people copy the behavior of others in a selective and asymmetrical manner. A bidirected graph with edges weighted by influence can be thought of as one's social network, though it may be more appropriate to call it an *influence network*. For practical purposes, $\mathbf{N^t}$ will generally be considered stable over time in our discussion, with only occasional shocks. Though the model itself can handle arbitrarily dynamic networks, such changes are exogenous and excessive dynamism in the network becomes cumbersome and complicates interpretation of the results.[8] The purpose of making the network time variant is to allow for investigation into the effects of isolated shocks such as a new community member joining or leaving, or a temporary connection between individuals.

Influence is expressed in terms of underlying factors that impact the transmissibility of support between agents.

Let

$$n_{i,j}^t = e_{i,j}^t l_{i,j}. \tag{10}$$

[8] We will briefly touch on such dynamism in section 5.8

Factors that affect the influence one actor has over another include: exposure $e^t_{i,j} \in [0,1]$ to the behavior of another, and the amount of social leverage $l_{i,j} \in [0,1]$ j has over i.[9]

Exposure indicates how frequently or intensely agents i and j interact. Social leverage quantifies how much agent i cares about aligning their behavior with agent j. Social leverage represents the motivation behind behavioral adjustment, while exposure facilitates the transmission of support. An agent may be highly exposed to another (e.g., through frequent contact) but the effect on support may be minimal if social leverage is low. Conversely, an agent may meaningfully modify their support in response to high social leverage even if exposure is limited. We discuss these factors in turn.

3.7.1 Exposure

In this model, exposure $e^t_{i,j} \in [0,1]$ refers to the degree to which agent i is socially exposed to j's relevant behaviors. It can be thought of as the frequency or intensity with which agents i and j interact during the time period, where more frequent or sustained interaction indicates higher exposure. Exposure is the mechanism by which agents observe and receive information about others' behaviors, enabling social influence to occur. It is context-dependent, asymmetrical, and, in practice, variable over time.

When agents interact, whether through direct contact, or indirect means such as through media, they share information about their behaviors, which in turn reveals evidence of their level of support. In this model, we assume that higher exposure corresponds to interactions that are more frequent, more intense, or richer in information about the agent's support level.

Exposure is not necessarily symmetric. An agent i may have high exposure to agent j (e.g., by watching a television broadcast or following them on social media), but the reverse may not be true if j has no reciprocal contact with i. In such cases, $e^t_{i,j}$ is non-zero, while $e^t_{j,i}$

[9]This definition of influence is associated with *social* norms. If, for example, one wished to include, or restrict their focus to, *legal* norms we could add to these terms or replace them with enforcement rates, likelihood of being observed by the authorities, etc.

is zero. This asymmetry highlights a fundamental feature of exposure: it allows for one-sided information flow.

In this model, exposure also adds time variance to influence. Exposure can fluctuate as individuals change social groups, work environments, or routines. For example, moving to a new neighborhood or joining a different social circle would alter the network of interactions, and thus, the exposure to different agents. In such cases, changes in exposure can lead to shifts in an agent's support as new social influences take hold. Even transient changes in the network, such as a temporary decrease in the frequency one has contact with a good friend, could plausibly have measurable impact on support and the associated behaviors.

The role of exposure in this model underscores the importance of social networks and the structure of interactions in shaping behavior. Agents who are more exposed to certain groups are more likely to be influenced by the behavior of those groups, provided that social leverage is also present. This is especially relevant in understanding how norms spread within a population: those who have high exposure to certain agents (e.g., leaders, celebrities, or close social contacts) will more readily adopt behaviors that align with those agents' support levels.

3.7.2 Social Leverage

Social leverage is a component of influence, reflecting the degree of personal or relational importance that one agent places on aligning their support with that of another agent. This alignment may be motivated by various factors, such as the desire to maintain favor, receive rewards, avoid sanctions, or simply to mirror influential social connections. For each ordered pair of agents, i and j where $i \neq j$, a social leverage value, $l_{i,j}$ exists. This value represents how much agent i values reducing the difference between their own support level s_i^t and the support level of agent j in the previous period, s_j^{t-1}.

Since the common theme of various types of social leverage is that one person has an advantage over the other, we can decompose leverage into endowments for further theoretical

depth.

$$l_{i,j} = \frac{v_j}{v_i + v_j}, \tag{11}$$

where v_i is the endowment of agent i. High leverage, in this view, comes from endowment imbalances between the parties. Endowments in the sense meant here can take many forms. Being monetarily wealthy, attractive, having a position of power, or even a great sense of humor.

The concept of social leverage operates independently of agent i's interest in the direct consequences of choosing any given support value s_i^t, which are instead captured in $\dot{s}_i^t$. In other words, agent i places a unique value on coordinating their support with each specific agent, irrespective of the other costs and benefits of that choice. For example, though agent i may not typically enjoy playing golf, if they place a high value on their relationship with agent j, who does enjoys playing golf, then agent i will increase their support for golf in proportion to the social leverage j has over them.

Social leverage encompasses both the risks of explicit sanctions and the potential rewards from a counterparty. It also accounts for any natural, evolutionary, or moral benefits arising from coordination, even in the absence of overt coercion. While the details of such interactions can vary considerably, their impact with respect to behavior remains consistent. The shared element is that agents derive value from mirroring the observed behavior of others, and the degree of mirroring is proportional to how important the other person is to the agent. This holds true regardless of the specific nature of that benefit.

In practice, this can manifest in various social contexts. In the workplace, for example, an employee may outwardly adopt the same values or behaviors as a coworker they depend on, even if they do not personally agree with those values. Similarly, a teenager might follow a physically uncomfortable fashion trend to signal alignment with their peers. In both cases, the value placed on reducing differences in support reflects the social leverage that the counterparty holds over the agent.

Social leverage has been defined in this paper as non-negative, meaning agents are motivated to reduce the gap between their support levels and that of others. However, one could extend this concept to include negative values, representing relationships in which an agent actively seeks to differentiate themselves from another. Such a feature could have interesting empirical implications. However, such an extension may result in disequilibrium in the model as currently specified and, as we will see in Section 5.4, such a feature is not required for polarized behavior in a network.

4 Macro-Properties of the Model

Having specified the model in terms of micro behaviors, we can now turn our attention to network-wide measures. The model affords the application of statistical methods to quantify and compare macrosocial properties of a society. In this section we define and discuss several such metrics. We begin with simple mean measures before moving on to less obvious measures. Of particular interest to later discussion will be the measures of target support and norm strength.

4.1 Mean Support

$$\bar{s}^t = \sum_{i=1}^{p} \frac{s_i^t}{p}, \tag{12}$$

Mean support is simply the average of the support values in any given time period. It is a point measure of aggregate support, and its calculation does not require awareness of network topology.

4.2 Mean Social Pressure

Mean Social Pressure is the mean influence exerted on a unit of deviance from one's neighbors.

$$\bar{n}^t = 1 - \sum_{i=1}^{p} \frac{n_{i,i}^t}{p}. \tag{13}$$

Recall that $n_{i,i}^t \in [0,1]$ is the coefficient on non-social factors. $1 - n_{i,i}^t$ is therefore the sum of social influence on agent i. The mean of this value across the population indicates how responsive the population is, as a whole, to normative pressure. At $\bar{n}^t, \bar{n}^{t+1}, ..., \bar{n}^{\infty} = 0$ agents are not subject to any external behavioral forces and all agents will behave according to $\dot{s}^t$ in the long run. At $\bar{n}^t = 1$ behavior is entirely dictated by the most recently observed support of neighbors in the network.

4.3 Target Support

Target support τ^t is the level of support that the population collectively considers ideal at a particular point in time. It is average individual support values weighted by their influence in the population in a single period.

$$\tau^t = \frac{\sum_{i=1}^{p} d_i^t s_i^t}{\sum_{i=1}^{p} d_i^t}, \tag{14}$$

where

$$d_i^t = \sum_{\substack{j=1 \\ j \neq i}}^{p} n_{j,i}^t. \tag{15}$$

Intuitively, τ^t is the level of support a typical agent in the network would have to choose in order to minimize conflict with the other support levels they have most recently observed.

Target Support can be approximated by Mean support under certain conditions, such as when a mature well-connected network has a high habit formation term (low native preference), or when the standard deviation of support is low.

4.4 Standard Social Deviance and Standard Native Deviance

Standard social deviance σ_τ^s is the standard deviation of individual support from target support at time t.

$$\sigma_s^t = \sqrt{\frac{\sum_{i=1}^{p} (s_i^t - \tau^t)^2}{p}}. \tag{16}$$

This is a measure of how much the support of a typical agent varies from what the collective network would find ideal at that time. Closely related, *Standard Native Deviance* is the same calculation with native preference support in place of actual support.

$$\sigma_{\dot{s}}^t = \sqrt{\frac{\sum_{i=1}^{p} (\dot{s}_i^t - \tau^t)^2}{p}}. \tag{17}$$

These should not be confused with standard deviations of support. The differences are not calculated against mean support, but target support. That is, they account for influence, whereas a conventional standard deviation would compare support values to mean support. Generally, σ_s^t and $\sigma_{\dot{s}}^t$ will be larger than their corresponding standard deviations.

4.5 Norm strength

$$\eta^t = \frac{\sigma_{\dot{s}}^t}{\sigma_s^t} \tag{18}$$

Norm strength is the ratio of the standard deviations of current native support over actual support, $\eta^t \in [0, \infty]$ and is undefined when there is no network. All else constant, if η^t increases in t, we might say an associated norm becomes "stronger." It captures the proportion of the change in variance which is due to accumulated social dynamics. Notably, η^t is a dimensionless value. For this reason, it can be compared across norms.

Also observe,

$$s_i^u = \dot{s}_i^t \; \forall i \in p \implies \eta^t = \frac{\sigma_s^u}{\sigma_s^t}. \tag{19}$$

In other words, if at any point in time behavior across the network was unaffected by the norm in question, behavior at that time could be used to calculate the strength of the norm today, assuming all other factors remain constant. For example, as the model is formulated in the initial condition (Equation 7), $\eta^t = \frac{\sigma_s^0}{\sigma_s^t}$ when support is time invariant.

4.6 Defining a Social Norm

At this point we can define a social norm and the conditions for its existence:

> *The pair ($\vec{W}$, τ) is said to be a social norm in influence network N^t at time t if and only if $\eta^t > \eta^t_{crit}$,*

where $\vec{W}$ is the mapping between behavior and support. τ indicates a target support value. η^t_{crit} is a critical norm strength which denotes the cutoff indicating when a norm is strong enough to be said to exist. The choice of η^t_{crit} is somewhat arbitrary, just as the cutoff for calling a person 'tall' is arbitrary. Notably though, when $\eta^t > 2$ more of the variation in support is accounted for by normative dynamics than native support. A less formal way to express this definition is: A norm is the (nontrivial) force, produced by agents' social embeddedness, acting to modify those agents' behavior from how they would otherwise act.

We will illustrate how this definition is applied by way of example. One's participation in football hooliganism can be measured by taking into account a variety of behaviors, attending football matches, chanting, frequenting certain pubs, engaging in post-match vandalism. All these behaviors and their relative contributions to the idea of hooliganism is captured by a $\vec{W}$. It could be that the norm in a given community is very low levels of football hooliganism, or it could be high, or anything in between. This target value is indicated by τ. These two values, $\vec{W}$ and τ, specify a norm conceptually, but this is only a hypothetical norm of hooliganism. To be relevant, that level of hooliganism must exist in the context of a group of people at a particular time. This time and group are identified by $\mathbf{N^t}$. Having chosen a norm and a group, we can turn our attention to testing for the

existence of the norm in the wild. It is not enough to notice that individuals in $\mathbf{N^t}$ typically, or on average, engage in hooliganism at the level we are interested in. High(low) hooliganism can be explained by non-social factors such as having few(many) other pastimes available or having low(high) police presence in the area. Such factors are ones of native preference. In order for such behavioral regularity to be caused by a norm for hooliganism, it must be transmitted socially. We test for this by measuring η^t, which is how much cumulative effects of social embeddedness have reduced support variance in the population compared to a world without such influences. We can observe this most simply by noting the changes in behavior of people who join the network. If we see their support levels changing in line with the target support levels, a high η^t is implied, and we say the norm exists in that network at that time. We have then specified and established the existence of the hooliganism norm.

It is worthwhile to briefly examine this definition through the lens of the classifications provided by Opp (2001). Oughtness is captured in the tuple ($\vec{W}$, τ), which identifies the set of behaviors relevant, and how they contribute to the norm. While $\vec{W}$ is universally known, τ is not and agents only experience it through their network connections. In this way, expectations are experienced differently depending on one's position in the network.

In an established population, some amount of behavioral regularity is a necessary consequence of, but not a sufficient condition for, the existence of a norm. In this sense, a support *regularity* is not what a norm *is*; support *regularization* is what a norm *does*. The component of regularity that is caused by the norm is captured by η^t. Whereas general regularity would be captured by a simple standard deviation over behavior.

The model makes social harmony a part of the utility function, as we will see in Section 6.2. This internalized need to conform allows for sanctioning that is largely passive and self-imposed. That said, external sanctioning certainly occurs, and much of the existing literature concerns itself with rationalizing the costly enforcement of such sanctions. While there is no doubt that sanctions exist, it is not necessary for them to be especially costly to society. In fact, most of the overt sanctions we observe in broader society (e.g., yelling at strangers)

are not the primary method of influence but the extreme and do not impose significant costs on either the sanctioner or the sanctioned in an economic sense. The model incorporates these overt sanctions through the behaviors that contribute to support. Such behaviors may be approved by the norm and transmit information to the sanctioned as well as anyone else observing the exchange. These observers then choose to adjust their future support levels based on the social leverage that person has. The model implicitly asserts, though, that such sanctions are not required for a norm to exist.

This definition offers a relatively rigorous and quantifiable way to identify and test for the presence of norms within a population. The distinction between behavioral regularity and normative dynamics is critical, as it highlights that norms are not merely patterns of behavior but forces that actively shape individual actions within social networks.

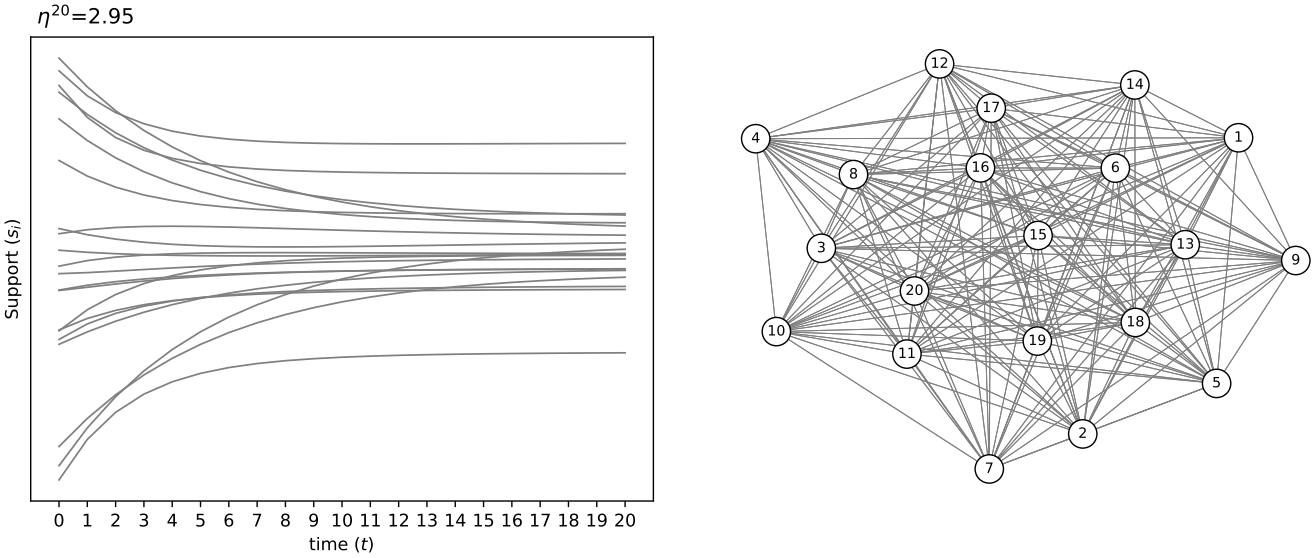
5 Model Dynamics

This section explores dynamic aspects of the proposed model, demonstrating how individual behaviors aggregate to produce evolving social outcomes. Through various simulations, it explores different network structures and their impact on the stabilization and diffusion of social norms. These illustrations offer insights into the conditions under which norms emerge, persist, or fade.

5.1 The Baseline Simulation

We begin by presenting a baseline demonstration of the model, which excludes shocks, large asymmetries, or complex network topologies. Each example in the following section modifies some aspect of the baseline simulation to illustrate its effect on support. The code used to generate this simulation, as well as all others presented in this paper, can be found in Appendix D.

In figure 5a, each line corresponds to the support value of an agent in the network



(a) Support over Time

(b) Network Layout

Figure 5: The Baseline Simulation

over time. Initially, all agents start with a support level equal to their time-invariant native preference level, $\dot{s}$. Each of these values is randomly determined by an independently and identically distributed (i.i.d.) draw from a uniform probability distribution over the range [0,1]. Over time t, the agents' support values converge, causing η^t to increase from its base value of 1. The η^t value in the last period is displayed on the upper left of the plot.

In Figure 5b, the network layout for this simulation is shown. The network is fully connected in this case, meaning there are two edges between every pair of agents, one in each direction. The connection weight of each edge is selected randomly and i.i.d. from a uniform distribution over the interval [0, 1]. The agents' self-influence values set to be 80% of the total influence. Habit formation terms are chosen from the range [0.75, 1.0]. These two ranges were explicitly selected to ensure that convergence and clustering occur at a rate suitable for the chosen number of time periods. Weights are then normalized to conform with Equation 5. The nodes are numbered in descending order according to initial support values ($\dot{s}^0$) values. Agent 1 has the highest $\dot{s}^0$, Agent 2 has the second highest, and so on.

Collective instrumentality is not a requirement for a norm to exist. It could be that

the t=0 states were better for all instrumentally. There is no welfare function applied or included in the model independently. Still, potential welfare gains should improve the likelihood of a norm emerging via $\dot{s}$. Collective instrumentality can enter through its impact on global native preferences as we will see in 5.5. Passive transmission of behavior allows for influential people to unintentionally sway τ^t and create a norm. In general, a norm will emerge when an influential subset of society has coordinated support. This coordination could be coincidental, intentional, or incidental.

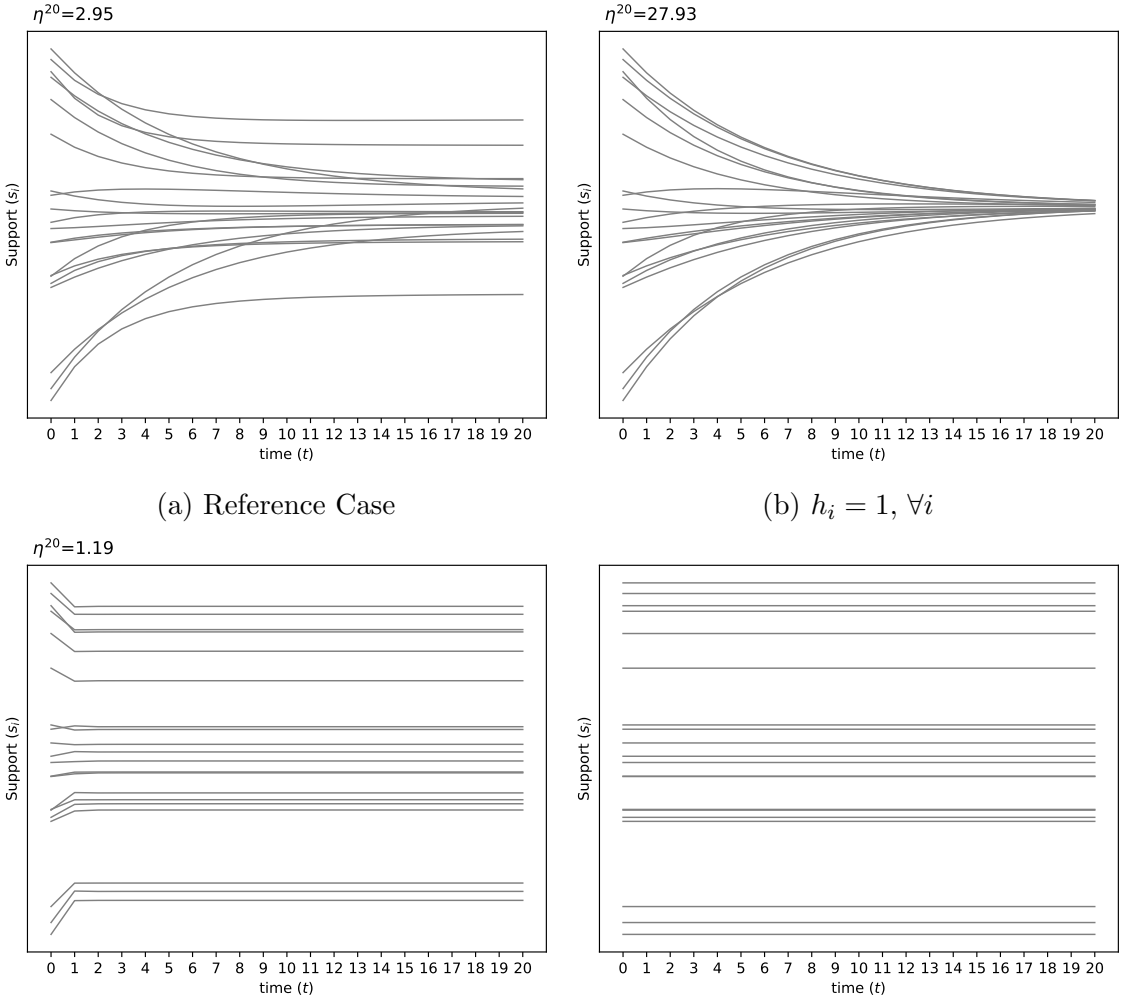
5.2 Degenerate Forms

Here we explore three degenerate forms of the model by removing each of the three primary components (native preferences, habit, and influence) from the model in turn. Studying these degenerate forms illustrates the impact of each term on the model. This approach also tests the robustness of the model by exploring how it behaves under extreme or simplified conditions, helping to reveal its limitations or potential weaknesses.

By setting $h_i = 1$ for all agents, we remove the effect of native support. That is, agent's self-influence is driven entirely by the desire to indulge in habit over any other self-interested factor. All agents are free to converge on a single support level because there is no long-term cost associated with deviating (See figure 6b). Since there are also no shocks in our example, the mathematical behavior of this simulation mimics that of DeGroot learning (DeGroot, 1974).

Figure 6c shows the effect of removing the habit formation term. Since agents have no affinity for their past behaviors, adjustments happen almost immediately. Since that necessarily implies a heavier weight on native preference, the result is lower conformity.

When there is no network, all agents are free to choose according to their native preference. Habit is irrelevant as there are no competing desires or shocks (Figure 6d).



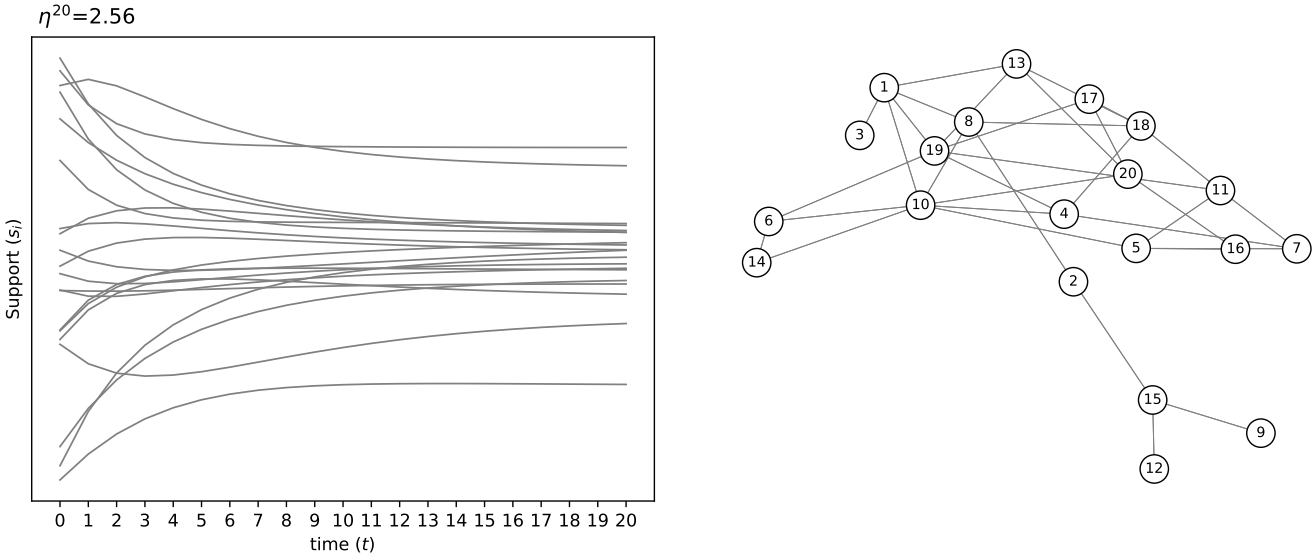
(a) Reference Case

(b) $h_i = 1, \forall i$

(c) $h_i = 0, \forall i$

(d) No Network

Figure 6: Degenerate Forms of the Model



(a) (b)

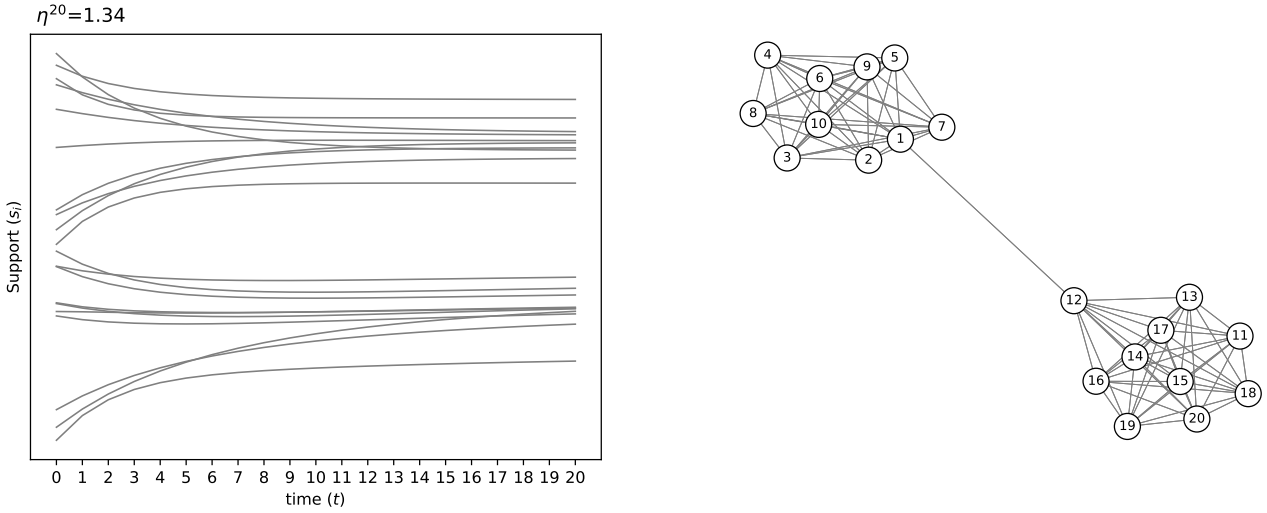
Figure 7: Sparse Network

5.3 A Sparse Network

In Figure 7, symmetric edges have been trimmed from the default network and self-influence has been adjusted in proportion to the outside influence lost. The model is then applied to the resulting sparse network. Early changes in support are more pronounced in some agents owing to the idiosyncratic variety of influencers on each agent. Normative effects are apparent, but norm strength is lower. This suggests that communities that are less tightly knit would have weaker norms, all else equal. Any factor that increases or creates influence (increased exposure or social leverage) across the population should result in more uniform behavior in each period and faster convergence toward the target support.

5.4 A Bridged Network

In Figure 8, we modify the baseline simulation by trimming the default network into two partitions which we leave connected by a single edge. Note that support values are collecting around at different values for each cluster. This illustrates how clustered networks can pool support at different levels. Note the lower η^{20} value compared to that of the baseline example



(a) (b)

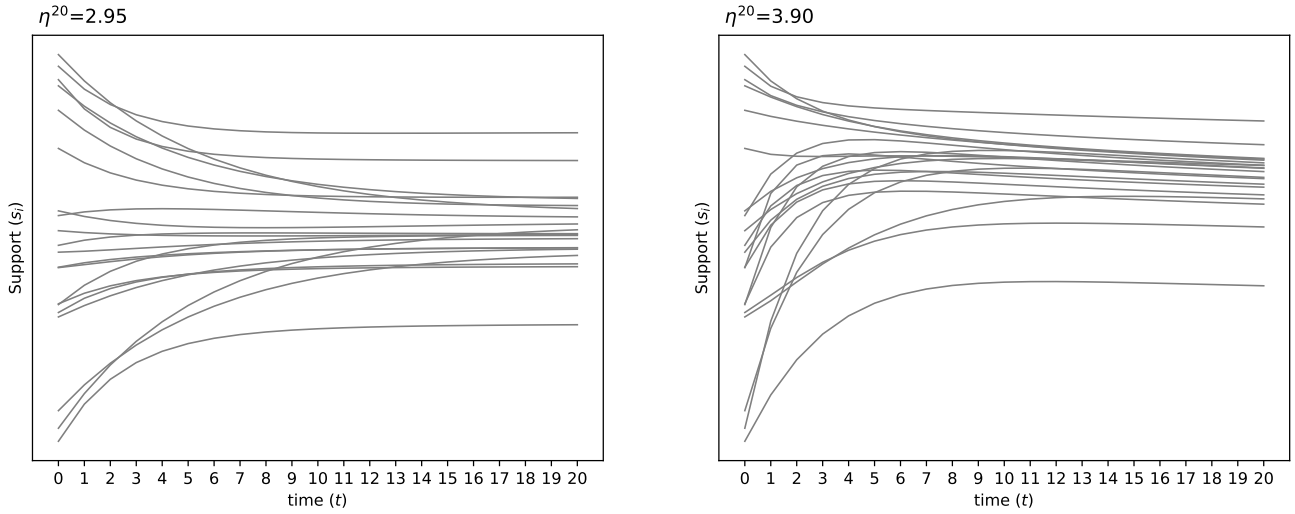
Figure 8: A Bridged Network

in Figure 5. Because η^{20} describes the collective behavior across the entire population, the norm is significantly weaker. However, if one were to restrict $\mathbf{N^t}$ to either of the sub-network clusters, the corresponding η^{20} would be significantly higher.

This example illustrates what may happen to two mostly isolated social groups, such as different insular religious or ethnic communities in a city, which are largely separated but have a few key individuals or organizations that function as bridges between them. These bridging individuals might facilitate the dissemination of norms, but because the two groups do not interact much beyond these limited connections, their social norms and behaviors would still be free to evolve quite distinctly.

5.5 A Single Influential Individual

The effect of a single individual can be dramatic even without complex network topology. Figure 9 shows the effect of modifying the baseline network to make agent one disproportionately influential to all other individuals. Agent 1's influence on each other agent is increased by a factor of roughly 10 on a re-normalized matrix. This results in more concentrated support levels and a shift in aggregate support level from baseline. In this way, large numbers

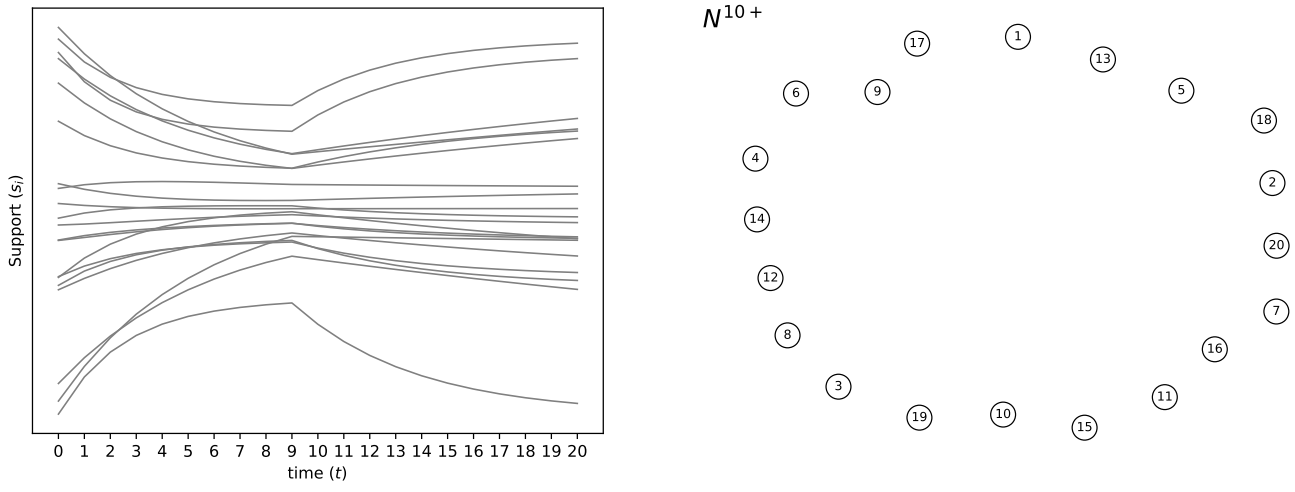


(a) Baseline (b) Influence of Agent 1 Increased

Figure 9: A Particularly Influential Individual

of indifferent people can be swayed by a relatively small number of well-known or important individuals.

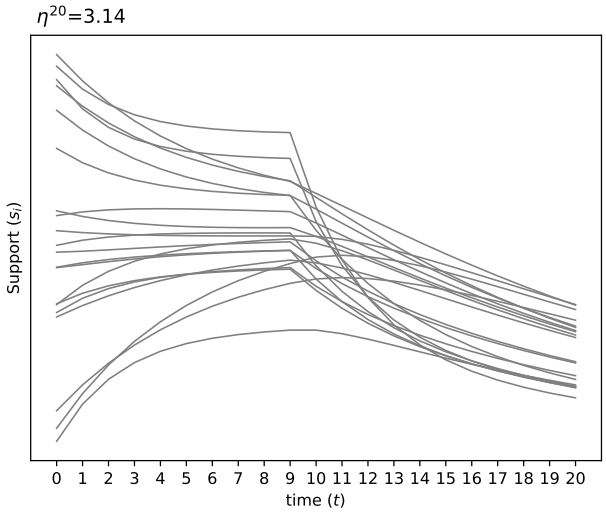
5.6 Disconnection from the Network



(a) (b)

Figure 10: Removal from the Network

Figure 10 shows that individual support persists when influence is suddenly removed.



(a) Shock to $\dot{s}^t$

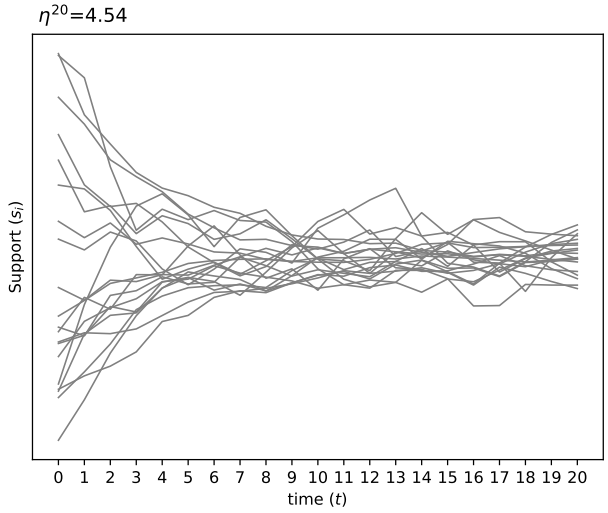
Figure 11: Shock to Native Preferences

After period 9 we disconnect each agent from the network. This captures the effect of norm internalization. If one were to become separated from the social context which imparted a normative behavior shift, its influence on their behavior would persist for a time before the long-run equilibrium of $\dot{s}^t$ is restored.

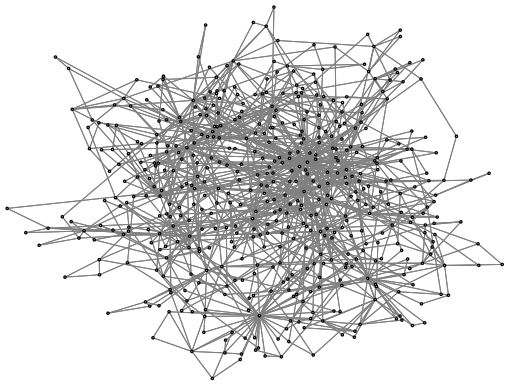
5.7 A Shock to Native Preferences

Figure 11 models a significant downward shock to native preferences $\dot{s}_i^t$ for all agents at time 10. This simulates a sudden change in the intrinsic value of the behaviors associated with support.

An example of such a shock might be a law that bans smoking in public places. For decades, smoking in restaurants, bars, and public spaces was widely accepted. In countries like the United States and across Europe, tobacco bans were introduced in public spaces. This increased the costs associated with smoking, which immediately altered native preferences regarding tobacco across whole societies. While many individuals and communities may have preferred to continue smoking in public, the external shock of legal changes contributed to a decline in smoking as a social norm.



(a) Random Sample of Agents



(b) Sample Network from One Period

Figure 12: Shocks Every Period to $\mathbf{N^t}$ and $\dot{s}^t$ in a Large Population

5.8 Continuous Shocks to $\mathbf{N^t}$ and $\dot{s}^t$ in a Larger Community

Finally, we examine the model on a large, complex population under constant shocks to both the network $\mathbf{N^t}$ and native preferences $\dot{s}^t$. Here, each edge in the network of 500 agents is randomized every period using the Holme and Kim algorithm for growing graphs with power law degree distribution and approximate average clustering (Holme et al., 2002). Connection generation favors certain popular individuals. In each period, the network is scale-free with clustering. Additionally, each period every agent's native preferences are re-randomized uniformly and i.i.d. to a new value in the range [0,1]. Such volatility in the parameters does not prevent a norm from emerging, though individual support values are now noticeably erratic over time (See figure 12).

6 Relationship to Rational Choice

This section aims to demonstrate that the model presented in this paper has rational foundations. Importantly, the model does not include explicit forward-looking elements. A case is made for this exclusion. Whether the resulting rationality is said to be bounded or not

depends on the theoretical perspective one adopts.

After describing some of the key assumptions, the model is derived from individual utility. Then, after a brief discussion on the role of information, we will then move on to describing the equivalence relationship between observable behavior and the construct of support.

6.1 Assumptions

Here we will declare some of the key assumptions embedded in this model.

Assumption 1 *Let s_α^t and s_β^t be two potential support levels at time t. Let s_j^{t-1} be an observed neighbor's support level. If $|s_\alpha^t - s_j^{t-1}| < |s_\beta^t - s_j^{t-1}|$ then $s_\alpha^t \succsim s_\beta^t$, all else equal. That is, agents prefer to align their support levels with those they observe in others.*

Assumption 2 *Agents prefer to minimize $\Delta_i^t = |s_i^t - s_i^{t-1}|$. That is, support is habit forming.*

Assumption 3 *The second derivative of each term in the utility function is strictly negative. That is, marginal deviations in $\tilde{s}_i^t$ become increasingly costly as one moves farther from ideal values.*

Assumption 4 *For any support s^* in the image of s there exists a set of behaviors $\vec{B}^*$ such that $s(\vec{B}^*) = s^*$ and $U_{instr}(\vec{B}^*) > U_{instr}(\vec{B})$ for all $\vec{B} \neq \vec{B}^*$ where $s(\vec{B}) = s^*$. That is, behavior can be inferred from support.*

Assumptions 1 and 2 incorporate influence and habit formation respectively. Assumption 3 ensures that preferences over potential outcomes remain ordered.

Assumption 4 ensures each scalar support value corresponds to a unique set of preferred behaviors. The existence of the function s establishes that any set of behaviors can be expressed as a single scalar value. See the Sections 6.4.1 – 6.4.3 for examples of utility

functions which meet this requirement. While assumption 4 is not required for the model to function, it is necessary to make predictions about specific behaviors from the model's outputs.

6.2 Derivation of the Model from Utility

We start by characterizing the relevant forward-looking utility U_{fl}, which we express in terms of four components, native preferences, habit formation, influence, and expected future utility given the current period choice. For the sake of mathematical convenience, we will make these terms additively separable. For notational convenience, we start from the perspective of a single agent.

$$U^t_{fl}(\tilde{s}^t) = U^t_{native}(\tilde{s}^t) + U^t_{habit}(\tilde{s}^t) + U^t_{influence}(\tilde{s}^t) + \sum_{i=1}^{\infty} \beta^i E_t[U^{t+i}_{fl}(s^{t+i}|s^t = \tilde{s}^t)]. \tag{20}$$

Here, β is the discount factor on utility for one period, $\tilde{s}^t$ is a measure of a basket of behavior choices, and s^t is a measure of the specific basket of behavior choices chosen at time t. Since s values are measurements of behavior rather than quantities of consumption, good-related assumptions such as non-satiation and diminishing marginal utility do not apply. Some, or even all, component behaviors of $\tilde{s}^t$ may be consumption items, but their consumption would be limited by budget constraints. Other behavioral choices, such as what color to paint the living room walls are simple preferences over an uncountable set of possibilities. To account for the innumerable variety and types of behaviors which make up support, we model utility as continuous. To ensure preferences remain ordered we make Assumption 3, that as one is forced to deviate in a particular direction from the ideal support value for each isolated term, the marginal cost of doing so increases.[10]

Next, we will also define immediate utility, that is, utility which is U^t_{fl} without the

[10]See Appendix B for a brief further investigation in solving for optimal support under this forward-looking agent.

forward-looking term:

$$U^t(\tilde{s}^t) = U^t_{fl}(\tilde{s}^t) - \sum_{i=1}^{\infty} \beta^i E_t[U^{t+i}_{fl}(s^{t+i}|s^t = \tilde{s}^t)] \tag{21}$$

$$= U^t_{native}(\tilde{s}^t) + U^t_{habit}(\tilde{s}^t) + U^t_{influence}(\tilde{s}^t). \tag{22}$$

There are three major arguments for the use of U^t over U^t_{fl} to model behavior.

The first is to declare that agents do not take such calculation into account because it is either too computationally intensive or the cost of gathering the necessary information to perform such a calculation is too high. To accept this argument is to accept bounded rationality.

The second is to declare that the forward-looking term is assumed to be incorporated into U^t_{native}. In this way the model becomes compatible with forward-thinking without explicitly modeling it. This is the strategy taken in a large majority of the game theoretic modeling on norms, where such considerations are exogenously incorporated through the strategy payouts.

The third is to argue that the effect of the forward-looking term is trivial and does not need to be included. This is a challenging argument to make definitively. The structure of local clusters within a broader network can amplify the influence of individual agents over time. It is also true that in extremely small scale isolated social environments strategic manipulations would likely have real effects. Despite such observations, good arguments remain for the elimination of this term in some or most cases. Most obviously $U^t = U^t_{fl}$ when $\beta = 0$. It's also true that U^t and U^t_{fl} are ordinally equivalent if, for some reason, the forward-looking terms are not influenced by current behavior. Furthermore, U^t approximates U^t_{fl} with increasing accuracy as $\beta \to 0$ and as current support has less influence on neighbors. In large, well-connected networks, the strategic influence of a single agent on the overall climate, and consequently their impact on future inputs to their own utility, is generally

quite limited. Even if one could have such impact, the risk of having such manipulations discovered could be quite costly to those involved. Also, if one frames the world as one of incomplete information, the agents may be unable to calculate expectation. Such limited visibility reduces an agent's ability to strategically leverage their influence to manipulate the network to their benefit. To accept this argument is to accept that the model may not work in every case, but that it works in most cases, particularly at large scales.

There is merit to each of these arguments and which one appears the dominant factor is a function of the temperament of the reader. For the purpose of demonstrating other ways in which the model is rational, the path forward is the same; we proceed under the assumption that U^t suitably approximates U^t_{fl}.

We will define each of the first three terms in U^t such that they are parabolic in s with axis of symmetry representing the preferred choice within that term. We will use coefficient $a_k \in \mathbb{R}^+_0$ to represent the relative rate of falloff in utility as $\tilde{s}^t$ deviates from preferred choice. Higher values of a_k indicate that deviation is more costly relative to low values.

First, we declare utility from matching the observed behavior other agents.

$$U^t_{influence}(\tilde{s}^t) = -\sum_{k=1}^{K} a_k(\tilde{s}^t - \ddot{s}^{t-1}_k)^2. \tag{23}$$

Here, K is the number of others who can be observed by the agent. $\ddot{s}^{t-1}_k$ is the last observed support level of neighbor k. We continue the convention of making terms additively separable.

Next, we declare the native utility.

$$U^t_{native}(\tilde{s}^t) = -a_{K+1}(\tilde{s}^t - \dot{s}^t)^2. \tag{24}$$

$\dot{s}^t_i$ represents the support level that the agent would choose in absence of any habituation or social influence. $\dot{s}^t_i$ is the optimal choice generated by any such utility function which generates a preference and, more importantly, includes an increasing and symmetric marginal

cost for deviations from that preference. Note that native utility is maximized when $\dot{s}_i^t = s_i^t$. In a sense, this declaration is simply an assertion that utility inputs unrelated to habit formation and social influence can be expressed in this form.

Finally, we declare a simple habit formation term.

$$U_{habit}^t(\tilde{s}^t) = -a_{K+2}(\tilde{s}^t - s^{t-1})^2. \tag{25}$$

Here we have asserted a preference to behave as one did last period with no consideration for behavior further in the past beyond its impact on last period.

Combining the above gives us the functional form of utility.[11]

$$U^t(\tilde{s}^t) = -a_{K+1}(\tilde{s}^t - \dot{s}^t)^2 - a_{K+2}(\tilde{s}^t - s^{t-1})^2 - \sum_{k=1}^{K} a_k(\tilde{s}^t - \ddot{s}_k^{t-1})^2. \tag{26}$$

Since U^t is a concave unconstrained function of a single choice variable, finding the utility-maximizing support s^t is a simple matter. The first order condition is

$$\frac{\partial U^t}{\partial \tilde{s}^t} = 2a_{K+1}(\dot{s}^t - \tilde{s}^t) + 2a_{K+2}(s^{t-1} - \tilde{s}^t) + 2\sum_{k=1}^{K} a_k(\ddot{s}_k^{t-1} - \tilde{s}^t) = 0. \tag{27}$$

$$\implies s^t = \frac{a_{K+1}\dot{s}^t + a_{K+2}s^{t-1} + \sum_{k=1}^{K} a_k\ddot{s}_k^{t-1}}{\sum_{k=1}^{K+2} a_k}. \tag{28}$$

After normalizing the weights such that $\sum_{k=1}^{K+2} a_k = 1$, we have

$$\implies s^t = a_{K+1}\dot{s}^t + a_{K+2}s^{t-1} + \sum_{k=1}^{K} a_k\ddot{s}_k^{t-1}. \tag{29}$$

This describes optimal support for a single agent given their previous period support choices and those of other agents with influence over them. Arranging the weights of a collection of agents into a matrix, we arrive at the model described by equation 3. This gives the optimal

[11] Some readers may notice that the value of the utility function can never take a positive value. This is of no consequence as the preference relation between possible choices is conserved. Choices that result in a utility closer to zero are preferred by the agent.

support levels for all agents who are influencing each other in an interconnected network:

$$s_i^t = n_{i,i}^t\big((1-h_i)\dot{s}_i^t + h_i s_i^{t-1}\big) + \sum_{\substack{j=1,\\ j\neq i}}^{p} n_{i,j}^t s_j^{t-1}. \tag{30}$$

Outside of the diagonal positions, the values in each row in a matrix $\mathbf{N^t}$ correspond to the set of a values in the influence term given in the single agent utility case above. They represent the weights of social influence that each agent j has over agent i. The diagonal of the matrix corresponds to the weight of influence each agent has on themselves, which is itself further divided into habit and native influences through $h \in [0, 1]$.

6.3 The Role of Information

The utility function described is driven by local observations of past behaviors (both the agent's own and others in their immediate network). Given this setup, and if one also accepts that forward-looking considerations are trivial, the agent's utility is not derived from knowledge of the entire network but rather from their own actions and the observed behaviors of those directly connected to them.

Since the utility function does not depend on unobservable parts of the network and only involves observable past behaviors and locally available information, this model is one of complete information. In this formulation, the missing information about the broader network is not crucial to the agent's well-being and the agent need not be concerned about the macrosocial consequences of their choices.

6.4 Connecting Support and Behavior

In Section 6.2 we made use of Assumption 4 to derive the model from a utility function. This assumption asserts that for a given agent and $\vec{W}$, a particular support value implies a unique set of behavior choices. This assumption is important because if it does not hold, then when the agent chooses a support level, it is unclear which actual behaviors have been

selected. Such ambiguity would create a theoretical problem as it would partially disconnect the model from the rational foundation outlined. While the model may still be testable, a failure of Assumption 4 would also mean that behavior could not be predicted directly. In that case behavior could only be translated into support through $\vec{W}$ to compare with the model's output. For this reason, it is worthwhile to demonstrate that such equivalence is possible under rational choice with reasonable assumptions.

The intuition underlying how support implies behavior can be understood as follows: The agent wishes to meet a given support level and can select from a variety of behaviors to meet it. Ideally, the natively desired behaviors coincide with the support level. Generally though, this will not be the case, and the agent will have to compromise. She does this at the margin by choosing the single behavior which adjusts support toward the target value at the lowest cost to her in terms of native utility. As she shifts this behavior, the marginal cost of doing so increases. Eventually, some other behavior can be shifted at a lower marginal cost. She then modifies the new lowest-cost behavior in the direction target support demands. This process is repeated until the target support is reached.

We will go through the trivial case of a single behavior support measure before moving on to two behaviors and then arbitrary numbers of behaviors.

6.4.1 The Single Behavior Case

Recall from the definition given by Equation 9 that support is defined as a sum of m products, where m is the number of behaviors being measured. If support is measured only in terms of a single behavior ($m = 1$), then $\vec{W}$ and $\vec{B}^t$ become single element vectors. Therefore, support is

$$s^t = \vec{W} \cdot \vec{B}^t = wb^t, \tag{31}$$

which implies

$$b^t = \frac{s^t}{w}. \tag{32}$$

Since $b^t(s^t)$ is a linear function, any given support level s^t implies a unique optimal behavior b^t. Since $\vec{W}$ is intended to capture the relative importance of behaviors with respect to a norm and there is only one behavior, the value assigned to w is arbitrary. Letting $w = 1$, we express Utility and optimal behavior in terms of other observed behaviors.

$$U^t(\tilde{b}^t) = -a_{K+1}(\tilde{b}^t - \dot{b}^t)^2 - a_{K+2}(\tilde{b}^t - b^{t-1})^2 - \sum_{k=1}^{K} a_k(\tilde{b}^t - \ddot{b}_k^{t-1})^2. \tag{33}$$

Where $\dot{b}^t$, b^{t-1}, and $\ddot{b}_k$ are the native behavior, last period behavior, and neighbor last period behaviors, respectively. Following the same steps used in Section 6.2 we arrive at a unique solution entirely in terms of behavior.[12]

$$b^t = a_{K+1}\dot{b}^t + a_{K+2}b^{t-1} + \sum_{k=1}^{K} a_k\ddot{b}_k^{t-1}. \tag{34}$$

6.4.2 The Two Behavior Case

When $m = 2$, the relationship between support and optimal behavior becomes less immediately clear. By definition, support in this case is

$$s^t = \vec{W} \cdot \vec{B}^t = w_1 b_1^t + w_2 b_2^t. \tag{35}$$

Here, support is precisely defined by behavior. However, given only this definition and $\vec{W}$, there are an infinite number of unique $\vec{B}$ vectors which produce a given support value. Only one basket of behaviors can ultimately be chosen, though. For a given support level, the agent will choose a $\vec{B}$ such that it minimizes the cost of deviating from some combination of behaviors $\dot{\vec{B}}$ which would be natively selected in the absence of social and historical considerations. Let the loss from deviating from native behaviors be

[12]It may that the lone behavior is measured in discrete terms. In this case, the b^t value given here may not be in the choice set, but between achievable values. In this case, evaluating utility at each of the two choices which bound the b^t value will determine optimal behavior.

$$c_1(\tilde{b}_1^t - \dot{b}_1^t)^2 + c_2(\tilde{b}_2^t - \dot{b}_2^t)^2, \tag{36}$$

where c_1 and c_2 indicate the relative costs of deviation in each dimension. This is a disutility function which specifies the relative cost of deviating from native behaviors.[13] Identifying precise behaviors becomes an optimization problem.

$$\arg\min_{\tilde{b}_1^t, \tilde{b}_2^t} \left(c_1(\tilde{b}_1^t - \dot{b}_1^t)^2 + c_2(\tilde{b}_2^t - \dot{b}_2^t)^2 \right), \tag{37}$$

subject to the binding constraint provided by support, which is known,

$$s^t = w_1\tilde{b}_1^t + w_2\tilde{b}_2^t. \tag{38}$$

The Lagrangian is

$$\mathcal{L}(\tilde{b}_1^t, \tilde{b}_2^t, \lambda) = c_1(\tilde{b}_1^t - \dot{b}_1^t)^2 + c_2(\tilde{b}_2^t - \dot{b}_2^t)^2 + \lambda(w_1\tilde{b}_1^t + w_2\tilde{b}_2^t - \tilde{s}^t), \tag{39}$$

giving us the first order conditions,

$$\frac{\partial \mathcal{L}}{\partial \tilde{b}_1^t} = 2c_1(\tilde{b}_1^t - \dot{b}_1^t) + \lambda w_1 = 0, \tag{40}$$

and

$$\frac{\partial \mathcal{L}}{\partial \tilde{b}_2^t} = 2c_2(\tilde{b}_2^t - \dot{b}_2^t) + \lambda w_2 = 0. \tag{41}$$

Solving for $\tilde{b}_1^t$ and $\tilde{b}_2^t$ gives optimal behavior choices b_1^t and b_2^t,

$$b_i^t = \dot{b}_i^t - \frac{w_i}{\frac{w_1^2}{c_1} + \frac{w_2^2}{c_2}}(\dot{s}^t - s^t), \tag{42}$$

which uniquely identifies behavior in terms of support and known parameters. One may substitute in Equation 29 and the definition of support to express the single agent model

[13] Note that the units are not the same as utility given by U^t.

entirely in terms of behaviors.

$$b_i^t = \dot{b}_i^t - \frac{w_i(w_1+w_2)}{\frac{w_1^2}{c_1}+\frac{w_2^2}{c_2}}\left((1-a_{K+1})(\dot{b}_1^t+\dot{b}_2^t) - a_{K+2}(b_1^{t-1}+b_2^{t-1}) - \sum_{k=1}^{K} a_k(\ddot{b}_1^{t-1}+\ddot{b}_2^{t-1})\right). \tag{43}$$

6.4.3 The n Behavior Case

For completeness, we will show that the method used to illustrate identification of behaviors extends to an arbitrary number of behaviors. Support is given by

$$s^t = \vec{W}\cdot\vec{B}^t = \sum_{k=1}^{m} w_k b_k^t, \tag{44}$$

and our relative cost function is

$$\sum_{k=1}^{m} c_k(\tilde{b}_k^t - \dot{b}_k^t)^2. \tag{45}$$

Specific behavior is found by solving the following optimization problem.

$$\underset{\tilde{b}_1^t,\tilde{b}_2^t}{\arg\min}\left(\sum_{k=1}^{m} c_k(\tilde{b}_k^t - \dot{b}_k^t)^2\right) \text{ s.t. } \sum_{k=1}^{m} w_k b_k^t = s^t. \tag{46}$$

The corresponding Lagrangian function is

$$\mathcal{L}(\tilde{b}_1^t, \tilde{b}_2^t, \ldots, \tilde{b}_n^t, \lambda) = \sum_{k=1}^{m} c_k(\tilde{b}_k^t - \dot{b}_k^t)^2 + \lambda\left(\sum_{k=1}^{m} w_k\tilde{b}_k^t - \tilde{s}^t\right), \tag{47}$$

resulting in the following first order condition for each $k \in \{1, 2, \ldots, n\}$

$$\frac{\partial\mathcal{L}}{\partial\tilde{b}_k^t} = 2c_k(\tilde{b}_k^t - \dot{b}_k^t) + \lambda w_k = 0, \tag{48}$$

and the constraint condition

$$\frac{\partial\mathcal{L}}{\partial\lambda} = \sum_{k=1}^{m} w_k\tilde{b}_k^t - \tilde{s}^t = 0. \tag{49}$$

Solving the system of equations for $\tilde{b}_k^t$ and making use of the definition of support gives us optimal behavior of a single agent in terms of support and native behavior.

$$b_k^t = \dot{b}_k^t - \frac{w_k}{\sum_{i=1}^m \frac{w_i^2}{c_i}} \left(\dot{s}_i^t - s^t \right). \tag{50}$$

Optionally, we can express this entirely in terms of observable behaviors.

$$b_k^t = \dot{b}_k^t - \frac{w_k}{\sum_{i=1}^m \frac{w_i^2}{c_i}} \left((1 - a_{K+1}) \sum_{i=1}^m w_i \dot{b}_i^t - a_{K+2} \sum_{i=1}^m w_i b_i^{t-1} - \sum_{j=1}^K \sum_{i=1}^m a_j w_i \ddot{b}_i^{t-1} \right). \tag{51}$$

7 Comparison of the Model to Existing Work

In this section, the proposed model is situated within the existing literature. Broadly speaking, the model differentiates itself by its level of adherence to methodological individualism; it scales to macro-social predictions more seamlessly than game theory, providing specific predictions about a wide variety of norms given parameter inputs. It also offers improved microfoundations over threshold and EGT models. The model is further distinguished from all others by its use of the concept of support which captures nuanced behavioral patterns. This approach allows for partial adherence and varying degrees of alignment with norms, offering a richer framework for analyzing how behaviors evolve in response to both individual preferences and social pressures.

Having highlighted these general differences, it is worthwhile to more closely examine how this new model differs mechanically from others. Such comparison highlights where this work has innovated in addressing the limitations of traditional approaches. Specifically, we evaluate how the model contrasts with popular and relevant works in game theory, EGT, and threshold models, which have each contributed to the understanding of norm formation and dynamics in distinct ways. Additionally, we explore how RAA and DeGroot Learning inform the model. Through this comparative analysis, the strengths and novelty of the proposed model will be underscored.

7.1 Traditional Game Theoretic Approaches

Game theory models, such as those introduced by Coleman and expanded by others, focus primarily on steady-state solutions for norms through strategic interactions. This approach is well-suited to explaining coordination and cooperation problems, particularly in small-scale interactions. The framework has offered crucial insights into how groups can reach equilibrium in such settings, where well-defined games accurately describe the incentives of each player and stable outcomes are quickly reached.

Coleman's modeling, and others which build upon it, largely center on the satisfaction of existing demands for norms through equilibrium-based games. These models explain how individuals respond to present incentives but do not explicitly incorporate the mechanics of rational expectations. While forward-looking strategic interaction is not ruled out, it is hidden in the assumed payoffs of the proposed games. Similarly, $\dot{s}^t$ in the model presented in this paper can be interpreted to exogenously include such strategic thinking. In this sense, game theory and the proposed model are similar.

Like Bicchieri's extension of Coleman, this model incorporates social embeddedness in decision-making but focuses on different mechanisms. Specifically, Bicchieri emphasizes beliefs and social expectations, while this work features habit and influence.

Broadly speaking, game theory models, though microfounded to roughly equivalent depth, do not incorporate dynamics and are not readily extendable to the macro level, insofar as they are formally specified. For this reason, the new model is distinguished by its capacity to make precise predictions about how the development of norms is influenced by changes in network structure and individual actions. Thus, while game-theoretic models provide general insights into long-run outcomes, the approach presented in this paper offers more precision in forecasting the dynamic evolution of norms in specific social systems.

Next, we will look more closely at how the model differs from Bicchieri and Coleman's work in turn.

7.1.1 Comparison to Coleman

Coleman's (1990) model approaches social norms through a rational-choice framework, where norms are treated as solutions to public goods problems or coordination challenges. He focuses on how norms emerge from the need to regulate behaviors that produce externalities. Norms, in this view, function as mechanisms of social control, where certain behaviors are incentivized and made stable through sanctions as necessary. Sanctions provide direct incentives by impacting payouts. Where stabilization through internalization is present, it is also externally imposed on the agent rather than being a behavioral assumption. The disjoint beneficiaries of the norm invest in instilling the behavior in the mind of the individual.

Mechanically, Coleman's framework employs prisoner's dilemmas and coordination games to explain individuals' decisions on norm conformity. In these models, agents choose to follow the norm, or not, on self-interested grounds. Norm stability in Coleman's model depends on the presence of social sanctions along with individuals' capacity to communicate or coordinate their actions to solve public goods problems.

The primary distinction between Coleman's framework and that of this paper lies in the incorporation of time and dynamics. Coleman's model is static, emphasizing equilibrium outcomes and assuming that norms emerge once rational incentives align. In contrast, the model presented in this paper captures dynamic processes by simulating continuous interactions over time, allowing for fluctuations, shocks, and gradual changes in norm adherence. This makes it better suited for understanding the temporal development and evolution of norms. This is important because, in practice, shocks prevent any real-world macro-social system from ever achieving equilibrium. This type of empirical noise makes the application of Coleman's work inherently imprecise. Still, producing such equilibria is valuable for predicting macrosocial movements and this paper's model does produces them reliably, as demonstrated in Appendix C.

Additionally, the new model is better equipped to describe network effects. While not incompatible with the existence of social networks, Coleman's model does not explicitly

incorporate such influence in norm diffusion as this model does.

While Coleman's model provides valuable insights into the emergence and enforcement of norms from a rational-choice perspective, the model in this paper offers a more dynamic framework for understanding the complex interactions and temporal processes involved in norm formation and change. It builds upon the strengths of Coleman's approach while addressing its limitations, offering a more nuanced and flexible tool for analyzing norms in constantly evolving social contexts.

7.1.2 Comparison to Bicchieri

Bicchieri's model (2005) offers a psychologically grounded approach to social norms, emphasizing how individuals conform to norms based on their social expectations. Her model declares that norms are adhered to only if certain cognitive conditions are satisfied. These include empirical and normative expectations. Empirical expectations are when one believes that others are following the norm. Normative expectations are when one believes that others expect them to follow the norm and may sanction them for failing to do so. Bicchieri's model is unique in proposing that norm-following behavior is contingent on both of these factors. If individuals perceive that others are not following the norm or do not expect them to conform, their motivation to adhere weakens. Additionally, Bicchieri draws a distinction between social norms and descriptive norms and handles them differently.

Mechanically, Bicchieri's model conceptualizes norms as conditional strategies in mixed-motive games. They emerge as stable outcomes when individuals align their behavior with social expectations, driven by a desire to avoid sanctions or gain approval. While her model offers precision in identifying cognitive conditions required for norm adherence, like Coleman's, it is unconcerned with temporal dynamics or the role of network structures in norm diffusion.

The above differences are owing primarily to the fact that Bicchieri is also applying game theory in much the same manner as Coleman. More uniquely, while Bicchieri's model

relies heavily on cognitive assumptions about expectations, it does not explicitly endogenize them in the model. The model in this paper does not make use of belief, which is difficult to observe empirically. In an effort to improve observability, it relies more directly on individual behavior and social relationships. In this regard, the contributions of this paper extend Bicchieri's approach to bridging behavioral observations and rational choice. A notable consequence of this extension is that the formal definition of a norm generated by this model is both simpler and more general.

While Bicchieri's model offers important insights into the cognitive mechanisms behind norm adherence, the model in this paper provides an important alternative framework for understanding long-term behavioral change and norm diffusion. The present model's incorporation of habit formation, network structures, and temporal dynamics makes it better suited for analyzing certain large-scale environments where norms adapt rapidly. This comparison highlights how the two models complement each other: Bicchieri's model excels in capturing situational adherence to norms, while this model offers promise in explaining norm evolution in macrosocial contexts.

7.2 Evolutionary Game Theory

Evolutionary game theory offers a distinct approach by modeling norms as emergent properties of population-level interactions over time. EGT uses a trial-and-error mechanism where agents' strategies evolve over generations based on their success in repeated interactions. EGT assumes agents have fixed strategies within their lifetimes and do not react directly to their social environment.

Such models lend themselves to computational experimentation and include time dynamics, however the immediate goal of such research differs from that of this paper. EGT models serve to justify why we cooperate or coordinate behavior, whereas this model is concerned with how such coordination occurs. As such, EGT models provide some amount of theoretical justification for assumptions included in the model, such as providing an evolu-

tionary basis for why we value coordinating with others. The model presented in this paper instead describes intra-generational dynamics, with evolutionary effects already having been established.

For a more specific discussion of differences, we will focus our attention on Axelrod, as his contributions remain some of the most widely recognized in the area of social norms and evolutionary game theory. The differences between this work and the broader research efforts in EGT are largely captured in this comparison.

7.2.1 Comparison to Axelrod

Axelrod (1986) applies evolutionary principles to explore how norms emerge and persist within populations. His model treats norms as strategies that agents employ repeatedly, with successful strategies replicated over time in the next generation of agents. The process is analogous to natural selection. The evolution of behavior is driven by payoffs from interactions, where individuals who adhere to beneficial norms are more likely to have their strategies adopted by others in subsequent iterations.

Axelrod's approach is well-suited for exploring macro-level, inter-generational dynamics, focusing on how cooperation and coordination evolve over long periods. Norms stabilize when they yield benefits that surpass individual incentives to deviate, ensuring that agents with cooperative strategies out-compete those who defect over multiple generations. However, agent strategies are fixed for each generation in Axelrod's model, with changes occurring only through replication in future generations. This treatment contrasts with the continuous, adaptive behavior in the model presented in this paper. Axelrod's model captures long-term, evolutionary change, while this paper's model instead emphasizes intra-generational dynamics where agents adapt their behavior in real-time through social influence and habit.

Finally, network structures play a central role in this paper's model, shaping how behaviors spread and stabilize through local influence and connectivity. Axelrod's model, while powerful in explaining population-wide trends, abstracts away from specific network

structures and focuses on aggregate outcomes and end-state equilibria. This paper's model offers a richer understanding of how specific social structures impact norm adoption, showing that network structure can influence specific norm outcomes.[14]

Axelrod's evolutionary model provides a well-generalized framework for understanding an evolutionary basis for cooperative behavior. Together, these two approaches complement one another well; Axelrod's model highlights the evolutionary roots of norms, while this paper's model provides a more detailed account of how individuals adapt within their social environments.

7.3 Threshold Models

Threshold models, which emphasize how individual behaviors cascade within social networks, provide additional insight into norm diffusion and persistence. In these models, agents typically make binary choices, such as whether to follow a norm, based on the behavior of others. In contrast, the model in this paper allows for decision gradients, incorporating the trade-offs between various behaviors rather than just a simple adoption or rejection of a norm. Through support, we incorporate the idea that one can fulfill one's normative obligations in a variety of ways and to a variety of degrees. Agents can demonstrate partial adherence or gradual adoption of a norm. As we have shown in Section 5, this is not incompatible with widespread behavioral shift.

Threshold models focus on structural explanations for why norms emerge rather than emphasizing individually rational grounds for individual behavior or norm emergence. The model being introduced with this paper incorporates the effect of social structure, individual agency, and complex heterogeneity. To illustrate such differences more concretely, we will next contrast our model with the work of Centola, whose recent efforts with threshold models are most closely aligned with those of this paper.

[14]It should be noted that Axelrod's work, and most EGT models, can be readily extended to include networks but such modification is not directly analogous to how this paper incorporates networks.

7.3.1 Comparison to Centola

Centola's work, particularly in his studies on norm diffusion through clustered and random networks, explores how different network topologies influence the spread of social norms. His concept of complex contagions highlights that certain behaviors, especially those that require reinforcement from multiple sources, spread more effectively within dense, tightly connected clusters. This insight contrasts with simple contagions like information, which travel more easily across weak ties in sparsely connected networks (Centola, 2018).

Centola's research emphasizes the importance of redundant social exposure in achieving behavioral change. In his simulations, individuals adopt new behaviors only after seeing them demonstrated repeatedly by several members of their close-knit social circle. This model explains how norms requiring social validation, such as lifestyle changes or political participation, spread through specific types of networks. Furthermore, Centola's model captures the contextual nature of adoption thresholds, illustrating how norms fail to diffuse in fragmented or weakly connected networks despite initial exposure.

While both the proposed model and Centola's emphasize the influence of network structures on norm diffusion, the core mechanisms driving behavior differ significantly. In Centola's model, the diffusion of norms relies on exposure thresholds that depend on multiple reinforcing signals, whereas the model in this paper introduces support as a continuous measure of norm alignment. This means intensive, not just repeated, interactions can lead to norm proliferation. This innovation allows for two advantageous features in this paper's model. First, it allows for partial adherence to a norm. Second, it allows for norms to be transmitted over sparse connections under certain circumstances. Namely, when a person adhering to a norm has strong influence in each connection despite having few connections.

Another key distinction is temporal adaptability. Centola's framework explores how norms spread across networks over time, but individual behaviors are static during the simulation. Agents either adopt or reject the norm based on their exposure threshold. In contrast, this paper's model allows agents to adapt their behavior dynamically, making use of

habit formation as a stabilizing force even when external social pressure changes. This feature enables the current model to simulate more gradual adoption processes and explain how behaviors persist beyond initial exposure or after removal from the social network entirely.

Finally, our model integrates elements of rational behavior. It assumes that agents' behaviors reflect a combination of native preferences, external social pressures, and the inertia created by habits. This allows our model to bridge network and rational-choice perspectives, offering a more comprehensive account of norm dynamics that considers both individual rationality and social influence. The absence of rational foundations in Centola's model makes it less effective at explaining persistent non-conformity or situations where individuals adopt norms selectively. For example, rational-choice approaches can model scenarios where individuals deviate from norms to maximize personal utility, even under social pressure; something Centola's framework is not designed to capture.

Together, these models offer complementary insights into the dynamics of social norms. Centola's model excels at explaining how complex behaviors spread through specific types of networks where redundancy is crucial. In contrast, the model presented in this paper provides a framework that can account for variability in individual behavior as well as network conditions, offering additional insights into how norms develop. While Centola's work emphasizes the conditions for behavioral diffusion under certain circumstances, the present model extends this analysis by showing how such diffusion can happen under rationality, more complex heterogeneity, and more diverse network structures.

7.4 Reasoned Action Approach

RAA, rooted in social psychology, posits that individuals' behaviors are driven by attitudes, perceived behavioral control, and perceived norms. The model includes these ideas by incorporating attitudes and control under native preference and by endogenizing the perceived norm through observations of support. RAA, in contrast, treats norms as external and fixed. Our model incorporates a mechanism for individuals to shape norms through their actions,

rather than just responding to them. This paper represents an effort to link micro-social theories like RAA to macrosocial studies.

7.5 DeGroot Learning

Mathematically, the DeGroot family of models is quite similar to the one outlined in this paper. Particularly the variant explored by Friedkin & Johnsen (1990). The structure of our model distinguishes itself from that variant in that both the network and anchor position (native support) of each agent are time variant. Variations in these values allow for shocks that create the desired long run dynamics as well as allowing for long run equilibrium behavior which is not bounded by $t = 0$ conditions. The additional abstraction layer of support is also novel in this context.

More significantly, DeGroot models are typically applied to learning information conveyed through a social network. The model in this paper, on the other hand, is designed to explain the formation and maintenance of social norms, which may not always lead to consensus but instead reflect the ongoing influence of personal and social factors on support values. To that end, this paper defines each term differently, includes justification for the novel application, and proposes rational foundations for such usage.

8 Empirical Considerations

Although this research is theoretical in nature, it also aims to provide a foundation for future empirical research by offering both conceptual tools and practical insights into how social norm dynamics might be observed in the real world. This section explores four complementary aspects of how the model connects to empirical work. First, some key propositions generated by the model are laid out, suggesting readily testable empirical predictions. Second, some potential advantages the model provides over previous work are discussed. Third, the challenges of dealing with large datasets are examined. Finally, guidance is provided on

the measurement of key constructs within the model.

This section demonstrates that the theoretical framework developed here extends beyond abstract concepts, offering pathways for future empirical validation and practical application. The topics presented highlight how the forces described in the model manifest in real-world scenarios. The propositions and examples provide empirical insights and suggest fruitful avenues for future research. The discussion of measurement suggests how researchers might operationalize the model's constructs, laying the groundwork for additional empirical testing. By addressing both measurement strategies and real-world applications, this section serves as a bridge between theory and empirics, showing how the model can inform research and guide inquiries into social norms.

8.1 Key Propositions

The proposed model offers a structured account of how individual decision-making gives rise to collective normative behavior. While it provides precise predictions, capturing the full resolution of its output requires extensive data collection. Such an ambitious effort is not always feasible using standard empirical methods. However, the model also generates a number of broad claims that lend themselves to more general testing.

This section presents a series of structured, testable propositions that distill the model's key theoretical implications into falsifiable claims. Each proposition highlights a specific normative dynamic and suggests empirical approaches for determining its validity. By focusing on these targeted claims, researchers can evaluate the major aspects of the model without needing to account for its full complexity in a single study. Each proposition is stated explicitly, followed by a discussion of its implications for normative behavior and an outline of potential empirical approaches for testing it.

8.1.1 Proposition 1: Substitution Under Behavioral Restriction

> **Proposition 1** *When a norm-supporting behavior is externally restricted, individuals will compensate in the short run by increasing complementary behaviors that serve the same normative function. The magnitude of this compensatory shift is proportional to the normative importance of the restricted behavior relative to the availability and cost of substitutes.*

This follows from the inclusion of habit and multi-behavior support in the model. When a behavior that is associated with a norm becomes restricted, whether by law or a dramatic increase in cost, it is effectively no longer available to signal support. Nonetheless, because habit reinforces prior support levels in the short run, agents will choose alternative behaviors in order to maintain support requirements. This leads to a temporary increase in complementary support behaviors before moving toward a new equilibrium.

The practical consequence of this is that individuals do not immediately abandon a norm when a focal behavior is restricted. Instead, they adjust by engaging in alternative behaviors that fulfill a similar normative function. For instance, if the right to vote were taken away, the immediate effect would not be a decrease in support for democratic ideals. Rather, there would be an increase in pro-democratic discussion and demonstration among those who value democracy as they make use of these alternative outlets to express normative support. The magnitude of this compensatory shift depends on the extent to which the restricted behavior is integral to the norm and whether viable substitutes are available.

On a smaller scale, Proposition 1 could be used to study the effects of workplace break restrictions. In this context, if socializing breaks for a company's employees were banned, any existing norm of informal workplace socializing would, at least temporarily, cause a shift toward other related behaviors. Suppose a company implements a stricter break policy, which severely hinders workplace socializing. According to the proposition, employees who previously relied on these breaks for social bonding and informal collaboration

will immediately compensate by increasing engagement in alternative forms of workplace interaction. Examples of this include sending more informal messages, lingering longer at the water cooler, or turning routine meetings into more social interactions.

A quasi-experimental study could analyze employee behavior before and after such a policy is introduced, tracking changes in messaging volume, meeting length, and informal in-office interactions. A difference-in-differences approach could then compare offices with and without such restrictions, measuring whether employees in restricted offices increase substitute social behaviors. Such a hypothesis might also be tested through an observational study. Researchers could track whether employees shift to alternative break-like behaviors by comparing organizations that recently implemented break-restricting rules to those that have not. If Proposition 1 holds, employees will find ways to compensate for the removal of informal breaks by increasing other behaviors that fulfill the same normative function, at least temporarily.

Beck et al. (2024) provides an existing study that, with slight modification, could be conducted to empirically test this aspect of the model. They examined compensatory behavior in adolescents' physical activity by tracking natural deviations from habitual exercise levels and identifying instances where individuals compensated for reductions in activity with increased alternative physical behaviors. Their study employed a mixed-methods crossover design, using self-reported habitual activity schedules, smartphone-based activity diaries, and follow-up interviews to analyze compensation patterns. To test Proposition 1, a similar methodology could be employed, but with a systematically imposed restriction on a norm-supporting behavior rather than relying on natural and idiosyncratic fluctuations.

8.1.2 Proposition 2: Long-Term Effects of Policy

Proposition 2 *In the long run, policies aimed at norm modification are most effective when alternative norm-supporting behaviors are costly or difficult to adopt. When substitutes are low-cost and accessible, individuals shift toward these alter-*

natives, sustaining the norm despite efforts to restrict focal behaviors.

Whereas Proposition 1 addresses short-term effects, Proposition 2 concerns itself with long-term or equilibrium consequences of compensatory behaviors.

The extent to which the norm ultimately persists depends on the availability and cost of alternative norm-supporting behaviors. This is because in the long run, native support dominates habit. If substitute behaviors are cheap and readily available, individuals can easily transition to these alternatives, mitigating the long-term impact of the policy on the norm. In contrast, if substitutes are expensive, inconvenient, or otherwise very costly, norm adherence will ultimately decline as individuals find it too taxing to maintain.

For example, researchers could examine smoking rates in communities that ban public smoking, chewing tobacco, and vaping versus those that banned only smoking, to determine whether restricting substitute behaviors better leads to a long-run decline in nicotine use. Even without direct restrictions on substitute behaviors, researchers can still explore alternative testing opportunities. A natural experiment could analyze consumer behavior in regions where a ban on a norm-supporting behavior was introduced, with and without easy access to substitutes. When smoking is banned indoors, people may be forced to smoke outside in designated areas, but for smokers in Florida, such an alternative is more attractive than it would be for smokers in Alaska, particularly during winter months. By tracking behavioral adaptation over time, researchers could assess whether very costly substitution is merely a temporary adjustment or whether it sustains the targeted norm in the long run, contradicting this proposition.

This suggests that empirical tests of long-run policy effectiveness should focus on whether norms shift or change over time, conditional on the cost of alternative behaviors. A robust empirical test would involve a longitudinal comparison of policy interventions that restrict a norm-supporting behavior, with variation in the availability and cost of substitutes. A strong research design would control for pre-policy trends and economic conditions over time to isolate the policy's effect on norm persistence.

A study like that of Akee et al. (2010) would be a strong candidate for testing Proposition 2. This longitudinal analysis leveraged a natural experiment in which an exogenous increase in household income from casino revenue distributions allowed researchers to observe long-term behavioral changes. For the relevant test, a comparable design could be used, incorporating the additional step of collecting data on the availability and cost of alternative behaviors associated with previously established norms.

The findings of such studies would have important implications for policymakers designing norm-diminishing interventions.[15] To ensure the long-term effectiveness of such policies, regulators must not only restrict certain behaviors but also consider whether alternative norm-supporting behaviors will emerge to replace them. If substitution effects are anticipated, additional policy mechanisms, such as raising the cost of substitutes through taxation or restricting their availability through complementary regulations, may be necessary to prevent the long-run persistence of the undesired norm. Understanding these substitution effects will help in crafting policies that not only create immediate behavioral change but also weaken or strengthen norm adherence in the long run, leading to more lasting social transformation.

8.1.3 Proposition 3: The Role of Influencers

> **Proposition 3** *Individuals with high social leverage who gain high exposure can shift population target support toward their own. Sustaining this effect in the long run requires continued exposure.*

In this paper's model, when a high-leverage individual receives substantial exposure, others who observe this directly shift their own support to more closely align with that individual, all else being equal. If this exposure is widespread enough, the macro-social effect can be substantial. The model further implies that this effect is time-sensitive. Without continued

[15] As opposed to behavior interventions, which are more straightforward. To diminish a behavior, effectively outlawing it would be sufficient.

reinforcement, the impact of a high-leverage individual's actions will diminish at a rate inversely proportional to habit.

This diminishing effect, illustrated in Figure 10, occurs because habit formation alone cannot sustain a new norm indefinitely. Once this exposure is removed, native preferences begin to reassert themselves, gradually reducing long-term adherence to the newly shifted support.

If a high-leverage individual demonstrates a specific norm through a single widely covered media event, for example, there is expected to be a temporary increase in the norm's adoption, which will decay unless reinforced through sustained exposure. Empirically, this should be clearest if the high-leverage individual is relatively uninfluenced by others. This is because if they are also subject to heavy influence themselves, the feedback influence will mute the difference between their support and that of the audience prior to exposure. For example, between two public figures of equal stature and sustained exposure, the one over whom the audience has less leverage will have a greater normative impact in their respective network.

Proposition 3 suggests that the news cycle can drive large changes in collective behavior. Media coverage enables widespread and unidirectional exposure of individuals who are frequently of high status. See the discussion of the "Angelina Jolie Effect" in Section 8.2.1 for a more detailed example of this effect. Possible tests include a longitudinal survey to track behavioral shifts in a population exposed to high-leverage influencers over time or a field experiment in which different social groups receive norm-promoting messages from either high-leverage influencers or low-leverage peers. Additionally, social media data could be analyzed to examine how changes in influencer advocacy impact public sentiment over time.

A study such as that of Jackson & Darrow (2005) could be modified to empirically test Proposition 3. This experiment exposed young adults to political endorsements from celebrities through controlled media presentations, including televised interviews and campaign

advertisements. Participants' political opinions were measured before and after exposure to assess the immediate impact of celebrity influence and strong evidence was found for the effectiveness of such exposure. However, the study examined only short-term reactions and did not explore whether these effects persisted over time. To test Proposition 3, a similar methodology could be employed with a more longitudinal design, where some participants are repeatedly exposed to norm-promoting messages from the same high-leverage individuals over an extended period. By tracking changes in participants' behaviors and support levels at multiple time points, researchers could assess whether continued exposure is necessary to sustain the initial influence of high-leverage individuals, thereby providing empirical validation for the proposition's emphasis on the importance of sustained visibility for long-term normative change.

8.1.4 Proposition 4: Endowment and Social Influence

> **Proposition 4** *In any interaction where exposure occurs ($e^t_{i,j} \neq 0$), the social influence of individual j over individual i increases in j's endowment.*

Given that influence is a function of exposure and leverage, and that leverage increases in endowment, we arrive at Proposition 4 directly from Equations 4 and 11. In the model, endowment amplifies an individual's ability to shape the normative choices of others by increasing the influence their previously expressed support carries. When two individuals interact, the one with greater endowment has a larger effect on the behavior of the other when controlling for exposure. While endowments salient to the norm can be expected to have the most practical impact, (e.g., fashion guidance would be more well-received from a successful fashion designer) this assertion applies to all forms of endowment. All else being equal, the impact one agent's support has on others increases in wealth, beauty, power, authority, or any other observable asset or right.

As this proposition is micro-focused, it may be well-suited to a lab experiment. A

researcher might pair participants with differing levels of perceived authority, knowledge, or material resources and assign them a task requiring norm formation, such as deciding on the appropriate response to a novel social etiquette scenario. By tracking changes in behavior after interactions, researchers could measure the extent to which the lower-endowed individual shifts toward the norm expressed by the higher-endowed individual.

A study such as that of Asch (1956) could be modified to empirically test Proposition 4. In the original experiment, participants were placed in a group setting with confederates who intentionally provided incorrect answers to simple perceptual tasks. The study measured the extent to which individuals conformed to the majority opinion, demonstrating the power of social influence. However, the experiment did not specifically examine the role of individual endowments, such as authority or expertise, in influencing conformity. To test Proposition 4, a similar experimental design could be employed where participants are paired with confederates who are presented as having varying levels of endowment relevant to the task such as differing levels of expertise or status. By systematically varying the perceived endowment of the confederates and measuring the degree of conformity exhibited by participants, researchers could assess whether individuals with higher endowments exert greater social influence during interactions, thereby offering empirical support for the proposition.

8.1.5 Proposition 5: The Role of Deviation Costs in Norm Strength

> **Proposition 5** *A social norm will be stronger when the cost of deviating from native preference toward behaviors that support the norm is lower.*

This proposition follows from the inclusion of a coefficient on the native preference term in utility, which represents the cost of deviating from intrinsically desired behaviors. Because an individual's native behavior preferences will typically be misaligned with the prevailing norm, adherence requires individuals to adjust their expressed behaviors at some personal cost. If the cost of deviation from native preferences is high, individuals are less likely to

make large adjustments to coordinate with the norm. Conversely, when the cost of deviation is low, individuals will more readily shift toward behaviors that reinforce the norm, making it stronger. At the micro scale, this suggests that if the norm is something a person would have followed in the absence of social pressure, or if the person does not intrinsically care much about the relevant behaviors, they will adhere more closely to the norm. At the macro scale, this means that a social group will adhere more strongly to the norm if there is generally little instrumental harm in aligning their behavior.

Empirical testing of this proposition could focus on contexts where behavioral shifts occur with varying levels of cost. For example, recycling norms are expected to be stronger in environments where waste sorting is simple, with clearly labeled bins and minimal effort required from individuals. However, in areas where recycling requires complex sorting, extra fees, or special drop-off locations, the norm is predicted to be weaker. This weakness is expected to be observable not only in focal behavior but also in support behaviors such as recycling advocacy.

An empirical study that supports Proposition 5 in this manner is Niu et al. (2023), which investigates the influence of personal costs on adherence to pro-environmental social norms. In this study, participants were asked to engage in environmentally friendly behaviors such as recycling, reducing energy consumption, and minimizing single-use plastics, with varying levels of personal cost imposed such as time, effort, and financial burden. The researchers found that participants were more likely to adhere to pro-environmental norms when personal costs were low, while higher costs led to reduced adherence. This aligns with Proposition 5 by demonstrating that lower costs of deviating from native preferences facilitate stronger adherence to social norms. To test the proposition more directly, future studies could manipulate the cost of deviating from self-interest in controlled environments, such as by varying the complexity of recycling tasks or the financial incentives for non-recyclable waste disposal, and measuring subsequent adherence to recycling norms. This approach would isolate the causal effect of deviation costs on norm strength, providing

direct evidence that when the cost of ignoring the norm is high, adherence diminishes, while lower costs encourage stronger norm-following behavior. Additionally, measuring not just overt behavior but also attitudes and advocacy for the norm under different cost conditions would capture the broader social reinforcement effects that the proposition anticipates.

Proposition 5 is also particularly relevant to voting norms because the costs related to participation are so low. See Sections 8.2.2 and 8.2.3 for further discussion of how these low costs aid in norm formation and adaptability.

8.2 Insights from the Model

The following topics illustrate some strengths of the proposed model by highlighting its capability to address certain social phenomena that other models struggle to capture. This is not to claim that other models cannot accommodate these observations with modification, nor was the primary aim of this research to resolve such issues. Rather, the idea is to convey that explanations for these observations emerge more naturally from this model than the alternatives. This indicates that the model may be especially well-suited for empirical study of such issues and can also be effectively tested and refined through them.

8.2.1 Celebrity Influence

In Section 5.5, we show how an influential individual can sway behavior across a society under the proposed model. This illustrates how community behavior can shift rapidly due to the involvement of celebrities or other prominent figures. Traditional game theory, EGT, and Threshold models do not natively account for such asymmetric influence in descriptive norms, though such effects are quite apparent empirically (Cialdini et al., 1990).

A specific example that demonstrates this issue is the so-called "Angelina effect". This refers to the surge in preventative healthcare behaviors, particularly genetic testing for cancer risks, following American celebrity Angelina Jolie's public announcement that she underwent a preventative double mastectomy in 2013. The number of women interested in BRCA1 and

BRCA2 genetic testing rose significantly after Jolie's announcement, demonstrating how celebrity behavior can influence societal priorities related to health practices. The influence occurred quickly and spread through media channels, reshaping health-seeking behaviors within months (Kosenko et al., 2016).

It is common knowledge that breast cancer should be screened for regularly, so arguments that Jolie's announcement constituted meaningful new information about cancer risk are not convincing explanations of such a shift. If the effect were primarily due to such learning, it would persist as the topic faded from public discourse, though that does not seem to be the case. Jolie herself has little to gain from others screening for cancer, so it also cannot be explained as a disjoint problem in the vein of Coleman. The only remaining conventional rational explanation for such shifts is that the public figure acted as a focal point to signal which strategy everyone should choose in a coordination game. If the benefit were entirely from such conventional norm effects, there would be little reason to shift behavior based on the actions of one individual over any established equilibrium for cancer screening in society at large.

Under the model presented, this shift in behavior is explained by a temporary shock to the network, where Jolie's experience with cancer received wide exposure. This, combined with her high social leverage, meant that the effort she made to protect herself from cancer led others to do the same shortly afterward. Of course, the social outcome was not that there was a massive increase in the focal action (double mastectomies). Rather, in line with the concept of support, individually appropriate measures were taken to combat cancer, the most common of these being to screen for breast cancer.

To illustrate the advantage more clearly, consider a hypothetical sub-population with a low baseline rate of cancer screening that is exposed to news on Jolie's announcement and health progress for some finite period of time. The proposed model predicts a shock upward in screening due to exposure which is sustained so long as exposure continues and decreasing gradually after exposure is removed. Game theory models generally predict no change due to

exposure unless the news transmits information which updates utility estimates. In that case, there would be a change in behavior that would be sustained even after exposure is removed. Threshold models would also predict a sustained increase, but only if Jolie's participation happened to be in a position to cause a network cascade. Otherwise, they would predict no change. EGT says nothing in this case. Only this paper's model is consistent with what occurred.

This example aligns strongly with Proposition 3 and illustrates one of the advantages this type of modeling has over other approaches. In this case, by explicitly modeling singular, asymmetric influences, such as those exerted by celebrities, the resulting volatility in norms is explained.

8.2.2 The Emergence of a Voting Norm

This section argues that alternative theoretical methods are less likely to predict the emergence of norms under certain conditions. We will use voter participation as an illustrative case.

Each theoretical framework we have discussed provides a different perspective on the emergence of widespread voter participation. Coleman's reasoning suggests that a coalition of individuals who benefit from high voter participation coordinate their efforts to sanction others into voting, thereby establishing voting as a norm. Bicchieri's model, on the other hand, focuses on expectations. In this view, a voting norm occurs because people expect others to vote and believe they are expected to do so themselves, leading them to vote as a behavioral response. EGT posits that the predisposition to vote could be bred into us through evolutionary processes. Threshold models propose that voting begins with a small number of individuals who vote because they enjoy it or have some other idiosyncratic motivation, which causes a cascade through the social network.

In contrast, the model presented in this paper tells a story of shared understanding, such as the belief that "democracy is good." Demonstrating appropriate support for this

value becomes socially important for each citizen. In the absence of any election, one can demonstrate such support in a variety of other ways, such as speaking about the benefits of democracy or participating in political protests. In cases where democratic elections are available, voting emerges as a cheap and accessible way to display one's alignment with this shared value. This aligns with Proposition 5, which suggests that norms become stronger when the cost of deviation is low.

The various explanations these different models provide for the emergence of a voting norm are different, but not fundamentally so. Determining which framework is most useful is partly an empirical question, but it is also true that the stories being told do not inherently contradict each other. Nevertheless, the nature of the model proposed in this paper does makes it better suited to predict the emergence of voter participation. It provides clearer guidance on how to measure factors related to the relevant norm and makes more precise predictions about both short-term and long-term outcomes. Unlike other frameworks, it better addresses the practical challenges of empirical application and provides greater specificity.

First, compared to alternatives, this model offers greater empirical applicability to voting norms. This stems from the greater detail in the proposed utility function. Game theory-based approaches often leave the underlying utility functions undefined to maintain generality, making them difficult to operationalize. Bicchieri's model, though somewhat specific, relies heavily on measuring internal beliefs, which are inherently harder to observe and quantify than behaviors or other external factors. While the model in this paper also generalizes, most notably in its treatment of native preferences, this factor is relatively unimportant in the context of voting, where direct self-interest plays a negligible role. This is because there are few reasons to vote which satisfy simple self-interest. This topic is discussed in greater detail in Appendix A. Instead, the habit formation and social influence components of this model are expected to dominate, and these components are given an explicit functional form. This reduces the burden on empiricists by offering clearer guidelines for measurement and analysis.

Furthermore, conventional game theory and EGT approaches introduce additional challenges by requiring empiricists to identify equilibrium states, a task that is as much art as science. There is no guarantee that such equilibria exist, and even when they do, their identification often involves significant uncertainty. In this sense, the model presented provides a prediction where many alternative models do not.

Alternative approaches provide limited insight into short-term behavioral shifts. Conventional game theory offers no dynamic framework, and EGT focuses on the evolution of coordination rather than the mechanisms driving it. Threshold models, while dynamic, often rely on overly simplistic mechanisms for explaining behavioral shifts. By contrast, this model provides precise predictions for individual agents at every point in time, rather than merely estimating the likelihood of a particular outcome or direction of movement, as game-theoretic solutions often do. By making predictions that are both precise and dynamic, this model is uniquely equipped to explain the emergence and stabilization of voting norms in ways that other frameworks cannot.

8.2.3 Variations in Voter Participation

Even in cases where high voter participation is already established, the proposed model appears to offer some additional empirical insight over the alternatives. One way the model distinguishes itself is by explaining the empirical observation that voters are more likely to participate in high-profile elections. In the United States, for example, voter turnout is significantly higher during presidential races than interim elections. The proposed model attributes this variability to dynamic changes in exposure across the network. That is, during presidential elections, there is significantly more discussion around the importance and stakes of voting in both the media and interpersonal communication. This difference is significant and regular even though the outcome of certain local elections may have more direct and practical impact on an individual voter and their community. Furthermore, a single vote in such elections is more likely to influence the outcome. This problem challenges

traditional rational-choice explanations, including Coleman's framework.

In the literature on norms, such variance would typically be handled as an exception by saying the norm is conditional, and that it is more important to vote if it is not an interim election. However, such conditionality is difficult to explain in this context as it is not clear why such a condition should exist, or how it comes about. Whatever established model one wishes to apply, the source of this conditionality is left purely to conjecture. The presented model, however, points to increased exposure to election topics from influential individuals as the driving force. The effect of the changes in exposure to the topic of voting pre- and post-presidential election could operate in a manner similar to that shown in Figure 10. During the interim election, the influence network is weak. Then, as the presidential election approaches and media around the importance of voter participation is maximized, conformity peaks. After the election, social influence on the topic decreases and so does the corresponding coordination on support until the next presidential election.

8.2.4 Harmful Behaviors

A central challenge in understanding apparently harmful norms is explaining why they persist despite their evident inefficiencies or detriments to individual and collective welfare. Traditional models, such as those grounded in evolutionary game theory or coordination games, often presume that norms emerge and persist because they optimize group welfare. These models struggle to account for norms like hazing rituals, smoking, or self-injury, which endure despite widespread acknowledgment of their harmfulness. This model addresses these gaps by incorporating the concept of support, which captures how behaviors align with a norm not solely through individual adherence but as part of a broader signaling process.

Unlike simpler models that view norm adherence as a binary comply-or-defect decision, this model recognizes that individuals engage in a constellation of interrelated behaviors that collectively indicate their support for a norm. For example, in the case of hazing rituals, enduring physical or psychological harm serves as a public signal of loyalty to the group.

These actions not only align with the group's expectations but also reinforce the individual's identity as a committed member. By emphasizing the interconnectedness of behaviors under a single norm, the model explains how even costly or harmful actions can accrue net personal benefit, such as increased status or belonging, even though the practice it is costly for all involved when considered in isolation.

Additionally, the model integrates habit formation and social influence to illustrate how harmful norms become stabilized over time. Habit formation ensures that a support level, once established, becomes a default response, reducing the likelihood of deviation even when external conditions change. Social influence can magnify this effect by creating a feedback loop, where observing others' adherence to the norm reinforces an individual's interest in conforming. Together, these forces create a mechanism for norm persistence that does not rely on the assumption of group-level optimality, setting this model apart from alternatives.

Unlike traditional frameworks, which often assume that norms dissolve when they no longer serve a clear utility, this model demonstrates how harmful norms can persist due to their role in signaling and social cohesion. This signaling reinforces the norm's legitimacy, even as external campaigns highlight its harms. The model's nuanced understanding of support as a dynamic, multidimensional measure enables it to capture these complexities, offering a superior explanation for the endurance of harmful norms. By addressing both individual motivations and network-level dynamics, this framework provides insights that other models fail to deliver.

8.2.5 Pluralistic Ignorance

This model potentially resolves a notable conflict between rational choice theory and the concept of pluralistic ignorance discussed in Section 2.6.1. Pluralistic ignorance, as defined by Perkins & Berkowitz (1986), suggests that individuals systematically misperceive others' behaviors and attitudes, resulting in erroneous beliefs about how others behave and biased choices. However, rational choice models typically assume that individuals should have no

such persistent biases.

This inherent tension arises from the belief that rational agents should, in theory, correct misperception through something like Bayesian updating. Even if the estimates made are not precise because of incomplete information, the expectation should at least remain unbiased. Yet, empirical evidence of persistent norms, such as alcohol use, suggests that these misperceptions are stable and endure, contradicting the assertions of conventional rational choice theory.

The model presented in this paper potentially reconciles these seemingly opposed perspectives by suggesting that the conflict may be the result of the difference between mean and target support. That is, the survey was interpreted as measuring estimates of mean behavior across all college students without accounting for how the influence network weights the behavior of each student differently. What was relevant to the students' desires to fit in, and what they were actually responding to, may have been behavior weighted by influence. If that is the case, there is no misperception on the part of the students. Given the survey results, the model predicts that students with higher social influence are, in fact, more likely to consume alcohol, and that in turn should drive a general elevation in consumption.

8.2.6 Persistence of Behaviors in Changing Social Contexts

When individuals move to a new cultural setting and sever ties with their previous social environment, they do not immediately shed old behaviors. As individuals make these transitions, established habits exhibit inertia that slows down adaptation to the new environment, influencing their interactions within the new context until new behaviors take root through repeated exposure.

Traditional game-theoretic models struggle to account for the persistence of behaviors during social transitions. These models often assume that behaviors are determined by immediate payoffs and fixed strategies, which makes them more suited to explaining equilibrium states or short-term coordination problems rather than dynamic, long-term behavioral

change. They do not adequately capture the stabilizing effects of habituation or the cumulative influence of social exposure over time. As a result, game-theoretic approaches fail to conclusively address why some individuals might continue to engage in familiar behaviors after a significant change in cultural context, particularly when those behaviors no longer provide clear utility or align with the new environment's norms. The model discussed here not only predicts such lag, it also shows how one's social placement in the new society predicts the speed at which such adaptation occurs.

8.3 Managing Large Parameter Sets

A significant difficulty in the approach presented in this paper arises in the large number of parameters involved. Given a population size p, a maximum number of time periods $t_{\max}$, a number of measurable behaviors m, as well as network density δ, suppose we aim to fully test the model on a real-world community. To do so, we may require up to the following:

p endowments ($\vec{V}$),

$\delta t_{\max}(p^2 - p)$ exposures ($\mathbf{E}^t$),

p habit terms ($\vec{H}$),

$pt_{\max}$ native support terms ($\dot{\vec{S}}^t$),

$mpt_{\max}$ behaviors ($\vec{B}^t$),

m support weights ($\vec{W}$).

Thus, the maximum number of measurements required to make a precise prediction with the model can be expressed as

$$[\text{Max measurement count}] = \delta t_{\max} p^2 + (2 + t_{\max} + m t_{\max} - \delta t_{\max})p + m. \tag{52}$$

For instance, in a system with 100 people, 10 behaviors, 10 time periods, and a network density of 0.1, the upper bound on the number of scalar values needed to input or compare is 21,110. This number can be reduced by about an order of magnitude if shocks are limited, but since at least one complete social network is required for the model to function, the parameter count grows at a rate of $O(p^2)$ for tightly connected networks.

If the influence network values are taken as given, the parameter growth rate becomes $O(pmt_{\max})$, which is an improvement as p would typically be the largest value. However, this situation remains far from practical. Ignoring the network variables, the maximum number of scalars in the 100-agent example above reduces to 11,200. In a system with few shocks, this number could be as low as 1,400. While the model is testable directly in principle, this large number of parameters makes such tests impractical on all but the sparsest of networks.

Still, less direct methods are available. Most obviously, agent-level testing is a crucial first step in model validation which does not require extensive data collection. By verifying that agents' actions at the micro-level reflect realistic decision-making and social dynamics, researchers can ensure that the foundation of the model is accurate before moving on to broader system-level validation. This step ensures that agent behaviors are grounded in empirical evidence. Tests of propositions derived from the model, such as those highlighted in Section 8.1 can also help validate and refine the model. From there, calibration and sensitivity tests can be conducted based on known data to test larger scale predictions.

8.4 Measurement of Parameters

This section focuses on the conceptual groundwork for measuring parameters in the model which are new to the literature. The purpose of this discussion is to highlight that these parameters can be measured in principle and offer guidance on how they can be operationalized.

8.4.1 Native Support $\dot{s}_i^t$

Native support can be approximated using any established utility function that excludes factors already accounted for elsewhere within the model. The process for doing so involves solving for optimal behaviors and applying the vector $\vec{W}$. The result is $\dot{s}_i^t$. If the underlying utility function has been validated, conforms to Assumption 3, and $\vec{W}$ has been validated, then the resulting $\dot{s}_i^t$ is valid. More directly, though $\dot{s}_i^t$ can be determined by isolating the agent from social influence, observing their steady-state behavior, and calculating the corresponding support value, which estimates $\dot{s}^t$. This can also be deduced from the movement of s_i^t values outside of equilibrium; an agent placed in a variety of environments s_i^t will converge more quickly on τ^t values which are closer to $\dot{s}^t$.

8.4.2 Habit Formation Factor h_i

Habit formation in an individual can be empirically measured by observing the speed at which an agent adjusts their support when transitioning between two stable network environments. An agent starts in an initial network where their support has reached equilibrium, meaning their support value remains steady over time. When the agent moves to a new network environment, they will begin adjusting their support to align with the new τ^t. The rate at which their support converges to the new steady state is indicative of their habit formation factor h_i. Agents that settle into a new steady state quickly have lower habit formation factors, while those who take longer to adjust have higher h_i. This movement caused by habit is distinguished from the movement caused by the difference between native support and influence in that it is symmetrical; the rate of convergence will be evenly paced irrespective of the direction of movement.

8.4.3 Support Weights $\vec{W}$

In the model, it is assumed that the vector $\vec{W}$, which represents the weights assigned to various behaviors, is known to all agents within the network. As a result, both the agents

and external observers (e.g., empiricists) are expected to have some understanding of the behaviors that contribute to support for a given norm. For example, in the case of a voting norm, the act of voting itself is a clear and significant contributor to support. Additional actions, such as encouraging others to vote, providing voting incentives, or advertising one's participation, are also behaviors that align with this norm. Relying entirely on such guesswork is not ideal, however. Given two competing $\vec{W}$ values, it is desirable to be able to determine which definition of $\vec{W}$ is most appropriate in capturing support for the norm.

In well-connected and mature networks, where a norm is strongly established, the support values across agents are expected to be relatively uniform within network clusters. However, individual support levels in those clusters will still vary, allowing competing definitions of $\vec{W}$ to be applied to generate differing support values. The preferred definition of $\vec{W}$ is the one that minimizes the variation in support within each cluster while also explaining the differences in support between clusters. In this way, the chosen $\vec{W}$ better reflects internal consistency within network clusters and accounts for broader differences across the network. In principle, the parameter space of $\vec{W}$ can be searched using this method to objectively calculate support weights.

8.4.4 Target Support τ^t

Target support τ^t is a macrosocial measure derived from micro-social properties. As such, it can be determined from estimates of the agent-level variables. Still, it may be useful to measure it more directly. It can be approximated by sampling individuals, measuring their behaviors, applying $\vec{W}$ and calculating mean support. While this loses the effect of the influence network, such a mean support measure should approximate well in large, mature, or well-connected networks. For more precise measurement where local exposure and leverage can be observed, one can weigh such means by their influence.

9 Conclusion

This study set out to develop a deterministic model of social norm dynamics that is both explicit in its mathematical formulation and broadly applicable across various social contexts. It integrates elements from several other works: The rational foundations of game theory, behavioral extensions inspired by Bicchieri, EGT, and RAA, and network considerations gleaned from DeGroot learning and threshold models. Unlike traditional game-theoretic models that focus on static equilibria and are best suited to small-scale interactions, this model emphasizes dynamic processes and scales seamlessly. These features give it the potential to bridge a crucial gap in understanding the temporal development and stabilization of norms in large-scale contexts.

The additional introduction of support as a measure of behavioral alignment with social norms provides a nuanced view of normative behavior over simpler models that focus solely on single behaviors. Despite this difference, graceful degradation of support into focal action allows it to generate directly comparable micro-level predictions to a variety of existing models. Computational simulations further demonstrate the model's potential to capture how network structures and social connectivity influence the emergence and persistence of norms.

However, as with any new theoretical model, limitations must be acknowledged. While simulations provide valuable insights, real-world testing is necessary to validate the model's applicability and robustness. Empirical challenges such as collecting reliable data, defining appropriate metrics, and accounting for complex, context-dependent variables must be addressed. Recent and future advances in computational tools, such as big data analytics and machine learning, offer opportunities to track social behaviors in real time, possibly providing the granular data needed to validate and enhance the model. Meanwhile, validation on a small scale can provide deeper insights into the applicability of this work or points of refinement. Voter participation appears to be a particularly relevant and important early empirical application as no convincing model has yet emerged.

An important next step is to empirically test key propositions derived from this model. By systematically evaluating these propositions through controlled studies and real-world applications, future research can validate its predictive power and refine its theoretical assumptions. Future theoretical research could examine the role of strategic thinking more extensively, specifically exploring cases where it becomes a critical factor in norm outcomes. Additionally, investigating the impact of more complex forms of habituation or nonlinear relationships among behaviors that contribute to support could enhance the model.

Despite such challenges, this preliminary research, rooted in methodological individualism, offers a compelling lens through which to view the emergence and evolution of social norms. By emphasizing how personal preferences, social pressures, and habitual behaviors aggregate into collective patterns, it offers a novel perspective of the reciprocal relationship between agents and their social environments.

References

Abrams, S., Iversen, T., & Soskice, D. (2011). Informal social networks and rational voting. *British Journal of Political Science*, *41*(2), 229–257.

Akee, R. K., Copeland, W. E., Keeler, G., Angold, A., & Costello, E. J. (2010). Parents' incomes and children's outcomes: a quasi-experiment using transfer payments from casino profits. *American Economic Journal: Applied Economics*, *2*(1), 86–115.

Akerlof, G. A. (1980). A theory of social custom, of which unemployment may be one consequence. *The quarterly journal of economics*, *94*(4), 749–775.

Anderson, J. E., & Dunning, D. (2014). Behavioral norms: Variants and their identification. *Social and Personality Psychology Compass*, *8*(12), 721–738.

Asch, S. E. (1956). Studies of independence and conformity: I. a minority of one against a unanimous majority. *Psychological monographs: General and applied*, *70*(9), 1.

Axelrod, R. (1986). An evolutionary approach to norms. *American political science review*, *80*(4), 1095–1111.

Aytaç, S. E., & Stokes, S. C. (2019). *Why bother?: Rethinking participation in elections and protests*. Cambridge University Press.

Barzel, Y., & Silberberg, E. (1973). Is the act of voting rational? *Public Choice*, *16*(1), 51–58.

Becher, M., & Stegmueller, D. (2015). *Rational mobilization of ideological group members in elections: Theory and evidence* (Tech. Rep.). Konstanz, Germany: University of Konstanz.

Beck, F., Swelam, B. A., Dettweiler, U., Krieger, C., & Reimers, A. K. (2024). Compensatory behavior of physical activity in adolescents–a qualitative analysis of the underlying mechanisms and influencing factors. *BMC Public Health*, *24*(1), 158.

Becker, G. S., & Murphy, K. M. (1988). A theory of rational addiction. *Journal of political Economy*, *96*(4), 675–700.

Berkowitz, A. D. (2005). An overview of the social norms approach. In *Changing the culture of college drinking: A socially situated health communication campaign* (pp. 193–214).

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bowles, S., & Gintis, H. (2011). *A cooperative species: Human reciprocity and its evolution*. Princeton, NJ: Princeton University Press.

Centola, D. (2018). *How behavior spreads: The science of complex contagions*. Princeton University Press.

Centola, D., Willer, R., & Macy, M. (2005). The emperor's dilemma: A computational model of self-enforcing norms. *American Journal of Sociology*, *110*(4), 1009–1040.

Chandrasekhar, A. G., Larreguy, H., & Xandri, J. P. (2020). Testing models of social learning on networks: Evidence from two experiments. *Econometrica*, *88*(1), 1–32.

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of personality and social psychology*, *58*(6), 1015.

Cialdini, R. B., & Trost, M. R. (1998). Social influence: Social norms, conformity and compliance. In *The handbook of social psychology* (p. 151-192). McGraw-Hill.

Coase, R. H. (1960). The problem of social cost. *Journal of Law and Economics*, *3*, 1.

Coate, S., & Conlin, M. (2004). A group rule-utilitarian approach to voter turnout: Theory and evidence. *American Economic Review*, *94*(5), 1476–1504.

Codd, E. F. (1968). *Cellular automata*. Academic press.

Coleman, J. S. (1990). *Foundations of social theory.* Belknap Press of Harvard University Press.

Compernolle, E. L. (2017). Disentangling perceived norms: predictors of unintended pregnancy during the transition to adulthood. *Journal of Marriage and Family*, *79*(4), 1076–1095.

Coppock, A., & Green, D. P. (2016). Is voting habit forming? new evidence from experiments and regression discontinuities. *American Journal of Political Science*, *60*(4), 1044–1062.

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

Dellavigna, S., List, J. A., Malmendier, U., & Rao, G. (2016, 10). Voting to Tell Others. *The Review of Economic Studies*, *84*(1), 143-181. doi: 10.1093/restud/rdw056

Downs, A. (1957). *An economic theory of democracy.* Harper New York.

Elster, J. (1989). Social norms and economic theory. *Journal of economic perspectives*, *3*(4), 99–117.

Elster, J. (2003). Coleman on social norms. *Revue française de sociologie*, *44*(2), 297–304.

Engelen, B. (2006). Solving the paradox: The expressive rationality of the decision to vote. *Rationality and Society*, *18*(4), 419–441.

Fallucchi, F., & Nosenzo, D. (2021). The coordinating power of social norms. *Experimental Economics*, 1–25.

Feddersen, T. J. (2004). Rational choice theory and the paradox of not voting. *Journal of Economic perspectives*, *18*(1), 99–112.

Festinger, L. (1957). *A theory of cognitive dissonance* (Vol. 2). Stanford university press.

Fieldhouse, E., & Cutts, D. (2018). Shared partisanship, household norms and turnout: Testing a relational theory of electoral participation. *British Journal of Political Science*, *48*(3), 807–823.

Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior: The reasoned action approach*. Psychology press.

Fowler, J. H. (2006). Altruism and turnout. *The Journal of Politics*, *68*(3), 674–683.

Frank, R. H. (1992). Melding sociology and economics: James coleman's foundations of social theory. *Journal of Economic Literature*, *30*(1), 147–170.

Friedkin, N. E., & Johnsen, E. C. (1990). Social influence and opinions. *Journal of mathematical sociology*, *15*(3-4), 193–206.

Fujiwara, T., Meng, K., & Vogl, T. (2016). Habit formation in voting: Evidence from rainy elections. *American Economic Journal: Applied Economics*, *8*(4), 160–88.

Galais, C., & Blais, A. (2016). Beyond rationalization: Voting out of duty or expressing duty after voting? *International Political Science Review*, *37*(2), 213–229.

Gelman, A., Silver, N., & Edlin, A. (2012). What is the probability your vote will make a difference? *Economic Inquiry*, *50*(2), 321–326.

Gerber, A. S., Green, D. P., & Larimer, C. W. (2008). Social pressure and voter turnout: Evidence from a large-scale field experiment. *American political Science review*, *102*(1), 33–48.

Gerber, A. S., Huber, G. A., Doherty, D., & Dowling, C. M. (2016). Why people vote: Estimating the social returns to voting. *British Journal of Political Science*, *46*(2), 241–264.

Goodin, R. E., & Roberts, K. W. (1975). The ethical voter. *American Political Science Review*, *69*(3), 926–928.

Granovetter, M. (1978). Threshold models of collective behavior. *American journal of sociology*, *83*(6), 1420–1443.

Hechter, M. (2008). The rise and fall of normative control. *Accounting, Organizations and Society*, *33*(6), 663–676.

Hechter, M., & Opp, K.-D. (2001). *Social norms*. Russell Sage Foundation.

Heider, F. (1958). *The psychology of interpersonal relations*. Psychology Press.

Hillman, A. L. (2010). Expressive behavior in economics and politics. *European Journal of Political Economy*, *26*(4), 403–418.

Holme, P., Kim, B. J., Yoon, C. N., & Han, S. K. (2002). Attack vulnerability of complex networks. *Physical review E*, *65*(5), 056109.

Horne, C. (2009). *The rewards of punishment*. Stanford University Press.

Horne, C., & Mollborn, S. (2020). Norms: An integrated framework. *Annual Review of Sociology*, *46*, 467–487.

Hussam, R., Rabbani, A., Reggiani, G., & Rigol, N. (2016). Handwashing and habit formation. *Yale University, Economic Growth Center*.

Jackson, D. J., & Darrow, T. I. (2005). The influence of celebrity endorsements on young adults' political opinions. *Harvard international journal of press/politics*, *10*(3), 80–98.

Jadbabaie, A., Molavi, P., Sandroni, A., & Tahbaz-Salehi, A. (2012). Non-bayesian social learning. *Games and Economic Behavior*, *76*(1), 210–225.

Jasso, G., & Opp, K.-D. (1997). Probing the character of norms: A factorial survey analysis of the norms of political action. *American Sociological Review*, 947–964.

Jindani, S., & Young, H. (2020). *The dynamics of costly social norms* (Tech. Rep.). Oxford: University of Oxford, Department of Economics.

Jones, P., & Hudson, J. (2000). Civic duty and expressive voting: Is virtue its own reward? *Kyklos*, *53*(1), 3–16.

Katz, D., Allport, F. H., & Jenness, M. B. (1931). *Students' attitudes; a report of the syracuse university reaction study.* Craftsman Press.

Keefer, P., & Knack, S. (2008). Social capital, social norms and the new institutional economics. In *Handbook of new institutional economics* (pp. 701–725). Springer.

Kosenko, K. A., Binder, A. R., & Hurley, R. (2016). Celebrity influence and identification: A test of the angelina effect. *Journal of Health Communication*, *21*(3), 318–326.

Kuran, T. (1995). *Private truths, public lies: The social consequences of preference falsification.* Harvard University Press.

Lally, P., Van Jaarsveld, C. H., Potts, H. W., & Wardle, J. (2010). How are habits formed: Modelling habit formation in the real world. *European journal of social psychology*, *40*(6), 998–1009.

Legros, S., & Cislaghi, B. (2020). Mapping the social-norms literature: An overview of reviews. *Perspectives on Psychological Science*, *15*(1), 62-80.

Mäs, M., & Opp, K.-D. (2016). When is ignorance bliss? disclosing true information and cascades of norm violation in networks. *Social Networks*, *47*, 116–129.

Myerson, R. B. (2000). Large poisson games. *Journal of Economic Theory*, *94*(1), 7–45.

Niu, N., Fan, W., Ren, M., Li, M., & Zhong, Y. (2023). The role of social norms and personal costs on pro-environmental behavior: the mediating role of personal norms. *Psychology Research and Behavior Management*, 2059–2069.

Opp, K.-D. (2001). How do norms emerge? an outline of a theory. *Mind & Society*, *2*(1), 101–128.

Opp, K.-D. (2018). Externalities, social networks, and the emergence of norms: A critical analysis and extension of James Coleman's theory. *Social Research: An International Quarterly*, *85*(1), 167–196.

Paluck, E. L., Shepherd, H., & Aronow, P. M. (2016). Changing climates of conflict: A social network experiment in 56 schools. *Proceedings of the National Academy of Sciences*, *113*(3), 566–571.

Perkins, H. W., & Berkowitz, A. D. (1986). Perceiving the community norms of alcohol use among students: Some research implications for campus alcohol education programming. *International journal of the Addictions*, *21*(9-10), 961–976.

Repacholi, B. M., Meltzoff, A. N., Rowe, H., & Toub, T. S. (2014). Infant, control thyself: Infants' integration of multiple social cues to regulate their imitative behavior. *Cognitive Development*, *32*, 46–57.

Riker, W. H., & Ordeshook, P. C. (1968). A theory of the calculus of voting. *American political science review*, *62*(1), 25–42.

Robalino, J. D., & Macy, M. (2018). Peer effects on adolescent smoking: Are popular teens more influential? *PloS one*, *13*(7), e0189360.

Schelling, T. C. (1960). *Strategy of conflict* (1st ed.). Harvard University Press.

Schelling, T. C. (1978). *Micromotives and macrobehavior*. W. W. Norton & Company.

Shachar, R., & Nalebuff, B. (1999). Follow the leader: Theory and evidence on political participation. *American Economic Review*, *89*(3), 525–547.

Sugden, R. (1986). *The economics of rights, cooperation and welfare*. Oxford, UK: Basil Blackwell.

Uhlaner, C. J. (1989). Rational turnout: The neglected role of groups. *American Journal of Political Science*, 390–422.

Ullmann-Margalit, E. (1977). *The emergence of norms*. Oxford: Clarendon Press.

Verplanken, B., & Aarts, H. (1999). Habit, attitude, and planned behaviour: is habit an empty construct or an interesting case of goal-directed automaticity? *European review of social psychology*, *10*(1), 101–134.

Villatoro, D., Sen, S., & Sabater-Mir, J. (2010). Of social norms and sanctioning: A game theoretical overview. *International Journal of Agent Technologies and Systems (IJATS)*, *2*(1), 1–15.

Weber, M. (1922). *Economy and society: An outline of interpretive sociology* (G. Roth & C. Wittich, Eds.). Berkeley, CA: University of California Press.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and social Psychology*, *51*(1), 110.

Yee, A. S. (1997). Thick rationality and the missing "brute fact": the limits of rationalist incorporations of norms and ideas. *The Journal of Politics*, *59*(4), 1001–1039.

Young, H. P. (2015). The evolution of social norms. *Annual Review of Economics*, *7*(1), 359–387.

A Appendix: Voting as a Theoretical Case Study

This brief essay primarily serves as a sort of case study. It highlights one of the many topics where the inclusion of a practical model of social norms is needed and could have consequential impact on our understanding of an influential social process.

Long ago, I wondered "Why is voter turnout in elections so high?" There was more to the question than I had imagined at the time. Though it seemed unlikely that there was a definitive answer, I did assume that there was a rather thoroughly developed mainstream consensus that would satisfy my curiosity. A search of the literature revealed some partial answers, but none that felt conclusive. As you will see, my findings suggest that social norms likely play an important role in voter turnout, I did not find evidence of rigorous integration of the literature from these two areas. The body of work on social norms made for enjoyable reading, but mainstream norm theory was difficult to incorporate neatly into existing closed form voting models at any reasonable scale. This challenge is perhaps the reason why the treatment of norms in voting literature is not fully developed. This ultimately led to making social norms the focus of my own research.

A.1 Voting and Welfare

Common wisdom suggests that the democratic voting process is an effective, if at times crude, attempt at maximizing group welfare. On a small scale, the process certainly makes some sense. When a handful of friends are trying to establish where they should meet, a short discussion followed by a simple vote offers an expedient, low cost, and practical method to arrive at a collective decision. Each friend's vote contributes meaningfully to the expected outcome, which provides the necessary incentive to organize, and participate in, the modest election. As we will see in the next section though, at large scales, this intuitive reasoning breaks down in the context of rational choice. We observe non-trivial levels of turnout in national elections as well. This suggests that something else is motivating participation.

If we can confidently explain why the citizenry votes so reliably, we can also say what factors might alter that turnout. More fundamentally, such understanding would clarify the relationship between individual voter interest and group choice. We might also be able to estimate the consequences of voting as a decision-making process in novel situations or compare democratic outcomes to those of alternative group choice strategies. For these reasons, advancing our understanding in this area is of great consequence.

A person arguing to downplay this point might assert that where voting is used, it represents the best method available to us on functionalist or social evolutionary grounds. From this perspective, it might be claimed that little would be gained by further development of abstract models. It is challenging to say one way or another if that is true without the knowledge that a better model might provide. Still, there are arguments to be made that simple voting is not an ideal system at the scales that we use it.

By way of illustration, suppose that in a large national election, a single would-be voter is selected and informed that they would not be permitted to vote. By the law of large numbers, the outcome of the election would be unchanged, but that individual would be spared the effort of voting. In this case, the cost society pays for the election is reduced and the same political outcome is achieved. This idea could be extended to a jury-style selection of voters: Suppose, 25% of the population of eligible citizens is selected at random each election and only that subset of the population is allowed to vote. Again, by the law of large numbers, the political outcome would be unchanged, now with a significant reduction in the total costs associated with holding that election. Restricting the voter base in this way is an almost trivial effort in countries that require voter registration and already exclude groups from voting (minors, non-citizens, etc.).

These outlandish proposals are not Pareto improvements in the strict sense. Still, such arguments make it difficult to justify an assumption that the democratic mechanisms we use operate in such a way as to achieve their means at minimum social cost. There is a much more likely explanation for the functionalist: Paying these additional costs serves a purpose

outside of election outcomes, a purpose that a well-developed theory can help identify.

Next, we survey the body of existing theory on the subject of voting.

A.2 The Paradox of Voting

Downs (1957) presents a paradox that can be summarized as follows: When we apply simple cost-benefit analysis to voter behavior, the costs of voting heavily outweigh the benefits for a citizen. We therefore expect voter turnout to be extremely low in national political elections. This conclusion conflicts dramatically with observed voter turnout around the world.

In their seminal paper, *A Theory of the Calculus of Voting*, Riker and Ordeshook (1968) elaborate on this argument, while establishing some additional formalization and terminology. They begin by claiming that a rational individual will evaluate the costs and benefits of voting and participate only if there is a net benefit in doing so. Voting, in this case, is best understood as casting a ballot in a political referendum, though the reasoning applies to all discrete choice elections, whether under plurality or majority rule, when the voter base is large. The individual costs of voting in a modern U.S. election may include registering to vote, completing forms, gathering enough information to reliably establish a preference, and traveling to the polling station. With respect to the election outcome, the benefit is stochastic; there is some probability that casting a vote will affect the outcome in the voter's favor.

Borrowing Riker and Ordeshook's notation, the expected utility R from voting is

$$R = PB - C. \tag{53}$$

Here, P is the probability that the agent's vote will change the outcome of the election in their favor. This happens only when the agent's vote would break or create a tie. B is the marginal benefit to that agent, should their favored candidate win over the less favored one. The PB term, then, represents the expected benefit of voting, while C indicates the sum

total of the costs involved in casting the vote. If $R \geq 0$ the rational agent votes, if $R < 0$ they do not.

When we attempt to apply this reasoning empirically, an issue with our P values quickly becomes apparent; the probability of affecting the outcome of a democratic election becomes infinitesimal as the electorate grows to any appreciable size. The mathematics of determining the value of P can get arbitrarily complicated, depending on the specific information set available to the agent and the structure of the election. It is apparent though that when millions of people are voting on an issue, the chance of a single vote affecting the outcome is quite small indeed. The most generous estimate to be found for the odds of a single vote changing the outcome of a U.S. presidential election are 1 in 60 million elections (Gelman et al., 2012). Estimates go as low as 1 in 8 billion for a population of only five million voters (Myerson, 2000; Feddersen, 2004). Let us take the optimistic estimate and also suppose the costs of voting amount to only one U.S. dollar. The value to the *individual* of having the preferred side win would necessarily have to be at least $60 million to justify the effort of participating in the election.[16] It does not seem reasonable to believe that a typical American would rather have their candidate win than receive this astronomical amount as direct compensation. I will refer to this issue later as the *net cost problem.*

To resolve the paradox, Riker and Ordeshook propose adding a term, D, to the utility function, which covers all benefits of voting that are independent of the election outcome.

$$R = BP + D - C. \tag{54}$$

When we remove the trivial BP term, the decision rule reduces to "Vote if $D > C$," where D is the sum total of all benefits associated with voting and C is, again, the sum of the costs of voting. The nature of D is not rigorously examined by the authors, though it is described as being one's "sense of citizen duty."

[16] Additionally, observe that the form of the PB term inherently assumes risk neutrality in the agent. Any amount of risk aversion on the part of the voter would only serve to further trivialize the expected benefit of casting a ballot.

Though BP can be effectively dismissed, there have been notable efforts to re-incorporate election impact in some alternate form. Models that include non-trivial consideration for the election outcomes are described as being *instrumental*. Note that, as we use the term here, it refers only to individual instrumentality, not collective instrumentality or welfare.

One instrumental line of reasoning is that altruism leads people to intrinsically value benefits to others in addition to their own (Goodin & Roberts, 1975; Coate & Conlin, 2004; Fowler, 2006). For an altruist, personal interest in the election includes significant consideration for these outside benefits. This altruism would necessarily extend beyond family and friends such that the welfare of strangers would be the dominant force at work. Evidence for such selflessness is scant. If altruism toward strangers were so prevalent, we should expect to see much greater acts of charity and more frequent small gestures toward strangers than we do.

Studies with data supporting instrumentality are quite rare. Barzel & Silberberg (1973) observe that turnout increases as election outcomes become close. This result, though consistent with instrumentalism, shows a small effect and does not rule out the possibility that some mechanism of collective rationality is at work. Shachar & Nalebuff (1999), for example, make the same empirical observation but develop a non-instrumental theory. We are left to conclude that instrumental reasoning is, at best, only a minor part of the story.

A.3 Expressive Voting

Theories that expand on Riker and Ordeshook's D-term have come to be called *expressive* explanations for voting. According to these, motivations are not tied to an individual's interest in impacting election itself. Expressive models, in turn, are further subdivided into

intrinsic and *extrinsic* types.[17]

Intrinsic explanations include the desire to participate as a fundamental component of the utility function. This is analogous to declaring that one gets simple pleasure from cheering for a sports team, supporting one's self-image as a supporter of a certain ideal, or simply adheres to their "honorable duty" (Galais & Blais, 2016; Jones & Hudson, 2000; Engelen, 2006; Aytaç & Stokes, 2019).

The intrinsic expressivist view has some intuitive appeal; many can relate to feelings of duty and responsibility with respect to voting. The weaknesses with this way of thinking are primarily epistemological. First, this sort of preference is not falsifiable. If someone votes, we say it is because they enjoy voting. If they do not vote, they simply do not enjoy it enough. We disregard how the preference gets there, why it differs across individuals, or how it may be expected to change in the future. This limits the usefulness of any theory built on such assumptions. We are, in essence, *forcing* rationality on the agent. In defining their utility in direct response to observed behavior, we sidestep any deeper understanding of the mechanism at work.

By analogy, to conclude that we vote for intrinsic reasons is akin to saying we eat because we are hungry. While such a claim is not incorrect, and there is value in recognizing that hunger exists, we cannot predict when someone is going to eat if we only know that they will do so when some unobserved feeling tells them to. Ideally, we also recognize that there are important underlying causes and indicators of hunger, for example, one's blood sugar may be low. Using blood sugar levels as an explanation, even if it is also an incomplete one, is more objective and highlights a more fundamental mechanism at work. It is true

[17]There is a semantic issue in the literature that should be noted. Initially, it was natural to refer to the election outcome component of the voting equation as instrumental, and the apparently non-instrumental "duty" component as expressive. However, as social scientists took a closer look at Riker's D-term, an issue became apparent. There may be benefits to voting that are not related to the outcome of an election and are none the less instrumental in the self-serving sense. As a general rule, a large majority of expressive models are intrinsic. So much so, that in many cases "expressive" is taken to imply intrinsic expressive, as in the case of Hillman (2010). Given the common definition of the word "expressive," it's easy to see how such confusion occurred. The term "extrinsic expressive" is used here with some reservation, as "extrinsic instrumental" may be more appropriate but I chose to keep as consistent as possible with the broad body of literature.

Descriptors		"I vote because..."
Instrumental		...I want to ensure my candidate wins.
Expressive	Intrinsic	...it feels good to do the right thing.
	Extrinsic	...others are giving me incentive to do so.

Figure 13: Voting Model Categories and Associated Reasoning

that modeling hunger in such detail may not be necessary for most purposes. Hunger is such a fundamental and universal human experience that modeling it intrinsically is often an acceptable simplification. This argument cannot be made in the case of voting because such activity does not appear to be a fundamental part of human nature. As democracy is a relatively recent invention on an evolutionary timescale, there is little reason to believe that the act of voting itself is hardwired into the human mind. Thus, intrinsic modeling leaves important questions unanswered.

Extrinsic reasoning, by contrast, involves linking voter motivations to more convincingly fundamental needs. If candidates, say, paid voters directly for their participation, we could easily create an extrinsic model that exploits the desire for economic gain to explain voter turnout. An extrinsically expressive explanation for voter participation has the potential to avoid the rational failings of instrumental arguments and the empirical dead end of intrinsically expressive models. The cost of having these features is the added challenge of directly observing such an incentive system, which has proven difficult.

The extrinsic literature is dominated by group-based mobilization (GBM) models. They, arguably, represent the mainstream view insofar as one exists. The common theme among GBM models is that individual actors are provided with incentives to participate by community leaders. Voters are not instrumentally driven in this case; they are merely responding to outside incentives. It is the leaders, with their command of a nontrivial number of votes, who then drive election outcomes (Uhlaner, 1989; Shachar & Nalebuff, 1999; Feddersen, 2004; Becher & Stegmueller, 2015).

In a sense, this is a hybrid of instrumental and expressive thinking. The instrumentality

is apparent in the characterization of leaders. They are the de facto instrumental voters, while the rank-and-file de jure electorate responds only to extrinsic incentives. Because of their popularity in the literature, I will highlight some issues with GBM models in particular.

1. The mechanism by which leaders monitor the actions of their community members is typically assumed to exist without further theoretical grounding.

2. To avoid falling back into the net cost problem, only voting blocs that are large enough to have a non-trivial chance of being pivotal are predicted to exist.

3. There are no established empirical, or even stylized, facts that demonstrate the existence of a compensation distribution network within voting groups.

4. If one views the benefit of an election outcome to a community as the summation of benefits to each of its members, then the welfare function is essentially equivalent to that of the individual utility function. When the sum of individual costs (which are effectively transferred to a leader in these models) exceeds individual benefits, the same would be true of the collective. This means there is little incentive to organize on rational grounds.

5. Perhaps most critically, such models do not explain the high turnout of voters who are apparently unaffiliated with any organized group.

A.4 Social Norms and Voting

Recent empirical research provides promising evidence that social pressure has a meaningful influence on voter participation. This suggests that social norms may play an important role in voter turnout (Gerber et al., 2008, 2016; Dellavigna et al., 2016; Fieldhouse & Cutts, 2018).

As a purely extrinsic explanation, this line of research avoids the challenges of implausibility that come with instrumental approaches and the tautological traps of the intrinsic models. Since social norms are decentralized, a theory based on them has the potential to avoid many of the problems GBM suffers from as well. Unfortunately, this type of explanation has not been well formalized. Abrams et al. (2011) make a noteworthy first theoretical effort at incorporating peer approval into voting; however, they treat social pressure as an exogenous variable. Their approach is also game theoretic, creating some of the scaling problems discussed elsewhere in this paper.

If norms are, in fact, the primary drivers of voter participation, a better understanding of norm dynamics promises to answer questions about voter motivation. Better theoretical understanding of why and when some choose to vote while others do not will help better connect individual interests, welfare, and outcomes. It may also shed light on how social network structures, media influence, and other seemingly orthogonal aspects of society may affect the institution.

In just this context, a more scalable closed-form modeling of the relevant social norm could ultimately be of great consequence to welfare. Considering this is just one of the many ways social norms may shape the world, the importance of understanding their development seems high, indeed. The research done to generate this dissertation is an attempt to make incremental progress in this area.

B Appendix: Optimal Support with Forward-Looking Utility

We can approximate optimal strategy under forward-looking utility with dynamic programming. The utility function is

$$U_{fl}(s_i^t) = -n_{i,i}(1-h_i)(s_i^t - \dot{s}^t)^2 - n_{i,i}h_i(s_i^t - s_i^{t-1})^2 - \sum_{\substack{j=1,\\ j\neq i}}^{p} n_{i,j}(s_i^t - s_j^{t-1})^2 + \sum_{i=1}^{\infty} \beta^i E_t[U_{fl}(s_i^{t+i})] \tag{55}$$

The corresponding Bellman Equation is

$$V(s_i^t) = \max_{s_i^t} \left\{ -\left(n_{i,i}(1-h_i)(s_i^t - \dot{s}^t)^2 + n_{i,i}h_i(s_i^t - s_i^{t-1})^2 + \sum_{\substack{j=1,\\ j\neq i}}^{p} n_{i,j}(s_i^t - s_j^{t-1})^2 \right) + \beta E_t[V(s_i^{t+1})] \right\} \tag{56}$$

where $V(s_i^t)$ is the value function representing the maximum attainable utility from period t onward.

Since we do not know the exact form of $V(s_i^{t+1})$, we will assume that the value function has a similar quadratic structure to intra-period utility.[18] Here we will choose to express future utility as

$$V(s_i^{t+1}) = -\alpha(s_i^{t+1} - \mu)^2 + c. \tag{57}$$

Now, the first order condition is

[18] This is a significant assumption.

$$-2n_{i,i}(1-h_i)(s_i^t-\dot{s}^t)-2n_{i,i}h_i(s_i^t-s_i^{t-1})-2\sum_{\substack{j=1,\\ j\neq i}}^{p}n_{i,j}(s_i^t-s_j^{t-1})+\beta\alpha(s_i^t-\mu)=0. \quad (58)$$

Solving for s_i^t and recalling that $\sum_j n_{i,j}=1$ we arrive at

$$s_i^t=\frac{2n_{i,i}(1-h_i)\dot{s}^t+2n_{i,i}h_is_i^{t-1}+2\sum_{\substack{j=1,\\ j\neq i}}^{p}n_{i,j}s_j^{t-1}+\beta\alpha\mu}{1+\beta\alpha} \quad (59)$$

The parameters β, α, μ approximate the influence of future utility on the current decision. The relative importance of future utility is dependent upon each of these parameters. If the relative future cost of deviating from optimal immediate behavior α is low or if the discount factor on future utility β is low, optimal forward-looking behavior is well approximated by static optimization. When $\beta=0$ or $\alpha=0$, static optimization is equivalent to forward-looking optimization.

C Appendix: Proof of Convergence of Support

We wish to demonstrate that in the absence of shocks, support under the model converges on a steady-state equilibrium. We will do so by applying the Contraction Mapping Theorem. Holding N^t and $\dot{s}^t$ static, and assuming $n_{i,i} > 0 \ \forall i$, we can rewrite the model:

$$s_i^t = n_{i,i}\left((1-h_i)\dot{s}_i + h_i s_i^{t-1}\right) + \sum_{j \neq i} n_{i,j} s_j^{t-1}. \tag{60}$$

We can further be rewrite this in matrix form

$$\vec{S}^t = \mathbf{N}_h \vec{S}^{t-1} + \vec{C}, \tag{61}$$

where

$\vec{S}^t \in \mathbb{R}^p$ is the vector of support values at time t,

$\mathbf{N}_h \in \mathbb{R}^{p \times p}$ is the matrix of social influence weights, identical to $\mathbf{N}$ except that the diagonal entries are given by $n_{i,i}h_i$, where $h_i \in [0,1]$,

$\vec{C} \in \mathbb{R}^p$ is a constant vector representing the native preferences and habit formation terms $n_{i,i}(1-h_i)\dot{s}_i$.

Note that because $\mathbf{N}$ is a stochastic matrix and values of $\mathbf{N}_h$ are less than or equal to the corresponding values of $\mathbf{N}$, $\mathbf{N}_h$ is a sub-stochastic matrix. We will show that this system converges to a steady state $\vec{S}^*$ as $t \to \infty$. We begin by iterating the equation for $\vec{S}^t$ to express it as a function of the initial state $\vec{S}^0$ and the constant vector $\vec{C}$.

Substituting $\vec{S}^{t-1}$ from the previous period and iterating, we have

$$\vec{S}^t = \mathbf{N}_h^t \vec{S}^0 + \sum_{k=0}^{t-1} \mathbf{N}_h^k \vec{C}. \tag{62}$$

$\mathbf{N}_h$ being a sub-stochastic matrix, implies that the spectral radius $\rho(\mathbf{N}_h)$, which is the largest absolute value of its eigenvalues, is less than 1. Because of this, the powers of $\mathbf{N}_h$

decay geometrically as $t \to \infty$,

$$\lim_{t\to\infty} \mathbf{N}_h^t = 0. \tag{63}$$

Thus, for any initial state $\vec{S}^0$, we have:

$$\lim_{t\to\infty} \mathbf{N}_h^t \vec{S}^0 = 0. \tag{64}$$

Next, we consider the summation term, this is a geometric series in the matrix $\mathbf{N}_h$. Since $\rho(\mathbf{N}_h) < 1$, the series converges as $t \to \infty$. Specifically, the infinite sum can be expressed as:

$$\sum_{k=0}^{\infty} \mathbf{N}_h^k \vec{C} = (I - \mathbf{N}_h)^{-1} \vec{C}, \tag{65}$$

where I is the identity matrix, and $(I - \mathbf{N}_h)$ is invertible because $\rho(\mathbf{N}_h) < 1$.

As $t \to \infty$, the term $\mathbf{N}_h^t \vec{S}^0$ vanishes, and the summation term converges to a finite value. Therefore, the steady-state solution $\vec{S}^*$ is given by:

$$\vec{S}^* = \lim_{t\to\infty} \vec{S}^t = (I - \mathbf{N}_h)^{-1} \vec{C}. \tag{66}$$

D Appendix: Simulation Code

This code in this appendix runs all simulations presented in this paper. It creates PDF files for each of the figures in the directory in which it is run and was written in the following environment:

Package	Version
Python	3.12.0
NetworkX	3.3
NumPy	2.1.1
Matplotlib	3.9.2

```
import matplotlib.pyplot as plt
import numpy as np
import networkx as nx
import random


def compute_normalized(matrix):
    """
    Provided a network matrix, this function returns a normalized
    version of that matrix such that each row sums to 1.
    Compensation for rounding error is added to the diagonal.
    """
    row_sum = np.sum(matrix, axis=1)

    # Calculate row sum for each element in the row.
    matrix = matrix / row_sum[:, np.newaxis]

    # Ensure each row sums to exactly 1.
    np.fill_diagonal(matrix,
                     matrix.diagonal() + (1 - np.sum(matrix, axis=1)
                                          ))

    return matrix


def compute_s(n, h, s_dot):
    """
    This applies the model to the exogenous variables and returns
    the resulting s values as a matrix.
    """
```

```
    if s_dot.shape[1] != n.shape[2]:
        raise ValueError("t values not compatible in s_dot and h")

    # Infer T and P values so they do not need to be passed in.
    T = s_dot.shape[1]-1
    P = s_dot.shape[0]

    # Intialize the return variable.
    s = np.zeros((P, T+1))
    s[:, 0] = np.copy(s_dot[:, 0])

    # Populate s by iteratively applying the model.
    for t in range(1, T+1):
        for i in range(0, P):
            cur_s = 0
            for j in range(0, P):
                if i == j:
                    native_term = (1 - h[i]) * s_dot[i][t]
                    habit_term = h[i] * s[i][t - 1]
                    cur_s += n[i][i][t] * (native_term + habit_term)
                else:
                    cur_s += n[i][j][t] * s[j][t-1]

            s[i][t] = cur_s

    return s


def compute_eta(sigma_vector, sigma_dot_vector):
    """
    Return vector of eta values for each time step t based on the
    formula: eta^t = sigma^0 / sigma^t.
    """
    return sigma_dot_vector / sigma_vector


def compute_sigma(tau, s):
    """
    Returns a vector of standard deviance (sigma^t) for each time
    step t based on the formula:
    sigma^t = sqrt((1 / p) * sum((T^t - s^t_i)^2)) for all i.
    """
    return np.sqrt(np.mean((tau - s) ** 2, axis = 0))


def compute_tau(n, s):
    """
```

```
    Returns the vector Tau (Target Support) of the population over
    time. See paper for equation relating tau to n and s.
    """
    tau_vector = np.zeros(s.shape[1])

    for t in range(s.shape[1]):
        # Copy current network without self-influence.
        n_no_diag = n[:, :, t].copy()
        np.fill_diagonal(n_no_diag, 0)

        # Calculate sums of influence over others.
        d = np.sum(n_no_diag, axis=0)

        # Check for division by zero.
        if np.sum(d) == 0:
            #set tau for that period to nan.
            tau_vector[t] = np.nan
        else:
            #calculate and set tau for that period.
            tau_vector[t] = np.sum(d * s[:, t]) / np.sum(d)

    return tau_vector


def compute_model_outputs(n, h, s_dot):
    """
    Runs the model and returns s along with vectors of
    macro-social measures as a tuple.
    """
    s = compute_s(n, h, s_dot)
    tau = compute_tau(n, s)
    sigma = compute_sigma(tau, s)
    sigma_dot = compute_sigma(tau, s_dot)
    eta = compute_eta(sigma, sigma_dot)

    return s, tau, sigma, sigma_dot, eta


def add_shock(matrix, submatrix, t):
    """
    Copies the provided n-1 dimensional submatrix over all indices
    greater than or equal to the specified index in the last
    dimension of the n-dimensional matrix, which should be the
    time dimension. Applies shocks at time t both n and s_dot.
    """
    # 2-d submatrices imply this is n, so be sure to normalize.
    if submatrix.ndim == 2:
```

```
125         adj_matrix = np.copy(compute_normalized(submatrix))
126     else:
127         adj_matrix = np.copy(submatrix)
128 
129     matrix[..., t:] = np.expand_dims(adj_matrix, axis=-1)
130 
131 
132 def adjust_self_influence (matrix, self_influence_factor):
133     """
134     Ensures that self-influence mantains a range of proportions
135     relative to outside influence for each agent between 50% and
136     100% of the self-influence factor.
137     """
138     for i in range(matrix.shape[0]):
139         matrix[i, i] = (    np.random.rand() \
140                             * self_influence_factor \
141                             + self_influence_factor \
142                         ) / 2 \
143                         * (np.sum(matrix[i, :]) - matrix[i, i])
144 
145 def init_n(P, T, initial_network):
146     """
147     Returns a set of indentical networks over time.
148     """
149     n = np.zeros((P, P, T + 1))
150     add_shock(n, initial_network, 0)
151 
152     return n
153 
154 
155 def display_matrix(matrix):
156     """
157     Formatted display of a time series matrix to the console.
158     """
159     for t in range(matrix.shape[-1]):
160         print(f"t = {t}:")
161         print(matrix[..., t])
162         print()
163 
164 
165 def draw_support(s, eta, file_name):
166     """
167     Generates a plot of support over time and eta if applicable.
168     """
169     # Initialize the plot.
170     plt.rc('lines', linewidth=1)
171     fig, ax = plt.subplots()
```

```
172
173     P = s.shape[0]
174
175     # Create the time axis values.
176     time_axis = np.arange(len(s[0]))
177
178     # Add up to 50 agents as a lines in the plot.
179     for i in random.sample(range(P), min(20, P)):
180         ax.plot(time_axis, s[i], color='gray')
181
182     # Display the last eta value on the plot.
183     last_eta = eta[-1]
184     if not np.isnan(last_eta):
185         ax.text(0.01, 1.07,
186                 rf'$\eta^{{{time_axis[-1]}}}$={last_eta:.2f}',
187                 transform=ax.transAxes, fontsize=12,
188                 verticalalignment='top')
189
190     # Modify the labels.
191     plt.xlabel('time ($t$)')
192     plt.xticks(time_axis)
193     plt.ylabel('Support ($s_i$)')
194     plt.yticks([])
195
196     # Save the figure to a file.
197     plt.savefig(file_name+".pdf", format="pdf", bbox_inches="tight")
198     print ("Created " + file_name + ".pdf")
199
200     plt.close()
201
202
203 def draw_network(n, file_name, t=0):
204     """
205     Generates and visualization of a directed weighted graph of the
206     network at time t.
207     """
208     # Extract the network from the time series of networks.
209     weight_matrix = n[:, :, t]
210
211     # The number of nodes in the graph.
212     num_nodes = weight_matrix.shape[0]
213
214     # Create the directed graph.
215     DG = nx.DiGraph()
216
217     # Add nodes to the graph.
218     nodes = range(1, num_nodes + 1)
```

```
    DG.add_nodes_from(nodes)

    # Add edges with weights based on the weight matrix.
    for i in range(num_nodes):
        for j in range(num_nodes):
            # Avoid adding edges with zero weight or self arrows.
            if weight_matrix[i, j] != 0 and i != j:
                DG.add_edge(i + 1, j + 1,
                            weight=weight_matrix[i, j])

    # Visualize the directed graph.
    plt.figure()

    # Position the nodes using a spring layout.
    pos = nx.spring_layout(DG, seed=42)

    # If lots of nodes, make the display more compact.
    if num_nodes > 20:
        with_labels = False
        node_size = 3
    else:
        with_labels = True
        node_size = 400

    # Draw nodes and edges.
    nx.draw(DG, pos, with_labels=with_labels, node_color='white',
            edgecolors='black', font_size=10, arrows=False,
            width=0.5, node_size=node_size, edge_color='gray')

    # Draw edge labels (weights).
    edge_labels = nx.get_edge_attributes(DG, 'weight')

    # Add the label "N^t" somewhere on the plot if t is not 0.
    if t:
        plt.text(0.01, 0.99, f'$N^{{{t}+}}$', fontsize=20,
                 ha='center', va='center',
                 transform=plt.gca().transAxes)

    # Display the graph.
    plt.savefig(file_name+".pdf", format="pdf", bbox_inches="tight")
    print ("Created " + file_name + ".pdf")
    plt.close()

def create_social_network(num_nodes, num_edges=2, closure_prob=0.5):
    """
    Create a directed, weighted network with scale-free and
    clustering properties.
```

```
    """
    # Create a network with clustering and scale-free properties
    G = nx.powerlaw_cluster_graph(n=num_nodes,
                                  m=num_edges,
                                  p=closure_prob)

    # Convert to directed graph
    G = G.to_directed()

    # Add random weights to the edges
    for u, v in G.edges():
        G[u][v]['weight'] = np.random.rand()

    # Convert the graph to a weighted adjacency matrix
    adjacency_matrix = nx.to_numpy_array(G, weight='weight')

    # add self-weights
    for i in range(num_nodes):
        adjacency_matrix[i, i] = np.random.rand()

    return adjacency_matrix

def main():
    P = 20   # Size of the population.
    T = 20   # Max number of time periods, not including t=0.

    #adjusts ratio of self-influence to outside influence.
    self_influence_factor = T / 3

    # Initialize the default randomzied, fully connected network.
    default_network = np.random.rand(P, P)
    adjust_self_influence(default_network, self_influence_factor)


    # Initialize default s_dot (native preferences).
    default_s_dot = np.empty((P, T + 1))
    add_shock(default_s_dot, np.sort(np.random.rand(P))[::-1], 0)

    # Intitialize default h (habit formation) values.
    default_h =  np.random.rand(P) * 0.25 + 0.75

    # Run a baseline example.
    n = init_n(P, T, default_network)
    h = np.copy(default_h)
    s_dot = np.copy(default_s_dot)

    s, tau, sigma, sigma_dot, eta \
```

```
313         = compute_model_outputs(n, h, s_dot)
314     draw_support(s, eta, 'baselineS')
315     draw_network(n, 'baselineN')
316
317
318     # A Clustered network example.
319     start_n = np.copy(default_network)
320
321     # Remove necessary nodes.
322     for i in range(0, int(P / 2)):
323         for j in range(int(P / 2), P):
324             start_n[i, j] = 0
325             start_n[j, i] = 0
326
327     # Add back a single node connecting the clusters.
328     start_n[0, int(P / 2 + 1)] = default_network[0, int(P / 2 + 1)]
329     start_n[int(P / 2 + 1), 0] = default_network[int(P / 2 + 1), 0]
330
331     adjust_self_influence(start_n, self_influence_factor)
332
333     n = init_n(P, T, start_n)
334     h = np.copy(default_h)
335     s_dot = np.copy(default_s_dot)
336
337     s, tau, sigma, sigma_dot, eta \
338        = compute_model_outputs(n, h, s_dot)
339
340     draw_support(s, eta, 'clusteredS')
341     draw_network(n, 'clusteredN')
342
343     # An influential agent 0.
344     start_n = np.copy(default_network)
345     start_n[1:, 0] += start_n[1:, 0] * 20
346     n = init_n(P, T, start_n)
347     h = np.copy(default_h)
348     s_dot = np.copy(default_s_dot)
349
350     s, tau, sigma, sigma_dot, eta \
351        = compute_model_outputs(n, h, s_dot)
352
353     draw_support(s, eta, "influentialS")
354
355     # Agents are removed from network at t=T/2.
356     n = init_n(P, T, default_network)
357     add_shock(n, np.eye(P), int(T / 2))
358     h = np.copy(default_h)
359     s_dot = np.copy(default_s_dot)
```

```
360
361     s, tau, sigma, sigma_dot, eta \
362         = compute_model_outputs(n, h, s_dot)
363
364     draw_support(s, eta, 'removalS')
365     draw_network(n, 'removalN', int(T / 2))
366
367     # A shock to s_dot at t = T / 2.
368     n = init_n(P, T, default_network)
369     h = np.copy(default_h)
370     s_dot = np.copy(default_s_dot)
371     add_shock(s_dot, np.zeros(P), int(T / 2))
372
373     s, tau, sigma, sigma_dot, eta \
374         = compute_model_outputs(n, h, s_dot)
375
376     draw_support(s, eta, 'nativeS')
377
378     # Degenerate: full habit formation.
379     n = init_n(P, T, default_network)
380     h = np.ones(P)
381     s_dot = np.copy(default_s_dot)
382
383     s, tau, sigma, sigma_dot, eta \
384         = compute_model_outputs(n, h, s_dot)
385
386     draw_support(s, eta, 'fullHabitS')
387
388     # Degenerate: no habit formation.
389     n = init_n(P, T, default_network)
390     h = np.zeros(P)
391     s_dot = np.copy(default_s_dot)
392
393     s, tau, sigma, sigma_dot, eta \
394         = compute_model_outputs(n, h, s_dot)
395
396     draw_support(s, eta, 'noHabitS')
397
398     # Degenerate: no network.
399     n = init_n(P, T, np.eye(P))
400     h = np.copy(default_h)
401     s_dot = np.copy(default_s_dot)
402
403     s, tau, sigma, sigma_dot, eta \
404         = compute_model_outputs(n, h, s_dot)
405
406     draw_support(s, eta, 'noNetworkS')
```

```
407
408     # Sparse Network.
409     sparse_network = np.copy(default_network)
410     for i in range(int(P**2 / 2 * 1.7)):
411         row = random.randint(0, P - 1)
412         col = random.randint(0, P - 1)
413         if row != col:
414             sparse_network[row, col] = 0
415             sparse_network[col, row] = 0
416
417     # Add a randomly selected value back to any empty rows.
418     for row in range(P):
419         non_diagonal_sum = np.sum(sparse_network[row, :]) \
420                            - sparse_network[row, row]
421
422         # If the sum of the non-diagonal values in a row are
423         # zero, set a random non-diagonal element back.
424         if non_diagonal_sum == 0:
425             non_diagonals = \
426                           [col for col in range(P) if col != row]
427             random_index = random.choice(non_diagonals)
428             sparse_network[row, random_index] \
429                                 = default_network[row, random_index]
430             sparse_network[random_index, row] \
431                                 = default_network[random_index, row]
432
433     adjust_self_influence(sparse_network, self_influence_factor)
434
435     n = init_n(P, T, sparse_network)
436     h = np.copy(default_h)
437     s_dot = np.copy(default_s_dot)
438
439     s, tau, sigma, sigma_dot, eta \
440        = compute_model_outputs(n, h, s_dot)
441
442     draw_support(s, eta, 'sparseS')
443     draw_network(n, 'sparseN')
444
445     # Full shocks to a full network.
446     P = 500
447     T = 20
448
449     # Initialize default s_dot (native preferences).
450     s_dot = np.empty((P, T + 1))
451     add_shock(s_dot, np.random.rand(P), 0)
452
453
```

```
454     h =  np.random.rand(P) * 0.25 + 0.75
455     n2 = create_social_network(P)
456     adjust_self_influence(n2, self_influence_factor)
457     n = init_n(P, T, n2)
458 
459     # Add shocks to every time period.
460     for j in range(T):
461         n2 = create_social_network(P)
462         adjust_self_influence(n2, self_influence_factor)
463         add_shock(n, n2, j)
464         add_shock(s_dot, np.random.rand(P), j)
465 
466     s, tau, sigma, sigma_dot, eta \
467        = compute_model_outputs(n, h, s_dot)
468 
469     draw_support(s, eta, 'shocksS')
470     draw_network(n, 'shocksN')
471 
472     print('[Execution Complete]')
473 
474 if __name__ == '__main__':
475     main()
```