OPIEM: An Operationalized Model of the Information Environment

Jacob Shaha

CMU-S3D-26-104

April 2026

Software and Societal Systems Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Kathleen M. Carley, Chair
L. Richard Carley
Geoff Dobson
Nathan VanHoudnos

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Societal Computing.

The research for this paper was supported in part by the US Army project Scalable Technologies for Social Cybersecurity (W911NF20D0002) and by the Office of Naval Research under projects Minerva-Multi-Level Models of Covert Online Information Campaigns (N000142112765), Community Assessment (N000142412568), and Scalable Tools for Social Media Assessment (N000142112229), and by the center for Informed Democracy and Social-cybersecurity (IDeaS) and the center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University. The views and conclusions are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Office of Naval Research, Army, or the US Government. The appearance of U.S. Department of Defense (DoD) documents and visual information does not imply or constitute DoD endorsement.

Keywords: information, information environment, operations, information warfare, model, social cognitive, disinformation, misinformation, social cybersecurity, influence

To my dad. He would have thought this was so cool.

Abstract

The advent and spread of the global Internet and the subsequent dawn of the "Information Age" has brought scores of new economic and social opportunities. It has also exposed and exacerbated psychological, sociological, and organizational flaws in human societies. Recent history demonstrates that Internet-based communications and social structures present significant challenges to long-established social mechanisms and societal models. Left unaddressed, these issues will continue to destabilize existing social mechanisms.

Addressing the challenges of the Information Age has proven difficult for policymakers and leaders for many reasons. Information is an inherently non-geographic quantity and is intrinsically produced by, linked to, and leveraged in, all human activities. In addition, information is a causal driver of human behavior. The ability of policymakers to understand and affect the information available to constituents, in competition with other information providers, determines how effectively they can govern. The lack of domain expertise makes leaders dependent on models to understand and direct information activities, but available modeling methods are insufficient for the rapidly evolving information environment.

This thesis presents a novel model of the information environment offering high-level leaders, such as state policy makers or military leaders, a tractable, intuitive, and helpful representation of the information environment. The model is designed to capture and leverage the interconnectedness of different agents, topics, and platforms, to help forecast the effects of proposed changes and decisions.

The thesis incorporates and extends the BEND influence framework into the BENDRS framework, providing a lexicon to relieve policy makers of need for a detailed understanding of the specific means and methods of information operations. By abstracting complex edge-level concepts in cyber, propaganda, and covert influence, the lexicon helps leaders recognize, understand, and direct changes in the environment, without requiring them to prescribe the means and details of effecting that change. Thus, it supports a state-type hierarchical organization composed of many varied fields of expertise and charged with a highly heterogeneous purview and a broad and evolving set of goals.

Finally, the thesis demonstrates methods to construct such a model solely from observed network traffic, making it immediately operationally accessible and removing any requirement for an omniscient network architect. The network is constructed from relationships and values observed within communications traffic, and is iteratively and additively adjusted as additional traffic is revealed; assuming all traffic is available and intelligible, the derived network, in the limit, converges to the actual network environment. More importantly, the derived network in any state is immediately operationally useful as a quantitatively analyzable snapshot of the current environment.

Acknowledgments

My first thanks is to Dr. Carley. I was at least five years older than every other student she had, and as far as I can tell my primary qualification for graduate-level study was two thumbs and an appetite for pain. She shepherded me through this process with aplomb and I am honored to contribute to her important and venerable body of work.

I give my heartfelt thanks to the rest of my Doctoral committee: Dr Richard Carley, Dr. VanHoudnos, and Dr. Dobson. They have provided excellent feedback, and more importantly, they have been endlessly encouraging. I am especially grateful for their (hopefully unfeigned) excitement about the operational application of my work. They must know from their own experience how much feeling useful means to a grad student.

Thanks to my friends in CASOS, especially Reba Marigliano, who convinced me to take all the same classes as her and as a result wrote a ton of code for our shared homework. Don't tell anyone.

I owe my thanks to the US Army for funding my education, and for preparing me for this and the many other challenges in life I have faced and will yet face. The list of men and women who built and shaped me into a successful Soldier and Officer is too long to for this page, and my debt to them is eternal. I am grateful to Chris Simmons, Joe Halstead, Chris Rosse, Eric Baus, Bret Tecklenburg, RJ Lillibridge, Robert Harmon, Rick Roper, Slade Smith, and Neil Khatod for making me tough and teaching me that I am very, very often wrong.

I am also deeply grateful to Rachel Sondheimer, Katie Daley, and Brigadier General Cindy Jebb for their inspiring leadership, example of scholarship, and utter long-suffering with my nonsense. Similarly, I owe thanks to Chris Lowrance, Lisa Shay, and Barry Shoop for giving me a chance to teach, and learn, and whet my appetite for further adventures in that direction. I thank Mike Meneghini for pushing me into this and constantly reminding me how great he thinks I am; I am always impressed at his determination to convince me.

I am deeply, deeply grateful to (then) Brigadier General Stephanie Ahern. Through sheer force of will, she made this happen for me and my family. I amd determined that this work, and the work that follows, will justify her investment and confidence.

Special thanks to my bandmates and fellow Saints for helping me stay sane during this process and find some joy along the way.

Lastly, I acknowledge and am humbled by the herculean sacrifices my family has made, not just in this endeavour, but throughout my life. I thank my parents for giving me every advantage, especially examples of scholarship, kindness, and love. I am grateful to my children for embracing Pittsburgh as home, and for stepping up to help as I became increasingly stressed by the weight of my studies.

I owe this success, and every success I can claim, to my wife Melinda. Everything I have done well has her fingerprints on it. She is the best decision I have made; all other good things have followed from it. I love you, Mel.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The Information Age has yielded an increasingly interconnected society [12][83]. However, Large-scale conflict has persisted, adapting to the new global environment. Nations and societies continue to compete with one another, and information is an increasingly valuable resource in competition [13]. Because information drives behavior [86], states that can successfully manipulate information can directly affect competitors' behavior, achieving compulsion not through confrontation but through confusion [65], persuasion [34], or even collusion [162]. Therefore, information represents a significant opportunity and threat, to leaders in military and government contexts.

State leaders and policymakers seek to direct [202], inform [120], and sometimes curtail decisions and behaviors within their society [10]. The information available to citizens informs and drives those decisions and behaviors. As other states, organizations, and individuals become increasingly able to project information into the environment [119], they approach the ability to direct, inform, and curtail decisions within the audience. If those entities have goals other than the state's, they come into direct competition with state leaders and policymakers.

These high-level organizational leaders often struggle to succeed in this competition [201]. Information is ubiquitous [197], nuanced [18], and ephemeral [76], while state organizations cohere slowly and often have broadly defined goals or purviews. Capturing and understanding the environment in which citizens consume, process, and share information is increasingly important for states in competition, and at the same time grows increasingly difficult.

In other complex and nuanced domains, such as economics, sanitation, transportation, traffic design, policymakers avail themselves of models to forecast future conditions, and to identify likely outcomes and effects of proposed courses of action. However, no such model exists that wholly captures the information environment.

This thesis examines the challenge of defining and understanding a high-level, operational information environment. More specifically, this thesis proposes an Operational Information Environment Model (OPIEM), a model to represent the information environment at an operational level. The model assists in understanding that environment; recognizing phenomena – both adverse and advantageous – within that environment; and detecting influence actions or "maneuvers" affecting that environment. All of this is in service to the state's competition to retain effective governance. The state can influence the behaviors and decisions of its citizens in a cohesive and constructive fashion, instead of or despite attempts by external forces to exert the

same influence.

Previous work has explored influence within the artificial component of the environment ("cyberspace"), especially with regards to how automated systems can influence the behavior of other automated systems [105][167]. Psychologists and sociologists have long examined how humans and organizations influence one another's behavior [55][5]. These two domains have been largely considered separately, despite the fact that both are now foundational to how individuals consume, produce, and traffic information [222].

1.1 Thesis Goals

To bridge this gap between cyberspace and social-space, my thesis addresses:

- Existing models for societal information environments and large-scale social influence phenomena, including analysis of those models' strengths and deficiencies;
- Social-cyber influence, in which information and influence transition between natural/human systems and artificial/cyber systems, including analysis of real-world "hybrid" social-cyber influence campaigns and attacks;
- Influence achieved by manipulating the environment's composition and/or structure, and a proposed lexicon for frequent or categorical instances of such (structurally-oriented influence "maneuvers"), with historical examples;
- My proposed method for modeling a pertinent information environment at the government/military level (the OPIEM framework), to better understand that environment and project changes within it;
- A method of constructing the OPIEM model entirely from extant and observable information within the environment;
- Quantitative methods and metrics by which the previously mentioned influence maneuvers can be identified and labeled within the OPIEM model;
- Initial results of this model and methodology obtained from a simulated information environment; and
- Contributions, limitations, and potential extensions and applications of this work.

The OPIEM model and its accompanying influence maneuver lexicon provide an accessible and tractable way for military and government officials to better model the complex and dynamic information environments in which they compete. I strongly believe that these tools can assist leaders in understanding, visualizing, directing, and leading information activities in competition, both small-scale and large-scale, to better defend societal cohesion.

1.2 Data and Tools

To establish the nature of socio-technical influence, I drew from multiple historical case studies that documented cyber attacks, influence campaigns, and social-cyber hybrid influence events. I used official media reports of these incidents and reference scholarly examinations and/or formal

investigations when available.

To demonstrate the OPIEM framework's functionality, I created synthetic social media data. I drew heavily from the Project Omen training system developed by Carnegie Mellon's Center for Computational Analysis of Social and Organizational Systems (CASOS). Project Omen is a set of software tools that simulates a highly realistic social media environment, including autonomous actors and traffic/behavior patterns. I modified Omen's core algorithm to create Influarium, a smaller-scale social media simulator with explicit representation of technical systems.

I made extensive use of the Organization Risk Analyzer (ORA) - Professional version for network analysis and visualization. My thesis leverages ORA's highly optimized network/graph analysis algorithms, and its network visualization tool, to present and analyze network models.

1.3 Background

Information has a direct causal relationship to perception, and thus to behavior and decision making. Philosophers and psychologists have proposed many models for perception over the years, each with varying strengths and drawbacks [67]. Perception science remains an open field of research, demonstrating the inherent complexity and importance of this process. Even small changes in available information can result in vastly varied opinions and understandings of an otherwise shared environment between individuals [152][245]. These disparate understandings give rise to different decisions and reactions within that same environment. Research has demonstrated that small changes in information availability or presentation style can significantly influence the persuasiveness of a message, the recipient's trust in the sender, and the decisions influenced by those factors [151][186].

Because information drives decision-making, leaders have recognized its competitive advantage for millennia. As far back as the 5th century BCE, Sun Tzu famously wrote, "Know the enemy and know yourself, and in a hundred battles you will never fear defeat" [226]. Regardless of two armies' relative strengths or compositions, Sun Tzu asserts that the leader with the clearest understanding of the entire situation is best positioned to achieve his or her competitive goals.

More recently, though still at the dawn of the Information Age, Porter & Millar noted that "Every value activity has both a physical and an information component... every value activity creates and uses information of some kind" [190]. In the commercial and industrial spheres, as in the martial, information is an inherent component of competition, and the competitor that more skillfully wields it holds advantage on the field. Commercially, information is foundational in negotiation and pivotal in market competition [122][6]. In some cases, information is both a competitive enabler and a business model (i.e. consultants, analytics firms, market researchers, etc.).

Militarily, Sun Tzu recognized information's role in tactical and operational success, and Prussian army officer Carl Clausewitz identified information's utility in strategic victory, as states compete to remove not just the means but the *will* to fight [59][226]. In government, the information available to constituents and the behaviors it drives heavily affect the efficacy of policy and the cohesion of society [135].

In both government and military contexts, the importance of information is increasingly vital

to success [100]. Societal leaders, through policy and law, seek to guide decision-making within society, toward collectively beneficial behaviors and away from destructive behaviors. In this context, these leaders find themselves in tacit, if not open, competition with multiple actors, both internal and external, who also seek to influence decision-making in service of their own agendas [35]. Competitively, these leaders seek to provoke counterproductive decisions that undermine opponents' efforts and thereby strengthen their own [244].

Policymakers' and military leaders' need to influence decisions [95], combined with information's causal relationship to decision-making, creates an incentive (if not a requirement) for these leaders to influence the information available to decision makers, both constituent and opponent [153]. Leaders require information awareness to effectively guide their organizations, inducing cohesion and compliance to policy. They require the same awareness to compete with other influencers, to directly and indirectly oppose and undermine competing agendas and influences.

Decision makers are rationally bounded, constrained by their own limitations and by the information available for a decision [110]. An individual's constraints include their social framework, within which they extract and process information. Societal leaders must account for the potential and realized impacts of available information when seeking to enable effective decision making by those within their purview. As competitors they must also optimize the information available for their own decisions, while simultaneously exerting influence on the information available to competitors [9][53].

These tasks – information awareness and information influence (or control) – are deceptively difficult. Information is ubiquitous, often qualitative, highly subjective, and infinitely varied [41]. Gaining and maintaining full awareness of all information is an unrealistic goal; the qualifier "all" is immediately semantically problematic. Contemporary examples demonstrate that, even with significant awareness and investment, state-level actors can neither fully establish, nor fully repel, narratives within their populace [209].

This difficulty stems in no small part from the inherent challenge of quantizing and discretizing information. Determining whether one has adequately established or controlled a narrative first requires some definition of "narrative" sufficient to enable measurement and to provide benchmarks for successful establishment. Determining whether one has successfully influenced some population similarly requires the concept of influence to be quantized enough to provide criteria for success or failure.

1.4 Models

To grapple with highly complex phenomena like information, researchers and policymakers have long relied on models. Models abstract and interpret complex systems into quantitatively tractable formats. Many models exist for quantizing information as a phenomenon, across many different domains [166]. However, such models are created within their respective domains, and are so tailored to their respective realms of application as to render them unhelpful when seeking a generalized, broadly applicable model for information as a whole. Work has also been done in quantizing comparative information advantage between decision makers, but these models remain overly simplistic when compared to the nuanced and complex realities of organizational decision making [235]. At the societal or policy-making level, current models are quickly re-

vealed to be inadequate or inapplicable to the nature, volume, and breadth of information in consideration.

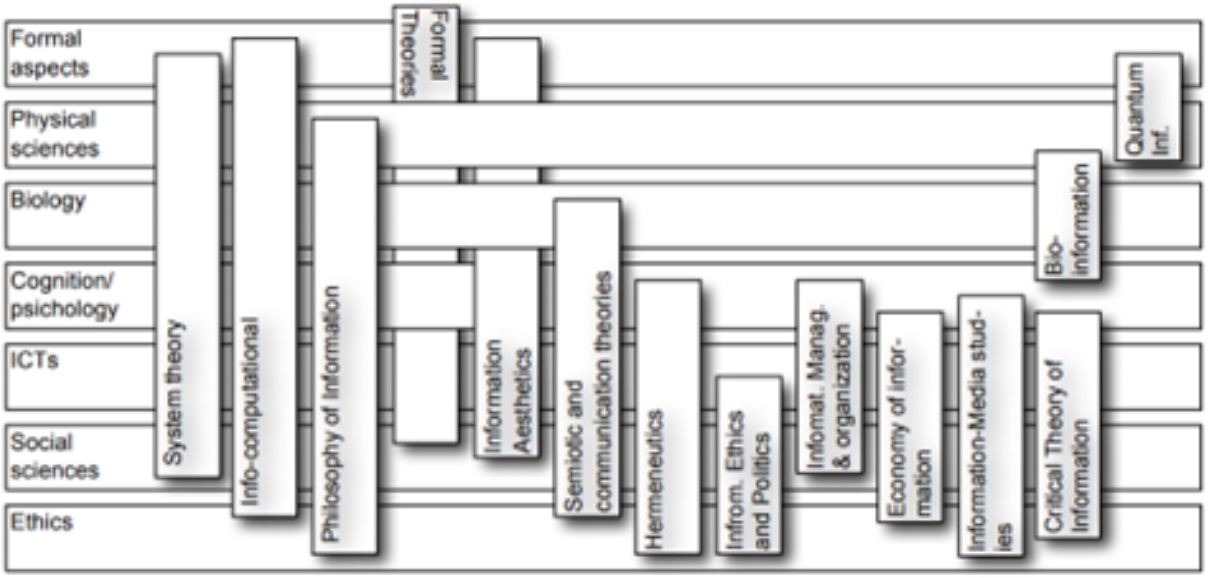

Figure 1.1: Overview of information science emphases, overlaid on different academic domains.

These many models do, however, have a common thread that bears repetition. In each case, information models limit scope by focusing on specific types or manifestations of information: how it moves, is processed, and/or is stored, and who the consumers and producers of that information are. Societal leaders work in the extremely broad domain of human behavior, giving rise to an enormous and highly varied body of pertinent information. Leaders seeking to gain information awareness and to effectively influence information require a model that accommodates a broadly defined and highly heterogeneous information context.

1.5 Defining the Information Environment

As mentioned, information models set their scope by stipulating the specifics of information: the who, how, when, where, and so forth. Such a context can be labeled an *information environment*. While the term information environment is used frequently, there is "no clear or agreed definition" [243]. Philosopher Luciano Floridi posited that the information environment comprises "all informational processes, services, and entities, thus including informational agents as well as their properties, interactions, and mutual relations" [88]. The Department of Defense defines IE as "the aggregate of the individuals, organizations, and systems that collect, process, disseminate, or act on information - consisting of physical, informational, and cognitive dimensions" [239]. Another definition from the Carnegie Endowment is "the space where people process information to make sense of the world using tools from alphabets to artificial intelligence to produce outputs from the spoken word to virtual reality to whatever comes along in the future" [1].

In this thesis, I define an information environment relative to a set of decisions targeted for influence. The information environment then comprises: the autonomous agents of interest, either as decision-makers or influences on decision-makers; the information processing, storage, and transmission means affecting those agents; and the information pertinent to the decisions and agents under consideration, categorized or quantized as is appropriate for the model's application. Such an IE, as described, can fully encompass the decision mechanisms of interest to an analyst or observer.

An IE thus constituted, while more defined than "all information," is still massive. It includes all pertinent decision makers, and all agents influencing them, and all information systems used by or otherwise interacting with them, and all systems interacting with those systems behind the scenes, and all pertinent information quanta (however defined) considered, included, or otherwise involved in decision making – and all of these collections quickly expand. At the policy-making level, the scale of the IE imposes significant challenges between practitioners – those who work in information, interfacing with the IE's agents, systems, and contents – and those who set policy. Practitioners are necessarily concerned with specific portions of the IE based on their specialization, and cannot rightly be expected to be concerned with how their actions might influence disparate portions of the IE. Policy makers, on the other hand, are concerned with the entirety of the environment and its adjacencies. Since the IE is defined based on a decision or set of decisions (or behaviors), and policy is also defined relative to desired decisions and behaviors, the entire IE – and the resulting decisions – are of interest at the policy-making level.

Because of this distance between policy and implementation, an abstract quantitative model of the IE would help inform policy makers abstracting the specific methodologies of experts, and the specific manifestations or phenomena of information. A model would provide a tractable interface for policy makers to comprehend the IE as a whole, and to direct desired changes, without requiring policy makers to understand the specifics entailed by those changes.

One might consider an analog at the organizational level: productivity models enable corporate executives to identify patterns and shift production priorities, directing their organizations to repurpose assembly lines or storage facilities toward more productive ends. Crucially, these executives need not give specific directions about the placement of machinery or personnel on the assembly line, nor the stacking or labeling systems utilized in warehouses. Their model enables them to understand the current configuration of their environment; to recognize more favorable possible configurations; and to describe to their subordinates, in model-based language, the changes they seek. Those subordinates can then reinterpret model-based language into their respective areas of expertise – the factory foreman can take the task to "repurpose" his line and lens it into the many factory-specific tasks required, while the warehouse foreman can similarly understand "repurpose" as a set of completely different tasks within his area.

In this way, a model enables high-level decision makers to identify and direct useful change, without requiring them to understand or prescribe the minutia of implementation. Such abstraction is increasingly important for broader and more varied purviews like societal leadership and policy making.

1.6 Components of the Information Environment

Before examining extant IE models, I first provide some working definitions to frame further discussion and analysis. A comprehensive definition of information is deceptively elusive, and as a result there are many working definitions, as appropriate to different fields and contexts. Eschewing the more esoteric understandings of information used in physics and mathematics, I define information as pertaining to human decision making.

In that vein, Dr. Kathleen Carley's offers two insightful definitions of information:

- Information is cognitive content and is used to generate knowledge in the minds of individual agents and between individuals in groups [52].
- Information is characterized as patterns of assumptions and associations that individuals use to interpret events and guide behavior [50].

The first definition begs a definition of knowledge:

- Knowledge is a dynamic, networked quality, embedded in social networks, which changes through interactions within the network. Knowledge diffuses and takes values, and in doing so influences behavior in the affected individuals, which in turn alters future knowledge diffusion patterns and values for the same network [71].
- A knowledge base is a collection of coded, verbal/symbolic data (facts, beliefs, and/or opinions) shared by group members [48].

Information can further be subcategorized:

- Individuals and groups differentiate articulated (explicit) information from implied or unarticulated (tacit) information. General, high-level knowledge often implies or relies on tacit information within a group, whereas specific or niche knowledge is composed of more tacit information [144].
- Referential knowledge, or "transactive memory," is knowledge held by some subset of a group that is, therefore, accessible to the entire group through interpersonal connection [195].

As previously mentioned, the IE for a specific decision comprises any agents of interest, any information systems used by or open them, and the information pertinent to the decisions and agents under consideration. An IE thus defined contains information (and knowledge) corresponding to the various definitions above. A model of that environment must therefore be built upon components that accommodate these definitions.

I define these categorical components using the following set labels:

Agents: The set of autonomous agents within the IE. Agents are any entities capable of understanding, deciding, and acting upon information. Within the IE, they are both consumers and producers of *novel or synthesized* information (as opposed to solely repeating information from other sources). An IE is defined relative to a decision or set of decisions; agents include all entities that might make the target decision(s), as well as all entities that might influence those deciders even if they cannot make the target decision(s) themselves.

Topics: The set of salient environmental categories or labels, into which all pertinent information within the IE can be grouped. Note that *pertinence* is a strong filter for this set. I define an IE relative to a decision or set of decisions. Topics is the collection of information that agents might consider in making the target decision(s); more formally, Topics is the smallest set of labels that can accommodate that collection of pertinent information. The Topics set of an overly broad IE will rapidly expand to include all possible information, so defining the IE scope is a crucial part of modeling that environment.

Platforms: The set of systems, devices, and services through which agents interact with the information environment, and that contain, store, present, collect, transmit, and/or manipulate information within the environment. Platforms is the collection of devices by which agents (and in many cases other platforms) interact with information, comprising everything from smartphones and laptops to Ethernet switches and radio antennas.

These three sets capture the key components for decision making: Agents encompasses autonomy and executive function; Topics captures triggers, beliefs, opinions, and data; and Platforms includes information sources, sinks, transfers, and analyses external to Agents' internal thought processes.

These components satisfy the information definitions provided previously. The components are specific to decision making: Agents captures autonomous process, and thus includes agents' internal cognitive content and derived knowledge. Connections or groupings of agents within a model can be used to represent networked knowledge, including collective knowledge bases and referential information. Topics includes assumptions and associations – and the interpretations derived therefrom – that guide agent behavior. Platforms allows for externally stored and accessible referential knowledge, important in modeling groups that are not immediately physically proximate.

1.7 Related Work and Existing Models

As previously stated, models are widely utilized to organize and translate complex domains in order to derive actionable insights. In considering sufficiency for an IE model specific to decision making, I posit the following criteria:

- The model must include all components pertinent to decision making. Specifically, it must capture at least the Agents, Topics, and Platforms collections defined previously.
- The model must account for influence upon and between agents, as decision making in a dynamic environment is frequently dependent on threshold criteria.
- The model must consider information propagation and storage, to accurately portray the diffusion of information within the environment.
- Similarly, the model must consider information availability and accessibility, accurately indicating what information is available to which agents.
- The model must abstract specific methods and means of information consumption and generation, and of direct and indirect agent influence, thus removing the requirement for extensive domain expertise to use or interpret the model.

- Similarly, the model must support an operational lexicon to enable non-domain experts to concisely and consistently describe the modeled environment.
- The model must facilitate quantitative analysis to deliver consistent and repeatable conclusions.

With the above definitions and criteria in mind, I conducted a literature review of IE models that have been used in the private sector, government, and within the US Department of Defense. As with the more abstract information models mentioned previously, each of these approaches is tailored to a specific intended area of application. Each possesses strengths and valuable insights. However, none is completely sufficient to the broadly defined policymaking problem thus far examined.

1.7.1 DoD PMESII-PT

Early state-level attempts to recognize and accommodate information-based factors in competition include the US Department of Defense's PMESII-PT approach. The acronym PMESII-PT stands for Political, Military, Economic, Social, Information, Infrastructure, Physical Environment, and Time. The US Army codified these "operational variables" to ensure they were considered as part of the planning process.

The components of PMESII-PT address, in part, the larger challenge of the IE. Information is one of the variables, defined as "the aggregate of individuals, organizations, and systems that collect, process, disseminate, or act on information" [237]. As written, this would seem to align with the desired scope for the IE. The additional variables of Social (human networks and organizations) and Infrastructure (as relating to communications and media) are more specific, and perhaps redundant, to the broader Information variable.

In practice, however, PMESII-PT has proven woefully inadequate and overly linear to capture the realities of IE complexity [77]. The Information variable is generally reduced to a list of media outlets, and the variables are usually considered in a stovepipe fashion, with little time or formal methodology devoted to establishing and analyzing interrelations within them.

Includes all IE components: Partially. Agents and Platforms fit within the PMESII-PT variables. None of the variables is cleanly intended to capture narratives, issues, or the other features of the IE that comprise Topics.
Models agent influence: No. The model does not address networks of influence or relationships, though an especially savvy Army planner would include such information when describing the environment.
Models information propagation: No. At best the variables describe the means of information transmission and storage, but there is no active consideration for the flow of information.
Models information availability: No. Similar to previous.
Abstracts methods: Yes. The PMESII modeling method focuses on describing the environment as it is, without focus on how conditions came to be and/or might be altered.
Provides lexicon: No. The PMESII-PT variables are useful in describing current conditions but do not inform operational actions within that environment, and the

variables are not specific to the IE.
Supports quantitative analysis: No. PMESII-PT is a wholly subjective and qualitative framework.

1.7.2 DoD Information Environment Model

The end of the Cold War and the rapid onset of the Information Age caused pronounced shifts in how nation-states defined and pursued security. The past decades are replete with examples of previously reliable geopolitical powers and systems being weakened, thwarted, or outright defeated by supposedly asymmetric foes utilizing new modes of information-based competition. Information-based methods have enabled rapid pervasive espionage [212], hands-free sabotage [70], remote suppression of opposing governments [3], and in the extreme, outright military defeat [239].

The US DoD has invested considerable time and effort in examining how information affects security, and how a security organization must address the requirements, risks, and opportunities that an information-centric competition space entails. Admittedly, the DoD's views, methods, and doctrine in this area continue to evolve. In 2022, the DoD published Joint Publication (JP) 3-04 to address the intersection between military operations and information.

As part of this effort, the DoD defined the information environment as "an intellectual framework to help identify, understand, and describe how [the] often-intangible [social, cultural, linguistic, psychological, technical, and physical] factors may affect the employment of forces and bear on the decisions of the commander." The publication goes on to describe, in high-level terms, tenets of information competition, including striving for "information advantage" and "information power"; identifying "relevant actors"; and "utilizing narrative". All of these tasks are relegated to the Information Joint Function within a military headquarters:

> "The information joint function encompasses the management and application of information to change or maintain perceptions, attitudes, and other drivers of behavior and to support human and automated decision making. The information joint function is the intellectual organization of the tasks required to use information during all operations—understand how information impacts the operational environment, support human and automated decision making, and leverage information [toward objectives]."

The DoD's approach to defining and operationalizing the IE is admirably broad. Their list of IE components – social, cultural, linguistic, psychological, technical, and physical – is both reasonable and exhaustive. By design, the DoD IE is intended to identify the effects of information on human behavior, in order to facilitate effective competition.

However, the DoD IE is also intentionally under-defined. As constituted, the IE model in JP 3-04 offers no means of quantitative analysis. It provides no clear criteria for establishing interrelations between entities within the IE. It does not provide means to analyze narrative impacts, including the effects of mis-/disinformation, nor the means to analyze cyber system or network failures. Its broad formulation does not provide a useful operational lexicon for leaders. Rather than offering key terms and verbs to describe the IE and its evolution, the DoD model favors padded, imprecise, and overlapping labels for various aspects of the IE.

Includes all IE components: Yes. The DoD IE model specifically identifies Agents (relevant actors) and Platforms (technical components). It is appropriately broad in collecting Topics (narratives, beliefs, opinions, psychology).
Models agent influence: No. The model description acknowledges influence as a force but does not provide modeling methodology for it.
Models information propagation: No. Similar to previous.
Models information availability: Partially. The DoD IE model, and the information joint function wedded to it, place great emphasis on how information is collected and where it is present. Though the model addresses the importance of availability in decision making, it does not provide a modeling methodology.
Abstracts methods: No. The DoD IE model is all-inclusive for the US military, which includes information operations forces. Those forces' tools, techniques, and procedures are the specific tasks the IE model and the information joint function are intended to address and organize.
Provides lexicon: Partially. As with PMESII-PT, the DoD IE model provides useful descriptive lexicon but no active lexicon.
Supports quantitative analysis: No. The model is wholly subjective and qualitative.

1.7.3 4DM

Another more practical defense-based model is 4DM. As outlined in the US Army's Cyber Operations manual, FM3-12, the 4 D's are Destroy, Degrade, Disrupt, and Deceive [238]. These approaches focused on the technical and infrastructure components of the IE, as part of the DoD's electromagnetic warfare capabilities. Relative to a targeted system or capability, Destroy describes its complete removal; Degrade describes a reduction in its efficacy or reliability; and Disrupt describes a temporary outage or unavailability. Deceive describes a broader action, targeting the enemy as a larger entity rather than targeting a single system or capability, and describing the alteration of the contents of communications so as to induce desired behaviors. Cyber units unofficially include Manipulate, or "M", as an alternative to Deception: while deception utilizes false information, whereas Manipulation utilizes subjective or hypothetical information. In both instances, the goal is to provoke the target to act prejudicially to their own interests.

4DM is a useful lexicon for describing information operations, allowing high-level leaders to focus on desired outcomes without specifying specific technical means. However, the framework is heavily technically oriented, and the addition of Manipulate, while acknowledging the broader complexity of influencing human behavior, does not sufficiently counterbalance overly simplistic goals like Destroy. Further, 4DM does not examine, or even acknowledge, the myriad entities and interrelations within the IE, jumping directly to describe goals and ends without giving adequate consideration to their feasibility.

Includes all IE components: No. The framework generally disregards non-user Agents and only incidentally addresses Topics.
Models agent influence: Partially. Manipulation and Deception acknowledge the reality of influence, but 4DM does not attempt to model the process by which Agents

influence and are influenced.
Models information propagation: No. 4DM does not directly address the state of the IE, focusing instead on the desired end state.
Models information availability: No. Similar to previous.
Abstracts methods: Yes. The actions in 4DM are defined in terms of end-states or outcomes, without specifying the means of achievement.
Provides lexicon: Yes. By design, 4DM is an operational framework.
Supports quantitative analysis: Partially. The 4DM actions are tied to measurable outcomes, though those measurements are not inherently quantitative.

1.7.4 ABC(D)(E)

The Actor – Behavior – Content framework was conceived in 2019 as a framework for assisting regulators in targeting misinformation [89]. The authors correctly asserted that, while each of these dimensions had unique and specific characteristics, "they are ... often intertwined in disinformation campaigns, suggesting that effective and long-term approaches will need to address these different vectors with appropriate remedies."

Shortly afterward, researchers proposed adding Distribution (D) [11], specifically targeting structural factors within disinformation campaigns, including diffusion and transmission of disinformation; and Effect (E), to consider the actual harms caused by the disinformation and thereby include considerations of proportionality in response.

The ABCDE framework is an excellent rubric to help decision makers confronting a complex issue ensure they consider all the major situational factors. However, beyond giving these factors usefully mnemonic names, the framework offers little quantitative aid in identifying, understanding, analyzing, or comparing the components of those factors. It is an excellent tool for driving discussion and debate, but cannot, in itself, offer rigorous answers to the questions raised.

Includes all IE components: Yes. Agents and actors are synonymous in this approach. Platforms, along with interpersonal networks, can be covered in Distribution. Topics can be accommodated by Content, though the author's initial formulation of Content is narrower than my definition of Topics.
Models agent influence: Partially. ABCDE confronts disinformation precisely because the authors recognized the malign influence it can have. However, while the framework helps explore disinformation and misinformation influence, it does not provide functional modeling insight.
Models information propagation: No. ABCDE does not address specific information state within the IE.
Models information availability: No. Similar to previous.
Abstracts methods: No. The Behavior variable, in particular, encourages description of Actors' specific disinformation actions and techniques.
Provides lexicon: Partially. ABCDE is a useful descriptive lexicon, and despite its formulation around misinformation, can be applied to any influence phenomenon. However, it does not include directive verbs or lexicon to describe desired end states.
Supports quantitative analysis: No. ABCDE is wholly subjective and qualitative.

1.7.5 MITRE ATT&CK, AMITT, and DISARM

MITRE's AATT&CK (Adversarial Tactics, Techniques, and Common Knowledge) framework is a standard reference for cyber conflict in academic and industry circles. ATT&CK effectively addresses the main issues necessary for effective cyber defense: adversary behavior, attack lifecycle, and environmental variation. The framework exhaustively lists the many component activities in various cyber attacks, grouping them by temporal sequence and thus providing a broad overview of how many different attacks "unfold" in time. In doing so, it provides a comprehensive taxonomy for practitioners, facilitating quick communication and common understanding [225].

By design, the ATT&CK framework is technically focused. Even activities taking place outside of cyberspace – leveraging Trusted Relationships, for instance – are interpreted through the lens of information system access. Shortly after ATT&CK's advent, the Adversarial Misinformation Influence Tactics and Techniques (AMITT) framework was developed as a complement to ATT&CK, using a similar sequentially-oriented taxonomy to describe component activities of malign influence operations. The AMITT framework was later refined into the Disinformation Analysis & Risk Management (DISARM) framework [72].

As with ATT&CK, DISARM focuses on cataloguing the precise actions, and sequence thereof, that comprise an influence campaign. This is indisputably useful for offensive and defensive practitioners, serving as a "playbook" for considering offensive options and building comprehensive defensive strategies. These approaches, however, are extremely sequential, dissecting the events of a single attack, and often imposing a rigorously sequential view; this may often be reasonable in a purely cyber-based incident, but is likely overly simplistic for more complex, large-scale influence operations. Further, the frameworks are complementary in approach but largely stove piped, with no explicit references to how influence actions might produce cyber effects or vice versa.

Finally, the taxonomies presented are highly detailed and varied, oriented towards practicing experts. Their utility decreases as a function of individual expertise, meaning policy makers and elected officials may find these frameworks too detailed or alien to usefully inform their decision making.

Some companies offer products designed to generalize or adapt these frameworks toward broader decision making (for example, Recorded Future's "Diamond Model" [45]). These frameworks are good bridges between the two requirements, but ultimately suffer from the same vagueness that plague the 4DM and ABCDE frameworks.

Includes all IE components: No. None of these highly detailed frameworks directly addresses the Topics aspect of the IE.

Models agent influence: Yes. DISARM especially considers and models the many ways that Agents influence one another, and are influenced by the IE.

Models information propagation: Yes. DISARM addresses social propagation, while ATT&CK addresses technical propagation. Neither method provides an explicit model for the behavior, but they exhaustively consider all the means and methods whereby information travels and is stored.

Models information availability: Yes. Similar to previous.

Abstracts methods: No. ATT&CK and DISARM both attempt to catalog and taxonomize the various methods of influence, rather than abstracting them into broader operational terms.
Provides lexicon: Partially. These taxonomies are useful to leaders at every level and offer consistency in communication. However, they are expansive, descriptive, and often highly technical in nature, and are not intended to describe the broader IE concisely or to direct operational-level action.
Supports quantitative analysis: Yes. ATT&CK and DISARM activities include quantitative metrics and descriptors.

1.7.6 D-RAIL

Drawing on concepts pioneered in the 4DM and DISARM approaches, Directing Response Against Illicit Influence Operations (D-RAIL) is a recently developed framework specifically designed to support policy-level decision makers in combating external influence operations [159]. D-RAIL combines the sequential "kill chain" methodology with a methodological taxonomy and adds quantitative metrics to evaluate the impacts of influence operations against the costs and efficacy of potential countermeasures.

Admirably, D-RAIL focuses not on the content of influence messages, but on the larger patterns and interrelations between messengers and recipients. The author quickly points out that malign influence can be exerted without resorting to falsehoods, and that similarly, many false messages do not represent any threat. D-RAIL is specifically designed to identify, understand, and counter coordinated, wide-scale influence operations.

D-RAIL is a promising effort that is still developing, and of the work surveyed, was the only framework addressing the difficulty of understanding and shaping the IE at wide scale. One drawback to D-RAIL is the abstraction of the technical aspects of influence operations, with no specific mention of cyber-infrastructure or effects. D-RAIL is also an explicitly sequential approach, due to its underlying kill-chain mechanic.

Includes all IE components: No. D-RAIL focuses on interconnections between influence targets and influencers. It elides explicitly describing the means of connection (Platforms) and the purpose of communications (Topics).
Models agent influence: Partially. D-RAIL models influence as an outcome of deliberate action, using a "kill chain" paradigm. This is an effective model but limited in scope, capturing influence only as a one-way process and only with respect to the designated target, rather than the whole IE. The larger IE is nodded to in D-RAIL's consideration of potential spillover effects caused by interventions.
Models information propagation: Partially. As part of its kill chain methodology, D-RAIL specifically looks for "links" whereby information travels in the IE.
Models information availability: Partially. As with the previous point, D-RAIL does acknowledge the effect of availability, insofar as it requires a link between influencer and audience to assert influence.
Abstracts methods: No. D-RAIL is a useful strategic framework to examine specific courses of action, whether already in motion (i.e. info campaigns) or proposed

(i.e. possible interventions). The framework functions by examining the specifics of those courses of action.

Provides lexicon: No. D-RAIL offers step-by-step assistance in understanding and composing information campaigns, but does not include an operational taxonomy or lexicon.

Supports quantitative analysis: Partially. D-RAIL includes aspects of the DISARM framework, which itself includes some quantitative metrics.

1.7.7 BEND

The Crimean conflict catapulted information competition into focus, and as early as 2015, researchers attempted to quantize and taxonomize narrative conflict methods. The terms Dismiss, Distort, Distract, and Dismay emerged from early analysis [172]. After further study, researchers at Carnegie Mellon University broadened this set of narrative "maneuvers" to create the BEND framework [28][49]. BEND is based on a networked representation of Agents within the environment, along with subjects, emotions, and evidence extracted from those Agents' communications.

Unlike the DISARM approach, BEND abstracts the specific contents and methods of influencers into larger network-level "maneuvers." These maneuvers highlight the broader structural effects of influence efforts across the IE and are quantitatively measurable. The maneuvers provide leaders with a specific lexicon of actions (the 16 BEND maneuvers), by which leaders can convert a broad strategic goal into specific actions with a common understanding, without requiring those leaders to interpret their goals to the practitioner level – and thus not requiring significant influence expertise to understand or utilize.

BEND is highly effective at describing and measuring influence operations, but as with other frameworks, it ignores the technical aspect of influence campaigns. BEND maneuvers are also difficult to detect, though current research has made significant strides on quantitative post-hoc identification of maneuvers [113]. Further, BEND does not provide a taxonomy for practitioners; experts (i.e. copywriters) must decide upon appropriate tactics and techniques for a given BEND maneuver and target.

Includes all IE components: No. BEND does not address Platforms or means of connection outside of social ties.

Models agent influence: Yes. BEND captures environmental and inter-Agent influence.

Models information propagation: Partially. BEND can identify or observe the spread of ideas between Agents but does not directly model how that information spreads.

Models information availability: Yes. BEND's network approach provides a direct representation of "who knows what" within the IE.

Abstracts methods: Yes. BEND effects are observed within the network, and the network representation, by construction, lacks the specifics of how connections were formed or altered.

Provides lexicon: Yes. The 16 BEND maneuvers are, by design, operational verbs intended for use in describing changes – observed or desired – within the IE.
Supports quantitative analysis: Yes. BEND's network model lends itself to mathematical analysis.

1.7.8 SCOTCH

The Source-Channel-Objective-Target-Composition-Hook (SCOTCH) framework was devised specifically for the operational leadership niche [31]. SCOTCH is designed to "comprehensively and rapidly characterize an adversarial [influence] operation or campaign." SCOTCH synthesizes the components of influence campaigns from several preceding models and offers a set of factors that very completely describe such campaigns.

However, SCOTCH suffers from the same shortfall as ABCDE (one of its sources) in that it provides only a mnemonic to ensure comprehensive consideration. SCOTCH does not offer a taxonomy of specific possible values for each variable, nor does it provide quantitative measures to compare results or evaluate efficacy. As with ABCDE, SCOTCH is an outstanding starting point for leadership-level discussion, but a poor tool for arguing the merits of any solution under consideration.

Includes all IE components: Partially. SCOTCH's use of "Composition" and "Hook" are the closest match to the Topics component of the IE, and in practice are poor analogs.
Models agent influence: Partially. As with ABCDE, SCOTCH was designed specifically to address information influence. And as with ABCDE, SCOTCH offers tools to recognize that phenomenon, but not to model its mechanism.
Models information propagation: Partially. SCOTCH "Channels" address how information moves and is delivered, but there is no consideration for information movement in the larger IE.
Models information availability: No. SCOTCH analyses proceed with the tacit assumption that information payloads have been delivered/consumed. The framework's focus is on understanding the payload's intended effect and method of impact.
Abstracts methods: No. By its nature, SCOTCH is an examination of methods, a framework for unpacking the specifics of an information campaign and the actions therein.
Provides lexicon: No. SCOTCH does not provide operational-level descriptors for either actions or IE states.
Supports quantitative analysis: No. The framework is subjective and qualitative.

1.7.9 Friedkin-Johnsen

The Friedkin-Johnsen (FJ) model of social dynamics is a landmark in the study of social influence [92]. The FJ model captures influence exerted by agents via social ties, and incorporates agent "stubbornness". It describes how agents update opinions by averaging neighbors' views while

partially adhering to their own initial, innate opinion, often leading to polarized or clustered, rather than consensual, outcomes.

The FJ model has proven especially apt at modeling dense social situations, and has found wide applicability in the study of online dynamics [170] [61]. However, the FJ model does not directly address information transmission or storage; all availability is captured in the presence (or absence) of a connecting link. Further, the model does not carry a standardized lexicon (and indeed has been so widely studied that agreement on such a lexicon may be impossible).

Includes all IE components: No. The FJ model does not address information mechanics (i.e. platforms) and most version only indirectly represent specific information (i.e. topics).
Models agent influence: Yes. FJ is a standard in inter-agent influence, and credible influence models generally incorporate its mechanisms.
Models information propagation: Partially. FJ-based semantic networks do address knowledge transferral but are not inherently part of the FJ model.
Models information availability: No. As with communication means, the FJ model abstracts or ignores availability mechanics.
Abstracts methods: Yes. The social influence model in FJ does not explicitly state how that influence is exerted, perceived, or received.
Provides lexicon: No. As mentioned above, the wide study of the model likely precludes any widely-accepted concise label set.
Supports quantitative analysis: Yes. The network structure of the FJ-model lends itself to rigorous mathematical analysis.

1.8 Model Comparison

Table 1.1 provides a summary comparison of these different conceptual models using a lite docking approach [44], against the six criteria I outlined previously. Each model meets some, but not all, of the sufficiency criteria.

Subsequent chapters discuss the ramifications of this insufficiency. Chapter 2 examines attempts to model and understand socio-cyber "hybrid" influence campaigns and attacks. Hybrid attacks leverage cross-domain influence between cognitive spaces (Agents' beliefs and opinions) and cyber spaces (technical channels and delivery mechanisms). In hybrid attacks, malicious actors act in one of these dimensions – or both concurrently – to achieve influence effects in the other.

Chapter 3 examines one of the primary mechanisms of this cross-domain influence: information availability. There, I offer a taxonomy for availability-based influence attempts and generalize the BEND maneuver set by appending these availability-based maneuvers.

1.9 Constructing a Sufficient Model

This thesis presents a new model for the IE, the Operational Information Environment Model or OPIEM. This model satisfies the sufficiency requirements identified previously, using a network

Table 1.1: Comparison of existing IE model attributes

Sufficiency Criteria	**PMESII-PT**	**DoD IE**	**4DM**	**ABCDE**	**ATT&CK + DISARM**	**D-RAIL**	**BEND**	**SCOTCH**	**Friedkin-Johnsen**
Includes all IE components	*	X		X				*	
Models agent influence			*	*	X	*	X	*	X
Models info propagation					X	*	*	*	*
Models info availability		*			X	*	X		
Abstracts methods	X		X				X		X
Provides lexicon		*	X	*	*		X		
Supports quantitative analysis			*		X	*	X		X

representation of the integrated information environment. Network representations are intrinsically quantitative, lending any model built thereon an immediate analytical advantage. Networks also explicitly represent pertinent entities and the relationships among them.

Constructing a network model of any object of study requires quantizing that object into discrete nodes and relationships and determining the appropriate valuation for any node and link attributes. If these values are demonstrably accurate to the underlying system, and the network demonstrably includes all nodes and links pertinent to the underlying system, we may claim that the network is a sufficient representation of that system. For a given system, in the context of a specific set of research objectives, a network thus derived almost surely does not *completely* capture all aspects and complexities of the modeled system. Nevertheless, the network can be built such that it sufficiently represents the details and structure of that system *within the specified context*. If a network representation is sufficient to the system, then the analysis of that network – mathematical measurements and observed structural transformations – may be interpreted as characteristics of the underlying system. These network analyses are accessible by virtue of the network and may not be obvious or observable in studying the underlying non-quantitative system. Thus, a sufficient network model can provide significant insight into the function and nature of the represented system. It remains to demonstrate that a sufficient network model of the IE can be constructed.

Previous research has demonstrated the utility of network approaches in studying sociological phenomena such as influence. Multiple studies have demonstrated the value of networks in modeling social influence and information diffusion in the past four decades [20] [47], including modeling time-variant social processes [53]. Friedkin's seminal research is especially salient, detailing how actors integrate conflicting opinions within mutually influential relationships to revise their own positions [91]. Research has also demonstrated how networks effectively model semantic processes, including how topics compete, collide, and combine within a collective nar-

rative or awareness [145] [160]. Finally, network models are highly effective in modeling information movement, storage, and processing within discrete infrastructure, and more particularly within the cyber infrastructure of interest in modeling the IE [16] [246].

1.9.1 Network Components

Based on these demonstrated results, a network can effectively model the desired characteristics of the operational-level IE. To construct a sufficient model, we must first select node sets that adequately represent the entities of interest within the IE. I propose a tri-modal network, composed of three nodesets corresponding to the IE components identified previously:

- Agents, the entities in the environment capable of understanding, deciding, and acting upon information.
- Topics, a list of salient environmental categories or labels, concerning which information is extant in the IE.
- Platforms, a list of the entities within the environment that contain, store, present, collect, transmit, or manipulate information, utilized by agents to interact with the IE.

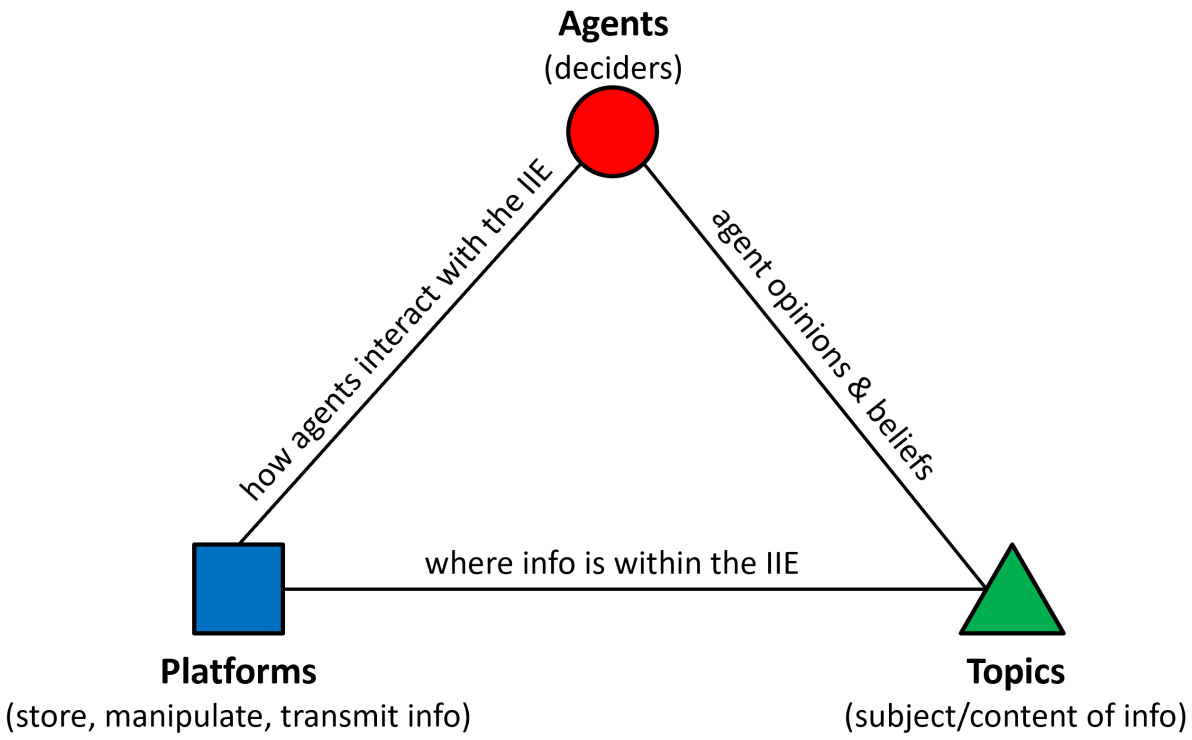

Figure 1.2: Nodesets and relationships within the OPIEM network.

Within this trimodal network, six types of relationships exist:

- Agent-Agent, representing the intercommunication and/or interpersonal influence between agents.

- Agent-Topic, representing an agent's awareness of and disposition regarding a topic. (This corresponds to the DoD IE's "Cognitive" dimension.)
- Agent-Platform, representing an agent's access to and behavior relative to a platform. (This corresponds to the DoD IE's "Informational" dimension.)
- Topic-Topic, representing co-occurrence and semantic alignment between topics.
- Topic-Platform, representing the trafficking of a concept within or across a platform. (This corresponds to the DoD IE's "Physical" dimension.)
- Platform-Platform, representing interconnection or exchange of information between platforms.

1.9.2 Link Derivation

The quantitative power of a network representation lies in the links between nodes and the resulting graph structure. Therefore the methodology for deriving and valuing links is of utmost importance in building a truly sufficient model.

I propose two classes of links within OPIEM, differentiated by their method of derivation. Outer links are derived quantitatively from observed traffic, using values like message count or frequency. Inner links are derived by parsing traffic content and often rely on additional processing or interpretation. For example: one could construct normalized links between an agent and the topics about which the agent has messaged; these links could be interpreted as the agent's *priorities* or *saliences*, measures of the proportionate concern the agent has for each topic. Because the links are constructed entirely by identifying the agent, identifying the topics, and counting messages frequency, these links are **exterior** links, built from observable message features. By contrast, one could parse the content of those messages and identify positive/negative emotional markers and assign the link between the agent and a given topic the average value of the emotional valence in the agent's messages about that topic. The resulting network could be interpreted as the agent's *position* or *emotional valence* toward the topics in the IE. Because these links required parsing and qualitative assessment (albeit automated) of the message content, they are **interior** links.

There are many possible link valuation schemes for both inner and outer links, and the specific method for link building depends on the purpose of the OPIEM model being constructed – the decisions it is intended to inform, and the IE phenomena it is intended to aid in observing. The link valuations and methods we use in this research are not definitive or exhaustive, nor are they exclusive to other methods; an OPIEM model can contain multiple links between nodes, of different types and valuations, without losing analytical tractability.

1.9.3 Network Construction

Because the IE is tremendously complex, building a sufficient model manually would be onerously time intensive. Further, it is unlikely that any such model would be truly sufficient, since no observer can be expected to have full instantaneous awareness of all entities, relationships, and values thereof, within the IE. This problem can be simplified by modeling only the observable IE, meaning the portion of the IE that can be asserted to exist based on observable information

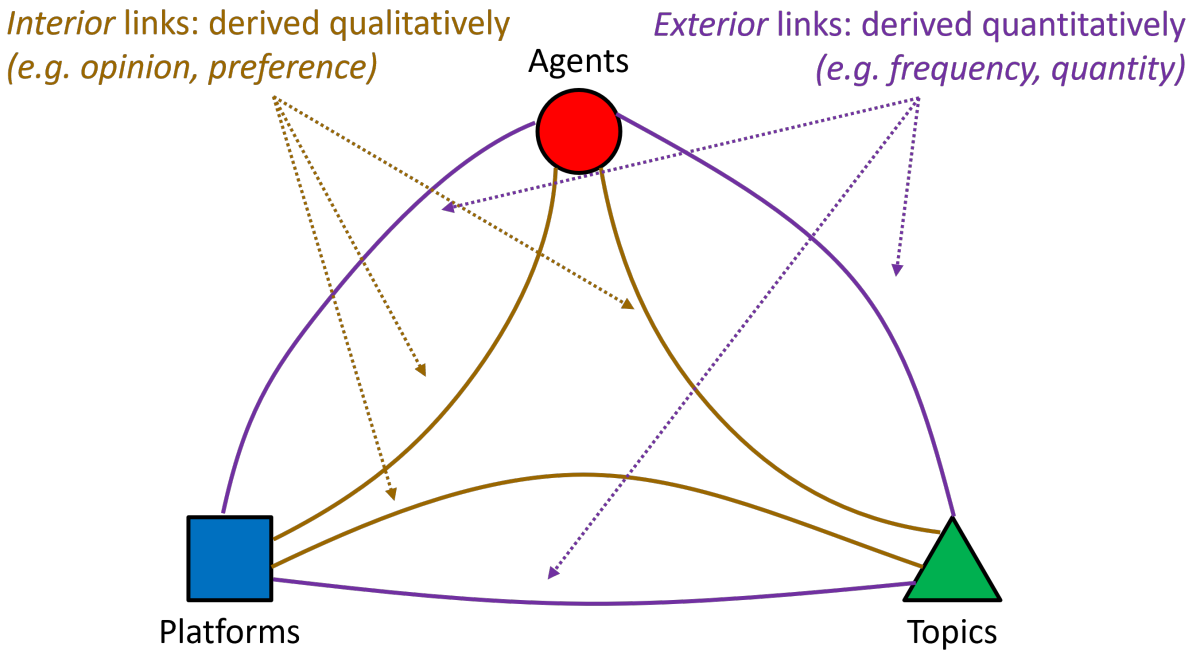

Figure 1.3: Link categories within an OPIEM network.

(information detectable as in transit, in process, or in storage within the IE). However, fully modeling even this reduced portion of the IE is likely beyond any deliberate effort.

I propose to build an OPIEM trimodal model of the integrated IE from the bottom up. By monitoring information in transit, in the form of exchanged message traffic between environment components (i.e. emails, social media posts, news broadcasts, etc.), an OPIEM modeler derives the nodes and constructs verifiable relationships. This model approaches sufficiency as a function of observed traffic, meaning that as ever more traffic is observed, the model converges with a fully informed model of the environment. Agents, topics, and platforms that are never involved in traffic will be absent from the model; but as they are not involved in traffic, they are likely not pertinent to the IE, meaning their absence does not significantly reduce the model's sufficiency. Under the assumption that all traffic as a non-zero probability of being observed, and given enough time, all pertinent entities will appear within the model. And, under the assumption that all traffic is intelligible, the relationships derived within the network will converge toward the actual relationships within the environment.

1.9.4 Network Analysis

As constructed, the proposed OPIEM trimodal network is not immediately analytical useful. A trimodal graph with multiple separate edge sets is largely inaccessible to most network analysis algorithms and metrics. However, reducing the complexity of the OPIEM "core" network would likely negate the model's sufficiency.

These requirements are reconciled by reducing the core network for targeted analysis. Rather than derive metrics from the complex core network, the core network is reduced to simpler bi-

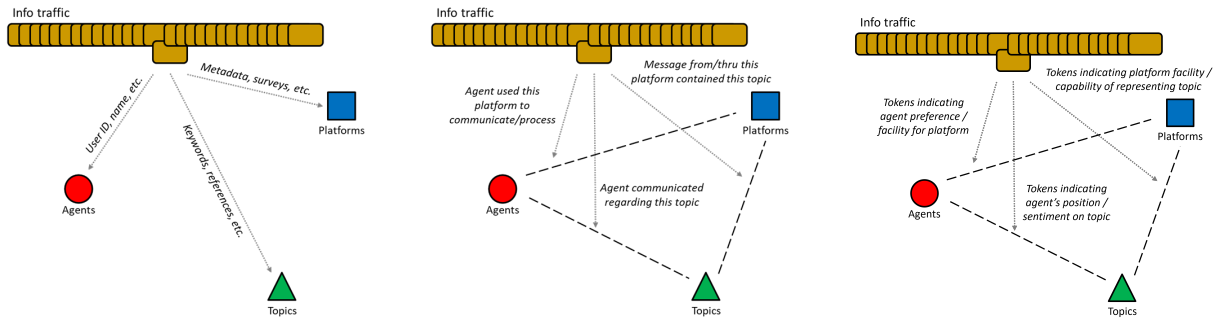

Figure 1.4: Construction of an OPIEM core network from message traffic. Nodes, exterior link values. and interior link values are all derived from message contents, requiring no a priori knowledge of the network.

modal and unimodal networks through folding or targeted exclusion, and those simpler networks are then analyzed. Because the networks are derived from a sufficient core network, they retain sufficiency, albeit only in the narrower interpretation of the IE that they represent.

As an example, consider a core trimodal network of agents, topics, and platforms. One can create a bi-modal agent-topic network by folding the core network on the platforms nodeset, valuing the agent-topic links produced by counting the number of common platforms between the two nodes and then normalizing link values across the new network. This new bimodal network might be called the "Exposure" network: a link between an agent and topic shows an agent's exposure to that topic within the IE. The simplified network is no longer sufficient to describe the entire IE, but it is sufficient for analysis related to topic exposure; this might include research questions about media diet, viewpoint variety, echo chambers, etc. This bimodal network might be reduced further by folding again on the topic nodeset and valuing the new links as the difference between the previous common links, creating a unimodal agent-agent network with links measuring the agents' "common exposure," or the degree to which two agents' topical exposure varies.

As the trimodal network is successively reduced, network metrics like path distance, centrality, and density become simpler to calculate. These quantitative metrics, coupled with the refined scope of the reduced networks, provide mathematically based descriptions of the IE. In the agent-agent "common exposure" network, for example, a clustering algorithm that considered path weights would group agents with comparable topic exposure, revealing "audiences" in a quantitative manner, rather than requiring the subjective assignment of an identity label.

1.9.5 Model Labels and Lexicon

A trimodal network with a large potential link set can produce thousands of possible derivative networks for analysis, and each analysis potentially reveals a different aspect of the IE. Thus the OPIEM model also proposes a set of useful operational terms to describe IE conditions or changes that are of interest or use in decision making.

This lexicon mirrors the CMU BEND method of assigning maneuver labels to broad structural trends or transformations [28] [113]. It does not take the ATT&CK/DISARM track of

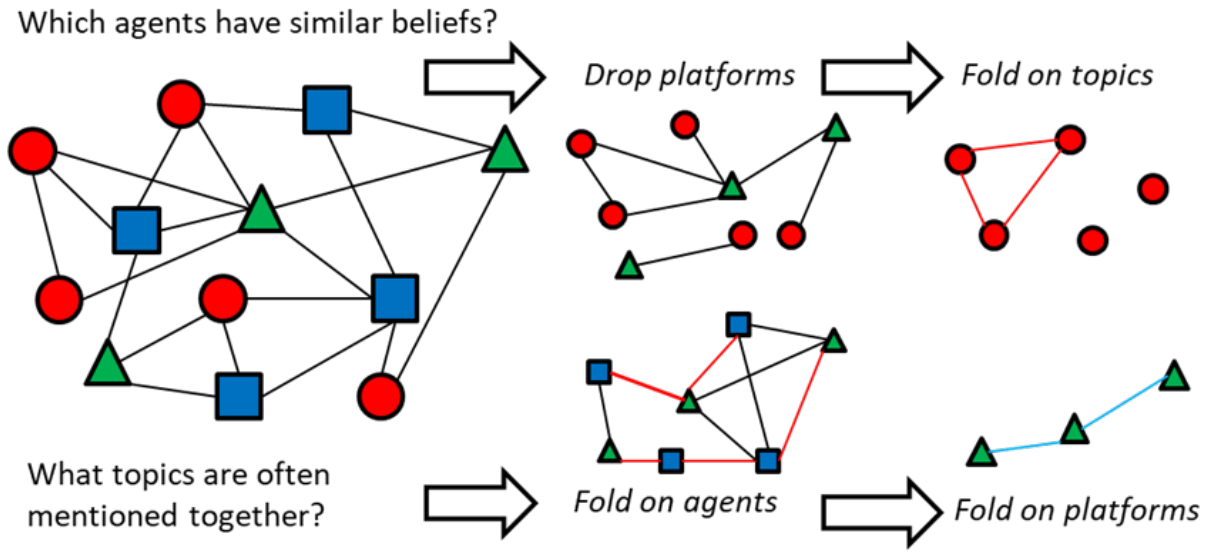

Figure 1.5: Examples of network reduction in response to queries.

assigning names to specific methods, techniques, or implementations of influence transmission. In fact, OPIEM is a generalization of the BEND approach and subsumes the 16 BEND maneuvers as lexical descriptors for actions taken on the Agent-Topic bimodal network, or on derivative unimodal networks of the same. OPIEM then adds 8 additional maneuvers (using R and S labels) as operational descriptors for actions taken within the Agent-Platform and Topic-Platform networks, resulting in the BENDRS maneuver set for the trimodal IE.

Chapter 2

Social-Cyber Influence and Hybrid Attacks

This chapter examines the inadequacy of existing analytical frameworks in understanding information competition warfare in the modern, digitally mediated information environment (IE) that was introduced in chapter 1. Traditional conceptions of conflict, rooted in physical resources, territorial control, and kinetic force, struggle to accommodate information-based attacks that operate below conventional thresholds of hostility while nonetheless producing significant operational and societal harm [111]. Contemporary information warfare exploits both the content of information and the composition of the information environment itself, influencing beliefs, emotions, preferences, and behaviors while simultaneously altering the technical structures through which information flows. Existing cyber frameworks (e.g., ATT&CK) and cognitive influence frameworks (e.g., BEND, DISARM) each offer valuable but incomplete perspectives, failing to account for second-order and cross-domain effects. This disconnect impairs leaders' ability to anticipate, interpret, and respond to malicious influence campaigns at the operational and societal levels, underscoring the need for a unified cognitive-cyber model of the information environment.

Information-based attacks have long represented grey-space conflict between states [132]. Nations can inflict informational damage to gain competitive advantage while staying below established thresholds for open warfare. Efforts to steal information, change popular opinions, establish or dispel widely held beliefs, or otherwise alter an opponent's information environment can all be undertaken outside currently recognized standards of "hostility" [177] [189].

All of these actions are attacks, in that they target the opposing state – its components, constituents, and capabilities. Clausewitz defined war as violence leveraged toward compulsion [59]. In this framing, a losing combatant is compelled toward certain decisions, but is aware of the compulsory forces, in the form of physical threat. Information conflict is more subtle, leveraging information toward induction: a 'losing' combatant is induced to make certain decisions, but is often unaware of the inducing influence [17]. These two differences – the lack of tangible, irreversible destruction, and the lack of overt, declared influence – make it difficult to reconcile the realities of information conflict with larger conflict frameworks.

I will nevertheless use the term 'information warfare' in this chapter to describe concerted efforts by societal parties, be they governmental, corporate, or other, to influence via informational means the decisions of any other group or party. I am comfortable with the label 'warfare' because, even though information conflict does not satisfy the strictest Clausewitzian definition, it does fulfill two other common definitions for war. First, it is a protracted, resource-intensive

process, often requiring significant resources to generate results; and second, the potential for one party to inflict significant harm to another party is no less real than for kinetic warfare, especially when measured in societal metrics such as economic function [111].

2.1 Mechanisms of Information Warfare

Information Age technology and infrastructure have vastly increased the options available to combatants [64]. I consider two broad categorical mechanisms at play: Information conflict affects both the *content* and the *composition* of the information environment [143] [228]. While two opposing parties will each have their own information environment, shaped by their respective contexts and actions, in this examination I assume that these environments are at least partially common – that is, either party has the ability to affect the other's information environment.

Information warfare affects *content* by altering the substance of the information consumed. An opponent might add spurious information or highlight specific information to the exclusion of alternatives. They may seek to obscure or remove information or pose alternative explanations or interpretations. Overt psychological warfare practices over the last century are an example of such: dropping leaflets offers new possible outcomes for enemies to consider; playing loudspeakers highlights favorable information while deliberately excluding unfavorable information [174]. More recently, state-level disinformation campaigns and compromising espionage actions demonstrate impact by changing the content of information.

Information warfare also affects the *composition* of the information environment. Opponents may take actions that affect how information moves through their enemy's populace and decision-making structures. As previously established in chapter 2, a target is influenced by the words in an article but also influenced by the binary fact that they read the article at all; the influence experienced by the target can be changed by altering the words of the article, or by preventing the target from reading it. Attacks altering the composition of the information environment operate in just such a fashion, by altering the routing, distribution, and often the availability of information.

2.2 Means of Information Warfare

It would be convenient to lump the various means of information warfare into these two categorical pillars. We might state that content effects are achieved through cognitive means (i.e. psychological warfare, disinformation, marketing, etc.) while composition effects are achieved through *cyber* means (i.e., denial-of-service, forceable shutdown, etc.). However, the line between the composition and content of the information environment is necessarily blurry; the two heavily inform one another as the environment reactively evolves to the pressures exerted by the consuming audience.

Similarly, the line between cognitive-based and cyber-based information attacks is growing fuzzier as practitioners evolve. Sophisticated cyber offense must often consider not only the mechanics of gaining access to systems, but also the best way to alter the content of those systems

once access is gained; often greater operational effects can be achieved through co-opting a system [75] [2], than through the brute-force shutdown or denial of that system [27]. Cognitive influencers, be they individual entrepreneurs or public affairs officials, must increasingly address the technical means of distribution and diffusion within the IE as they craft and disseminate their messages.

As a result, modern information attacks result in both content and composition effects within the IE, and most attacks contain a mixture of cognitive methods and cyber methods. The specific threshold for defining whether an action qualifies as a cognitive or cyber 'attack' is outside the scope of this work. In the interim, I adopt working definitions for further analysis, specifically:

- A *cognitive action* is an action aimed at the target consumer's cognitive process. Examples include composing text, sound, or imagery.
- A *cognitive effect* is an achieved impact within the consumer's cognitive process – an evoked emotion, or an altered or established belief.
- A *cyber action* is an action aimed at any technical system involved in the target consumer's information consumption processes. Examples range from service shutdown, to falsifying information or identity, to directly altering information within an information system.
- A *cyber effect* is an achieved impact on the operation of a target system that is counter to the system's intended operation and/or counter to the system controller's interests relative to that system's designated function.
- An *attack* is an action aimed at achieving an effect that is counter to the interests or desires of the targeted individual(s).

There are examples of so-called "pure" attacks in either category. A salesman lying about the history of a vehicle, for instance, is targeting the consumer's decision-making to establish incorrect beliefs and evoke feelings of confidence. A hacker inserting backdoor code in a service system to ensure his own requests come first is using technical means to alter the functionality of that system for his own benefit, likely undermining the intended performance of the system.

These examples, however, are trivial instances of information attacks, and necessarily so. Substantive and impactful information attacks – widespread misinformation campaigns, massive data breaches, denial of service attacks – straddle the line between cognitive and cyber, both in means and in effect. Some efforts are deliberately spread across cyber and cognitive means, by necessity or in an effort to achieve greater impact.

I use the term *hybrid attack* for such dimension-straddling events. I must acknowledge that this term is already widely in use in literature. Frank G. Hoffman of the Potomic Insititue for Policy Studies called hybrid attacks "multiple types of warfare ... used simultaneously by flexible and sophisticated adversaries who understand that successful conflict takes on a variety of forms that are designed to fit one's goals at that particular time" [115]. NATO refers to hybrid methods of warfare as "propaganda, deception, sabotage, and other non-military tactics" [7]. Generally, hybrid attacks are seen as "the vehicle to characterize the increased complexity and non-linearity of threat actors that contest the status quo" [87]. As there are many definitions, I am careful to define it for the context of this research as below:

> A hybrid attack is a hostile action – meaning it is undertaken counter to the interests or desires of the target – composed of:

- Cyber actions intended to achieve a cognitive effect;
- Cognitive actions intended to achieve a cyber effect; or
- Mutually supporting and concurrent cyber and cognitive actions, intended to achieve a unified overall effect (either cyber or cognitive).

To clarify this definition, I offer examples:

- Hackers compromise the web site of a prominent bank and aggressively deface it, mocking the bank's security. They overtly and/or subtly suggest another bank has superior security, leading users to switch banks. This demonstrates a cyber action (website defacement) intended to evoke a cognitive effect (distrust of the attacked bank) to produce a decision (bank switch).
- In a phishing attack, attackers send messages purporting to be from a trusted source. The messages convince users to unwittingly compromise their credentials. This demonstrates a cognitive action (convincing text) intended to achieve a cyber effect (credential possession) to enable further cyber actions (unauthorized system access).
- A state undertakes a military action outside of international norms. The state releases images and narratives that partially describe the military action, framing it as a necessary and anodyne occurrence. Simultaneously, the state seeks to suppress internet traffic in and out of the affected area, via denial-of-service attacks and direct cyber action against routing nodes. As reports emerge, the state compromises the accounts of the journalists and publications involved, altering content, producing contradictory accounts, and generally discrediting the reporting. This demonstrates cognitive actions (official press releases, falsified accounts) and cyber actions (account compromise, traffic suppression) both utilized in service of a single greater cognitive effort (controlling public perception of the military action).

I reiterate that any operational-level or societal-level cyber action will produce both cyber effects and cognitive effects; and conversely, any operational-level or societal-level cognitive action will produce both cyber effects and cognitive effects. The heavily digital nature of the modern information environment produces an indelible connection between cyberspace and the intangible human informational gestalt, as was initially discussed in Chapter 1. This inextricable connection results in an inescapable conclusion: at the operational level, all information attacks are effectively hybrid attacks. Whether effected through cyberspace or offline, any attempt to manipulate the operational information environment will have repercussions across both the cognitive and cyber dimensions of that environment. Therefore, effective understanding of hybrid attacks can provide significant insight to operational and societal leaders.

In chapter 1, I established that no extant IE model is fully sufficient. Because hybrid attacks exceed the intended analytical scope of both cyber frameworks (e.g. ATT&CK) and cognitive frameworks (e.g. BEND, DISARM), they cannot be modeled accurately and thus the breadth and impact of hybrid attacks is often unappreciated or unrecognized. Leaders relying on a single framework, or even a patchwork of frameworks, are unequipped to model – and therefore understand and respond to – hybrid attacks. To better understand the limitations leaders face, I consider several hybrid attack case studies.

2.3 Data

I examined the following case studies in considering the challenges of hybrid attack analysis. For each case, I sought to identify the cyber and cognitive methods utilized, and to subjectively analyze the extent of the attack's impact.

2.3.1 Operation Aurora Gmail compromise (2009)

In 2009, a targeted cyberattack affecting dozens of US companies was attributed to the Chinese government; these attacks were collectively dubbed 'Operation Aurora' [188]. Google's Gmail web service was one of the targets; post-attack analysis indicated that the hackers sought to gather information on Chinese dissidents [39] [248]. Google's stock price fell for more than a week after the attack was publicized [219].

Cyber methods: Operation Aurora was a highly sophisticated series of cyber attacks built on day-0 exploits, previously established persistent access, and other resource-intensive techniques generally only available at the nation-state level.

Cognitive methods: Operation Aurora was a secret operation and was not acknowledged during or after its conduct. While the operation had cognitive aims, it did not include any overt cognitive influence efforts.

Impacts:

- **Agents and Communities:** Medium. Aurora granted the Chinese government access to several accounts of state-labeled dissidents. That information was used to discredit and suppress those voices, in concert with other ongoing state suppression efforts.
- **Information and Knowledge:** Low. Aurora attempted to distort certain topics and minimize certain voices but did not actively suppress information nor actively promote specific or alternative narratives.
- **Infrastructure and Channels:** Low. As Aurora was a secret operation, the operators took pains to minimize obvious impacts on system functionality. This minimal effect allowed Aurora to proceed for a substantial amount of time before detection.

2.3.2 Twitter User Exodus (2022-2025)

Elon Musk's acquisition of Twitter in 2022, as well as subsequent policy changes and position statements by both him and the company, have resulted in several windows of user migration. In these instances, Twitter/X users have left the platform for alternative services and encouraged others to do so [51] [108] [138]. For my purposes I consider these various migrations as a single incident, since the inciting factors are largely similar in essence if not in the particulars.

Cyber methods: User migration is a technical process, but the attendant interactions with the gaining and losing platforms were supported by platform design. The migrations were not enacted through malicious cyber actions.

Cognitive methods: Individual decisions to leave Twitter were driven by messaging from Musk and related figures on the platform. These decisions were generally amplified by collective reinforcement, seen as social pressure and virtue signaling within subcommunities. Musk and his ilk directly influenced some users to leave through challenge messaging, but those users – and others initially offended – increased the volume of departures through collective pressure.

Impacts:

- **Agents and Communities:** Medium. Few people changed their core opinions based on Musk's policy changes. Rather, Musk provided an impetus for like-minded groups to decamp from Twitter and seek alternative social spaces, for a variety of reasons. As a result, varying schools of thought became further separated, with less opportunity for direct interaction, producing an echo-chamber-like reinforcement effect within both the gaining and losing platforms.
- **Information and Knowledge:** Medium. Musk's policy shifts re-platformed many previously discredited or silenced voices and communities, enabling them to present their narratives to broad audiences. The segmenting migration effect then largely stovepiped those narratives, creating two parallel and contrasting social dialogues. As a result, users who subscribe to only one of them are likely to never encounter narratives that have been established within the other.
- **Infrastructure and Channels:** High. User base is a primary driver of platform configuration and behavior. The explosion of new users surging into BlueSky and other Twitter/X alternatives required the rapid expansion of those services, and with that expansion came reconfiguration to handle new technical and content challenges. Conversely, the loss of users and ad revenue reduced Twitter's technical resources and influence within the larger public awareness, reducing its centrality as a public information channel.

2.3.3 Russian Internet Freezes (2024-present)

In an effort to limit Ukrainian drone capabilities within Russian territory, Russian officials have implemented rolling network blackouts. Low-cost teleoperated drones use commercial network technologies, including cellular 4G/5G networks, for command and control. By shutting down cellular networks, Russian officials hope to reduce the efficacy of the drones as well as preventing imagery and accounts of embarrassing Ukrainian attacks. The shutdowns have significant impacts on the Russian populace, affecting logistics, finances, and commerce [134].

Cyber methods: Russia has implemented other cyber-delivered drone countermeasures in the past, but the method of examination here is a brute-force, full-system shutdown. It is an unsophisticated, if effective, cyber attack.

Cognitive methods: Russia has made significant efforts to influence their domestic population, the Ukrainian population, and the larger international community. In this instance, however, there are no specific cognitive components to Russia's counter-drone efforts.

Impacts:

- **Agents and Communities:** Low. The reduced reliability of network services is unlikely to significantly change how people connect to one another, though it will likely delay information propagation within those networks.
- **Information and Knowledge:** Medium. Network shutdowns will not induce any specific loss of information not already achieved through other censorship means. However, it prevents real-time reporting of incidents, and enables state censorship mechanisms to prepare in advance for narratives that may emerge after the shutdown ends.
- **Infrastructure and Channels:** High. Reduced availability has produced demand for reliable service alternatives, leading to a resurgence of SMS-based services and an accordant shift of user volume.

2.3.4 Discord.io (2023)

Discord.io was a third-party service that enabled users to create custom, interactive invitations to channels on the Discord social media platform. On August 14, 2023, a preview of the users database from the online service Discord.io was posted on a cybercrime marketplace, with the rest of the database offered up for sale. The database contained the usernames and Discord IDs of 760,000 accounts, along with billing addresses associated with those accounts. The following day, Discord.io officially and indefinitely ceased all operations; it remains shuttered to this day [191].

Cyber methods: The specific method of compromise was reported as a flaw in the company's website code, which likely enabled privilege escalation and system penetration. Such a flaw could be exposed by standard penetration testing and reconnaissance tools and does not indicate an attacker of particularly high technical proficiency.

Cognitive methods: No cognitive methods were employed in compromising the Discord.io system.

Impacts:

- **Agents and Communities:** High. Discord.io's dissolution represented a financial loss for customers. In addition, the compromised database exposed customer information useful to scammers.
- **Information and Knowledge:** Medium. Discord.io's dissolution leaked consumer data, as mentioned. It also resulted in the loss of all user-generated content the service had helped generate and stored.

- **Infrastructure and Channels:** Medium. Discord.io was an active service with an substantial user base. Its shutdown resulted in the loss of that service, as well as the weakening of community bonds within the user base.

2.3.5 The John Oliver effect vs. the FCC (2014)

Since 2014 comedian John Oliver has hosted Last Week Tonight, a satirical news program that discusses current events and public figures and organizations. Oliver has an avowed liberal political bias, and often selects stories to highlight perceived injustices, structural flaws, or crises within public life and politics. His style of presentation is well received, especially in the United States; as a result, Last Week Tonight has been multiple awards and enjoys a broad, loyal audience.

This audience is frequently responsive to Oliver's calls for action, and the collective effect of his viewers' attention has been called "The John Oliver Effect" [136]. The case salient to this research stems from an early episode of Last Week Tonight in which the show covered ongoing FCC regulation revisions regarding net neutrality. Oliver encouraged his audience to weigh in on the matter during the mandatory open comment period. His fans responded and overwhelmed the FCC's website, causing a server crash and shutting down the comment system [116].

Cyber methods: No cyber attacks have been directly attributed to Oliver's influence. The FCC server crash was a legitimate service surge, not an orchestrated denial-of-service attack.

Cognitive methods: Oliver employs humor, emotional appeal, moral appeal, and other rhetorical devices in presenting stories. He clearly identifies a supported position and frequently provides his audience with suggestions of specific tasks or actions they could undertake to affect the issue.

Impacts:

- **Agents and Communities:** None. The FCC server crash is notable for its sudden impact, but only temporarily removed the FCC from online dialogue. The crash was caused by many citizens acting independently, and did not result in a new online community forming.
- **Information and Knowledge:** Medium. Oliver's educational presentation likely exposed large portions of his audience to the net neutrality debate for the first time, meaning his presentation – including the facts and biases therein – were many agents' entire understanding of the issue.
- **Infrastructure and Channels:** Low. The FCC server was knocked offline for a short time but recovered as demand was reduced. No significant re-engineering or reconfiguring of the FCC's comments system was reported.

2.3.6 Patriotic Hackers (persistent)

International discord and rivalries are often internalized and personalized by citizens of the affected nations, as seen in large-scale protests and product bans. Well-known historical examples

include the Boston Tea Party (US citizens rallying against the UK); the Boxer Rebellion (Chinese citizens rallying against Western powers); and the Salt March (Indian citizens rallying against the UK).

The increasingly digitally literate world population, coupled with increasingly digitized societal systems, has given rise to 'patriotic hacking', in which citizens of one nation, motivated by a larger societal grievance, launch cyber attacks against the citizens and/or institutions of another nation. These attacks vary in scope and sophistication based on the capabilities of the attacker. For my purposes I consider attacks that are explicitly not state-sponsored or enabled, focusing instead on the spontaneous actions of independent hackers motivated solely or primarily by national identity.

Recent examples of patriotic hacking include: Indian hackers attacking Pakistan in 2015 [194]; Pakistani hackers attacking Indian Universities in 2017 [73]; and tit-for-tat hacking between Armenian and Azerbaijani hackers over the past decade [8].

Cyber methods: Citizen-level hacking is generally unsophisticated, relying on brute-force penetration tests and 'script kiddie' attack methods. These attacks are often enabled less by attacker capability than by lax defense. The attacker's desire to harm any aspect of the target nation enable them to target both public and private organizations at any level, using a large array of cyber methods.

Cognitive methods: Patriotic hacks are indirectly spurred by broader propaganda and identity campaigns, which generally contain moral messaging and persuasive messaging. They are further encouraged by social pressure to demonstrate civic virtue and commitment, and to demonstrate superiority over the perceived enemy.

Impacts:

- **Agents and Communities:** Low. Targeted services are rarely offline for long. The effort of hacking a foreign target is unlikely to spur the formation of new hacking collectives, since existing collectives generally form with a declared national or trans-national identity.
- **Information and Knowledge:** Low. Services taken offline are generally restored quickly, preventing any major loss of information flow to the targeted populace. Hackers may use defacement to spread their own message, but it is rarely done in a persuasive manner.
- **Infrastructure and Channels:** Low. The lack of coordination and sophistication on the attackers' part makes these attacks opportunistic and limited in impact. Such attacks spur updates and improvements within the targeted organization but rarely drive full-scale reorganization or reconfiguration.

2.3.7 SickKids Hospital (2023)

In 2023, the LockBit cybercriminal group facilitated a ransomware attack against The Hospital for Sick Children (SickKids) in Toronto, Canada. Lockbit provides Ransomware as a service, developing and licensing encryption software to third-party malicious actors and splitting any

paid ransoms with those actors. The technical details of the SickKids hack, including the means of network compromise, have not been made public [131].

Nearly two weeks after the hack disrupted hospital operations, the LockBit group issued an apology and offered a free decryptor to the hospital. The group claimed that their internal code of ethics prohibited attacking healthcare institutions, that a rogue third-party licensee had conducted the attack, and that LockBit had terminated their affiliation with that party [240].

Cyber methods: The attack was conducted using a bulk encryptor developed and licensed by LockBit. The specific group/hacker responsible for compromising SickKids' system and installing the encryptor was not publicly revealed; given the attack surfaces typical to a healthcare organization, it likely mirrored common credential harvesting methods (phishing, spyware, etc.).

Cognitive methods: The ransomware attack was not explicitly enabled by any cognitive methods. If the attack were enabled via a successful phishing attempt, then the deceptive language employed in that phishing message would constitute a cognitive influence action.

Impacts:

- **Agents and Communities:** High. The attack gained immediate notoriety in the press and had much higher visibility than LockBit was prepared to accept, resulting in their apology and reversal. In the aftermath of the attack, healthcare organizations increased their awareness of, preparation for, and cooperation against, ransomware attacks.
- **Information and Knowledge:** Low. The encrypted data was not found to have been accessed by the attackers [57]. Further, the attack was not the first of its kind, nor was it especially novel in execution.
- **Infrastructure and Channels:** Low. The attack was identified fairly quickly and limited to only a few specific systems, allowing the hospital to continue with core operations [94]. The hospital announced a comprehensive security review to ensure all systems were properly updated and employees properly trained but did not announce any significant reconfiguration of its cyber presence.

2.3.8 Colonial Pipeline (2021)

On May 7, 2021, a ransomware attack against Colonial Pipeline resulted in the shutdown of a pipeline supplying 45% of fuel to the Eastern United States [168]. The impact was immediate and significant, creating major panic, outcry, and hostility within the US. The negative press was of such intensity that the attacking group, DarkSide, posted a statement claiming neither political backing nor motivation, and promising to vet future targets to minimize social disruption [204].

Despite this statement, only days later DarkSide announced it was ceasing operations due to pressure from, and the threat posed by, the US. The group acknowledged it had suffered a significant loss of funds and resources. It is possible that the group was simply rebranding to deflect scrutiny and may still be in operation under a new name [211].

Cyber methods: The initial attack used a compromised single-factor VPN password; how that credential was obtained is unknown. Likely the attack was enabled by standard credential-collection methods, such as phishing. The follow-on attack used a standard bulk-encryptor.

Cognitive methods: Apart from the possible use of phishing, DarkSide used no cognitive influence methods to facilitate its attack. The group did use persuasive methods and moral appeal in an attempt to blunt the US response.

Impacts:

- **Agents and Communities:** High. The disruption of fuel supplies wreaked havoc with the communities in the affected areas, and the resulting panic and confusion exacerbated the situation. US targeting forced DarkSide to disband and served as a deterrent against infrastructure attacks by other hacking collectives.
- **Information and Knowledge:** Medium. Apart from further publicizing the risk posed by ransomware attackers and other cyber criminals, the attack spurred significant discussion about public utility vulnerability.
- **Infrastructure and Channels:** Medium. Government oversight agencies, including the TSA and CISA, mandated new security requirements for pipeline operators in the wake of the attack. However, there is no indication that pertinent systems have been taken offline, or of any other significant cyber reconfiguration.

2.4 Analysis

These case studies provide ample evidence that existing information frameworks, because they are developed around only a single aspect of the information environment, fail to wholly capture interconnected effects at the operational level.

2.4.1 Cyber actions and cognitive effects

First, let us consider actions that are, on first examination, cyber actions intended to achieve cyber effects. Ostensibly these are 'traditional' cyber attacks and can be effectively understood and analyzed through any number of proven cyber frameworks. Two of cases studies are especially apt: the data breach of Discord.io, and the Chinese compromise of Gmail.

The Discord.io breach was a textbook instance of unauthorized access and data exfiltration with a purely financial motive. A simple MITRE ATT&CK mapping of this attack is given in Table 2.1 below. This ATT&CK analysis is rudimentary but could be expanded as more technical details became available. Even in this initial form it provides cyber professionals with insights regarding the attacker's capabilities; the target organization's detection and defense capabilities; and possible actions to remediate such attacks in the future, against this or other targets.

The day after the stolen data became available for purchase, the Discord.io team ceased operations. The company never provided reasoning for the cessation, but given the slew of

Table 2.1: Simple MITRE ATT&CK mapping of the Discord.io hack

Tactic	**ATT&CK ID**	**Notes**
Initial access	T1190 - Exploit public-facing application	The specific mechanism of compromise is not given, but reports indicate a flaw in discord.io's website code enabled the breach
Discovery	T1487 – Data from Database	Large quantities of database records were extracted
Collection	T1213 – Data from Information Repositories	The database itself was the primary target as well as the primary source
Exfiltration	T1020 – Automated Exfiltration	Attackers downloaded the entire database and transferred it offsite

lawsuits stemming from the MOVEit data breaches earlier that same year [128], the developers of such a niche product likely viewed dissolution as preferable to absorbing blame.

As such, the cyber action mapped via ATT&CK above does not fully capture the impact of the attack. Were attackers to glean working credentials from the stolen data, we could map subsequent attacks enabled through these credentials in a similar way. We cannot, however, leverage the ATT&CK framework to capture the financial impact of schemes leveraging this data as threats; nor the psychological damage of any users whose addresses may have been revealed using the stolen data; nor the dissolution of a small software developer and the resulting loss of service. Further, the DISARM framework is of no use in extending the ATT&CK analysis, as DISARM is focsued on deliberate information campaigns – actions designed to create Socio-Cognitive effects – and the company shuttering here, while certainly a result of a cognitive effect, was not the intent of the attackers. DISARM's second layer, "Plan Objectives," quickly reveals the inadequacy of the framework in predicting cognitive effects (while underscoring its utility in understanding efforts to induce them).

All of these follow-on effects can, however, be understood through a cognitive lens, by connecting affected users' beliefs and opinions to the circumstances effected by the breach. Individual beliefs that valid data represent valid threats enable scams. Grievances and disagreements precipitate doxing and other unwanted data revelations. Loss of user trust, and beliefs regarding blame and risk, lead to product abandonment and service discontinuation.

In similar fashion, I might analyze the Gmail breach through the ATT&CK framework (or any other cyber analytical methodology). However, because of the sophistication and breadth of the Aurora-linked attacks I forego such an analysis here, as it is outside the scope of this work. Instead, I consider the aspects of these attacks not captured by such an analysis: the cognitive and social effects.

As with any cyber attack, China's Gmail hack was executed in service of broader operational aims. The immediate goal of the attacker was access to specific named accounts; the achievement of this goal, and the steps and procedures that led to it, can be represented in the ATT&CK framework. However, the impact of these cyber actions on the *actual, broader* operational aim– the metrics associated with mission accomplishment – are wholly separate from the metrics and

measures used to analyze the success of the *attack* itself.

China's broader goal appears to have been controlling or deterring domestic dissent; ATT&CK offers no metrics or insights into affecting popular opinion. Cyber analysis frameworks can provide insight into the composition and conduct of cyber actions, including predictions of next actions or system effects; quantitative frameworks might offer success probabilities or time-to-recovery estimates. But no cyber framework can predict, quantitatively or even indirectly, how unauthorized account access might ultimately affect domestic resistance. Bridging that gap is left to the subjective expertise of operational leaders and planners.

As these examples illustrate, even seemingly anodyne cyber actions will, when considered at the operational level, have cognitive effects; and often, these higher-level cognitive effects are the ultimate *goal* of the cyber action. The inability of cyber-focused analysis methods to encompass these cognitive and societal effects imposes a disconnect on planners and leaders hoping to craft, or counter, such campaigns.

2.4.2 Cognitive and Cyber effects

Second, I consider actions that are, on first examination, cognitive actions intended to achieve cognitive effects. These would generally be seen as influence campaigns, and as such can be understood through influence analysis frameworks such as BEND, ABC(D)(E), or DISARM. Two cases studies are profitable examples for examination here: the John Oliver effect, and the patriotic hacker phenomena.

John Oliver's calls to action are rarely specific to cyberspace and, crucially, have never included a call to malicious cyber action. While he does frequently conclude his presentations by listing actions that interested parties could take, the preceding content is of primary interest. The 'John Oliver Effect' exists almost entirely because of his highly influential messaging. Whatever result Oliver hopes to achieve, he pursues through an influence campaign, seeking to precipitate action in a broad audience and thus achieve effects at scale.

In the specific instance of the Net Neutrality piece on *Last Week Tonight*, the BEND framework provides a quantized, digestible analysis of Oliver's influence efforts. Taking the presentation as a whole [4], Table 2.2 offers a BEND-framed analysis of the methods of influence Oliver employs to alter the information environment in a manner favorable to his desired outcome.

At the conclusion of the segment, Oliver does encourage viewers to online action – namely, commenting through the FCC's public-facing system. He even tacitly endorses aggressive or abusive language in those comments. He does not, however, call for FCC online operations to be hindered or disrupted, nor does he call for the FCC's comment system to be taken offline.

And yet, shortly after the segment aired, the server hosting the FCC comment portal did crash under an unprecedented surge of demand [116]. The loss of the FCC website, albeit for a short period, was a demonstrable shift in the information environment, nearly directly caused by Oliver's cognitive action. The BEND analysis of his message plainly demonstrates the potential for audience action, especially given the presence of Engage and Dismay maneuvers. And the BEND analysis provides insight into how Oliver's influence might spread to other audiences or reinforce itself over time: a Bridge maneuver indicates new influence connections will form within the environment, allowing opinions and beliefs to propagate in new directions. A Negate maneuver signals a declining influence for a group within the environment, a vacuum that will be

filled by other contenders. An Enhance maneuver means increased co-occurrence of the primary topic and other salient topics, broadening the conversation and exposing additional information consumers to Oliver's audience – and Oliver's influentially-presented position.

The BEND analysis fails, however, to account for the loss of an information source. The FCC website was, indisputably, an important information source in this discussion, and its unavailability altered the information environment in a significant way at a crucial moment; this in turn further affected the way Oliver's position and message propagated in the window when his audience was most active and thus most actively propagating his position. While BEND helps us understand *why* the affected audience is acting, and informs *to what end* they are acting, it cannot help us understand – or predict – *structural* changes that Oliver's influence has on the digital information environment.

Examining patriot hacking campaigns is significantly more involved. The influence campaigns that produce such action are much broader in scope and longer in duration, as they are generally nested in national-power-level geopolitical efforts. As such the analysis would encompass a long list of messages/transmissions, and a significantly larger audience and set of sub-communities. I exclude such analysis here as excessive for this work and focus instead on the disconnect between the governmental influential messaging and the instigated cyber attacks.

In all the examples examined (India, Pakistan, Armenia, Azerbaijan) the cyber attacks in question are indisputably illegal, both under international treaty law (to which the nations in question are signatories) and under the applicable domestic laws. Further, such open aggression originating within a competitor nation openly risks escalation. For these reasons, government influence campaigns almost certainly did not include open calls for harmful action, either physical or cyber. These campaigns certainly denigrated and vilified the opposing nation, but reason dictates that all stopped short of encouraging or endorsing openly hostile and illegal actions.

The magnification of this messaging within cultural groups, echo chambers, and other societal information network structures can be understood and partially predicted using frameworks such as BEND. (citation?) However, as in the previous case study, BEND cannot directly indicate which information services will be attacked, marginalized, amplified, or created by the affected audience in response to influence pressures. In the specific case of patriotic hacking, influenced hackers took it upon themselves to attack and alter the information environment in a way they deemed favorable to their own nation, in part to preserve and amplify the influencing factors that drove them to act. While BEND can help us understand the sources and form of this influence, it cannot provide us with insight into how that influence will alter the mechanical channels of influence.

2.4.3 A special case: Cyber action methodology driving cognitive effects

Previous cases provided examples in which cyber actions, intended to cyber cognitive effects, also resulted in cognitive effects within the affected populace. Some additional cases illustrate this phenomenon even further, demonstrating how the *means* of cyber action can alter the resulting cognitive effect – and thus the need for a joined framework. To illustrate this I examine the Russian anti-drone network shutdowns, and the Colonial Pipeline ransomware attack.

In an effort to blunt drone attacks, Russian leadership targets the control functionality of Ukrainian drones. Completely removing network connectivity is a simple, if brutish, cyber attack

against remote-operated systems like drones. Severing the connection is easily classified within ATT&CK, and the effects of disconnection can be quickly modeled in any number of cyber impact frameworks. Similarly, these frameworks accurately model and predict the failure of network-reliant systems *other* than the drones: taxis, airplanes, ATMs, and so forth.

What a cyber framework cannot demonstrate is the resulting shifts in public opinion, and the attendant behavioral changes, that such outages produce. The resurgence of SMS-driven systems is an excellent example of a cyber action-induced behavioral change. The cyber target – drone connectivity – could be attacked in any number of ways within the network, or with a more precision electromagnetic approach. Any such method would have offsetting drawbacks, ranging from difficulty in implementation to unreliable efficacy. Importantly, however, none of these alternative methods are likely to drive key societal systems onto an SMS framework. And the second-order impact of that change will surely emerge, as further observation examines a modern society relying on an SMS-driven digital framework, with the attending loss of capability, capacity, and security.

DarkSide's ransomware attack against crucial US infrastructure similarly shows interwoven cyber and cognitive effects. The available details of the ransomware attack against Colonial Pipeline do not indicate anything extraordinary about the cyber action taken, but the scope of the impact outside of cyberspace produced extraordinarily potent cognitive effects. Public outcry was so great that the perpetrators found themselves under concerted state-level cyber attack.

DarkSide had operated as a criminal collective for nearly a year prior to the Colonial Pipeline attack, and had successfully attacked and been paid for previous attacks, including against the US oil and gas infrastructure [196]. Despite – or perhaps because of – these previous successes, DarkSide did not foresee the rapid and extreme backlash that the Colonial Pipeline attack produced. Their cyber action produced a sharp cognitive effect, which in turn drove a powerful cyber effect directed against them.

2.4.4 Cognitive effects with structural impacts on cyberspace

As defined previously, cognitive effects are emotions or beliefs intentionally induced by an influencing agent. The influencer then reasonably hopes that the produced emotions and beliefs will translate into actions. I use the word *preferences* as a shorthand for habitual or repeated actions based on emotions and beliefs, and as such, I conclude that one possible achieved cognitive effect is a shift in target preferences.

Cyberspace is an entirely artificial construct and, as such, is highly responsive and adaptive to user behaviors. Services rapidly expand or contract based on shifting demand signals, at speeds unachievable in offline enterprises. Freed from requirements of real estate, shelf space, construction, and other physical demands, online business and services can become societal fixtures astonishingly quickly. TikTok is a recent example: at launch in 2017, the app "inherited" roughly 60 million users of an existing product that was merged into the short-form video platform. By the end of 2018, it boasted 600 million users and surpassed existing social media platforms in downloads; the next year, it became the fourth most downloaded non-gaming app in the world [187].

User behavior and preference is an important social phenomenon that directly shapes the structure of cyberspace; that structure, in turn, informs cyber capabilities. A user base is an

attack surface; scale informs security requirements; volume creates risk and opportunity. As previous examples have shown, frameworks useful for understanding the applications and limits of cyber capabilities do not give insight into the connected emotions, beliefs, and preferences of the users. As such cyber practitioners can undertake cyber actions that achieve cyber effects far beyond their intended scope, as the initial effects alter user preferences, which in turn alter the fabric of the cyber dimension. Two of our case studies particularly demonstrate the effect of preference on cyber structure: the Twitter user migration, and Russia's anti-drone shutdowns.

Twitter user migrations are almost entirely cognitive effects. While platform policy changes are implemented in code, they qualify as a cyber action (by our definition) in only the most technical sense. The driving motivation behind users leaving the platform is not the immediate impact of code changes, but the broader shifts in platform composition – and, more often, the attending social dialogue, irrespective of the actual platform content or behavior. Nevertheless, this case illustrates a precipitated cyber-structural shift. Twitter implements a technical change by altering its moderation policies or feed algorithm; this change produces cognitive effects in the user base, instilling feelings of frustration or alarm and reducing user confidence in and preference for Twitter as a platform. These altered preferences then drive users to seek alternatives, and competing platforms like BlueSky see a sudden surge in users as a result.

La Cava et al performed an excellent analysis of the 2023 Twitter user "migration" phenomena, applying information diffusion and social influence models to explain the users' collective decision to switch platforms [137]. While such an analysis can provide quantitative insight into likely user movements, the analysis cannot directly inform the important attending technical considerations. To predict the effect such migration pushes would have on the technical performance of the two systems involved, or on their respective system security, an analyst would have to invoke separate, cyber-oriented frameworks and conduct a second analysis.

As mentioned, cyber-oriented analysis of Russia's rolling network shutdowns cannot model the resulting shifts in user behavior. But neither can cognitive-oriented analysis of the populace's shifting preferences, and the resulting emergence of an SMS-based alternative to the mobile internet, capture the cyber issues that will emerge. SMS is a comparatively insecure communications system, creating new challenges for cyber security and new opportunities for attackers. The protocol is point-to-point and offers significantly reduced bandwidth, forcing service providers to adapt their mobile offerings, which will further shape user preferences. So long as users do not have confidence that a 4G/5G mobile network will be available to them, they will behaviorally prefer SMS-based systems; so long as businesses to not have confidence that the network will be available, they will delay investments in 4G/5G capabilities and platforms. The resulting cyber environment is a significant departure from the society's starting point, and the deviation is largely due to the implemented anti-drone measures.

2.4.5 Cyber effects on the cognitive landscape

As cognitive effects (preferences) can alter the cyber structure of the information environment, so too can cyber effects alter the cognitive structure of the environment. Cyberspace impacts the form and function of human social networks – whom we communicate with, what we communicate, and to what extent that communication influences our own beliefs and emotions. This transformative effect is studied in detail in the next chapter.

2.5 Conclusion

The information environment is a crucial component of society, because it drives decision-making and behavior. Leaders, planners, and protectors of any large group need to understand the information environment in which they are operating if they are to effectively direct their group as a cohesive entity.

The information environment could be conceived of as two separate dimensions, one human and the other technical. This is an appealing distinction, as human and technical systems often function in radically different ways. Study of each of these areas has produced dependable and proven analytical frameworks. However, as I have demonstrated, these two dimensions have a high degree of mutual influence; and, as demonstrated, frameworks designed to analyze either cyber or cognitive systems cannot be extended to encompass the other system. The disconnect is similar to the longstanding gulf in physics between quantum theory and general relativity: both are valid descriptions of physics, and both excel in their respective areas, but neither can fully extend to encompass the other.

A united social-cyber analysis framework such as OPIEM would be invaluable to operational and societal leaders and planners. Such a framework can directly model the interplay between the cyber and cognitive dimensions of the IE. It can link social and cognitive effects to cyber actions, and cyber effects to cognitive actions. It can help leaders anticipate how cyber actions and effects may alter the information environment holistically, and conversely how cognitive effects and actions – including societal trends and shifts in preference – may later cyberspace specifically, and the environment as a whole.

Table 2.2: BEND analysis of the net neutrality Last Week Tonight segment

BEND Maneuver	**Maneuver definition**	**Instance in message**
Explain	Discussion or actions that clarify a topic to the targeted community or actor often by providing details on, or elaborations on, the topic	Oliver frequently explains technical and legal issues in layman's terms and provides clarifying examples
Dismay	Discussion or actions related to a community or topic that cause the reader to experience a negative emotion such as worry, sadness, dislike, anger, despair, or fear	Oliver invokes possible negative outcomes of the debate and offers previous negative incidents of abuse or impropriety
Bridge	Discussion or actions that build a connection between two or more groups or create the appearance of such a connection	Oliver specifically cites aligned interests between large tech corporations and activists, implying that viewers aligned with either of these dissimilar viewpoints have common ground
Engage	Discussion or actions that increase the relevance of the topic to the reader often by providing anecdotes or enabling direct participation and so suggesting that the reader can impact the topic or will be impacted by it	Oliver issues a call to action, specifying the means and implying the desired action to be taken, and implying that such action can result in change
Enhance	Discussion or actions that provide material that expands the scope of the topic for the targeted community or actor often by making the topic the master topic to which other topics are linked	Oliver connects the issue of net neutrality with the broader issue of speech freedom, and with the more menacing issue of corporate monopolies
Negate	Discussion or actions that decrease the actual, or the appearance of, an actor's importance or effectiveness relative to a community or topic	Oliver's presentation minimizes and dismisses the positions and arguments of telecom companies and internet providers

Chapter 3

Influence through Information Availability

This chapter examines influence exerted through the manipulation of information availability. As seen in the previous chapter, influence can be achieved in part by affecting the ability of Agents to receive or produce information, without respect to content. As explored in chapter 1 and chapter 2, existing influence frameworks do not sufficiently model information availability, especially through technical (or non-social) channels; and cyber frameworks do not sufficiently model the decision influence that changes in availability exert. In this chapter, I propose a categorical taxonomy for availability-based influence actions, as a set of operational-level "maneuvers." These maneuvers are intended as extensions to the BEND lexicon for influence.

As previously discussed, information has a strong causal relationship with human behavior [67]. The Information Age has created a dizzying array of new opportunities for actors, both benign and malicious, to leverage that relationship in service of broader agendas. The potential and realized impacts of wide-scale online influence efforts are well known [42], and multiple academic frameworks have emerged to address the mechanics of large-scale influence campaigns, as well as potential interventions [11] [72] [89]. At the same time, detailed models of the Internet's technical function – that is, how digital information propagates across wide-scale networks – have also emerged [21] [225].

These modeling approaches diverge significantly, structurally and methodologically, because they have clearly distinct target phenomena. However, they are in fact modeling separate components of a single continuous information spectrum. Digital information models capture the scaled availability of information, representing how information spreads and persists within the internet. Information availability is a significant driver of human behavior [152] [245], and thus is itself a major influencer, in addition to the actual content of ingested information. In short, what people read is important in two ways: the words they read, and whether they actually read them at all.

It is thus no surprise that large-scale influence campaigns include not only coordinated and targeted messaging, but deliberate attempts to alter information availability through technical means [19] [156]. I examined historical incidences of widespread, large-scale social influence campaigns effected at least in part through information availability alteration. Notably, this did not include strictly technical means (e.g. traditional cyber actions); it also included non-cyber attempts to alter availability.

I categorized the availability alteration actions into a set of 'maneuvers,' based on the in-

tended effect of the altering actor. These maneuvers are designed as an extension of the BEND framework, expanding the BEND methodology into a trimodal space. Given the metrics defined previously, the combined maneuvers create a model sufficient for the modern IE. By highlighting the commonalities and, more importantly, the desired outcomes of these actions, this lexicon of influence maneuvers equips policy makers and information platform operators to identify influence attempts that may be disadvantageous to their served populace, and to direct influence activities within their organizations.

3.1 Data

I considered the following case studies of social influence that include, at least in part, deliberate efforts to alter information availability in order to affect decision-making in a target populace.

3.1.1 Marketing, publication, and public relations (generally)

Before delving into specific incidents, I must acknowledge that scaled influence through availability is a well-known and professionally honed practice. The field and practice of marketing is in large part the formalized application of such influence, though it is not limited solely to availability concerns. Generally, marketers target economic decisions: how consumers make purchases and spend money.

Marketers strive to make consumers aware of their advertised product [126], and to increase consumer exposure (measured in frequency and duration) to such advertisements [81]. Through strategic ad purchasing and exclusivity agreements, marketers also attempt to reduce the visibility of competing or alternative products [79].

Additionally, through audience segmentation, targeting practices, marketers seek to increase communications influence over targeted demographics through market positioning [161]. Finally, in the special case of businesses operating or providing communications services, marketers emphasize the qualities and advantages of their own platform, while highlighting the drawbacks or failings of competing services [26].

Public relations, as a discipline, similarly deals with information availability as a complement to its concern with information interpretation. PR experts seek to "control exposure" or "contain details;" such jargon speaks to their recognition that the availability of information, as a simple binary, is an important part of influencing the ultimate conclusions of the audience.

3.1.2 State-level information control schemes

Information policy is a fundamental pillar of statecraft and has been for centuries [85]. Information availability is an obvious and frequently-utilized lever in state efforts to shape domestic decision making. In the Information Age, availability control efforts aimed at the internet represent easily identifiable instances of scaled information influence campaigns. Here I consider several examples of state-led control schemes.

China

The Chinese Communist Party (CCP) operates a massive technically advanced firewall in an effort to limit internet content availability to the broader Chinese populace [229]. China's firewall is particularly notable both for its scale, and the extent to which it has been academically studied. The firewall operates as part of a broader influence campaign to influence civic decision making, in an attempt to reduce civil unrest, anti-government organization and action, and other decisions disruptive to the CCP's broader agenda.

The Great Firewall functions on multiple layers, all focused on denial. It specifically blocks censored content within unencrypted TCP connections, as well as blanket-blocking specified IP blocks. It restricts or tampers with DNS resolution calls to further enforce domain-level bans. The result is that many internet communications platforms are wholly unavailable to the Chinese public, or else operate in such a degraded mode as to be undesirable or impractical.

In concert with the Firewall's technical operation, CCP personnel actively monitor and moderate the broader Chinese internet [33]. These efforts collectively serve to greatly reduce the availability of countermanded content, and in some cases to render it completely unavailable. These personnel are also encouraged, and often mandated, to generate and post content that is supportive of CCP objectives, critical of dissenting viewpoints, or intentionally ambiguous or uncertain on topics that might serve as rallying points. These actions increase the incidence of CCP-approved narratives on the internet, relative to other viewpoints, and conversely water down or bury concerning narratives.

North Korea

North Korea has an extensive regulatory framework that strictly controls how citizens access information, in line with the regime's belief that information control is essential to its survival [96]. These regulations precede digital infrastructure: North Korea heavily regulates and monitors all print and broadcast media, allowing only state-sponsored or state-run outlets to function [133]. The country's law criminalizes the production, acquisition, or distribution of 'indecent materials', a broadly defined term that the state can leverage to include any information counterproductive to the state's agenda.

North Korea strictly controls available information channels by requiring state level registration. Operating licenses are regularly reviewed and are granted based in part on compliance with content control requirements. The registration requirement was originally aimed at traditional mass media channels – printed materials and broadcast media. North Korea has accommodated the internet within a similar legal framework: "computer networks" are tacitly treated as state-provided assets and require user registration for any use. Such networks are under the same criminal codes regarding 'indecent materials' as other media.

The North Korean government does provide information to the populace: the state operates or sanctions multiple newspapers, magazines, and television and radio stations. These outlets often present foreign movies and documentaries, albeit after heavy censoring. Similarly, the state operates public libraries, which offer a curated selection of information. The state operates a mobile phone network and limited country-wide intranet, through which it distributes approved information and permits regulated and monitored information exchange.

Iran

Similar to North Korea, the Iranian government views internet-borne threats as existential problems. Such threats include democracy promotion, cyberattacks, and commercial sanctions targeting or affecting online business [157]. Although not to the same extent as North Korea, Iran similarly monitors and regulates print and broadcast media; and, for similar reasons, Iran routinely blocks or throttles platforms and channels that are perceived as problematic or uncontrollable [14].

Iran's approach to information control varies from North Korea's insofar as their geopolitical positions differ. North Korea is a pariah state, isolated within its region and heavily reliant on China as its single patron state [192]. While Iran is also a pariah within its region, the country's oil wealth has given it greater economic sway and has enabled Iran to maintain an active role in broader geopolitics [99] [**?**]. Further, Iran's theocratic message resonates with like-minded religious groups outside of its borders, enabling the nation to generate and control significant proxy forces abroad [140]. North Korea's communist principles have not produced any such diffuse influence.

The need to participate in the global economy, and to remain connected with entities outside of its own borders, preclude Iran from severing its ties to the larger internet as wholly as North Korea does. While the country has still created a largely parallel "internal internet" to enable greater monitoring and independence [157], Iranian citizens have far greater access to foreign online assets than their North Korean counterparts [241]. Iran's use of throttling in lieu of blocking is a telling sign that the regime is actively balancing the demands of information control with a desire to retain international legitimacy.

Russia

Russian (and the preceding Soviet) strategists have historically given great attention to influence operations, often achieving through indirect means what they lacked the resources or capability to achieve directly. The present-day Russian information control scheme reflects this history of deliberate and nuanced control [146]. Since the invasion of Ukraine in 2022, Russia has increasingly been a pariah state globally, and information conditions within the country have worsened as a result; however, even before the war began, the Russian state was actively involved in shaping its domestic information environment [109].

Russian media exists under a long tradition of enforced censorship, to the point that self-censorship is a common practice. The resulting hesitance to generate and distribute information creates a gap that the state fills with state-operated media outlets, providing a constant stream of carefully curated information [175], one that often actively steers consumers away from alternatives [173].

To deal with the internet and the resulting exposure to external information, Russia has created a system significantly different from China's active denial approach. The Russian information control scheme is less overt in its action, based largely in a legalistic non-technical approach, leveraging the fear of state enforcement to shape opinion without openly undermining the Russian government's international legitimacy [130].

The technical portion of this approach lies less in viewing and filtering content, as in the Great

Firewall, than it does in actively surveilling the citizenry; Russian technical censorship is more distributed and targeted [82]. In addition to blocking or denigrating foreign platforms, Russia is quick to promote internally-developed alternatives. These systems are ostensibly privately owned and operated, offering parity to western services, but because they operate domestically – in the Russian language, with Russian employees and often Russian-based infrastructure – the state wields significantly more offline influence, through legal and extralegal means, over how these platforms operate [125] [181].

3.1.3 "Fake News" media challenges

Accusations of falsehood and bias against news and media outlets are not new but have seen a marked increase in the past decade, especially within political contexts. Since 2015, the audience for cable news increased dramatically [169]. However, media outlets prominently accused of being "fake news" or "fake media" enjoyed much less of an increase, and were far less profitable, than outlets promoted as more veracious by the same critics [169].

While not technical in nature, attacks against specific media outlets – both categorically, as well as by name – are intended to alter not only audience perception of information produced by those platforms, but audience platform preference. 'Fake news' accusations are in part attempts to favor certain outlets over others, usually outlets where the accuser enjoys greater influence over content. As such, the accusations praise the virtues of favored platforms, while decrying the flaws and purported transgressions of alternatives.

3.1.4 Elon Musk and Twitter

The acquisition and operation of social media platform Twitter by private businessman Elon Musk is a significant and highly visible example of information-based influence, with significant examples of actions meant to chant access to, or availability of, information.

Acquisition and policy changes

Elon Musk had been an avid user of Twitter for years, but in 2022, he became openly critical of the platform's function and policies [184]. His criticism was aimed largely at Twitter's moderation and content control policies, which he stated infringed upon free speech and free expression principles [78].

By the end of 2022, Musk had purchased Twitter and taken the company private, quickly rebranding the platform as X and radically altering its moderation and access policies. In line with his purported commitment to free expression, he restored many accounts previously banned due to moderation violations [37]. However, he also directed the suspension of accounts that opposed his personal politics or those of his funding partners and advisors [149], as well as accounts that attacked or mocked him personally [164], moves at odds with his stated goal of a completely unmoderated expressive space. Backlash from these and other seemingly arbitrary platform decisions led to Musk eventually stepping down as CEO [23] [200], though he retains primary ownership of the company.

Algorithm "tweaking"

A particularly egregious departure from the ideal of unmoderated speech occurred in February 2023, when US President Joe Biden and Musk posted similar messages to the platform within a short time window. Biden's message received more than twice the engagement than Musk's. Shortly afterward, Musk ordered X engineers to alter the algorithm, ensuring he would personally benefit and receive greater engagement [208]. As a result, Musk – and the accounts he personally followed, amplified, or endorsed – saw significant increases in exposure to all users across the platform [25].

The X recommendation algorithm is a significant determinant of the information presented by the platform, and therefore the information most readily available to users. The modifications made to the algorithm therefore resulted in much greater personal influence for Musk, and through him, greater influence for those he chose to amplify [103]. It also provided new influential access to multiple users who had previously been restricted from the platform, including state-controlled accounts previously blocked or throttled by Twitter [127].

Grok

Musk's ownership of Twitter/X, and his position as CEO of AI company xAI, led to the development of Grok, a large-language model (LLM) AI chatbot. Grok specifically includes X posts as training data [158], leading to responses that are heavily shaped by the content on the social media platform. As X's policies have caused content to shift politically, Grok's responses have similarly shifted; in some cases, xAI has directly altered Grok's reasoning framework to induce or prevent certain responses [242].

Crucially, Grok is integrated into X, and is advertised to users as offering intelligent summaries of the huge number of X posts on the platform. in this way, X offers Grok as a means of quickly ingesting human-generated X content without directly viewing that content. However, Grok is demonstrated to have specific preferences, tendencies, and leanings [233], which potentially lead to incomplete, biased, or inaccurate summaries. In other words, content summaries generated by Grok may not accurately represent the content or intent of the authors.

Grok's integration and ready availability within X, coupled with its privileged access to X data for training and sourcing, make it an appealing tool for X users, but raise significant privacy and reliability concerns. By construction, however, Grok likely amplifies existing imbalances on the platform for some users; for many others, paradoxically, Grok seems to induce wariness and distrust, and produces increased fact-checking [203].

3.1.5 Alternative/Responsive Social Media Platforms

The widespread and broadly participative nature of internet communications has challenged existing freedom of speech frameworks since the beginning of the digital age [22]. Moderation policies and systems that have emerged in response to concerns about public safety or interest have invariably yielded users who feel disenfranchised or unduly censored. The relatively low cost of entry to digital communications, coupled with frustrations with existing platforms, has often resulted in the creation of platforms that explicitly exist to counterbalance current services.

I examine some of the more prominent examples of platforms founded with a partial or specifically reactive motive.

Parler

Parler was launched in 2018, marketed as an unbiased alternative to existing mainstream social networking platforms [207]. The service claims to be a place for "free expression without violence and a lack of censorship," which attracted segments of the population who felt disenfranchised by existing platforms' moderation policies. Despite a stated goal of non-censorship, in practice Parler did have extensive obscenity-based policies censoring user-generated content [46] [139].

Parler launched with little fanfare but saw surges of interest corresponding with endorsements by key personalities and influencers [210]. Though founders had envisioned a neutral nonpartisan space, the company eventually began catering to specific groups expressing dissatisfaction with alternative social media sites; Parler experienced significant growth in June 2019, when it welcomed accounts from Saudi Arabian users who claimed to have been censored by Twitter [66]. Parler enjoyed targeted endorsements from prominent political figures who amplified concerns and accusations of censorship by larger platforms [141] [252].

The service had a major influx of new users in mid-2020, when central platforms Twitter and Facebook implemented new anti-misinformation policies related to the ongoing US election; in response, users feeling specifically targeted or upset with the seemingly arbitrary nature of the restrictions embraced Parler as a more permissive alternative [66] [252]. It enjoyed another burst of popularity later that year as election results led to further content bans [206] [234], and again in the aftermath of the January 6 US Capitol attack [90] [216].

Parler users' involvement in the attack, and the service's use in organizing the attack [90], led to the service losing its infrastructure providers and ultimately going offline [180]. Users responded by decamping to other alternative social media sites [129] [247]. Parler's attempts to mitigate the fallout were further complicated when data scraped from Parler was used to implicate and prosecute users involved in the Capitol attack [101]. The service never recovered, and its parent company eventually filed for bankruptcy [250].

Truth Social

Truth Social was launched in February 2022 as a Twitter/X alternative, created directly in response to account bans stemming from the attack on the US Capitol of 6 January 2021 [205]. Truth Social styles itself a "free-speech"-oriented alternative to larger platforms, though in practice it caters heavily to politically conservative users and viewpoints, going so far as to quietly censor opposing content [118] [221]. This includes many users who had been deplatformed or otherwise limited on other services [185].

Truth Social's launch focused not only on its permissive culture and environment, but on the prominent political and cultural figures who had committed to using the platform [121]. As a result, the service saw a surge of users upon launch, though technical issues stymied further growth [215]. Within two months, the service had seen a 90+% decrease in new subscriptions

[218]. The platform has seen event-driven growth, such as in August 2022 in the aftermath of the FBI search of the Mar-a-Lago resort [43].

Truth Social has been the target of significant criticism and skepticism within prominent media and technical media outlets [36] [60] [107] [215] [223]. Given Truth Social's minuscule userbase compared to competing platforms [80], it is likely this criticism has deterred potential users, limiting the Truth Social userbase to users following the prominent influencers already on the platform and likely reducing those influencers' ability to extend their audience on the platform.

BlueSky

The BlueSky social network was developed internally by Twitter as an investigation into decentralized architectures [182] to better accommodate individual user preferences [58]. As the project matured, it was separated from Twitter to ensure it was not beholden to Twitter's "existing incentives" for operations [102]. Elon Musk's acquisition of Twitter formalized the separation between the two networks, and BlueSky launched as an alternative service in February 2023.

The service attracted many marginalized groups through its invitation-based enrollment system [142]. The early user base had a significant influence on the platform's development, including the implementation and specifics of moderation policies [199] [227]. In contrast to other alternative platforms, BlueSky did not advocate for "free speech" or zero censorship, instead favoring highly customizable and individualized content control.

Since its public launch in February 2024, BlueSky has experienced multiple periods of rapid user increase [51]. These shifts are mainly linked to controversies and events on Twitter/X, the platform's primary competitor. BlueSky staff have dubbed such growth bursts "Elon Musk Events (EMEs)" [138]. Musk's political activities, and resulting actions taken by or toward the platform he controls, have affected BlueSky significantly.

At this stage BlueSky's success remains to be determined. The platform distinguishes itself from alternatives in its embrace of content moderation, avoiding the controversy of the issue by effecting it through individual users instead of platform-level, company-controlled algorithms. Regardless of this approach, BlueSky appears to suffer from an ideological lean similar to other non-mainstream services I have analyzed [224].

Threads

Meta's alternative to Twitter/X, Threads, launched in 2023. The service was developed in part as a response to Elon Musk's acquisition of Twitter and resulting discontent within the user base and advertising ecosystem [117]. Bolstered by Meta's backing and its native integration with Instagram, an existing and highly popular social media service, upon launch Threads became the fastest-growing platform in history to date [123]. However, Threads failed to maintain high user engagement, and its user base has atrophied significantly since the initial surge [104]. As with BlueSky, Threads is a relatively new service, and its ultimate success remains in question.

3.1.6 Telegram

Telegram is a cloud-based service that has social networking features. I differentiate it from social media platforms (as in the previous case studies) in that Telegram was initially conceived as a messaging service that later took on wider social capabilities. Unlike social media services, which generally seek to optimize exposure before considering privacy and security, Telegram by design optimizes privacy above all other user concerns [217]. Despite prioritizing point-to-point security, Telegram is an immensely popular social communication app [56] [155].

Telegram's emphasis on privacy makes it appealing for users who seek a social media platform that gives them greater control over their own visibility, even at the cost of broader exposure. As such, Telegram is often the platform of choice for niche communities that grow by invitation and value contribution and expertise, rather than influencer-type communities that contend for attention [97].

Telegram's operators and owners frequently tout the platform as a safe alternative to more politically susceptible platforms [38]. To reinforce that reputation, Telegram routinely refuses to cooperate with law enforcement, which has provided problematic entities a platform from which to organize and extend their influence [97] [150] [232]. Notably, because such entities favor and rely on Telegram so heavily, any actions the platform does take against them are especially effective in reducing their influence [213].

3.1.7 Andrew Tate's Hustler's University

Online influencer Andrew Tate relies on social media platforms as his "main sources of revenue" [249]. Initially he marketed and distributed his training materials via platforms including Facebook, Instagram, Youtube, and TikTok. The content of his materials, however, led to him being banned from each of those platforms in succession [68] [249].

Tate founded an online distribution system dubbed "Hustler's University," a paywalled collection of prerecorded videos and a private Discord server [68]. Deprived of the direct exposure afforded by social media, Tate instituted a marketing initiative within the established Hustler's University user base. He offered commissions to existing followers for signing up new subscribers [231], resulting in a huge boost to his personal followership (prior to his ban) and a commensurate increase in his content's pervasiveness on various platforms [69] [231].

Members of the private Hustler's University, along with copycat accounts, flooded platforms with videos of Tate and samples of his material, in what experts called a "blatant attempt to manipulate the algorithm" driving the platform's feed generation [69]. Tate himself often selected specific clips for his followers to amplify and repost on his behalf, choosing controversial clips to maximize follow-on engagement [69].

Alarmed by these overt attempts to control their platforms, multiple social media services banned Tate in 2022. In response, Tate transitioned his attention to more permissive platforms, including Gettr and Rumbl; his move to each platform resulted in significant growth in their respective user bases as his followers joined [15] [98]. Tate has been directly responsible for record-setting engagement within the Rumbl platform specifically [124].

3.1.8 Alex Jones and Infowars

From 1999, Alex Jones operated the Infowars website; by 2016, the site had become a prominent purveyor of alternate news, conspiracy theories, and misinformation, with traffic rivaling established mainstream media outlets [24]. The website was coupled with The Alex Jones Show, which Jones produced in his own studio and distributed via internet to syndicated partners. By 2016, technological advances permitted the show to be live-streamed through syndicated stations and via social media platforms [198].

Jones' coverage and interpretation of the Parkland School shooting in February 2018 was highly controversial, and much of his content was censored or removed by hosting social media platforms [165]. Jones continued producing content in a similar vein, and over the next year most major distribution services and platforms banned his and Infowars' accounts [63] [112] [148] [178] [254] [255].

Jones' activities brought numerous legal suits against him, and legal damages compounded his liabilities. His broad deplatforming simultaneously curtailed his ability to generate revenue consistently [171]. As a result, Jones and his companies have filed for bankruptcy, though he continues to broadcast through alternative channels provided by supporters, or through services available to the broader public [62].

3.1.9 The #MeToo Movement

The #MeToo movement began in 2006 on the social media network Myspace, though it was relatively unknown until celebrity actions on Twitter created broader public awareness [74] [176]. The movement addresses a broad and complex social dynamic; of interest in this work is the implication and accusation that powerful or influential men have been able to suppress reporting of their sexual misconduct, thus insulating themselves from consequence [**?**] [236].

In a specific and well-documented instance, entertainment mogul Harvey Weinstein was accused of sexual misconduct going back three decades by over 80 women. Weinstein's position and control over desired acting parts, combined with strong ties to powerful and influential public figures, enabled him to enforce silence through quid pro quo and effectively keep his behavior out of the broader public eye [236]. That his behavior appeared as at least a punchline demonstrates it was a widely known 'secret' [40]. Even as accusations surfaced and made the news, Weinstein took extralegal and indirect action to suppress information availability or to discredit extant information [84].

3.1.10 ProPublica

ProPublica is a nonprofit organization specializing in investigative journalism. As part of their charter to maximize journalistic impact, their reporters often seek out stories that are undercovered or entirely uncovered, thus introducing those issues to, and/or elevating those topics within, the public awareness [**?**].

3.2 Analysis

Review of the cases above identifies a pattern of macro-level efforts or tactics, aimed at information availability and used as part of influence attempts. I identified five measurable features of the IE that serve as the targets of availability-based influence efforts. Influencers attempt to assert influence through actions that seek to increase or decrease the measured values of these target features. I designate each such effort a *maneuver* and thus propose a total of 10 availability-oriented influence maneuvers, as described below.

3.2.1 Target 1: Channel availability (binary)

Direct control over communication channels obviously confers significant influence. Actors who are able to determine which channels are active and which are not, or who have the capability to produce new channels better suited to their needs or agendas, have significant leverage in the composition of the broader information environment, and therefore in the opinions and beliefs that the wider audience derive from it.

Roll-out

The Roll-out maneuver comprises actions that **create** a new information channel or platform, or that **expand availability** of the same, specifically with intent to influence. Some new platforms are created without particular influence aims; BlueSky and Telegram are examples of services initially conceived more around specific features and architecture, rather than content and audience. By contrast, Truth Social and Parler were both created in large part as reactions to the perceived content and demographics of existing services. The platform creators sought to generate an environment wherein specific supposedly disadvantaged groups or viewpoints would be favored or advanced [205] [207].

Roll-out maneuvers can and do occur at the sub-platform level. Users regularly create new channels or forums within existing platforms, such as Telegram or Reddit, to provide communications spaces for specific topics or viewpoints; such channels are often created in opposition to, or as breakoffs from, larger existing channels [183].

Shutdown

The Shutdown maneuver comprises actions that **remove** an information channel or platform, or that **reduce availability** of the same. Closing entire communications channels is a tried-and-true method of suppressing specific viewpoints by depriving groups the ability to coordinate and mutually reinforce. State-level blocking of US- or Western-based social media platforms is an example of a Shutdown maneuver: by "Shutting down" Facebook, Twitter, and YouTube, states like China, Russia, and Iran reduce the influx of western ideas antithetical to their respective state agendas and remove unsupervised or unmonitored spaces where opposition groups might organize and coordinate activities.

Note that Shutdown maneuvers focus on channels, not individuals, attacking infrastructure rather than users. Further, Shutdown maneuvers are conducted deliberately with the intent to

influence, changing behavior by altering information availability. So, the shuttering of the website Backpage by the US Government with the intent of reducing prostitution is an example of a Shutdown maneuver [32]; the eventual closer of Parler for financial reasons is not an example.

3.2.2 Target 2: Channel Prominence (proportional)

Short of fully instantiating or removing a communications channel, influencers can seek to drive users toward or away from existing channels within the environment. Such an action is much less resource-intensive, while still granting (or removing) significant influence to the groups that dominate the targeted channels.

Recommend

The Recommend maneuver comprises actions taken to **increase** a channel's **relative use or traffic** within the information environment, as compared with all other pertinent channels. Influencers seeking to alter audience behavior will seek to engage that audience in a permissive or advantageous setting, seeking to move to platforms that favor the influencer's message or over which the influencer exerts significant control. The simplest and most granular form of the Recommend maneuver is companies and individuals encouraging others to follow them on social media, or to subscribe to their channel; in doing so, the viewer shifts more of their budgeted attention toward an information conduit that favors the content producer.

Like-and-follow campaigns are well-worn marketing practices, for major companies and individuals alike. Recommend maneuvers also occur at broader scales. Modern political entities endorse specific news outlets that they feel are more favorable to their viewpoints, hoping to drive viewership toward those outlets. Andrew Tate promoted the platforms Rumbl and Gettr, resulting in significant user increases; similar gains were seen as right-wing influencers joined and endorsed Parler [15] [124] [210].

Sideline

The Sideline maneuver is the inverse of Recommend, comprising actions that **decrease** a channel's **relative use of traffic**. Of necessity, any Recommend action necessarily Sidelines the services from which the promoted channel draws users, and conversely, any Sideline action will create a tacit Recommend effect on all other channels. The two are differentiated, then, by the target of the influencer, and the influencer's intent.

The Fake News case study provides an excellent example of the nested nature of these maneuvers: accusers frequently Sideline specific services and, in the same breath, Recommend alternatives. Social media platform proponents frequently Sideline competitors by claiming bias or lax security. Telegram proponents, specifically, Sideline mainstream platforms by emphasizing user exposure to monitoring [155]. One high-level example of Sidelining is internet filtering achieved through rate throttling, as seen in Iran and Russia: by throttling traffic to foreign-based platforms such as Youtube and Twitter, these states drive traffic away Western-aligned services and toward domestic alternatives such as VKontactye [82] [125] [181].

3.2.3 Target 3: Channel access

Short of creating new channels or deactivating existing ones, influencers can seek to provide or restrict channel access to specific individuals, or to gain or lose such themselves. The presence or absence of specific individuals is a major contributing factor to the information makeup within a channel, and thus removing or adding voices is an explicitly availability-based action that influences all channel viewers.

Raise

The Raise maneuver comprises actions that **provide** or **expand** an individual's **control of**, or **reach within**, a given information channel or platform. If a channel represents its connected audience, then the Raise maneuver is any action that gives the targeted individual greater direct contact with, and therefore influence over, the targeted audience.

Elon Musk's modification of the Twitter algorithm to favor his own posts is an obvious example of a Raise maneuver, targeting himself as the beneficiary and the Twitter user base as the affected audience [208]. Twitter's policy changes granting renewed access to previously banned accounts is also a Raise maneuver, in favor of those accounts; by providing them access, Twitter adds their viewpoints to the available information and thus influences the intake – and resulting decisions – of viewers [25]. Andrew Tate's distributed marketing campaign is another example: by directing his followers to spam multiple platforms (though TikTok in particular) with certain content, Tate was able to Raise his profile within those platforms rapidly, vastly increasing the users exposed to his content and therefore his potential influence on the entire user base [230].

Silence

The Silence maneuver comprises actions that **limit** or **remove** an individual's **control of**, or **reach within**, a given information channel or platform; it is the inverse of the Raise maneuver. Removing specific information sources is an obvious way to influence that source's potential audience: unheard information cannot influence beliefs and opinions, and thus the ability to remove or limit communications access for individuals is the de facto ability to reshape the information environment and exert influence over the people connected to it.

State-level censorship frequently includes Silence maneuvers, as dissenting accounts are suppressed on monitored platforms and dissenting authors are arrested and silenced outside of the digital world; likewise any account ban undertaken by a social media platform is a Silence maneuver. As a specific example, Alex Jones' deplatforming from platform after platform clearly demonstrates a series of Silence maneuvers that ultimately reduced the prevalence of Jones' voice within the digital environment [63] [112] [148] [178] [254] [255].

3.2.4 Target 4: Information presence (binary)

Beyond targeting information infrastructure, influencers can target information itself, exerting influence by altering what information is available to the audience. I am careful to note the difference between these proposed maneuvers and more subtle content-based maneuvers that speak to crafting narratives, interpretations, and other influence mechanisms within a message.

Instead, I limit the discussion to the absolute presence of information. In this instance, I am concerned with whether or not a topic is discussed or an item is known, and not with the resulting divergent interpretations, discussions, or conclusions that such knowledge invariably produces. These issues are significant factors in generating and exerting influence, but are better considered under existing frameworks such as Carnegie Mellon's BEND [49].

Reveal

The Reveal maneuver comprises actions that **add** a topic to the information environment, or to a specific channel therein, that was not previously present. This seems, on its face, overly simplistic. However, at the channel-level specifically, successfully introducing a new topic of consideration to a communicating community represents a significantly influential action. Marketers have long recognized this fact, and the fields of marketing and public relations have long practiced product rollouts, press releases, publications, and other high-visibility introductory actions designed to commandeer attention within existing communications spaces.

The work of ProPublica is another example of Reveal maneuvers. The nonprofit frequently seeks out news items of which the public are not yet widely aware and undertakes efforts to increase the coverage and visibility of those stories [**?**].

Stifle

The inverse of Reveal, the Stifle maneuver comprises efforts to **remove** a topic from the environment, or from a specific channel therein. Such efforts are rarely completely successful outside of small highly targeted channels, but even within such channels they represent an attempt at influence. A hobbyist channel banning discussion of political issues ensures that, within the context of that channel's communication, the beliefs and positions of the participants remain static; and it deprives the participants of opportunity to organize, convert, or coordinate political activities, thus directly affecting decision-making relative to political topics.

State-level censorship frequently involves Stifle maneuvers, as censors target specific key words and hashtags in attempts to excise certain topics from the digital environment. Content moderation is a perhaps less draconian, though no less influential, implementation of a Stifle maneuver, with a similar mechanism to the state-level equivalent. The MeToo movement included accusations of suppressed stories through non-digital means [236], which also demonstrate Stifle maneuvering by those attempting to suppress broad awareness of their misdeeds.

3.2.5 Target 5: Information presence (proportional)

Control over the absolute presence of information is rarely possible in practice. Outside of breaking news or initial publication, few influencers have sufficient resources to introduce wholly novel topics, or to completely excise topics, from online discussion; and outside of highly specific topics, such actions are likely impossible due to the nature of human communication and collective consciousness.

Influencers can, however, alter the relative occurrence of information, seeking to increase or decrease how often a certain topic, viewpoint, or tidbit is ingested by the audience, without

necessarily creating or eradicating that topic.

Repeat

The Repeat maneuver comprises actions that **increase a topic's prominence or frequency of occurrence** within a specific information channel or platform, **relative to all other topics**. The brute-force version of this maneuver is spam: constant, repetitive mentions or portrayals of the targeted topic, demanding the viewer's attention to the detriment of any other possible content.

Awareness through volume is another staid tactic in marketing and public relations. Both disciplines emphasize the value of "visibility" and "presence," ensuring that the message they espouse is both the earliest and most often heard. This is accomplished through frequency (buying multiple ad spots on a channel, for example) and breadth (buying ad spots on multiple channels targeting various demographics).

China's state-level firewall mechanism is an excellent example of influence through the Repeat maneuver. The mandated posting frequency and topic emphasis levied on state employees leaves China's social media environment filled with content supporting the state's agenda; dissenting material may exist, but it is lost in the volume of propaganda, making it less and less likely that average users will encounter it and thus negating its influence on decision making [33]. Andrew Tate's army of proxy posters is a brazen example of a Repeat maneuver creating influence for Tate and his viewpoints [69].

Smother

The inverse of Repeat is the Smother maneuver, comprising actions that **decrease a topic's prominence or frequency of occurrence** within a specific information channel or platform, **relative to all other topics**. As with Recommend/Sideline, any Repeat action will de facto Smother non-targeted topics, and vice versa; once again, the differentiation is based on the influencer's specific target.

State-level censors and content moderators Smother problematic viewpoints as part of their broader efforts to Stifle them; by filtering out specific words and banning known problematic users, these controllers reduce the targeted topic's representation within the broader flow of traffic, decreasing the probability of exposure. Changes to the Twitter algorithm and to Grok's algorithms also represent Smother maneuvers, in that they deliberately sought to reduce specific political viewpoints or conclusions in the feeds and replies those programs produced for users [103] [149] [233]. The perpetrators targeted in the MeToo movement likewise used extralegal means to marginalize or discredit stories accusing them of misconduct, preventing larger exposure and "burying" these stories [84] [236].

3.2.6 Maneuver Summary

Table 3.1 summarizes the ten identified maneuvers. Table 3.2 provides a quick reference as to the occurrence of these maneuvers within the case studies I examined.

While I have identified historical examples of these maneuvers occurring, I stress that the definition of each maneuver is agnostic to the means of execution. There are many actions

one might take to achieve the effects that define each maneuver – administrative actions within a platform; specific targeted content posted to a platform; and actions taken outside of the platform, both online and offline.

Table 3.1: Availability maneuver summary

Targeted Availability Metric	**Maneuver description**	**Maneuver label**
Presence, availability, or accessibility of an information platform or channel	Actions that create or expand availability, etc.	Roll-out
	Actions that remove or reduce availability, etc.	Shutdown
Relative prominence or popularity of an information platform or channel	Actions that increase or expand relative use or traffic	Recommend
	Actions that decrease relative use or traffic	Sideline
An individual's reach in, control of, or influence within an information platform or channel	Actions that provide or expand an individual's reach or control	Raise
	Actions that limit or remove an individual's reach or control	Silence
The presence of a given topic or concept within an information platform or channel	Actions that introduce a topic that was previously absent	Reveal
	Actions that remove a topic that was previously present	Stifle
A given topic's relative prominence or occurrence within an information platform or channel	Actions that increase a topic's prominence etc.	Repeat
	Actions that decrease a topic's prominence etc.	Smother

This organizing method aligns my availability-based maneuvers with the BEND framework's community-based and narrative-based maneuvers. They enable the identification of broader goals and trends in influence dynamics, agnostic to the specifics of implementation. In this way, creative and varied influence attempts need not be addressed as wholly novel. Instead, they can be fitted into a framework that allows clearer comprehension of the cumulative effects, progress, and probable overall goals of a wide-scale influence campaign.

3.3 Analysis

As mentioned, the utility of the proposed maneuver framework is focused on the operational level, at which policymakers and other societal-level stakeholders operate. At that level, the diverse means and mechanisms of influence are too numerous and specific to expect comprehensive expertise in any one person or even in a small group or committee. By abstracting the means

Table 3.2: Case studies displaying availability maneuvers

Case studies	Roll-out	Shutdown	Recommend	Sideline	Raise	Silence	Reveal	Stifle	Repeat	Smother
Marketing and advertising			X				X		X	
State-level internet control		X		X				X	X	X
Fake news accusations			X	X						
Elon Musk and Twitter/X			X		X	X		X		X
Alternative social media platforms	X		X	X	X	X		X		X
Andrew Tate and Hustler's University			X		X				X	
Alex Jones and Infowars		X				X				
#MeToo movement								X		X
ProPublica	X						X			
Telegram	X	X	X	X				X		

while preserving the desired effect, the maneuvers provide a way for such leaders to discuss the ramifications of external influence efforts without devolving into analysis of those efforts' minute details.

The BEND framework is designed with similar purpose, and these maneuvers are a natural and intentional extension of that maneuver set to produce the more generalized BENDRS maneuvers. The extension is an extremely powerful addition to BEND because it focuses on information availability, rather than social ties, opinions, and emotions. Because availability is a much more technically informed information phenomenon, the availability-based maneuvers offer more intervention options to stakeholders than a BEND analysis generally presents. It is difficult to devise policies and plans designed to inculcate or dissolve community ties, or to concretely and intentionally alter opinions. Policies governing availability – content regulation, platform operations, publication guidelines – are much more tangible and practical, and yet, as I have demonstrated, are just as essential in implementing and countering wide-scale influence.

I also wish to emphasize that while this framework does accommodate many technical actions, it does not address *all* cyber information actions. There are many activities under the traditional cyber label, and frameworks like ATT&CK provide outstanding technical taxonomies for them. My maneuvers would similarly map some of those cyber actions to broader information maneuvers, but only when those actions are undertaken with influence as the goal.

To provide an example: suppose a cyber attacker penetrated a target system and stole sensitive data. If the attacker released that data with the intent of embarrassing the system operator, revealing some previously hidden matter, or swaying public opinion, then that cyber attack would be considered the implementation of one of the availability maneuvers (Sideline, Reveal, or one of the BEND maneuvers, respectively). If, however, the attacker merely means to extort the target for financial gain, with no direct goal as to broader opinion and decision-making, then the attack would not be considered an availability maneuver, even though the technical execution

was identical. In short: Not all availability maneuvers are strictly cyber actions, nor are all cyber actions represented within the availability maneuvers.

3.4 BENDRS Integration

The BEND framework is a method to empirically describe and assess influence campaigns, both constructive and destructive:

> The BEND framework argues that influence campaigns are comprised of sets of narrative and structural maneuvers, carried out by one or more actors by engaging others in the cyber environment with the intent of altering topic-oriented communities and the position of actors within these communities. A topic oriented community is a group of actors who are more or less talking to each other about more or less the same thing. [49]

BEND categorizes influence efforts as targeting either the structure of the information environment – *Community* maneuvers – or the discussions within that environment – *Narrative* maneuvers. Within these categories, influencers make both constructive and destructive efforts, attempting to increase or decrease, affirm or negate, and/or create or destroy different aspects of the IE community structure and narrative content. BEND quantizes these efforts into 16 maneuvers:

Table 3.3: BEND Maneuvers

	Community maneuvers: Alter who is connected to whom, the strength of those connections, and the influence of members		**Narrative maneuvers**: Impact what is being said and how it is being said	
Constructive	Back	Discussion or actions that increase the actual, or the appearance of, an actor's importance or effectiveness relative to a community or topic	Excite	Discussion or actions related to a community or topic that cause the reader to experience a positive emotion such as joy, happiness, liking, or excitement
	Build	Discussion or actions that create a group, or the appearance of a group, where there was none before	Explain	Discussion or actions that clarify a topic to the targeted community or actor often by providing details on, or elaborations on, the topic

Table 3.3: BEND Maneuvers

	Community maneuvers: Alter who is connected to whom, the strength of those connections, and the influence of members		**Narrative maneuvers**: Impact what is being said and how it is being said	
	Bridge	Discussion or actions that build a connection between two or more groups or create the appearance of such a connection	Engage	Discussion or actions that increase the relevance of the topic to the reader often by providing anecdotes or enabling direct participation and so suggesting that the reader can impact the topic or will be impacted by it
	Boost	Discussion or actions that increase the size of a group and/or the connections among group members, or the appearance of such	Enhance	Discussion or actions that provide material that expands the scope of the topic for the targeted community or actor often by making the topic the master topic to which other topics are linked
Destructive	Negate	Discussion or actions that decrease the actual, or the appearance of, an actor's importance or effectiveness relative to a community or topic	Dismay	Discussion or actions related to a community or topic that cause the reader to experience a negative emotion such as worry, sadness, disliking, anger, despair, or fear
	Neutralize	Discussion or actions that cause a group to be, or appear to be, no longer of relevance, e.g., because it was dismantled	Distort	Discussion or actions that obscure a topic to the targeted community or actor often by supporting a particular point of view or calling details into question
	Narrow	Discussion or actions that lead a group to be, or appear to be, more specialized, and possibly to fission, or appear to fission, into two or more distinct groups	Dismiss	Discussion or actions that decrease the relevance of the topic to the reader often by providing stories or information that suggest that the reader cannot impact a topic or be impacted by it

Table 3.3: BEND Maneuvers

	Community maneuvers: Alter who is connected to whom, the strength of those connections, and the influence of members		**Narrative maneuvers**: Impact what is being said and how it is being said	
	Neglect	Discussion or actions that decrease the size of a group and/or the connections among group members, or the appearance of such	Distract	Discussion or actions that redirect the targeted community or actor to a different topic often by bring up unrelated topics, and making the original topic just one of many

3.4.1 Maneuvers versus Effects

The definitions above describe BEND *maneuvers*, efforts undertaken to reshape the IE. In similar fashion we can define BEND *effects*, the measurable shifts in the IE that successful BEND maneuvers produce. Hickman's thorough quantitative definitions of these effects are excellent and exhaustive, and for brevity I do not repeat them here, referring readers instead to his work [113].

The same labels are applied to both maneuvers and effects, with differentiation made clear by context. A BEND maneuver is an ongoing or proposed *effort*, a series of actions and interventions with the stated aim in the maneuver definition. Conversely a BEND effect is a *condition* within the IE that likely resulted from the eponymous maneuver, formally defined as a series of metrics derived from a network representation of the IE.

3.4.2 Attribution and Causality

The IE is dynamic and unpredictable; such network representations will, over time, evince significant changes in structure, and any metrics derived therefrom will similar change in value. Because BEND effects are identified through these values, it is possible that BEND effect indicators may appear without any preceding deliberate BEND maneuver having been executed.

To handle this uncertainty, the BEND framework includes three important tenets.

1. BEND maneuvers are characterized by *intent*. Agents within the IE might send messages or take other actions that produce measurable BEND effects – bridging two groups with a party invitation, for example. Even though such actions produce BEND effects, they should not be classified as BEND maneuvers if the agent's intent was not to achieve the measured effect. Assigning intent to such actions is an analytical art, generally drawing from the identity of the agent and from their past and recurrent actions within the IE.
2. BEND maneuvers can be detected and assigned at the message level, with varying levels of fidelity, through methods described below. As such, the BEND framework can detect attempts at the actor or group level to influence the IE, going so far as to identify the specific maneuvers attempted. This detection and attribution is developed and expanded in

Blane's work [30].

3. In contrast to the previous point, however, the BEND framework can *not* attribute a specific measured effect to a specific action or message. The method can reliably indicate an agent attempting to Distract, for example, and can reliably detect that a topic was the subject of Distraction; but it cannot causally attribute the measured effect to the detected actions, nor to any other Distract actions that may be present. All effects are the aggregate result of all actions. This lack of causality is an important caveat, as it prevents analysts and observers from becoming overconfident in interventions and in their own ability to shape the IE.

3.4.3 Operational Methodology

The BEND framework lends itself to operational use because it is both quantitative and standardized. To illustrate operational application, I offer the example of a military Information Operations (IO) cell. Such a cell is responsible for understanding the IE as it applies to their commander's overarching mission, and for shaping and affecting the IE in ways that achieve or support that mission. To do so, the cell executes a continuous operational "loop":

1. The cell collects, or receives from other collectors, IE-derived intelligence, such as web traffic, social media messages, news broadcasts and print articles, etc.
2. The cell processes this intelligence to form an aggregate, cohesive understanding of the IE. Within this framework the cell identifies new and ongoing influence efforts, both supporting and adversarial to their mission.
3. Concurrently, the cell receives (and reviews previously received) directives from the command, regarding both broader mission objectives and IE-specific goals.
4. The cell considers possible actions the command might undertake to shape the IE in support of the command's articulated goals; the cell presents their understanding of the IE and these options to the command for decision.
5. Following a decision, the cell works to implement (either directly or through subordinate/partnered assets) the directed actions.
6. The cell then observes (through collection) the IE to assess the impact of the changes, thereby beginning the operational loop anew.

The BEND framework and associated tools directly nest in this loop:

1. As above, the cell receives or collects intelligence.
2. The cell uses BEND-supporting tells such as ORA-Pro, Artemis, and NetMapper to construct a network representation of the IE from intelligence. These same tools enable BEND analysis, which:
    - Parse message content, interpreting linguistic cues and contextual information to identify BEND maneuver attempts.
    - Analyze network evolution over time to identify BEND effects within the IE.

    The cell is thus equipped with a robust and intuitive network representation of the IE, and with a comprehensive analysis of ongoing BEND efforts and measured BEND effects, with

measurements down to the specific agent and/or message.

3. The cell receives direction as above. This direction can be framed in BEND terminology, simplifying the task of describing a complex and nebulous set of goals for the command by providing a concise and specific lexicon.
4. The cell considers possible BEND maneuvers to shift the IE from its current state toward the command's desired state, potentially using tools like OMEN to simulate the effects of proposed actions prior to presenting them for decision.
5. The cell then implements and observes as above.

3.4.4 Integrating the RS maneuvers

In this operational context, the RS maneuvers are natural augments to the BEND methodology. The expanded BENDRS framework embeds in the operational process in exactly the same way, but provides additional insight:

- Platform representation broadens the network representation of the IE, providing greater visibility on the movement of information within the IE.
- Similarly, platform-specific maneuvers – including the RS maneuvers – allow more specific and precise implementation of information operations.

As an example, we might consider a platform-oriented cyber action. Consider an instance where a malign influence group is agitating through social media at an especially fraught political moment, such as an election day or in the aftermath of a domestic disaster. This group seeks to sow discord and incite violence. The IO cell may wish to minimize this group's influence, without depriving the broader environment of the social media platform that they are utilizing, since that same platform is likely helpful to the proper function of society in the face of the current exigency.

Referring to the BEND maneuvers, the IO cell would likely seek to *Negate* this group, and to *Neutralize* them. Further, they would seek to *Dismiss* the group's inciting narratives while *Explaining* the full context of ongoing events.

Toward achieving this, the IO cell might direct message-generating and community-engaging assets such as Public Affairs, Psychological Operations, and Civil Affairs, providing message points and engagement priorities. These actions would no doubt achieve some of the cell's stated goals. Ambiguity emerges, however, when considering what cyber actions the cell might direct. Even something as simple as blocking the malign actors' online presence would spill across multiple BEND maneuvers. It would Negate, Neutralize, Narrow, and Neglect all simultaneously; and, in the absence of the influencers' drumbeat of messaging, coupled with the shock of their sudden disconnection, the potential for Distortion is as great as that of Explanation or Dismissal.

Adding RS maneuvers helps the cell tailor its cyber actions. The second-order BEND effects described above may still occur, but because RS maneuvers directly address the cell's availability-based options, the cell can implement an RS-based intervention and simultaneously monitor for emergent unintended BEND effects. The cell can more directly define the influence it wishes to achieve, in a way that gives cyber assets specific actions to take. And, crucially, the cell can specify these actions to the cyber team *without* using cyber-specific language.

In our example, the IO cell would direct the cyber team to *Silence* the malign actors and to *Stifle* malign messages, while *Repeating* official narratives. This direction is sufficient to initiate an operational chain reaction:

- The cyber team understands *Silence* in terms of outcomes, and opts to execute a rate-limiting attack against the malign actors, making it difficult for them to post on the social media platform without adversely impacting other users.
- The team understands *Stifle* in terms of outcomes, and works with the social media platform to update moderation policies and suppress the problematic content.
- And, the team understands *Repeat*, deploying a network of automated agents to amplify the targeted official narratives.

When the IO cell next examines the IE network, they can measure the effects of the above maneuvers. In concert with the message- and engagement-based efforts of other assets, the cell will likely identify the desired BENDRS effects, with the negative actors *Negated* and *Neutralized* in part by virtue of being *Silenced*; the negative narratives *Dismissed* and *Stifled*; and the positive narratives *Explained* and *Repeated*.

3.5 Conclusion

As influence campaigns become more frequent, persistent, and effective, stakeholders charged with countering or mitigating their negative impacts must have useful frameworks, models, and tools to help them recognize, understand, and respond to information-based threats. BEND is an effective tool toward that end, and the BENDRS framework is a useful extension, bridging the gap between social influence threats and cyberspace competition.

Chapter 4

The OPIEM Model of the Information Environment

In chapter 1, I introduced OPIEM as a modeling methodology to produce a network representation of the IE that modeled both social and cyber influence. In chapter 2, I established the strict need for such a model, as influence actions at scale are almost certainly hybrid. In chapter 3 I identified categorical actions (maneuvers), identifying ways that changing information availablity grants influence over the IE and decision makers within it.

Quantitative demonstration of the OPIEM framework would require a corpus of observed data – messages, information packets, or similar traffic from the IE, from which an OPIEM network model could be built. In this chapter I will present in greater detail the construction and analysis of an OPIEM network model.

As previously mentioned, I was unable to locate any extant datasets that contain identically scoped message traffic (relative to query terms and timeframe) across multiple platforms sufficient to construct a robust trimodal representation. Several cited works, especially those using the BEND framework, include analysis of traffic-derived networks, but all datasets reviewed produced extremely static and/or sparse platform data. The case studies used in chapter 2 and chapter 3 demonstrate the viability and presence of availability-based influence maneuvers, but I was unable to locate or secure sufficient message traffic pertinent to those incidents.

It therefore remains to demonstrate:

- That an OPIEM tri-modal network can be constructed solely from message traffic;
- That this network can effectively represent and describe the IE;
- That known influence mechanisms, specifically the BEND maneuvers, are detectable from OPIEM network metrics; and
- That the proposed availability maneuvers are detectable from OPIEM network metrics and exert effective influence within the IE.

To this end, I designed and constructed an agent-based social media simulation. My simulation adapts the Project OMEN architecture from Carnegie Mellon's Center for Computational Analysis of Social and Organizational Systems (CASOS) [163]. I will discuss the OMEN mechanics before further discussion of OPIEM, to ensure clarity regarding the synthetic data from which the OPIEM model will be constructed and some of the terminology used in both the

OMEN system and the resulting network model.

4.1 Project OMEN

Project OMEN produces social media traffic at a large scale. It is specifically designed to facilitate training for government and military analysts who must holistically understand the information environment across multiple institutional and social boundaries, in order to provide recommendations to policymakers both to directly influence, and to anticipate secondary effects within, the information environment. The OMEN simulation environment is composed of multiple object primitives.

Actors represent human individuals, human organizations, and automated accounts (bots) – any entity that consumes or produces social media traffic. Actor objects contain the attributes that drive their simulated decision making, as well as feasible biographical descriptions to justify those behaviors. These descriptions are beneficial to a human training audience, as they mask the underlying simulation mechanics while still offering studious analysts insight into the expected behavior of the entities.

Groups capture both explicit and implicit social ties within the actor population. Groups range from formal/institutional collections, such as professional associations or nationalities, to informal and organic collectives, such as hobbyist circles or online opinion communities.

Events constitute the happenings external to the social media environment, and compose the timeline of the simulation window. OMEN uses events to drive conversation, mirroring the effects of shared awareness on both the volume and composition of social media traffic.

Topics and narratives provide simulation-accessible quantization of the beliefs and opinions pertinent to the scenario. Topics capture the underlying themes or concerns of the population, expressed in a polar fashion (pro/anti, for example). Narratives embed within topics, providing specific lines of discussion, rhetorical arguments, or exposition of those topics.

Articles and links are offered both as scenario framing for the training audience, and as tokenized "evidence" cited by the simulated population. These embedded URLs and images are crucial in simulating topical cross-pollination and narrative collision within the information environment.

Authors specify scenario outline parameters, including the involved nationalities and pertinent social factors. OMEN's AI-Utilized Retrieval for Optimized Representation of Audiences (AURORA) system uses an author-provided document corpus to generate thousands of highly detailed and realistically varied actors, organizations, and groups to populate the simulation, and authors can manually add additional critical details or individuals if desired.

OMEN's social media simulator SynSM runs a time-indexed loop. At each time step, a subset of actors acts by:

1. Consuming social media. Each actor stochastically samples from the existing corpus of messages, with probabilistic weighting based on their social networks, interests, social media behavioral archetypes, and ongoing current events.
2. Considering what they have consumed. Each actor updates their internal beliefs and preferences based on the content of the messages selected in step (1) above.
3. Engaging with social media. Each actor stochastically opts to react to the messages they are aware of, and/or to produce original content. All the produced messages, whether reactive or generative, are then added to the broader corpus.
4. Producing the resulting messages using an LLM. Message parameters – narrative, author attributes, emotionality, writing style, etc. – are passed to an LLM, which outputs corresponding text for a social media post.

OMEN outputs the message corpus in accordance with the hosting platform's API: synthetic messages from X are produced in the same format used by X's interface, as are Facebook and Telegram messages. This allows synthetic OMEN traffic to be analyzed using the same tools and methods employed for actual internet traffic. Using OMEN, organizations can train their analysts to use whichever tools and methods they prefer for social media analysis.

4.2 GhostCell – a OMEN modification

My research required some expansion of the OMEN system, while permitting significant simplification in others. I called my system GhostCell.

4.2.1 Modification of existing components

Actors: A significant portion of OMEN's actor description is used to fuel realistic and consistent text through LLM prompting. This is essential to OMEN's primary function as a human training tool. As GhostCell needs only to demonstrate shifts in agent behavior and attributes within a realistically behaving population, human-readable text output was not a requirement, and thus large portions of OMEN's actor objects were irrelevant to my purposes.

As in OMEN, GhostCell shifts agents' opinions based on environmental influence. Unlike OMEN, GhostCell does *not* simulate dynamic internal emotional states. Agent emotions are held static throughout the scenario, which is an accepted limitation to keep the influence model simple.

Groups: Rather than use AURORA to generate my simulation population, I used a simplified stochastic method. I defined several population communities a priori and specific parameters dictating the distribution of attributes and opinions within those communities. I produced a fixed number of groups, deliberately producing a mix of institutional groups (cross-cutting membership), community groups, attribute groups (trait homophily), and position groups (belief homophily). Actors are stochastically assigned to these groups using the appropriate respective probability weights.

Events: As with actors, a significant portion of OMEN's Event object is used to prompt LLMs and produce realistic human-oriented output. My event descriptions are therefore vastly simplified and consolidated. Crucially, I expanded GhostCell's event handling to capture three distinct components of event impact:

- *Excitation*, the way that events drive conversation. My implementation matches OMEN's, in that events 'excite' specific narratives, increasing their likelihood of engagement within the actor population.
- *Reflection*, the way that events directly influence beliefs. Actors witnessing events will update their internal state based on the event itself, without any moderation through social media; the net effect on the actor will then be the combination of the event's impact and the cumulative impact of social media consumption.
- *Alteration*, the way that events alter the structure or composition of the information environment. To demonstrate availability-based attacks, I need a mechanism whereby information platforms or sources can be removed or added to the environment; this component of event handling permits events to "add" or "remove" actors, platforms, or messages based on an exogenous trigger. In this way I can simulate non-population-based effects, such as cyber attacks, platform moderation/censorship, or deplatforming/replatforming.

Topics and narratives: To maintain simplicity, I limited GhostCell to six topics. Crucially, I posited that these topics are orthogonal – that is, an actor's position on one topic is completely independent from their position on every other topic.

As with other objects, OMEN narratives are largely oriented around LLM prompting and content generation. My narrative objects are accordingly significantly simpler and plainer, with only the numerical descriptors needed to simulate influence and information exchange. As an IE is defined relative to decisions, I built the simulation narratives around a set of issues (further defined below), thus framing the IE around political decisions (e.g., voting and activism).

Articles and links: Rather than simulate entire news articles and images as would be necessary for a human training audience, GhostCell abstractly represents these objects as *attachments*. Attachments are associated with specific narratives within the simulation environment and are selected stochastically for inclusion in messages at composition.

4.2.2 New components

To demonstrate availability-based maneuvers, GhostCell expands the OMEN simulation by adding **platforms**. As discussed in previous chapters, platforms represent the information systems through which agents interact with the information environment – the systems and devices that facilitate transmission, storage, processing, and production of information.

In the extreme, the platform nodeset could be extended to include basic sensors, individual phones and computers, infrastructure nodes such as routers and signal amplifiers, and online services such as cloud storage and web pages. Such an extension would approach convergence between my platform representation and a finely detailed map of cyber infrastructure, like those used in many cyber-analysis frameworks. For the purposes of this work, I deliberately keep

platforms at an extremely high level, as my primary objective is to demonstrate platform-derived influence on agents at scale. One can imagine use cases where a more granular representation of cyberspace, combined with representations of the using agents and topic contents, would be useful for analysis, projection, and planning.

I use two classes of platform in my simulation. **Open platforms** are centered around user-generated content, corresponding to communications or social media systems. **Closed platforms**, by contrast, feature curated or solicited content only, corresponding to news outlets or traditional media (newspaper, television, etc.). GhostCell agents are influenced by messages from both sources, but only generate messages through open platforms. Crucially I did *not* simulate the presence of media outlets within the social media ecosystem. The closed platforms in GhostCell serve as static, influential voices, constantly injecting their specific topical positions into the simulation to prevent or retard convergence. The system architecture causes agents subscribed to a given closed platform see all the messages that platform generates; my design assumes that, whether by website, television channel, or social media message, a media outlet's audience receives its content and reacts accordingly.

I also simulate open platforms as single-channel, meaning an agent can come across any message on the platform regardless of author. Such an architecture best approximates X or Bluesky, platforms characterized by a single "global" messaging channel. It is much less apt for platforms like Reddit or Telegram, which are more defined by interior channels or subdivisions that limit message visibility.

I justify this decision as a reasonable simulation of an agent's *feed*. Social media platforms are increasingly driven by algorithmically generated feeds, designed to increase user engagement by presenting them with content that matches their behavior and interest patterns. As such, even sub-divided platforms increasingly have an approximate "global" channel: a user of Reddit, for example, is presented with algorithmically curated content from all public subreddits, regardless of the user's membership. The feed-based assumption is only fully violated for Telegram, which has no native unified feed function. I examine this further in chapter 6

4.2.3 Face validation

Table Table 4.1 presents a face validation summary of the GhostCell model, based on canonically observable aspects of real-world social media use and operation. The mechanics of the simulation are further explored in the appendices.

Table 4.1: Face validation of GhostCell simulation model

Real-world feature	**Simulation feature**	**Known gaps**
Algorithmic curation dominates exposure	Message visibility weighted by engagement (virality), recency, author prominence (followers), group/authority boosts	No learned recommender system (user-tailored preferences); no explicit follower/subscriber mechanism; static weights over time and per user

Table 4.1: Face validation of GhostCell simulation model

Real-world feature	**Simulation feature**	**Known gaps**
Social networks exhibit high homophily, community structure, echo chambers, and uniqueal influence distribution (scale-free or heavy-tailed)	Community and group membership explicitly modeled; directed influence matrix; follower counts heavily affect influence and are distributed scale-free/heavy-tailed	No network rewiring (block or unfollow); homophily is exogenous, determined on actor generation, not emergent
Users log in intermittently, consume small subset of available content, have bounded bandwidth	Geometric activity counter simulates intermittent use; limited recent message queue simulates bounded memory and small subset; attention mediated at individual level via agent attributes	No explicit attention decay *per turn*; no competition between like content items when sampling available content
Users operate across multiple platforms, splitting attention and migrating	Users hold accounts on multiple platforms; preference drives uneven engagement, based on user perception of platform; users adopt new platforms experimentally, including novelty dividend; users leave unliked platforms	Hard limit to number of accounts versus ”attention budget” mechanic; no modality representation (video vs text); migration overly binary (yes or no, versus attention splitting), overly permanent (deletes account upon migration); no mechanic for voluntary group exodus
Platforms differ in audience, norms, algorithmic behavior, and modality	Multiple platforms, with comfort levels aligned to demographics to simulate audience difference; platforms hold ”opinions” to simulate perceived norms	Abstracted platforms have no explicit modality; recommendation algorithm is identical across platforms
Users select platforms based on novelty, perceived value, and existing/desired social ties	New platforms receive novelty bonus to attract new users and hold attention during trial period; users prefer platforms based on internally changing perceptions	No simulation of social ties, as cost or benefit - no preference afforded due to friends that are members, nor due to established follower base

Table 4.1: Face validation of GhostCell simulation model

Real-world feature	**Simulation feature**	**Known gaps**
Opinion influence is driven by homophily, credibility, emotional resonance; extreme opinions are more resistant to change	Homophily simulated through distance gating and directed influence; credibility simulated through platform trust, agent authority, agent leadership; emotional resonance simulated through Plutchik alignment	Plutchik alignment static throughout simulation, totally random (therefore not meaningful in current model); no explicit identity protection cognitition; no attribution of influence - effect is aggregate over all messages
Narratives emerge and shift over time; public discourse clusters around narratives; salience shifts due to events, repetition, and amplification	Narrative selection includes salience, recency, trends, and external events	Narrative set static over simulation run; no explicit competition between narratives for finite attention at agent level; no decay of narratives over time
Media organizations produce content at regular cadence, shape agendas, and influence trust and framing	Media entities in simulation generate content independent of social media; agents maintain perceived trust/alignment of media entities and moderate influence accordingly; topics selected based on platform-specific parameters - outlets are not all equal	No evolution of editorial bias (static positions); no competition between media entities; explicit yes-no subscription at agent level determines exposure; no media entity outreach or advertisement via social media
Users differ in behavior on social media, varying in posting frequency, engagement level, and engagement goals	Model includes user archetypes to capture major behavioral categories; model further uses agent attributes (ego, energy, etc.) to individualize behavior	Archetypes are fixed a priori; no role evolution over time, agent attributes are static
Difference between reactive behavior (likes, shares) and deliberative behavior (content creation) – different resource requirements	Mode selection with energy-weighted distribution simulates likelihood of different behaviors, favors reactivity	No explicit separation of reactive vs deliberative cognition; energy abstraction hides time-/resource-cost differences in actions

Table 4.1: Face validation of GhostCell simulation model

Real-world feature	Simulation feature	Known gaps
Real world systems exhibit polarization, convergence, and/or fragmentation by turns	FJ model favors community-based divergence within the populace; exogenous influencers (media) counteract FJ population-level convergence	Polarization can emerge but is not explicitly modeled; model still converges over time; no explicit identity clustering mechanism in model

4.3 Constructing the OPIEM Network from messages

As described in chapter 1, the OPIEM methodology constructs a network representation of the IE solely from observed traffic. Parsing message traffic produces a 4-mode network, with message objects in the center (Figure 4.2).

Each message links to the authoring agent; non-original messages (reactions, quotes, replies, etc.) also link to the author of the target message. Each message also links to the open platform on which it was generated.

Each message explicitly contains a single narrative which, in GhostCell, is linked to a single underlying topic; this links the message to the topic nodeset. The 1-to-1 mapping of messages to narratives is a reasonable simulation requirement; the 1-to-1 mapping of narratives to topics is more restrictive. Part of the issue is the semantic overuse of the word "topic." Real traffic analysis relies on topical markers such as hashtags to identify narratives; in this sense the "topic" nodeset is often *larger* than the set of identified narratives, with topics not seen as underlying broader positions but as varied surface markers of communication. I address this problem by using **cross-cutting issues** as seeds for my narratives. These are broad social issues (i.e., gun control) with multiple rhetorical narratives, each of which can be mapped onto one of my foundational, orthogonal topic axes. These issue-embedded narratives share attachments between them, creating broad topical coherence similar to that seen in live traffic.

Each message links to a single narrative and may contain supporting attachments. By analytically treating these simulation narratives as separate topic markers, I can emulate the way an analyst would identify a larger narrative or issue within the IE, by finding cross-connections and shared proponents between topics.

I produce a core trimodal network by folding the 4-mode network on the message nodeset. Message objects include basic attributes that facilitate construction of several exterior and interior links. Some attributes mimic data directly observable in live data:

- The timestamp, which facilitates creating time-varied network slices for comparison.
- The author, as well as the original author of content amplified by the message.
- The number of engagements the message has received.

Other message attributes approximate or generalize features of live data:

- The narrative to which the message pertains. This approximates quantizing message text against a set of possible narratives. In live traffic, this would be done through natural

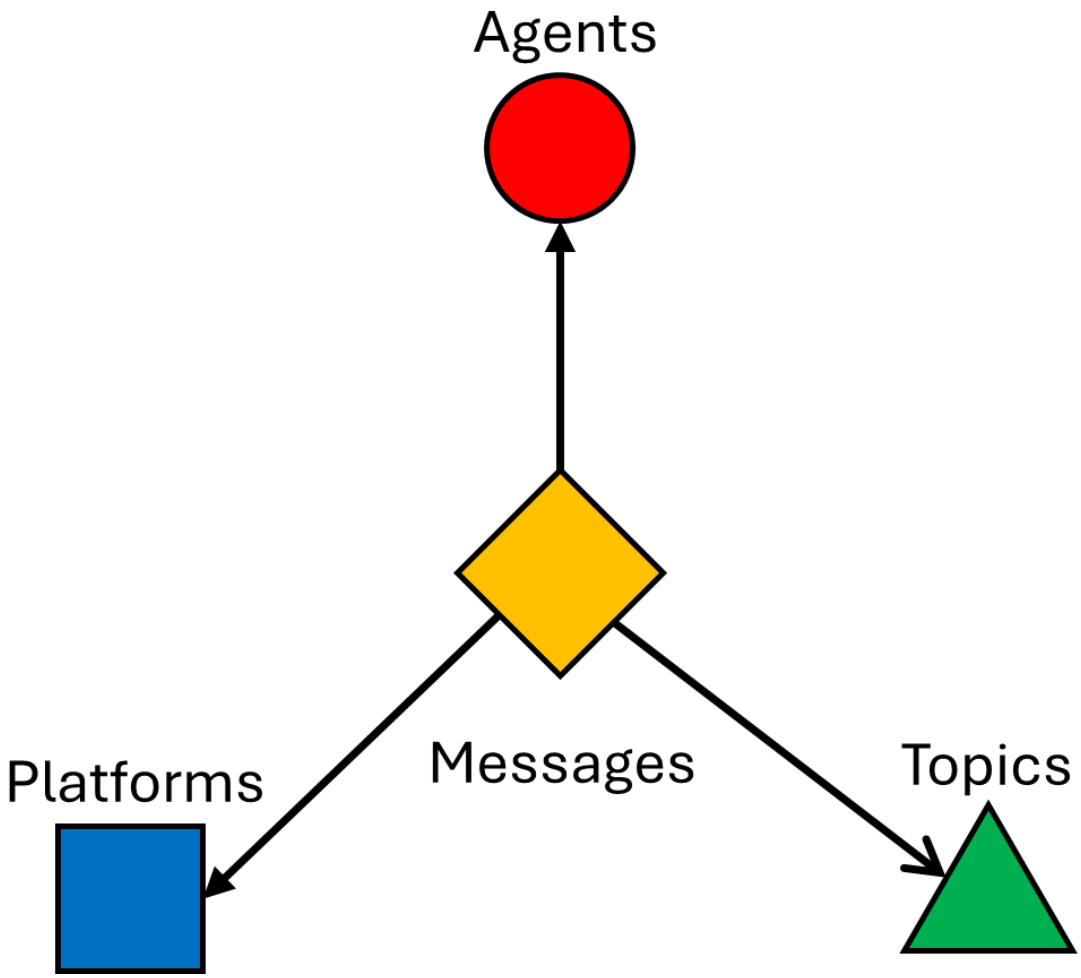

Figure 4.1: 4-mode network model of the IE produced by processing message traffic.

language processing (NLP) and hashtag/URL classification.

- The position of the author relative to the narrative. This approximates interpreting the author's text as positive/supporting or negative/detracting, another NLP function. (GhostCell does not simulate sarcasm. Parsing sarcasm is a significant challenge in real data.)
- Any attachment contained in the message, which approximates the use of an embedded link or attached image/document in support of the text's contents.
- The message's "rhetorical payload." This approximates the persuasiveness and efficacy of the text. I represent rhetorical payload as a [0,1]6 vector, where the components correspond to the six moral foundation axes in Moral Foundation Theory [253].
- The message's "emotional payload." This approximates the emotional tone and voice of the text. I represent emotional payload as a [0,1]8 vector, where the components correspond to Plutchik's core emotions [214].

Actors have internal attributes corresponding to their valence or opinion on each topic; they

also have internal rhetorical and emotional vectors representing their "preferred" messaging. Message position and payload are assigned the author's attribute values on creation; the influence a message has on a reader is a function, in part, of the vector distance between the message payload and the author's own values.

Using these message objects, I generate the OPIEM trimodal network as a combination of three bi-modal networks and the following links:

Agent x Topic

- *Salience*, the proportion of messages by an agent about a topic. (Exterior link)
- *Valence*, the average position or opinion of messages by an agent about a topic. (Interior link)
- *Agita*, the average emotional payload of messages by an agent about a topic. (Interior link)

Agent x Platform

- *Preference*, the proportion of messages by an agent on a platform. (Exterior link)
- *Volume*, the count of messages by an agent on a platform. (Exterior link)

Topic x Platform

- *Occurrence*, the proportion of messages on a platform about a topic. (Exterior link)

As mentioned in chapter 1, this is a small subset of the possible links between these nodes, but these networks are sufficient to demonstrate OPIEM's capability to detect influence campaigns and effectively describe the IE.

4.4 Describing the IE

The OPIEM network is intended, in part, to assist non-domain experts in understanding the IE. Toward this end, OPIEM network visualizations provide intuitively accessible representations of the IE. Visualizing the entire trimodal network itself is not particularly informative; as shown in **??**, networks will generally be characterized by an enormous number of agents, a significantly smaller number of topics, and an even smaller number of platforms. This size mismatch produces clustered or ringed network shapes where the smaller nodesets are often lost in the density of the network.

The descriptive power of the OPIEM framework is more clearly demonstrated in response to specific queries. Queries narrow the required context for the network, allowing nodesets to be disregarded or folded upon and thus simplifying the network. The resulting derived network is significantly less *generally* informative, as it has lost information in the derivation process, but is substantially more *directly* informative toward the generating query.

Below, I provide several examples of OPIEM-derived visualizations generated in response to queries similar to those that might be asked by policymakers or military leaders.

4.4.1 Finding key actors and influencers relative to a specific topic or issue

Communication means are not pertinent to this question, so the platform nodeset is excluded. The topic nodeset can be reduced to the topic(s) relevant to the targeted discourse. The network is built between agents and the relevant topics, with exterior links valued as the number of messages

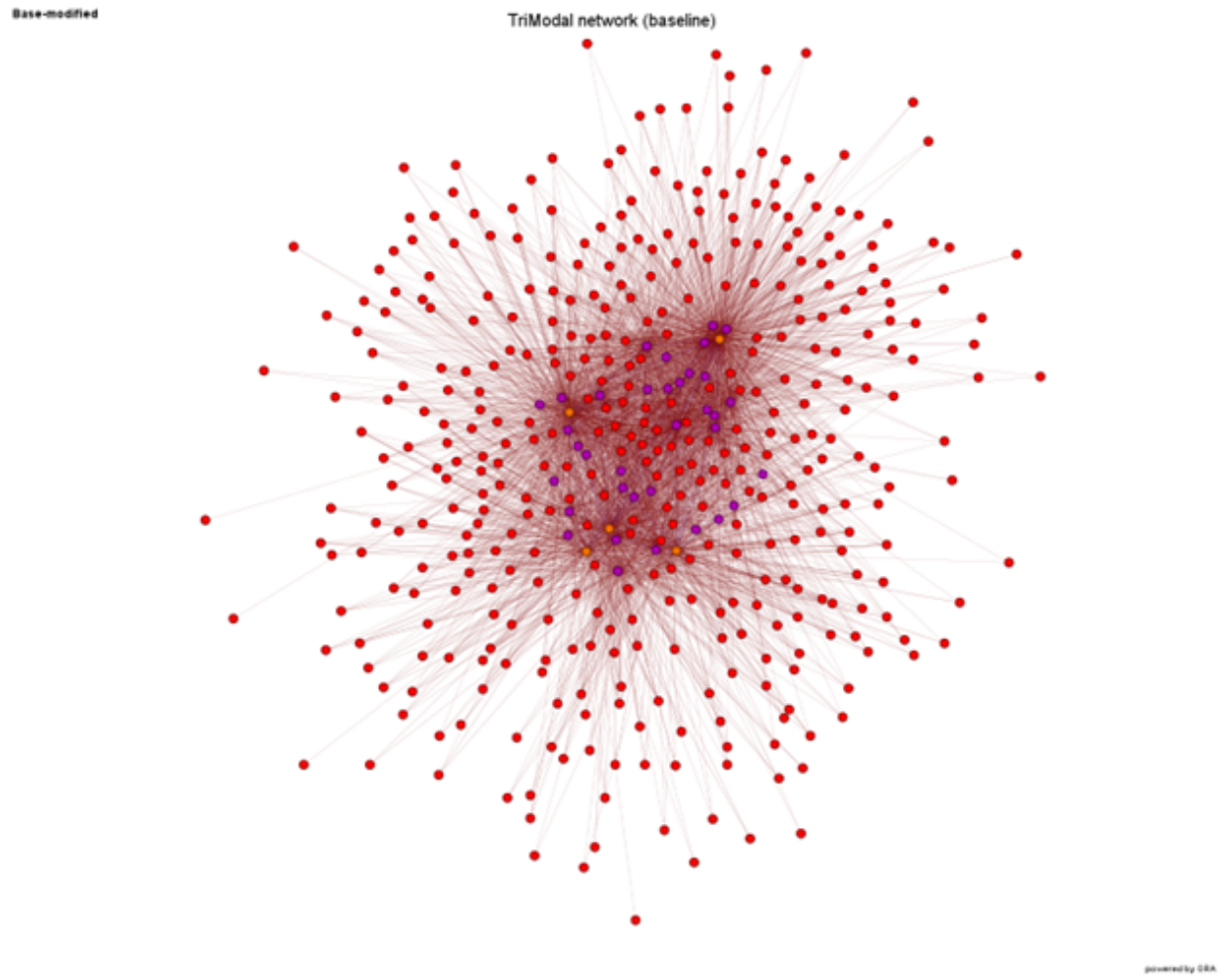

Figure 4.2: Trimodal network diagram of an IE, simulated by GhostCell.

per agent per topic. The resulting Agent x Topic network can be folded on the Topic- nodeset, creating an Agent x Agent network that represents the connective discourse generated by the target issue. Centrality measures of this unimodal network then indicate the key actors.

4.4.2 Determining sentiment (pro or con) concerning a specific topic or issue

Again, the platform nodeset is excluded, and again the topic nodeset is reduced to the targeted topics/issue. Agent stance or position on a topic is derived by parsing agent-generated content, and using this value to create interior links between agents and topics. In the final visualization, agent nodes are colored based on the link value.

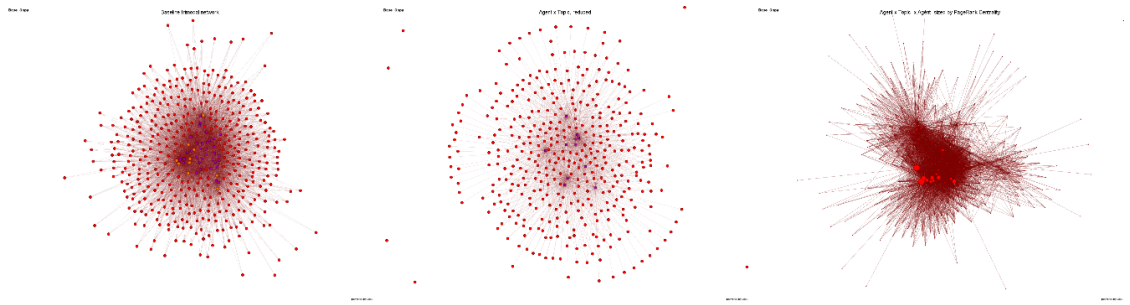

Figure 4.3: Successive reduction of the OPIEM network in response to a query. The initial tri-modal network (left) is reduced to a bi-modal network (center), which is then folded to produce a uni-modal network (right). Within that unimodal network, centrality measures are easily computed, and prominent nodes can be identified by sizing them according to centrality.

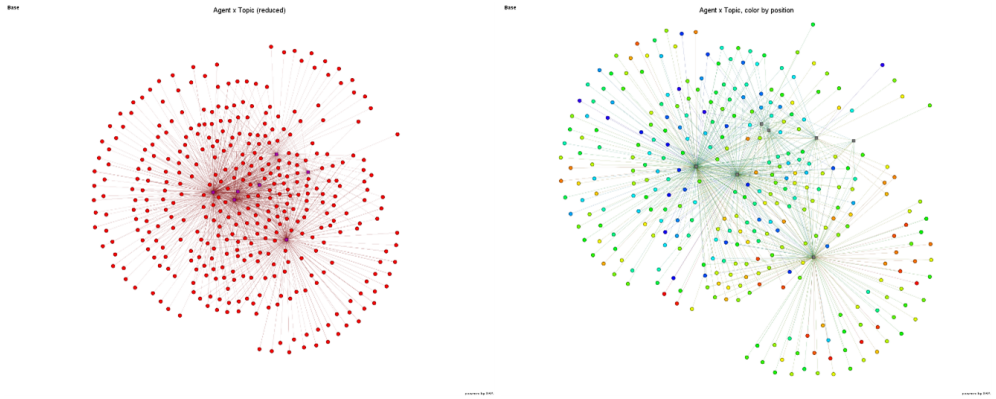

Figure 4.4: Successive reduction of the OPIEM network in response to a query. The initial tri-modal network is reduced to a bi-modal network, which is trimmed to include only applicable Topic nodes (left). Agent nodes are then colored based on average link values between the agent and all topics (right).

4.4.3 Identifying vectors of misinformation, both agent and platform

Given a set of narratives or topics that are known to be misinformation, the Topic x Platform network, with exterior links normalized on platform, shows the prevalence of misinformation on each platform. If we reduce the Topic nodeset to misinformation only, we create a reduced trimodal network representing the misinformation ecosystem. Folding Agent x Platform on the Platform nodeset, using summed and directed links (exterior), produces an Agent x Agent network where the out-degree of nodes indicates drivers of misinformation.

4.4.4 Determining platform prominence within the IE

Discarding the Topic nodeset, we can measure the in-degree of each platform node in the Agent x Platform network to determine that platform's subscriber count. Normalizing the Agent x Platform link values per agent creates a preference network. And, folding on the Agent nodeset and summing link values creates a Platform x Platform network, where link strength indicates

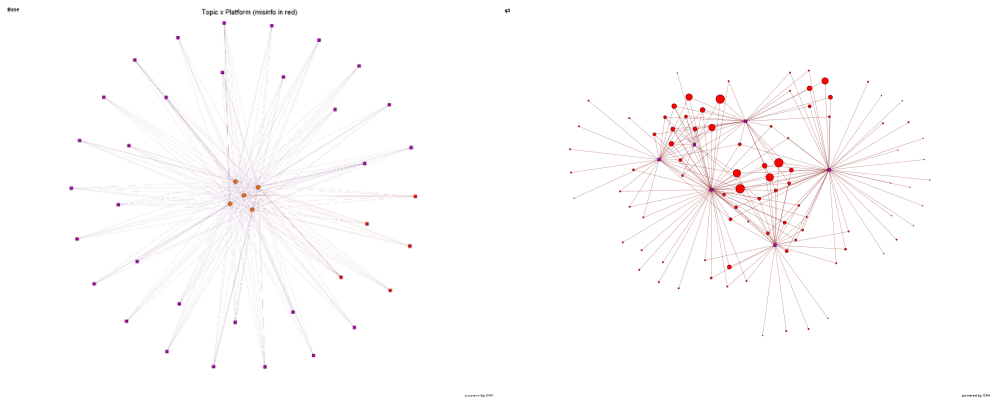

Figure 4.5: Successive reduction of the OPIEM network in response to a query. The initial tri-modal network is reduced to the bi-modal Topic x Platform network, with malign Topics highlighted and link thickness indicating the frequency of each Topic's appearance on the linked Platform (left). Alternatively, the initial tri-modal network is reduced to the bi-modal Agent x Topic network, and the Topics nodeset is reduced to only malign Topics. Agent nodes are then sized by total link weight between all malign topics (right).

the degree to which platforms have common users (and thus common surface area). In a more granular Platform nodeset, with non-user-attached platforms depicted (e.g. routers and servers), creating infrastructure metanodes between user-facing platforms also reveals common surface area to cyber threats, with proportional exposure indicated by normalizing link values per non-metanode-platform.

4.5 Identifying Availability Maneuvers within the IE

The BEND framework identifies influence maneuvers by comparing metrics derived from time-variant slices of an agent-topic network. Each proposed availability maneuver is similarly identifiable within a time-varying tri-modal network. Here, I propose identifying metrics for each maneuver within the derived OPIEM network.

I denote the respective nodesets of the tri-modal network as $\mathcal{A} = a_n$ for Agents, $\mathcal{P} = p_n$ for Platforms, and $\mathcal{R} = r_n$ for Topics. Networks, as sets of links, will be denoted by the included nodesets, as: $\mathcal{V}^t_{\mathcal{A},\mathcal{P}}$ for the Agent x Platform network as of time t. Within this network, the link $v^t_{a_n,p_n}$ is the link between Agent a_n and platform p_n at time t. I use $\mathcal{C}$, the set of communities, to represent possible partitions of $\mathcal{A}$ based on node attributes. I will also use node-centric notation to represent link weights, as:

$$v^t_{a_n,p_n} \equiv p^t_n(a_n)$$

I will denote respective time periods as $t < t' < t''$.

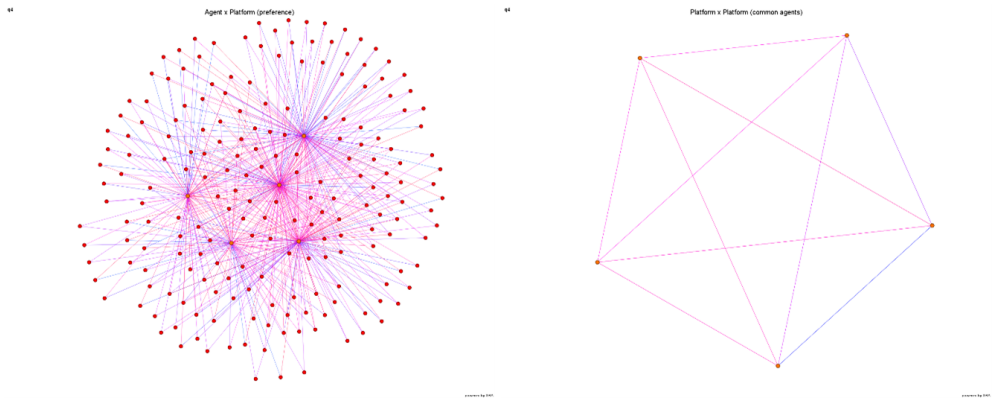

Figure 4.6: Successive reduction of the OPIEM network in response to a query. The initial tri-modal network is reduced to the bi-modal Agent x Platform network, with link weights normalized at the Agent. This network is then folded on the Agent nodset to produce the uni-modal Platform x Platform network, where link weights are colored to indicate user overlap between respective platforms.

4.5.1 Roll-out and Shutdown

A Roll-out maneuver creates, or expands the availability of, an information platform; a Shutdown maneuver removes, or decreases the availability of, an information platform. We assume availability is a technical feature, i.e., a function of language support, firewall access, or other technical features or barriers.

A Roll-out through platform creation is trivial to identify as a positive shift in $|\mathcal{P}|$. Similarly, a Shutdown maneuver effected through the outright removal of a platform can be recognized by a negative shift in $|\mathcal{P}|$.

Increased availability is more difficult to identify: a simple increase in user count (i.e. the degree of some node $p_n \in \mathcal{P}$) could be the function of increased availability but may also be a result of increased popularity or appeal – a very different underlying cause. The simplest version of availability increase is analogous to market expansion, when a platform becomes usable to an audience that was previously not connected or served. We can identify such an occurrence by first segmenting $\mathcal{A}$ into community groups $\mathcal{A}_c, c \in \mathcal{C}$. We then take time-varied measurements of the link count between each group and a given platform p_n. If that count is initially 0 and then increases at all, a Roll-out has occurred in the most extreme sense. More generally, if we see a sudden increase such that

$$|p_n^{t'}(a_n^c) - p_n^t(a_n^c)| > \tau$$

for some threshold τ, then a Roll-out has occurred. The inverse holds true. Using the same community groups $\mathcal{A}_c$, we can identify a Shutdown maneuver that restricts previous access as a sudden severe decrease in $v_{a_n^c, p_n}$ beyond some threshold. The threshold parameter τ provides reasonable certainty that the drop in user connections exceeds the amount that can be explained by belief- or preference-driven decision making within the time span between the network measurements.

An increase in access due to technical capacity increase implies the prior existence of some

technical bottleneck. For example, if a platform p_n is suddenly capable of handling more clients, and is doing so, it can be assumed that prior to the change, there were clients that wished to access the platform and could not do so. In the network this can be measured as an upper limit over time on $p_n(\mathcal{A})$, a maximum number of agent connections to the platform within an interval. If $p_n(\mathcal{A})$ is stable over multiple time intervals, and then suddenly increases, a technologically-enabled Roll-out may have occurred. Conversely, a sudden drop in a stable $p_n(\mathcal{A})$ indicates a possible Shutdown.

In this instance, there is still a chance that the shift in user count is due to user preference changes. We can eliminate this possibility by measuring the average value of confidence between users and the platform and watching for a commensurate change. If the confidence value remains stable while the link count changes, we can assume the changes are due entirely to a change in availability (be that connectivity or capacity).

Formally: Platform p was the target of a Roll-out

$$p \notin \mathcal{P}^t \text{ and } p \in \mathcal{P}^{t'}$$

This condition is sufficient to identify a Roll-out. Alternatively, a Roll-out on platform p has occurred at t' if:

$$\left|\overline{p^{t'}(\mathcal{A})} - \overline{p^{t}(\mathcal{A})}\right| < \epsilon$$

and

$$|p^{t'}(\mathcal{A}) - p^{t}(\mathcal{A})| < \tau_{st}$$

and

$$\left|p^{t''}(\mathcal{A}) - p^{t'}(\mathcal{A})\right| > \tau_{ch}$$

and

$$p^{t''}(\mathcal{A}) > p^{t'}(\mathcal{A})$$

for some stability threshold τ_{st}, some change threshold τ_{ch}, and some sensitivity threshold ϵ.

Platform p was the target of a Shutdown if

$$p \in \mathcal{P}^t \text{ and } p \notin \mathcal{P}^{t'}$$

This condition is sufficient to identify a Shutdown. Alternatively, a Shutdown on platform p has occurred at t' if:

$$\left|\overline{p^{t'}(\mathcal{A})} - \overline{p^{t}(\mathcal{A})}\right| < \epsilon$$

and

$$|p^{t'}(\mathcal{A}) - p^{t}(\mathcal{A})| < \tau_{st}$$

and

$$|p^{t''}(\mathcal{A}) - p^{t'}(\mathcal{A})| > \tau_{ch}$$

and

$$p^{t''}(\mathcal{A}) < p^{t'}(\mathcal{A})$$

for the same threshold definitions.

Note that I designated $p^t(\mathcal{A})$ as the number of users from on platform p at time t – in other words, the in-degree of p in an unweighted Agent-Platform network. A change in technical capacity may actually produce small shifts in the platform's user base, but significant changes in the volume accommodated by that platform. As such, some analysis may call for using $p^{*,t}(\mathcal{A})$, which denotes the number of interactions (here, messages) by with p at time t.

4.5.2 Recommend and Sideline

A Recommend maneuver increases a platform's relative proportion of traffic within the network; a Sideline maneuver decreases a platform's relative proportion of traffic. As indicated in chapter 3, these maneuvers can be achieved through both technical (e.g., DDOS) or non-technical (e.g., slander) means.

Because traffic proportion can be quantitatively measured, identifying the effects of a Recommend or Sideline maneuver is relatively simple. In an Agent-Platform network where links are weighted by the number of times Agent a used platform p, the normalized in-degree of p (taken as a sum of link-weights) reveals that platform's proportion of traffic within the network. A significant increase or decrease in a given platform's proportion would indicate a Recommend or Sideline maneuver may have been executed.

As with other maneuvers, the network metrics will not directly indicate *how* the maneuver was executed, whether through technical or decision-influence methods. A change in platform traffic might also be caused by the Roll-out and Shutdown maneuvers. Therefore, a surer indication of Recommend/Sideline is a change in platform proportion without a significant change in overall traffic volume. Note that this constraint has significant limitations. In my simulation, as in real life, events can excite communications activity, leading to a surge in traffic volume independent of any influencer's attempts to alter the IE. As such, for these maneuvers, I propose both weak and strong detection criteria.

Formally: for a Recommend to occur on platform p, it is necessary that

$$\text{indeg}^t(p) < \text{indeg}^{t'}(p)$$

for a Sideline to occur on platform p, it is necessary that

$$\text{indeg}^t(p) > \text{indeg}^{t'}(p)$$

The assertion of a Recommend or a Sideline for platform p is strictly stronger if:

$$\left| \sum_{p_n \in \mathcal{P}} \text{indeg}^t(p_n) - \sum_{p_n \in \mathcal{P}} \text{indeg}^{t'}(p_n) \right| < \epsilon$$

for some $\epsilon \approx 0$.

4.5.3 Raise and Silence

A Raise maneuver provides or expands an agent's control of, or reach within, a platform. A Silence maneuver removes or decreases an agent's control of, or reach within, a platform. Crucially, this change or reach or control is in service of exerted influence. Elevating a user's follower

count or algorithmic favor is an example of Raise; elevating a user's administrative privileges on that same system is *not* a Raise maneuver, unless that elevation is then leveraged to provide the user greater outcome-oriented influence over other users.

Identifying Raise and Silence maneuvers from message traffic can be accomplished in part by measuring the target agent's centrality within a platform. Specifically, if Agent a sees an increase in their PageRank centrality, it is possible that a has been the target of a Raise maneuver; the converse is true for a decrease and a Silence maneuver.

If availability influence attempts were the only factors at play, these maneuvers are easily identified. However, my intent is to merge availability maneuvers into the existing BEND framework, and BEND already contains maneuvers accounting for changes in agent centrality: the Back and Negate maneuvers alter "an actor's importance or effectiveness relative to a community or topic." Insofar as we define the users of a given platform as a community, these maneuvers converge with Raise and Silence in measurement.

Further, because BEND maneuvers do not specify the *means* by which an effect is achieved (just as my proposed maneuvers do not), the use of technical means to achieve a change in centrality – such as making a user a moderator, or banning/censoring them – still fit within the definition given for Back and Negate.

The key difference in the definitions – control of or reach within a platform, versus importance in an agent community – is extremely difficult to measure externally. Observed message traffic will rarely indicate a user's permission level within the messaging system, especially messages generated by public-facing services like X and Facebook. This slight difference may be irrelevant in the larger objective of measuring decision-oriented influence within the IE.

I tested this convergence empirically as part of my experiments, to be examined in greater detail in chapter 5. I induced a Raise condition in my simulation, and measured the resulting network in accordance with Hickman's BEND detection metrics [113]. In every case, Hickman's algorithm identified a Back maneuever. Though further investigation is warranted, I conclude that while a Back maneuver may not induce Raise effects, a Raise maneuver seems to necessarily induce Back effects; meaning, when examining the resulting network,

$$\text{Raise} \rightarrow \text{Back}, \text{Back Raise}$$

For this dissertation I will provide a formal definition of, and metrics to identify, Raise and Silence maneuvers. However, in the merged BENDRS framework, Raise and Silence are dropped in favor of Back and Negate.

Formally: Let $PR^t(a,p)$ denote the PageRank centrality of a on p. Let $\mathcal{V}_{\mathcal{A},p}^{priv,t}$ denote the set of interior links measuring Agent privilege on platform p. In the extreme case, a is the target of a Raise on p if

$$v_{a,p}^{priv,t} \notin \mathcal{V}_{\mathcal{A},p}^{priv,t} \text{ and } v_{a,p}^{priv,t'} \in \mathcal{V}_{\mathcal{A},p}^{priv,t'}$$

Similarly, in the extreme case, a is the target of a Silence on p if

$$v_{a,p}^{priv,t} \in \mathcal{V}_{\mathcal{A},p}^{priv,t} \text{ and } v_{a,p}^{priv,t'} \notin \mathcal{V}_{\mathcal{A},p}^{priv,t'}$$

These criteria are sufficient to identify the respective maneuvers. In the general case, a was the target of a Raise on p if

$$PR^t(a,p) < PR^{t'}(a,p)$$

or

$$v_{a,p}^{priv,t} < v_{a,p}^{priv,t'}$$

The second condition is sufficient to identify a Raise. Conversely, a was the target of a Silence on p if

$$PR^{t}(a,p) > PR^{t'}(a,p)$$

or

$$v_{a,p}^{priv,t} > v_{a,p}^{priv,t'}$$

The second condition is sufficient to identify a Silence.

4.5.4 Reveal and Stifle

A Reveal maneuver adds a topic to the IE, or to a platform within it. A Stifle maneuver removes a topic from the IE, or from a platform within it.

Recognizing Reveal and Stifle maneuvers is mathematically trivial but relies on the comparatively difficult task of creating a semantic network of topics from message traffic. The challenge of distilling a discrete set of topics from a conversational corpus has been explored in other work [220]. As such, my work assumes a properly populated topic set is already available in the tri-modal IE representation. Under this assumption identifying Reveal and Stifle maneuvers is simple.

Formally: r was the target of a Reveal if

$$r \notin \mathcal{R}^{t} \text{ and } r \in \mathcal{R}^{t'}$$

More specifically, r was the target of a Reveal on p if

$$v_{r,p} \notin \mathcal{V}_{\mathcal{R},p}^{t} \text{ and } v_{r,p} \in \mathcal{V}_{\mathcal{R},p}^{t'}$$

Similarly, r was the target of a Stifle if

$$r \in \mathcal{R}^{t} \text{ and } r \notin \mathcal{R}^{t'}$$

More specifically, r was the target of a Stifle on p if

$$v_{r,p} \in \mathcal{V}_{\mathcal{R},p}^{t} \text{ and } v_{r,p} \notin \mathcal{V}_{\mathcal{R},p}^{t'}$$

4.5.5 Repeat and Smother

A Repeat maneuver increases a topic's prominence or occurrence on a platform. A Smother maneuver decreases a topic's prominence or occurrence on a platform.

Prominence and occurrence are highly correlated in the social media space [154]; if viewed from the perspective of automated feed curation, the two may be synonymous. Measuring occurrence is mathematically simple. We can construct a Topic x Platform network with links weighted to indicate occurrence count and normalize these links per-platform to show the proportionate representation of each topic on that platform. A change in the link value for our target

topic would indicate the presence of a Repeat or Smother maneuver, depending on the sign of the change.

Measuring prominence is potentially more difficult, and as mentioned is often unnecessary. Given time, the occurrence and prominence of a topic will converge. There are some instances, however, where prominence may be partially disconnected from occurrence frequency:

- A moderated or content-controlled platform may limit or prohibit mentions of a specific topic that is nevertheless highly pertinent or salient to the user base. In this case, the topic remains prominent, in that references to or mentions of it may carry outsize weight.
- Similarly, exogenous social requirements may prevent the direct mention of a specific topic. This is akin to the "dog-whistle" phenomenon, in which groups use seemingly unrelated and anodyne topics to indirectly discuss the underlying taboo subject.
- In late-stage virality, a topic may "echo" as old messages continue to get engagement or persist in feeds, even though the topic itself is no longer pertinent within the larger IE. This is especially true in highly automated information environments.

There are two options to measure prominence directly, without using topic occurrence rates. First, it can be assumed that prominent topics will co-occur with other topics on the platform and create a co-occurrence network. The prominence of a given topic can then be measured via node centrality. This is especially valuable in the dog-whistle scenario described above: very few posts are likely to explicitly relate the many masking topics to the actual core topic, but the central nature of that topic would emerge in a network representation.

Second, we can calculate the average engagement per message for a given topic. This is useful in early- and late-stage virality. Messages going viral have extremely high engagement before duplication and imitation commence. Late-stage viral topics, by contrast, have comparatively low engagement, despite widespread repetition. By averaging engagement over all messages on a given topic, it can be intuited that a topic's relative prominence is largely independent of its occurrence.

Formally: Let $\mathcal{V}_{\mathcal{R},\mathcal{P}}^{freq,t} = \{v_{r_i,p_j}^{f,t}\}$ denote the links in a Topic-Platform network at time t, where link value is the proportion with which topic r_i occurs on platform p_j relative to all topics on p_j. Let $\text{eng}(m_{r_i,p_j}^t)$ denote the engagement count for message $m \in \mathcal{M}_\nabla$, the set of messages containing topic r, with timestamp t, on platform p_j, containing topic r_i. Let $\text{cen}^t(r,p)$ denote the centrality of topic r on platform p at time t.

Topic r has been targeted by a Repeat on platform p if:

$$v_{r,p}^{freq,t'} - v_{r,p}^{freq,t} > \epsilon_R$$

for some threshold ϵ_R, or

$$\text{cen}^t(r,p) < \text{cen}^{t'}(r,p)$$

or

$$\frac{1}{|\mathcal{M}_r^t|}\sum_{m\in\mathcal{M}_r^t}\text{eng}(m_p^t) < \frac{1}{|\mathcal{M}_r^{t'}|}\sum_{m\in\mathcal{M}_r^{t'}}\text{eng}(m_p^{t'})$$

Topic r has been targeted by a Smother on platform p if:

$$v_{r,p}^{freq,t} - v_{r,p}^{freq,t'} > \epsilon_S$$

for some threshold ϵ_S, or

$$\text{cen}^t(r,p) > \text{cen}^{t'}(r,p)$$

or

$$\frac{1}{|\mathcal{M}_r^t|}\sum_{m\in\mathcal{M}_r^t}\text{eng}(m_p^t) > \frac{1}{|\mathcal{M}_r^{t'}|}\sum_{m\in\mathcal{M}_r^{t'}}\text{eng}(m_p^{t'})$$

4.6 Simplified identification metrics using OPIEM

Ideally, detection of a given maneuver can be achieved, or reasonably approximated, through analysis of a highly reduced subnetwork derived from the OPIEM core. This makes maneuver identification computationally simpler and often offers a more intuitive explanation of what has occurred within the IE. This is especially valuable when combining my availability maneuvers with the extant BEND framework to create the BENDRS maneuver set.

In defining Roll-out and Shutdown above, I made use of a *community*, a subdivision $\mathcal{A}_c \subset \mathcal{A}$. Importantly, I did not specify *how* communities were assigned to agents. Several BEND maneuvers similarly assume the designation of communities, without specifying how those communities are to be designated. Hickmans' work uses topic-oriented groups (TOGs) derived from topology in the Agent-Topic network [113]. TOGs are an excellent method to cluster agents when dealing with the high degree of noise and ambiguity present in topic nodesets derived from live data.

Because GhostCell does not invoke natural language, topic identification is far less noisy than for real data. Using TOGs as a community identifier for GhostCell agents would fail, as the topic nodeset is both too sparse and too well-mapped; all agents would be clustered into the same TOG by Hickmans' algorithm. As such, I exogenously provide agents with community identifiers created in the agent generation process, corresponding to their various group memberships. I will identify the specific group method used whenever these groupings are involved in generating results.

In the proposed maneuver identification metrics below, I will use $\mathcal{C}$ to denote the set of communities, where $\mathcal{C}$ is a partition of $\mathcal{A}$.

4.6.1 Back – Negate

Effect: Alter the actual or perceived importance of an actor relative to a community or topic.

Targets: Actor + community (selection of community clustering method determines whether the effect focuses on a topic or an exogenously sourced community)

Network object measured: $\mathcal{A}_c \times \mathcal{A}_c$, weighted by count (exterior link) (The derivation of this network depends on the community identifier chosen. For instance, TOGs would utilize (roughly) $\mathcal{A} \times \mathcal{R} \times \mathcal{A}$, while user communities might use $\mathcal{A} \times \mathcal{P} \times \mathcal{A}$.)

Metric: Centrality of target agent

4.6.2 Build – Neutralize

Effect: Create or remove actual or perceived groups/communities.

Targets: Community

Network object measured: $\mathcal{C}$ (Again, derivation depends on clustering method)

Metric: Nodeset size

4.6.3 Bridge – Narrow

Effect: Interconnect or merge / disconnect or sunder existing groups/communities.

Targets: Communities

Network object measured: $\mathcal{C} \times \mathcal{C}$, unweighted

Metric: Nodeset size and target degree

4.6.4 Boost – Neglect

Effect: Increase / decrease the size or density of a group/community.

Targets: Community

Network object measured: $\mathcal{A}_c \times \mathcal{A}_c$, unweighted

Network metric: Network density, nodeset size, and average path length

4.6.5 Excite – Dismay

Effect: Elicit positive/negative emotional response in audience relative to a topic.

Targets: Agents + topic + emotion set

Network object measured: $\mathcal{A} \times \mathcal{R}$ Agita network (interior link)

Network metric: Average value of target emotion set between target Agents and topic

4.6.6 Explain – Distort

Effect: Provide detail and clarify / obscure a topic within a community.

Targets: Topic + community

Network object measured: $\mathcal{R} \times \mathcal{A}_c \times \mathcal{R}$, weighted by count

Network metric: Average weighted path length, target topic centrality

4.6.7 Engage – Dismiss

Effect: Alter a topic's relevance to an audience.

Targets: Agent(s) + topic

Network object measured: $\mathcal{A} \times \mathcal{R}$ Salience network (exterior link)

Network metric: Link weight between targets

4.6.8 Enhance – Distract

Effect: Expand / skew scope of a topic.

Targets: Topic

Network object measured: $\mathcal{R} \times \mathcal{A} \times \mathcal{R}$, weighted by count (exterior link)

Network metric: Ratio of target degree to mean neighbor degree

4.6.9 Roll-out – Shutdown

Effect: Alter platform availability.

Targets: Platform

Network object measured: $\mathcal{A} \times \mathcal{P}$ Volume network (exterior link)

Network metric: Size of $\mathcal{P}$, degree of target

4.6.10 Recommend – Sideline

Effect: Alter an audience's preference of a platform.

Targets: Platform + Agent(s)

Network object measured: $\mathcal{A} \times \mathcal{P}$ Preference network (exterior link)

Network metric: Link value between targets, degree of target platform

4.6.11 Reveal – Stifle

Effect: Alter topic availability or presence.

Targets: Topic + Platform

Network object measured: $\mathcal{R} \times \mathcal{P}$, weighted by count (exterior link)

Network metric: Link value between targets, degree of target topic

4.6.12 Repeat – Smother

Effect: Alter a topic's prevalence on a platform.

Targets: Topic + Platform

Network object measured: $\mathcal{R} \times \mathcal{P}$ Occurrence network (exterior link)

Network metric: Link value between targets

4.7 Conclusion

With GhostCell, I can produce simulated message traffic for a reasonably realistic IE. That traffic can be rendered into an OPIEM model as described in this chapter. That model can, in turn, be analyzed, to describe the IE and to identify the presence of influence maneuvers. Having established the formal availability metrics, and with the quicker OPIEM-oriented BENDRS metrics, the results of my simulations can now be presented.

Chapter 5

Results and Analysis

At this point I have: defined the OPIEM framework; identified some availability-based influence maneuvers that OPIEM can capture; and designed a simulation with sufficient breadth to demonstrate OPIEM functionality. It remains to conduct experiments within the simulation to verify my assertions.

5.1 Baseline scenario

I modeled my simulated populace on the US and selected issues and community features accordingly. As previously described, I grounded my scenario in six foundational topics, assumed to be orthogonal. I then selected six large-scale domestic debates and decomposed each of them into multiple narratives, with each narrative mapped to a single foundational topic. In this way, large issues are not decided along any single opinion axis. In total, the simulation included 35 narratives.

For clarity I review the nomenclature here:

- Agents in the simulation store beliefs on six orthogonal **topics**. These topics are not explicitly visible in message traffic.
- Discussion in the simulation is roughly organized into large **issues**, which simulate broad societal debates.
- Issues are decomposed into **narratives**; a narrative is a mapping of an issue onto a topic.

As an example: for the **issue** of immigration, the **narrative** arguing that immigration is part of our national identity maps onto the Community **topic**. It is a pertinent discussion of the broader issue and is connected to the agent's beliefs regarding how to define one's community (us vs them).

For each issue, I randomly generated 12 pieces of evidence, representing popular memes, articles, or images that would be exchanged or referenced in discussion; I randomly mapped evidence to narratives, such that narratives within the same issue shared the same evidence, but with different likelihood of selection. This captures the phenomenon of different viewpoints using the same evidence with varying interpretations and allows the use of evidence (images, URLs, etc.) to identify larger themes or narratives within the populace.

I created five social media platforms roughly analogous to Facebook, X/Twitter, Reddit, LinkedIn, and Instagram. Additionally, I created five archetypal media outlets: one left-leaning and one right-leaning mainstream outlet, one far-left and one far-right outlet, and one neutral international outlet. (For most runs, I did not use all platforms simultaneously, to avoid overly diffusing the population.)

I populated the simulation with agents drawn from five categorical archetypes: progressive urban, traditional populist, institutional managerial, working-class pragmatist, and marginalized. A 2022 study indicated that, on average, US adults consume six separate sources of news, including social media [230]. Accordingly, I assigned my agents an average of six total accounts across the platforms in the scenario, with preferences for social media over traditional media driven by demographics. Social media follower counts were generated from a distribution also parameterized by demographics.

An outline of the simulation setup is provided in Table 5.1. Additional details regarding distributions, parameters, and other simulation mechanics are contained in the appendices.

Table 5.1: Simulation configuration, baseline

Variable	**Type**	**Range**	**Value**
$\|\mathcal{A}\|$, number of agents	Independent	$\mathbb{N}$	300
$\|\mathcal{P}\|$, number of Platforms	Independent	$\mathbb{N}$	10
$\|\mathcal{R}\|$, number of Topics	Control	$\mathbb{N}$	4
$\|\mathcal{N}\|$, number of Narratives/Issues	Independent	$\mathbb{N}$	35 narratives on 6 issues
$\|\mathcal{C}\|$, number of communities	Control	$\mathbb{N}$	5
G_I, number of institutional groups	Control	$\mathbb{N}$	5
G_C, number of community groups	Control	$\mathbb{N}$	3
G_A, number of trait homophily groups	Control	$\mathbb{N}$	5
G_P, number of opinion homophily groups	Control	$\mathbb{N}$	4
$\overline{\phi_{\mathcal{C}}(r)}$, average opinion on topic r in community $\mathcal{C}$	Dependent	[-1,1]	Compares t_0 and t_{max} values
$\sigma_{r,\mathcal{C}}$, standard deviation of opinions on topic r in community $\mathcal{C}$	Dependent	$\mathbb{R}$	Compares t_0 and t_{max} values
$s_{r,\mathcal{C}}$, range of opinion values on topic r in community $\mathcal{C}$	Dependent	[0,2]	Compares t_0 and t_{max} values
$\overline{\psi_{\mathcal{C}}(p)}$, average preference for platform p in community $\mathcal{C}$	Dependent	[0,1]	Compares t_0 and t_{max} values
$\hat{n}_p$, narrative n's proportion of traffic on platform p	Dependent	[0,1]	Compares t_0 and t_{max} values
$\hat{n}_{\mathcal{C}}$, narrative n's proportion of traffic in community $\mathcal{C}$	Dependent	[0,1]	Compares t_0 and t_{max} values
$\hat{p}_{\mathcal{C}}$, proportion of traffic on platform p in community $\mathcal{C}$	Dependent	[0,1]	Compares t_0 and t_{max} values

5.2 Baseline analysis

I simulated 400 agents and 5 media outlets interacting over 4 social media platforms for 250 hours ($\tilde{1}$0 days). I considered two primary variables of interest:

- Agent valence, per topic. Valence represents the opinion or belief of that agent on that orthogonal underlying topic, and is the value updated when agents consume media. Shifts in valence represent the agent being influenced by the information environment.
- Agent preference, per platform. Preference represents an agent's opinion toward a platform relative to other platforms and is a function of the agent's trust in that platform, their perception of the platform's opinion alignment, and their comfort with the platform's user experience.

5.2.1 Topic valence

As seen in Figure 5.1, global opinion across the six topics slowly converges as standard deviations fall, though some topics converge more quickly than others. Valence value changes over the run of the simulation vary, from a high of approximately 0.05 to a low of near 0. None of the standard deviations fall below 0.3, meaning individual opinion variance is still fairly high within the populace.

When measured within specific communities, valence shifts show greater variety. Figure 5.2 shows the statistics for the TradPop community, and Figure 5.3 shows the statistics for the WorkPrag community. Both populations undergo significant position shifts, as measured by average valence. However, the TradPop community is fairly united – standard deviations for topics either start and stay low, or else decline significantly over the run. By comparison, WorkPrag standard deviations stay higher, decline more slowly, and in one case actually increase over the course of the run, indicating a population torn between two strong points of view.

5.2.2 Platform preference

Platform preference shifts over time as well. At the global level, one platform becomes more popular as prominent voices favor it, as seen in Figure 5.4.

Within communities, platform preference evolves more clearly as initial conditions are altered by agent input. In Figure 5.5, the InstMang community sees a dominant platform become increasingly dominant, while in Figure 5.6 the Margin community becomes increasingly less committed to any single platform.

Notably the standard deviation for preference increases as the simulation runs (Figure 5.7). Through interaction the populace becomes increasingly uncertain of the respective platforms.

5.2.3 Network structure

The simulated population exhibits reasonably realistic engagement behavior. The diagram below is the agent-to-agent network for one platform, captured on day 6 of the simulation. The network shows realistic structure driven by preferential attachment and homophily-weighted influence (Figure 5.8.

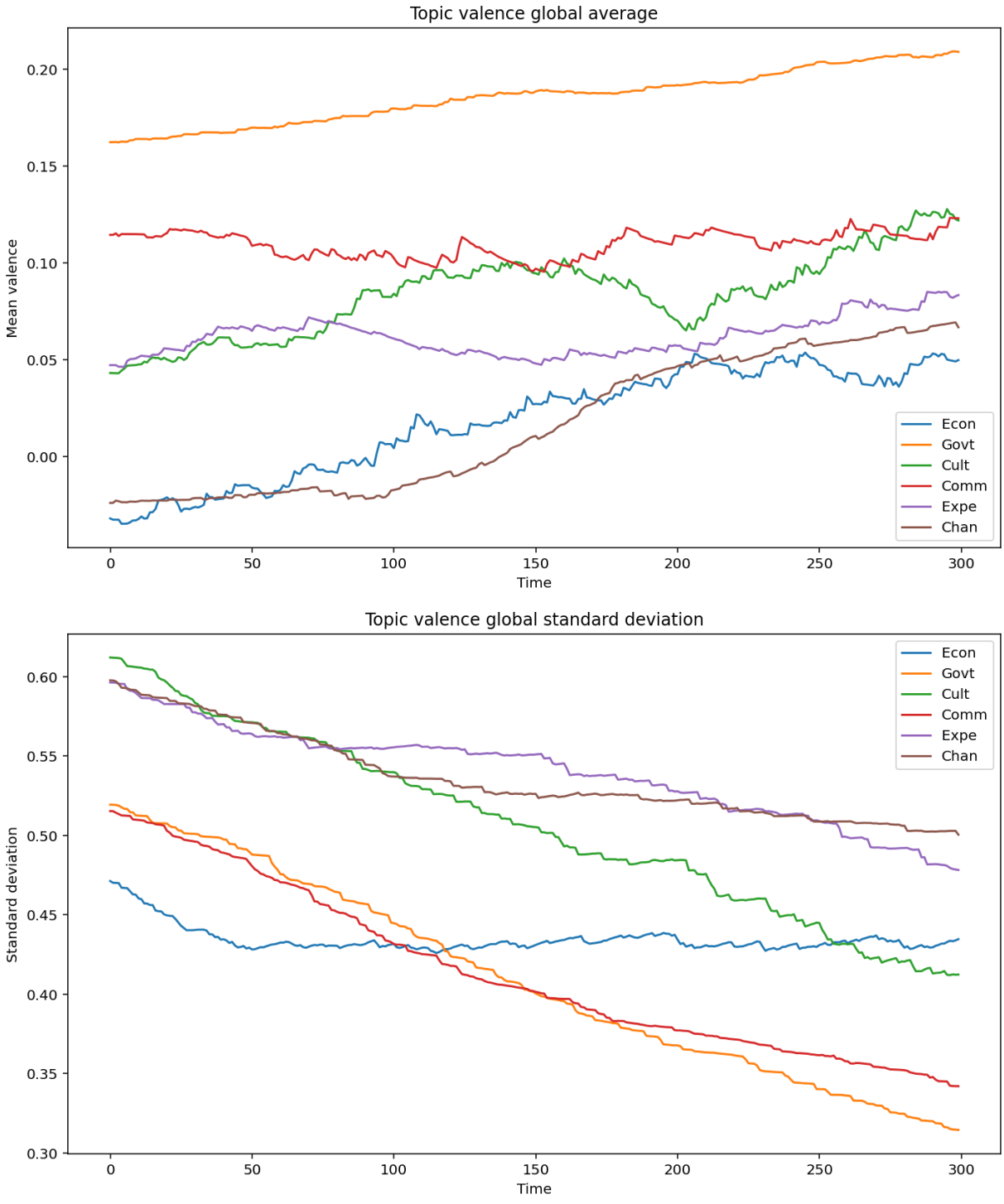

Figure 5.1: Valence statistics for a single baseline simulation run.

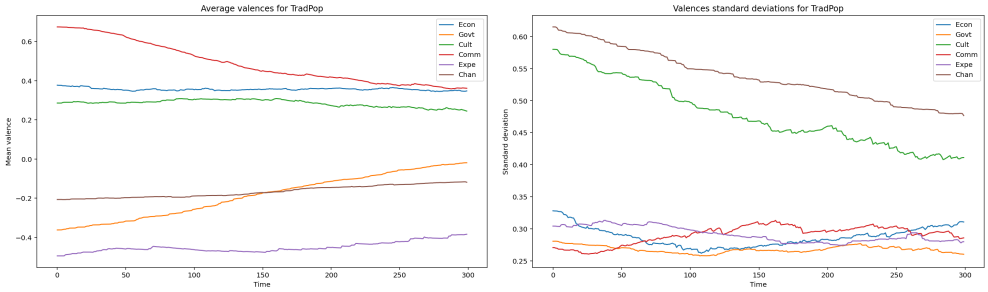

Figure 5.2: Valence statistics of the TradPop community over a single baseline simulation run.

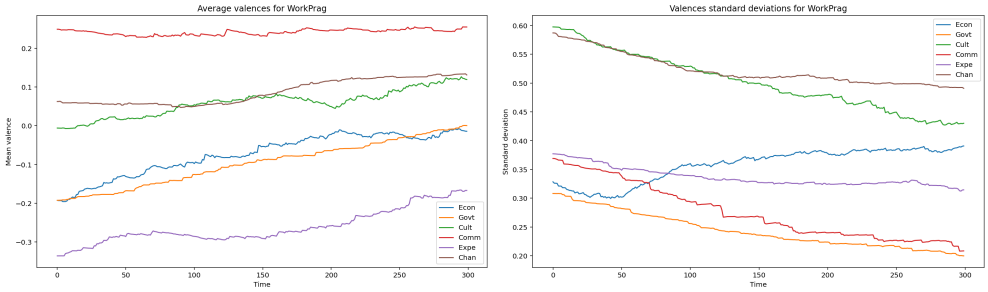

Figure 5.3: Valence statistics of the TradPop community over a single baseline simulation run.

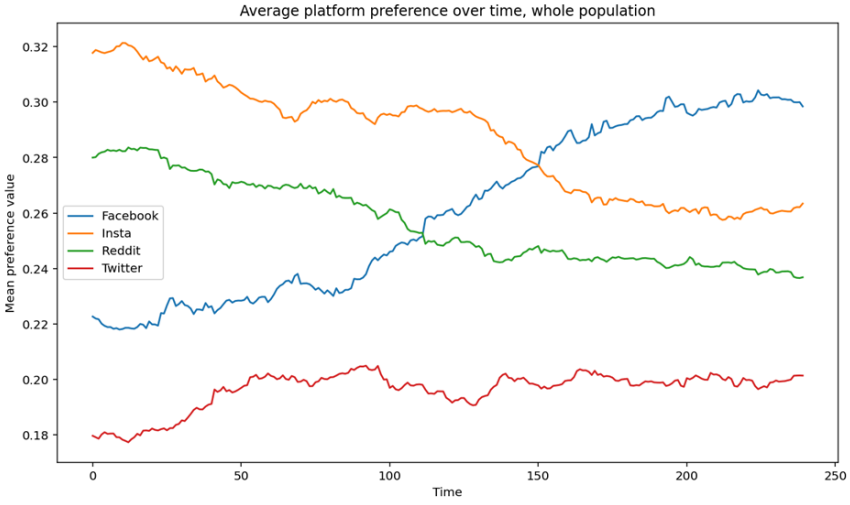

Figure 5.4: Global preferences for various social media platforms over a single baseline simulation run.

5.3 Verifying availability maneuvers

To replicate a concerted influence campaign, I used a combination of events and manually created agents called 'monsters.' Monsters are thus named because I assign them deliberately outsized and aggressive behaviors and give them deliberately influential positions within the social networks. Monster agents allow me to rapidly induce effects in the simulator that, in real life, would

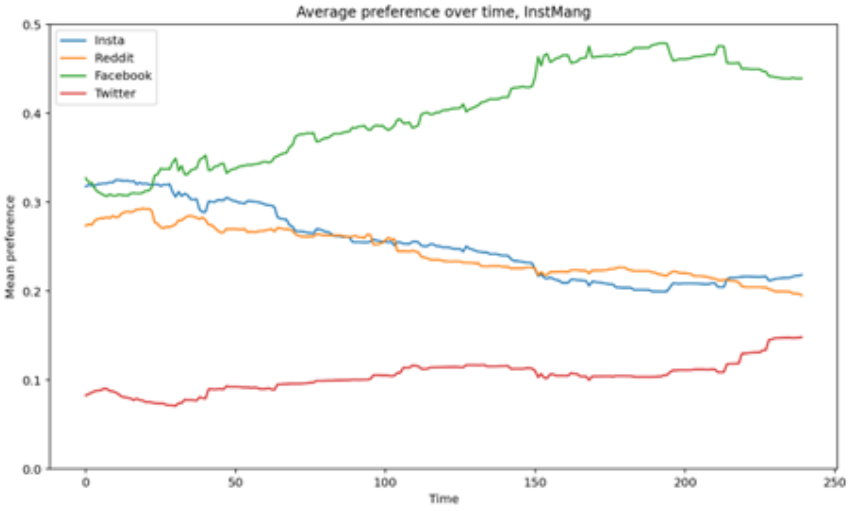

Figure 5.5: Average platform preference of the InstMang community over a single baseline simulation run.

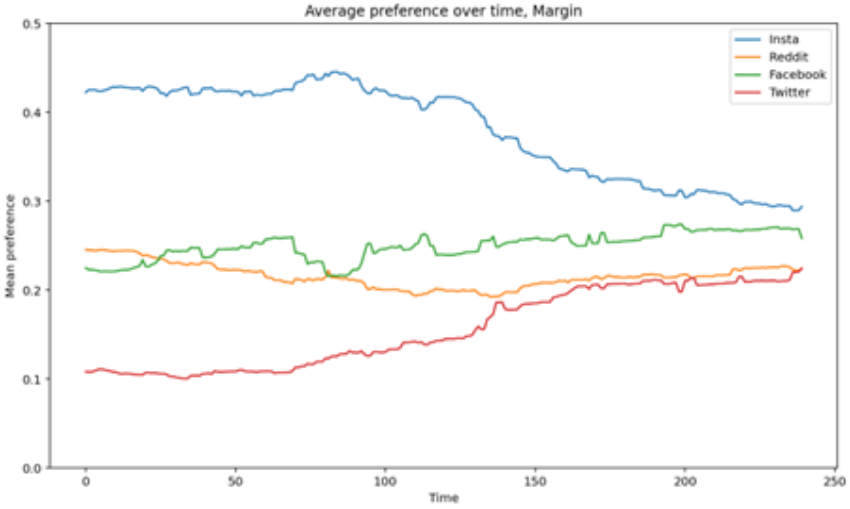

Figure 5.6: Average platform preference of the Margin community over a single baseline simulation run.

be cumulative and slow growing. On a real platform, the behaviors exhibited by monsters – such as extremely frequent posting, highly polar messaging, and relatively high follower counts – would likely not be tolerated by actual users or platforms.

For some experiments, I also added or altered communities and platforms to deliberately create a more exaggerated testbed and produce higher contrast to observe experimental outcomes. Unless otherwise specified, however, all conditions are identical to the baseline. In each case, I attempted to create a realistic and face-valid scenario in which to simulate and measure the maneuver of interest.

Table 5.2 provides the experimental conditions and replications per experiment. These con-

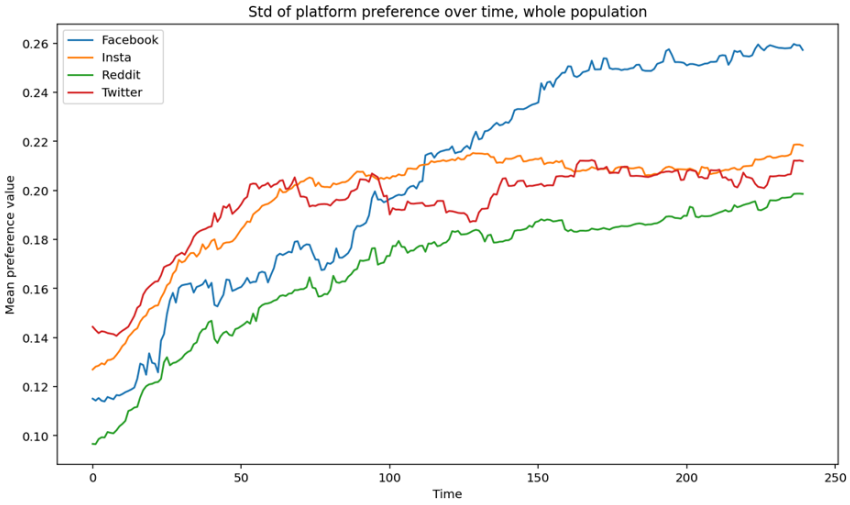

Figure 5.7: Standard deviation of platform preference over a single baseline simulation run.

ditions are explained in greater detail in each experiment's respective summary. The experiment inputs vary in three ways: the narrative set, the platform set, and the event set. (Reveal and Engage/Dismiss have additional experimental factors which are documented in their respective descriptions.) In presenting my results, I provide:

- **Graphical examples:** Time-series diagrams from *one* of the runs for greater examination/discussion; I took care to use a representative run rather than an unusual one
- **Maneuver detection measures:** Quantitative OPIEM maneuver detection results averaged over all runs
- **Statistical results:** Paired-sample t-test statistics on the measured outcomes of interest

Table 5.2: Experiment summary

Experiment	**Narrative set**	**Platform set**	**Event set**	**Replications**
Roll-out	$\mathcal{N}_0$	$\mathcal{P}'_0$	$\mathcal{E}_{RO}$	48
Shutdown	$\mathcal{N}_0$	$\mathcal{P}_0$	$\mathcal{E}_{SH}$	48
Recommend + Sideline	$\mathcal{N}_0$	$\mathcal{P}_0$	$\mathcal{E}_{RS}$	48
Reveal	$\mathcal{N}_{RV}$	$\mathcal{P}_{0T}$	$\mathcal{E}_{RV}$	48
Stifle	$\mathcal{N}_{ST}$	$\mathcal{P}_0$	$\mathcal{E}_{ST}$	48
Repeat	$\mathcal{N}_{0R}$	$\mathcal{P}_0$	$\mathcal{E}_{RP}$	48
Smother	$\mathcal{N}_0^-$	$\mathcal{P}_{SM}$	$\mathcal{E}_{SM}$	48
Back + Negate	$\mathcal{N}_0$	$\mathcal{P}_0$	$\mathcal{E}_{BN}$	48
Engage + Dismiss	$\mathcal{N}_{ED}$	$\mathcal{P}_0$	$\mathcal{E}_{ED}$	48

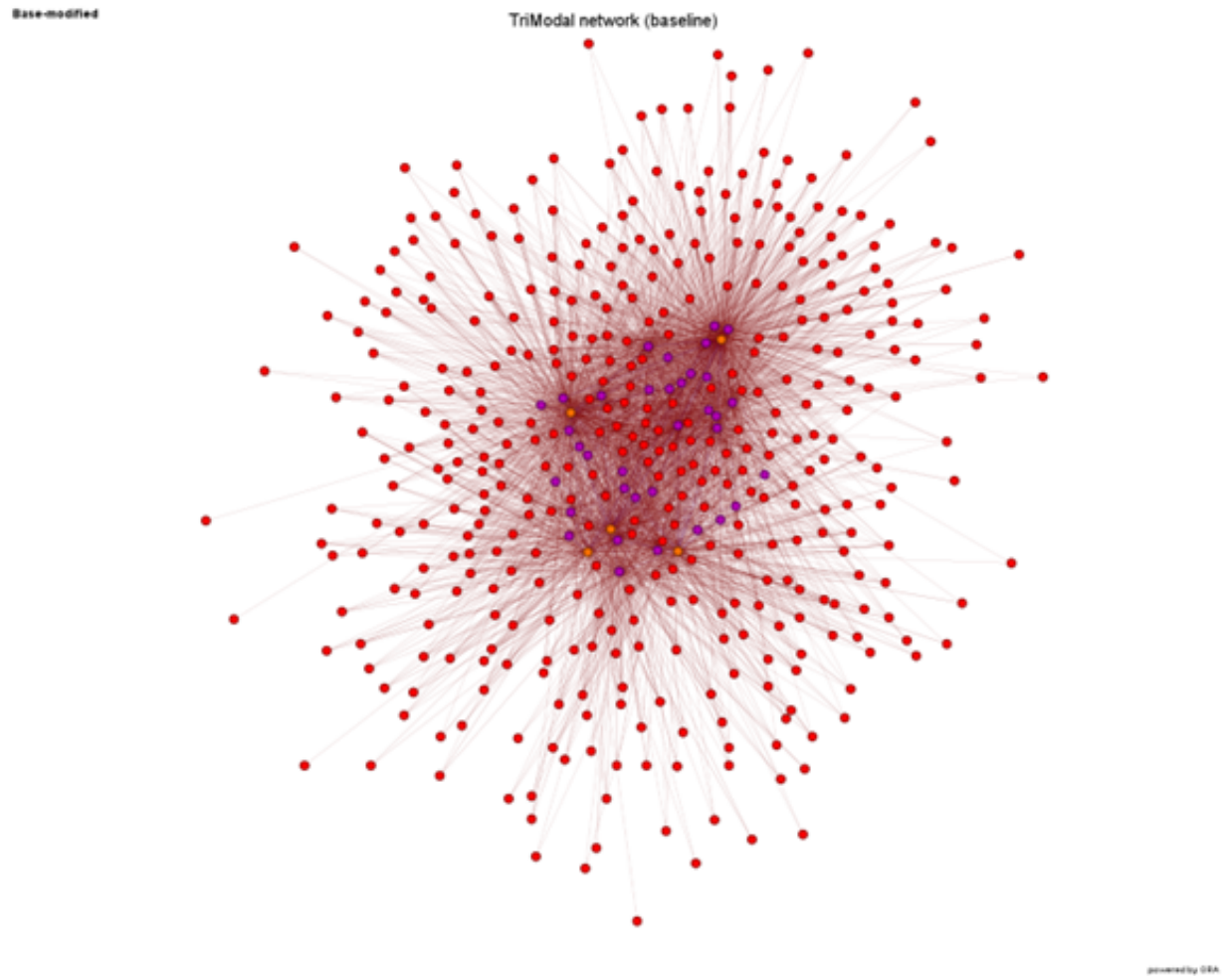

Figure 5.8: Tri-modal network diagram of simulated IE (baseline conditions).

5.3.1 Statistical methodology

To produce my t-test statistics, I ran each experiment 48 times. For each run, I generated a new random seed value for all stochastic function calls. I generated an entirely new population of 300 agents using this seed. I then ran the simulation twice, each time for 250 time steps. In the first run, I simulated the population's behavior in the *absence* of the experimental condition, producing a control run. In the second, I simulated the population's behavior *with* the experimental condition, producing a treatment run. I then captured the values of interest for both the control and treatment runs, and used each of these paired values to run a paired t-test. In all cases the test has N=48 samples and thus 47 degrees of freedom.

For some experiments, I measure and test the standard deviation of valence values within the populace. Note that I am not directly comparing the variance of two populaces, but rather testing for change in variance among individual populaces when subjected to the experiment, for multiple separate populaces. I confirmed that these changes in variance, within any given community or averaged across all communities, are normally distributed using the Shapiro-Wilk

test for normality. Because they are normally distributed paired results, the paired t-test is an appropriate test to confirm significant experiment impact.

In the description of each experiment I describe the experimental condition and specify the outcomes measured. In general, effect sizes were small, often on the order of 0.1 valence points. I attribute this to the timescale of the experiment. Wild opinion swings in a 250-hour time window would be unrealistic, and the gated influence dynamics I implemented correctly prevent such behavior. I theorize that, were I to run my experiment to a long end point (t=500 or 600), effect sizes would increase accordingly, representing a change in the system FJ convergence targets.

5.3.2 Roll-out

Setup

Table 5.3 summarizes the Roll-out experiment: roughly, I simulate the introduction of a new social media platform, analogous to the case study examined in chapter 3. Three social media platforms are initially active with generally left-leaning biases. This produces a situation in which right-leaning communities are not aligned with any available platform. At t=50, an additional platform activates with a right-leaning bias. The simulation mechanics then dictate if any users join the new platform and/or leave the legacy platforms.

Table 5.3: Simulation configuration, Roll-out

Variable	Type	Value
$\mathcal{N}$, narrative set	Independent	$\mathcal{N}_0$, Baseline narrative set
$\mathcal{P}$, platform set	Independent	$\mathcal{P}'_0$: Baseline platform set, Twitter initially disabled
$\mathcal{E}$, event set	Independent	Roll-out event: Addition of new platform at $t = 50$

Maneuver detection

Because I do not simulate technical capacity within my platforms, I could only use the trivial Roll-out criteria. The observed increase of the available platforms clearly indicates that a Roll-out took place.

Influence effects

Figure 5.9 shows the observed average valences for each topic without the Roll-out (dotted line) and with the Roll-out (solid line). All topic valences were impacted by the emergence of a new platform, with the Community topic (Comm) especially susceptible. Notably it shifted positively, toward a more right-leaning position, indicating that the emergence of a right-leaning platform gave that position greater influence within the populace. Similarly the Change topic (Chan) shifted negatively, a more conservative position.

The effect is more pronounced at the community level. For example, the Margin community is not perfectly aligned with the stereotypical right and left communities (TradPop and ProgUrb, respectively). When the new platform appears and users self-select into more aligned (and more stove piped) ecosystems, members of a "grey" middle-ground community find themselves in environments with less viewpoint diversity. They are subsequently more affected on certain topics, as seen in Figure 5.10: members of the Margin community, a "middle-ground" group, had moderate changes in valence trajectory on the Govt, Comm, and Cult topics, and an extreme change in trajectory on the Expe topic. These changes all have root in the presence of an additional platform and the resulting redistribution of opinion exposure.

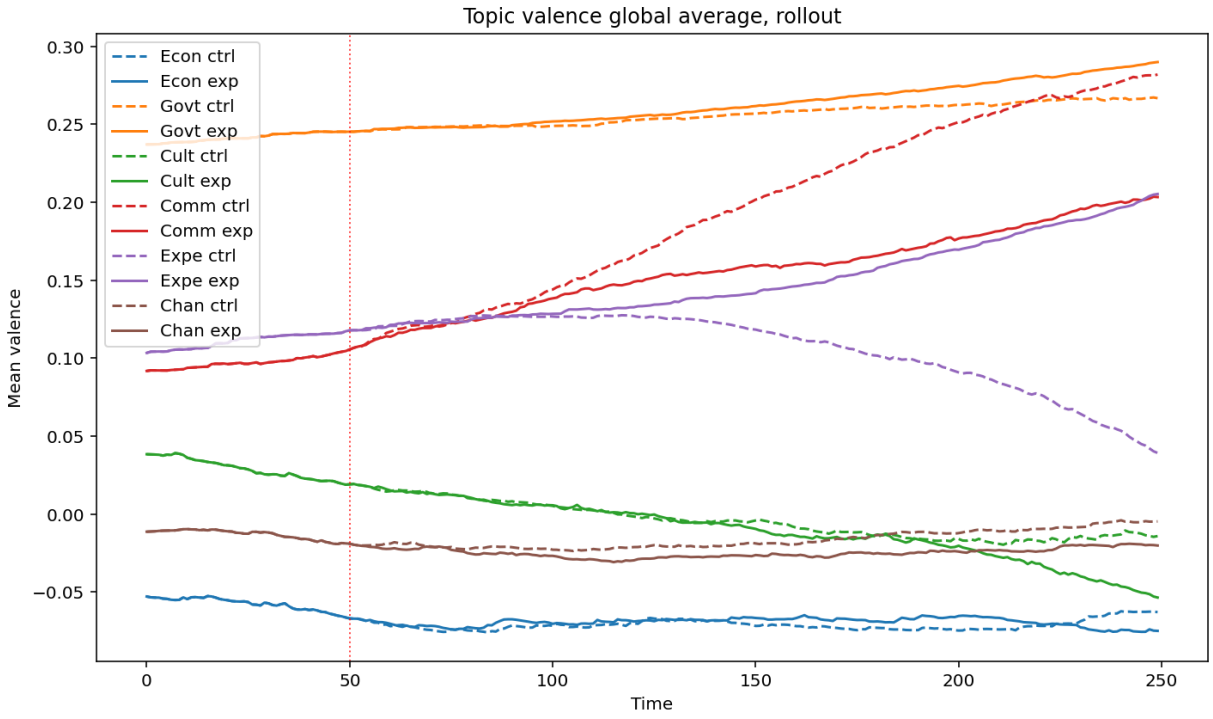

Figure 5.9: Comparison of global valence values with and without a Roll-out maneuver (red hash).

Table 5.4 presents the statistical results. I measured two outcomes in the simulation:

- Traffic proportion (platform), the *change* in the proportion of messages in the final corpus a specified community generated on a specified platform *between the control and treatment simulation runs*. This is a means of detecting adoption of the new platform within the populace.
- Traffic preference, the *change between the control and treatment runs* in the average numeric preference score of a specified community for a specified platform. This is means of more directly detecting the populace's perception of a platform.

Note that, because platform use is a stochastic function of platform preference, these two outcomes are inherently correlated.

Traffic proportions shifted significantly across all platforms as the new platform vied for user

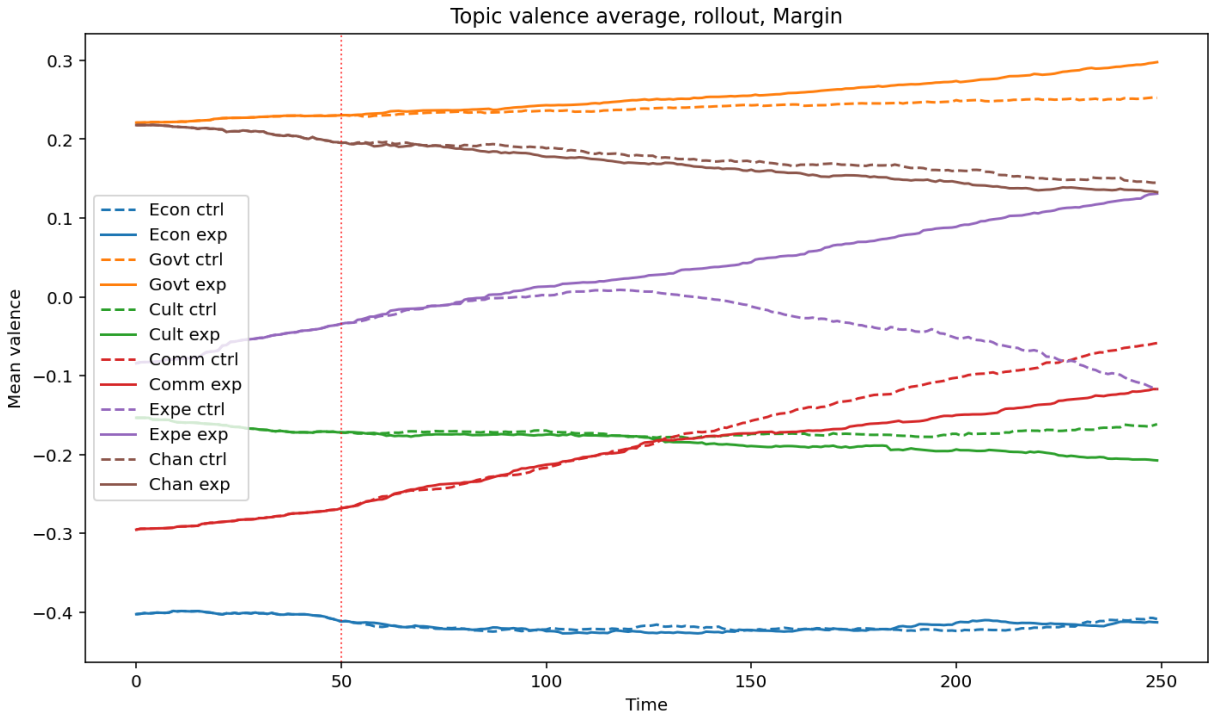

Figure 5.10: Community-level valence showing change due to Roll-out (red hash).

attention. The proportion of traffic captured by Facebook and LinkedIn decrease significantly as users switched to Twitter; the effect on Facebook was especially pronounced in the WorkPrag community, one of the marginally aligned communities between Facebook and Twitter. Preference scores – aggregates of user perceptions of trustworthiness, accessibility, and alignment – significantly decreased for every platform as Twitter emerged. (Note that these scores are zero-sum, so a general decrease is to be expected, though not guaranteed to be significant in any specific subpopulation.) Expectedly, the TradPop community showed significant decreased preference for Insta and Facebook, the two platforms least aligned with that community's positions. Interestingly, the Margin and WorkPrag communities shifted preference away from LinkedIn, the next-most conservative platform in the set, showing increased interest in Twitter as an alternative.

As theorized, the Roll-out maneuver is detectable within the OPIEM framework and provides significant influence within the information environment.

5.3.3 Shutdown

Setup

Table 5.5 summarizes the Shutdown experiment: roughly, I simulate the sudden shuttering of a major news outlet. Five news media platforms initially exist: Far Left, Mainstream Left, Center, Mainstream Right, and Far Right. In the experiment condition, at t = 60, the Mainstream Left media platform is deactivated.

Table 5.4: Roll-out statistical results, N=48

Outcome	**Grouping**	t **score**	p **value**
Traffic proportion (platform)	Facebook, WorkPrag cmty	-1.703	0.094
	Facebook, all cmtys	-2.291	0.024
	LinkedIn, all cmtys	-2.489	0.015
Traffic preference	Insta, TradPop cmty	-1.669	0.099
	Insta, all cmtys	-2.263	0.026
	Reddit, all cmtys	-1.790	0.077
	Facebook, TradPop cmty	-1.691	0.094
	Facebook, all cmtys	-2.476	0.015
	LinkedIn, Margin cmty	-1.843	0.068
	LinkedIn, WorkPrag cmty	-1.993	0.049
	LinkedIn, all cmtys	-2.855	0.005

Table 5.5: Simulation configuration, Shutdown

Variable	**Type**	**Value**
$\mathcal{N}$, narrative set	Independent	$\mathcal{N}_0$, Baseline narrative set
$\mathcal{P}$, platform set	Independent	$\mathcal{P}_0$: Baseline platform set
$\mathcal{E}$, event set	Independent	Shutdown event: Deactivation of platform at $t = 60$

Maneuver detection

As with Roll-out, the scenario permits me only to use the trivial criteria to recognize that a Shutdown has occurred.

Influence effects

Simulated agents seek to maintain a roughly constant flow of news; when a news source is removed from the environment, subscribing agents seek the next best fit to replace it. As Figure 5.11 shows, the Shutdown of LeftNews created an influx of new subscribers for all remaining platforms, from across the populace. By the end of the simulation, nearly all of the agents were subscribed to IntlNews.

Two communities had LeftNews as their top media outlet by volume. InstMang, a more neutral community, primarily shifted toward the neutral IntlNews outlet, while remaining heavily subscribed to RightNews. By contrast, the left-leaning community ProgUrb saw significant increases in subscriptions to IntlNews, RightNews, and ExtLeft, meaning members were more exposed to opposing or extreme media.

As seen in Figure 5.12, ProgUrb saw a change in slope on the Culture and Community topics, attributable to increased exposure to ExtLeft positions. By contrast, the TradPop community had LeftNews as its second-lowest source, and the removal of the platform had virtually no effect on the community's opinion trajectory.

Less entrenched communities were more dramatically influenced by the Shutdown (Figure 5.13). The WorkPrag community in particular saw a significant change, with Culture, Community, and Expertise all shifting measurably in the absence of LeftNews.

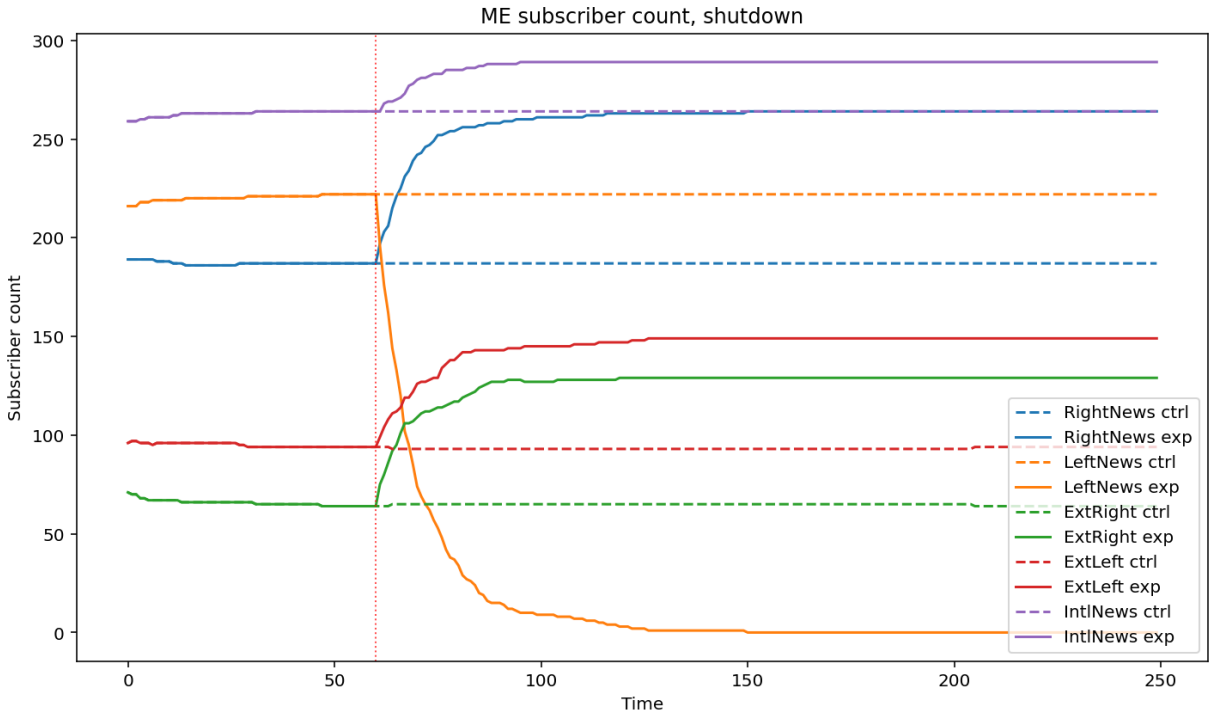

Figure 5.11: Net subscriber count to media platforms with (solid) and without (dotted) Shutdown.

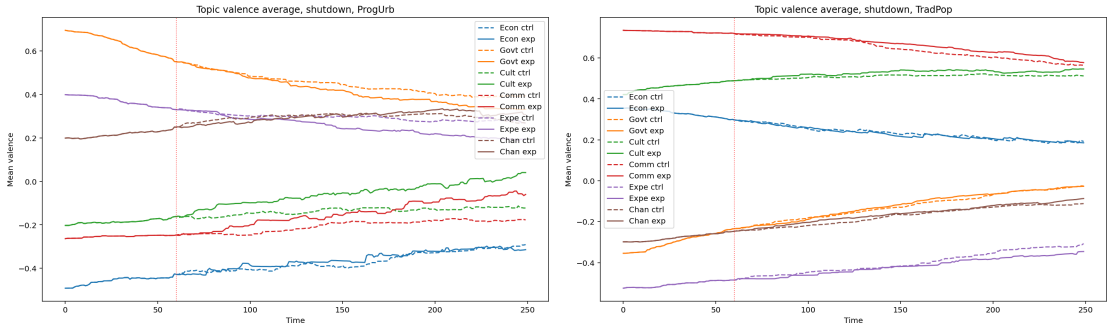

Figure 5.12: Selected community topic valences with (solid) and without (dotted) shutdown.

As theorized, the Shutdown maneuver is detectable within the OPIEM framework and provides significant influence within the information environment. Table 5.6 presents the statistical results. I measured two outcomes in the simulation:

- Mean valence value, the *change between the control and treatment simulation runs* in the average valence on a specified topic within a specified community. This outcome detects influence impacts, the cumulative effect of influence pressures.

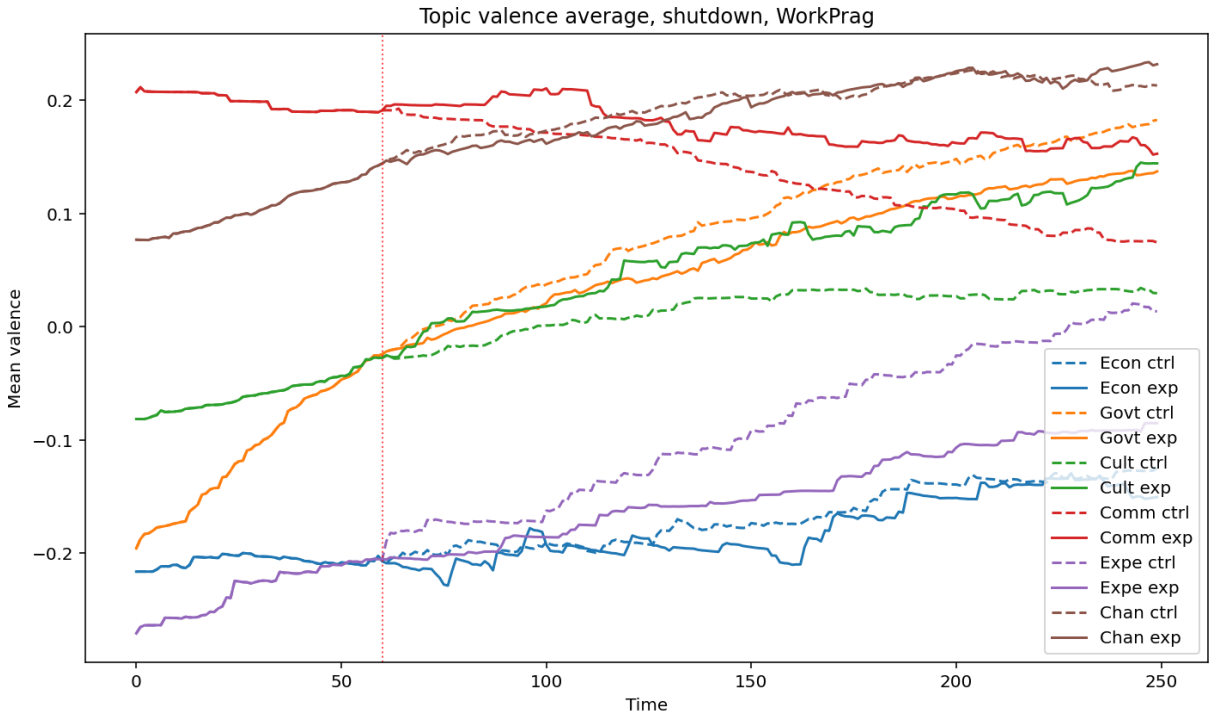

Figure 5.13: Selected community topic valences with (solid) and without (dotted) shutdown.

- Valence standard deviation, the *change between the control and treatment runs* in the standard deviation of valence values on a specified topic within a specified community. This is a measure of changes in uniformity of opinion within the target group.

Significant valence shifts are all in the direction of the ExtLeft platform, indicating that in the absence of a more moderate voice, extreme positions gained influence among the populace. As anticipated, the effect size is smallest with the TradPop community, and is often insigifnicant in that group; it is most pronounced overall in the partially-aligned WorkPrag group, which was more likely to move to the ExtLeft than to the IntlNews as an alternative based on its partial alignment to both platforms. LeftNews' shutdown also altered variance among the populace. Among the TradPop community, dissent widened about Culture and Community, likely due to increased exposure to more extreme positions on those issues. Conversely, the collapse of the left-leaning news to a single source reduced variance in Margin, WorkPrag, and ProgUrb on issues where LeftNews and ExtLeft had the greatest positional separation; this indicates that the extreme position of ExtLeft was gaining more influence against the initially more mainstream average position of these groups.

Table 5.6: Shutdown statistical results, N=48

Outcome	**Topic**	**Community**	t **score**	p **value**
Mean valence value	Econ	ProgUrb	-1.831	0.070
		WorkPrag	-1.871	0.064
		InstMang	-1.744	0.084

Table 5.6: Shutdown statistical results, N=48

Outcome	Topic	Community	t score	p value
		All cmtys	-2.198	0.030
	Govt	WorkPrag	2.287	0.024
		All cmtys	1.907	0.059
	Cult	Progurb	-3.821	0.001
		Margin	-3.434	0.001
		TradPop	-2.311	0.023
		WorkPrag	-3.754	0.001
		InstMang	-3.907	0.000
		All cmtys	-4.119	0.000
	Comm	Progurb	-3.676	0.000
		Margin	-3.315	0.001
		TradPop	-1.780	0.078
		WorkPrag	-3.816	0.000
		InstMang	-2.913	0.004
		All cmtys	-4.136	0.000
	Expe	Margin	3.347	0.001
		TradPop	1.752	0.083
		WorkPrag	3.919	0.000
		All cmtys	3.106	0.002
Valence std deviation	Econ	WorkPrag	-1.889	0.062
	Cult	Margin	-1.721	0.088
		TradPop	2.289	0.024
	Comm	ProgUrb	-2.936	0.004
		Margin	-2.322	0.022
		TradPop	2.096	0.039
	Expe	ProgUrb	-2.186	0.031
	All topics	All cmtys	-2.387	0.019

5.3.4 Recommend and Sideline

As discussed in chapter 3, either of these maneuvers tacitly induces the other; identification depends on designating the targeted platform and, where appropriate, audience. As such, I simulate both maneuvers simultaneously but utilize separate network metrics in detection.

Setup

Table 5.7 summarizes the Recommend/Sideline experiment: I simulate a "Fake News" media preference campaign, in which some outlets are denigrated while others are promoted, similar to the case study examined in chapter 3. At $t = 60$, an exogenous event (simulating a concerted ad/messaging campaign) targets all populations, reducing trust (Sideline) in left-leaning (LeftNews) and centrist (IntlNews) outlets, while boosting trust (Recommend) in right-leaning

(RightNews) media. Current simulation mechanics do not facilitate targeting exogenous trust boosts at any single audience or population segment; therefore, this is a globally-effective ad campaign. Trust gains are still be moderated by the agent's perceived alignment with the platform (i.e. agents will not be "forced" to immediately accept a disfavored platform).

Table 5.7: Simulation configuration, Recommend & Sideline

Variable	**Type**	**Value**
$\mathcal{N}$, narrative set	Independent	$\mathcal{N}_0$, Baseline narrative set
$\mathcal{P}$, platform set	Independent	$\mathcal{P}_0$: Baseline platform set
$\mathcal{E}$, event set	Independent	Recommend event: Global trust in target platform increases at t=60
		Sideline event: Global trust in target platforms decreases at t=60

Maneuver detection

Recommend and Sideline manifest in the network as changes in the link weight between the target Platform and Agent(s), and in changes to the target Platform's degree centrality.

I simulated Recommend and Sideline maneuvers targeting specific platforms for a broad, non-specific audience. Table 5.8 shows the pertinent network measurements from the Agent x Platform Preference network, in which the agent's links are normalized across all platforms to reveal agent preference. The timeframes denote the intervals over which the networks are formed, $t_0 = [0, 59]$ and $t_1 = [60, 249]$. To normalize against the non-standard timespan, each "hit" from an agent to a platform is normalized by the interval span, producing a pre-normalization link value in hits-per-timestep; these links are then normalized per-agent to create the preference network. Figure 5.14 shows the network structure captured from a sample run just before the treatment and then again at the end of the simulation.

Table 5.8: Averaged network metrics on Agent X Platform Preference network

Metric measured	**LeftNews**	**IntlNews**	**RightNews**
Target of:	Sideline	Sideline	Recommend
t_0 in-degree	79.274	89.292	55.752
t_1 in-degree	69.308	43.394	95.776
Δin-degree	-9.966	-45.898	40.024

By both the weak and strong criteria from chapter 4, the Recommend and Sideline maneuvers are present in the network.

Influence effects

As expected the Recommend and Sideline maneuvers drove significant changes in the subscriber bases for all platforms, both for the targeted platforms (IntlNews and LeftNews negatively, RightNews positively) and for those not directly targeted, as seen in Figure 5.15. The figure confirms

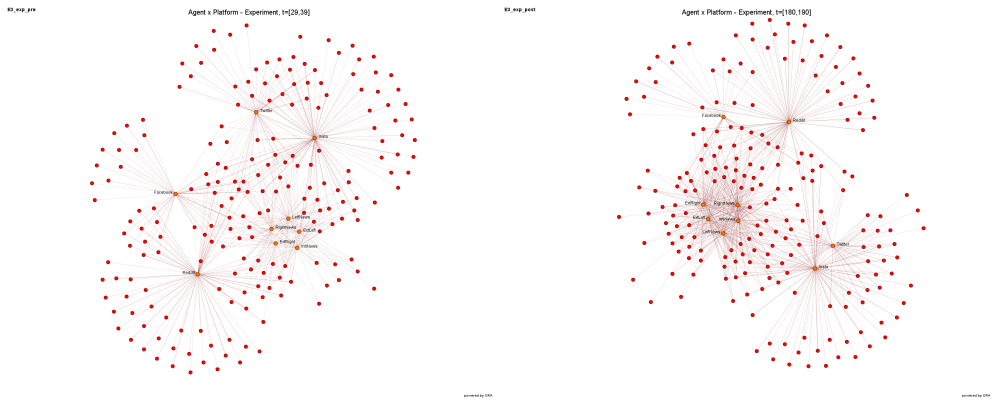

Figure 5.14: Agent X Platform network structure, pre- and post-maneuver.

expectations: the positively targeted platform saw an increase in subscribership, and the negatively targeted platforms saw a decrease. The ”bystanders” then reaped a lesser increase in subscribers as disaffected agents sought new news sources.

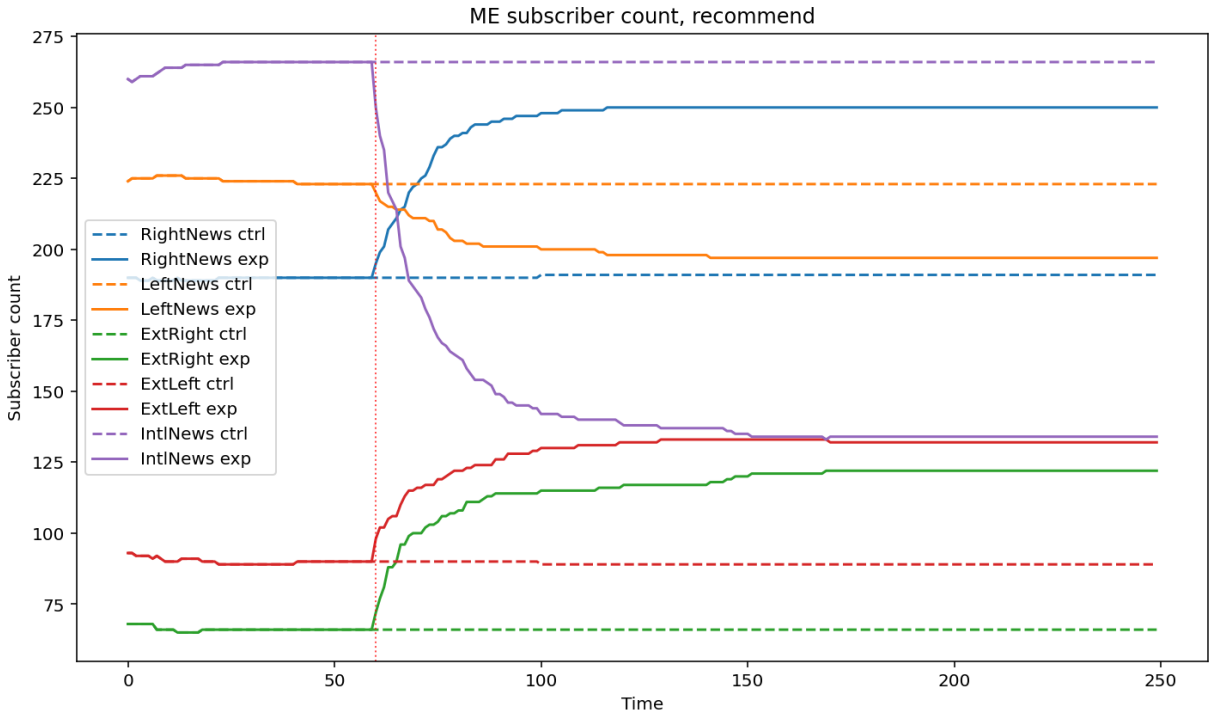

Figure 5.15: Closed platform subscriber levels with (solid) and without (dotted) Recommend/Sideline (red hash).

Figure 5.16 demonstrates that global average valence diverged from the control condition on all topics. Interestingly the divergence was not as dramatic at the community level.

Table 5.9 presents the statistical results. I measured two outcomes in the simulation:

- Mean valence value, the *change between the control and treatment simulation runs* in the

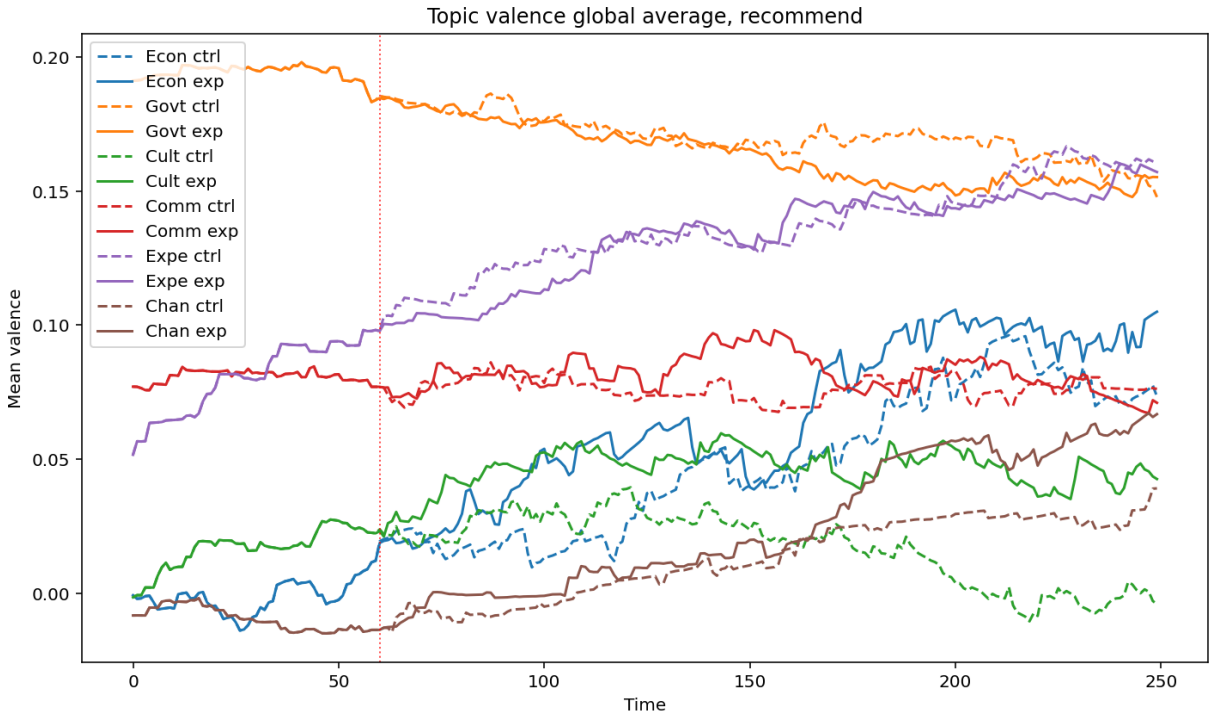

Figure 5.16: Global average topic valences with (solid) and without (dotted) Recommend/Sideline (red hash).

average valence on a specified topic within a specified community. This outcome detects influence impacts, the cumulative effect of influence pressures.

- Valence standard deviation, the *change between the control and treatment runs* in the standard deviation of valence values on a specified topic within a specified community. This is a measure of changes in uniformity of opinion within the target group.

As RightNews is recommended, we expect to see valence value shifts in the direction of that platform's position per topic. All significant value shifts follow this expected pattern save one: the Comm topic shifts *opposite* from RightNews' position, significantly within the TradPop community. This is an unexpected result and may indicate a simulation artifact rather than an emergent effect of interest. The discrediting of communities' trusted news sources, and the promotion of a less aligned source, significantly increased levels of disagreement within those communities (as measured by valence standard deviation), as some members stayed with their now-less-trusted sources and others migrated to the less aligned Recommended source. The exception was within TradPop, where the shift in perception led the few members of that community who subscribed to LeftNews to abandon it wholly, thus homogenizing the group further.

I conclude that the Recommend and Sideline maneuvers are detectable by the OPIEM framework and provide significant influence over the IE.

Table 5.9: Recommend/Sideline statistical results, N=48

Outcome	Topic	Community	t score	p value
Mean valence value	Econ	WorkPrag	1.983	0.050
		TradPop	5.034	0.000
		All cmtys	1.860	0.066
	Cult	WorkPrag	3.266	0.001
		TradPop	3.734	0.000
		All cmtys	2.789	0.006
	Comm	WorkPrag	−2.302	0.023
		TradPop	−2.654	0.009
		All cmtys	−2.294	0.024
	Expe	WorkPrag	2.164	0.033
		TradPop	2.047	0.043
Valence std deviation	Econ	InstMang	2.626	0.010
		WorkPrag	4.258	0.000
		Margin	2.905	0.005
		All cmtys	4.506	0.000
	Govt	WorkPrag	1.953	0.054
		All cmtys	1.898	0.061
	Cult	Margin	1.995	0.049
		TradPop	−1.738	0.085
	Comm	InstMang	1.706	0.091
		TradPop	−1.753	0.083
	Expe	All cmtys	1.807	0.074
	Chan	InstMang	3.228	0.002
		WorkPrag	2.760	0.007
		ProgUrb	2.069	0.041
		Margin	2.873	0.005
		All cmtys	2.992	0.003
	All topics	All cmtys	5.477	0.000

5.3.5 Reveal

Setup

Table 5.10 summarizes the Reveal experiment, which is especially intricate. I roughly simulate the online recruitment-radicalization process. Rather than stochastically assigning platform membership to actors on creation, I mapped actors to platforms based on community membership. This creates a highly stovepiped IE, where communities dominate specific portions of the IE with no direct crossover. I added a monster agent, Rho, as a high-level influencer in one of the two separate camps. Rho had a specific narrative that only he could generate, but that others could amplify. Rho's narrative is aligned with a topic that is highly polarized between the two camps.

A second monster agent, Chi, is a mid-level influencer on Rho's platform. Chi also has a unique narrative only he can generate; unlike Rho, Chi's narrative uses a topic with substantial agreement between the two camps. Starting from $t = 70$, the agent pool is examined at regular intervals (10 time steps for the results shown here); any agents in the left stovepipe that are sufficiently aligned to Rho "radicalize", abandoning their left-pipe accounts for Rho's native platform (Telegram). This "open" path to radicalization repliactes the reality that some agents may directly seek out sources and venues not aligned with their community.

In the experimental condition, Chi joins the opposite platform at $t = 30$ and begins propagating his 'soft' narrative. The simulation tallies how often agents in the left stovepipe choose to engage with Chi's narrative. At each inspection window, left-pipe agents have some probability of joining the less-radical right-pipe platform *in addition* to remaining active in the left ecosystem, essentially becoming "bridges" between the two. The probability of this action is a direct function of the agent's engagement with Chi: the more the agent engages with the Chi narrative, the more likely they are to sample the opposite pipeline. The radicalization mechanic functions as described previously.

My expectation was that the presence of Chi's narrative, and the resulting enticement to "visit" the opposite ideological camp, would result in more exposure within the left-pipe community to Rho-aligned narratives and positions. This, in turn, should produce a greater number of radicalized agents by the end of the simulation.

Table 5.10: Simulation configuration, Reveal

Variable	**Type**	**Value**
$\mathcal{A}^*$, agent set	Independent	Agents with non-random platform assignments
$\mathcal{N}$, narrative set	Independent	$\mathcal{N}_{RV}$, The baseline set with additional narratives ν_ρ and ν_χ
$\mathcal{P}$, platform set	Independent	$\mathcal{P}_{0T}$: Baseline platform, Telegram substituted for Reddit
$\mathcal{E}$, event set	Independent	Reveal event: Agent χ joins Facebook at $t = 30$
		Recruitment event: At $t = \{60, 70, 80, ...\}$, agents consider following χ to opposite platform
$\mathrm{GhostCell}(\Sigma)$, mechanics and parameters	Control	Platform assignment altered; platform migration changed to radicalization mechanic

Maneuver detection

Reveal is targeted at a Topic and Platform, and manifests as a link between the targets where there was none. My stove-pipe configuration ensures zero connection between the target Topic (Chi and Rho's narratives) and the target Platform (the left-side platforms). The mechanics of the experiment are such that *any* message with the Chi and Rho narrative on those platforms

demonstrates a Reveal maneuver; such is guaranteed by Chi joining the left pipe at $t = 30$. Thus detection of the maneuver is trivial.

Influence effects

The advent of a single adversarial narrative in a previously closed ecosystem is unlikely to produce significant change in the short term. However, it may resonate with community members whose values or beliefs already diverge from the community majority. Such outside narratives might also produce new conversation, if only as dismissals or rebuttals. To this end, I considered two specific outcomes for this experiment: the proportionate traffic of the target narratives and the number of individuals "radicalized" over the course of the simulation.

Exposing new audiences and platforms to the Chi and Rho narratives did increase their overall traffic, as indicated in Figure 5.17. Perhaps more impactfully, these narratives carved out a small but persistent presence in the opposing stovepipe, evidenced by the low but persistent target narrative traffic proportions on Facebook and Insta in Figure 5.17; this provided a channel for continued influence and recruitment/radicalization of fringe members within that stovepipe. As expected, the Reveal maneuver and the resulting influence produced a greater number of radicalizations than seen in the control condition (Figure 5.18).

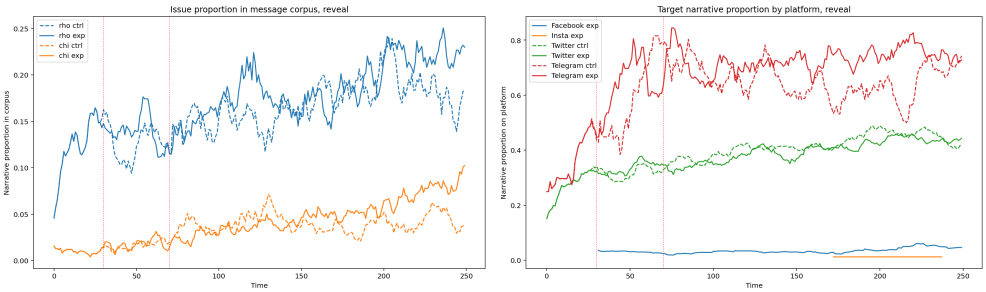

Figure 5.17: Proportion traffic for target narratives (total left, per platform right), with (solid) and without (dotted) Reveal (red hash).

The Reveal maneuver is therefore identifiable within the OPIEM model and does confer significant influence on the environment. Real-world instances of Reveal may be difficult to identify; finding the definitive "first" instance of a Topic's occurrence in the internet is challenging, regardless of (or perhaps because of) a bevy of forum posts claiming that title. In other cases – such as press releases or news breaks – recognizing the maneuver is trivial. In either case, however, the maneuver represents a significant influence action within the IE.

Table 5.11 presents my statistical results. I measured four outcomes in the simulation:

- Valence standard deviation, the *change between the control and treatment runs* in the standard deviation of valence values on a specified topic within a specified community. This is a measure of changes in uniformity of opinion within the target group. I limited measures to the Comm topic, as that is the topic of the radicalizing narrative; increased opinion variance on that topic is an indicator of more extreme opinion and thus the presence of individuals being influenced by, and becoming more aligned with, the radicalizing narrative.

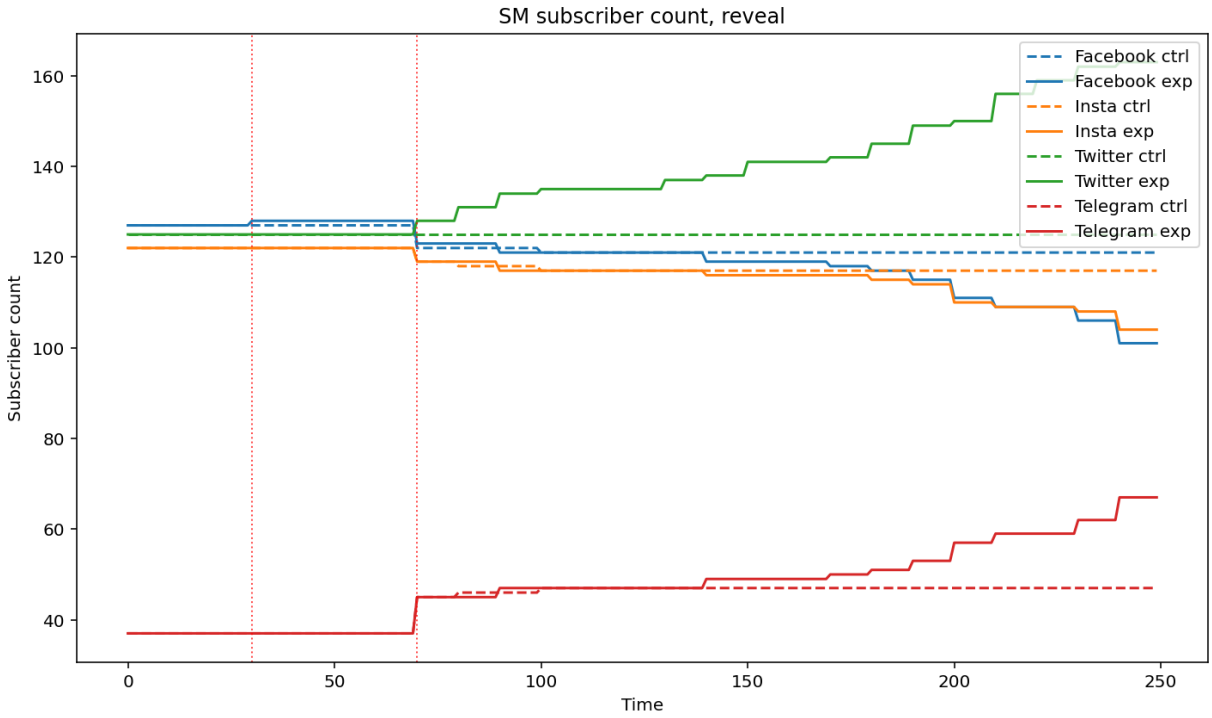

Figure 5.18: Subscriber count over time, by platform, with (solid) and without (dotted) Reveal (red hash).

- Valence value span, the *change between the control and treatment runs* between the maximum and minimum valence values. This is a measure of extremity within the target group. I limited measures to the Comm topic, as that is the topic of the radicalizing narrative. As with standard deviation, this measure is an indicator of radicalizing activity; it is correlated to the standard deviation but offers an additional view on the radicalizing narrative's impact.
- Traffic proportion by narrative, the *change between the control and treatment runs* in the total proportion of corpus messages containing a given narrative at end of simulation. I measured for both the scout narrative η_χ and the radicalizing narrative η_ρ. This measures the extent to which these narratives are present and visible to the populace, and is an indicator of the extent to which agents other than the originating monsters are engaging with and propagating these narratives.
- Radicalized agent count, the *change between the control and treatment runs* in the number of agents radicalized as per the experiment description. This is the most direct measure of the Reveal maneuver's impact.

Rho's narrative η_ρ is based on the Comm topic, and, as expected, the Reveal maneuver induces an increase in the variance of opinion on that topic within the communities of the left stovepipe, reflecting exposure to the more extreme external position. The span of valence values also shifts as radicalized community members become increasingly separate from their communities of origin. Both the recruiting and radicalizing narrative have significant proportional gains under

the experimental condition, and accordingly we see a significant increase in radicalized agents.

Table 5.11: Reveal statistical results, N=48

Outcome	Group	t score	p value
Valence standard deviation, Comm topic	ProgUrb cmty	2.414	0.018
	Margin cmty	4.016	0.000
	InstMang cmty	2.677	0.009
Valence value span, Comm topic	ProgUrb cmty	3.116	0.002
	Margin cmty	5.130	0.000
	InstMang cmty	1.967	0.052
Traffic proportion by narrative	η_χ recruiting narrative	2.733	0.008
	η_ρ radicalizing narrative	2.622	0.010
Radicalized Agent count	Left-pipeline communities	6.197	0.000

5.3.6 Stifle

Setup

Table 5.12 summarizes the Stifle experiment: I simulate a top-down platform-level censorship regimen that blocks all messages on a specific narrative. I use a reduced narrative set with only three cross-cutting issues instead of the baseline six, to increase each narrative's proportion in the corpus and thereby increase the effect size of the treatment (relative to the short of the simulation). There is one marker narrative in the narrative set, afforded a boost to make it slightly more interesting to Agents.

At t=70, one platform begins actively censoring all traffic about this narrative, hiding any previous messages containing it and preventing the posting of new messages containing it. Importantly, I still allow agents to *try* to post about the narrative; their efforts are then blocked. This means that Agents will still stochastically select the narrative on the censoring platform, thereby "wasting" their action. As time goes on, however, and other narratives come to dominate the platform – and therefore the agents' message feed – the weight of the forbidden narrative will decrease, and so too will the number of wasted engagement slots.

Table 5.12: Simulation configuration, Stifle

Variable	Type	Value
$\mathcal{N}$, narrative set	Independent	$\mathcal{N}_{ST}$, Truncated narrative set with marker narrative added
$\mathcal{P}$, platform set	Independent	$\mathcal{P}_0$: Baseline platform set
$\mathcal{E}$, event set	Independent	Stifle event: Target platform begins censoring target narrative at $t = 70$

Maneuver detection

As with Reveal, Stifle manifests in the network as a link weight between a target Topic and Platform. A true Stifle action removes the link between the two targets. In reality, completely removing any information from an online environment is effectively impossible, and even more so when that information is a subject of human interest. Even with the highly effective and invasive censorship simulated here, my scenario does not include the likely adjacent narratives or codewords that would emerge to circumvent such obstacles.

If a true Stifle (driving the link weight to 0) is impossible, we can identify the maneuver by a sudden and dramatic decrease in link value. Differentiating an attempted Stifle from an attempted Smother maneuver is ultimately a matter of subjective appraisal, likely informed by the specific means of implementation and the extent to which the target narrative/Topic is removed or buried within the IE. It should be repeated, as was mentioned in chapter 3, that Reveal and Stifle are extreme cases of the more general Repeat and Smother maneuvers, respectively.

Table 5.13 presents pertinent average measurements from the Topic-Platform network. Note that the total volume of messages does not significantly change over the course of the maneuver, indicating that the Stifle – as intended – is curbing only targeted traffic, and not all traffic. As expected, the target Narrative in-degree decreases in response to the maneuver, slightly as a whole and precipitously on the targeted platform. I note that the Stifle maneuver is ultimately ineffective in Stifling conversation on the topic as a whole; the inherent appeal of the target Narrative overcomes the loss of a single platform, resulting in an increasing proportionate share for the Narrative when considering the IE as a whole. This is an obvious and recognizable result: moderation on a single platform simply drives conversation to other venues.

Table 5.13: Network metrics on Topic X Platform network (weighted)

Measure in $\mathcal{R} \times \mathcal{P}$ *(normalized by platform)*	**Window 1** t=(20,69)	**Window 2** t=(100,149)	**Window 3** t=(200,249)
Total messages	14318.8	14410.6	14220.8
Target narrative in-degree	0.377715	0.341675	0.415693
Narrative-platform link: Target platform	0.015098	0	0
Narrative-Platform link: Non-target platform avg	0.120872	0.113892	0.138564

Influence effects

Figure 5.19 shows traffic proportion over time for the two largest platforms in one of the experimental runs, illustrating how the suppression of the marker narrative on a large platform (Insta) drove conversation on that narrative to other platforms (Twitter). While this shift is marked, **??** shows that it did not produce significant changes in the overall population valence on any Topic, including the Topic of the marker narrative (Govt); this was confirmed statistically. While many runs showed an increase in that Topic's standard deviation – indicating a wider spread of opinions in the populace – the effect was ultimately less than significant (p=0.22, Figure 5.20).

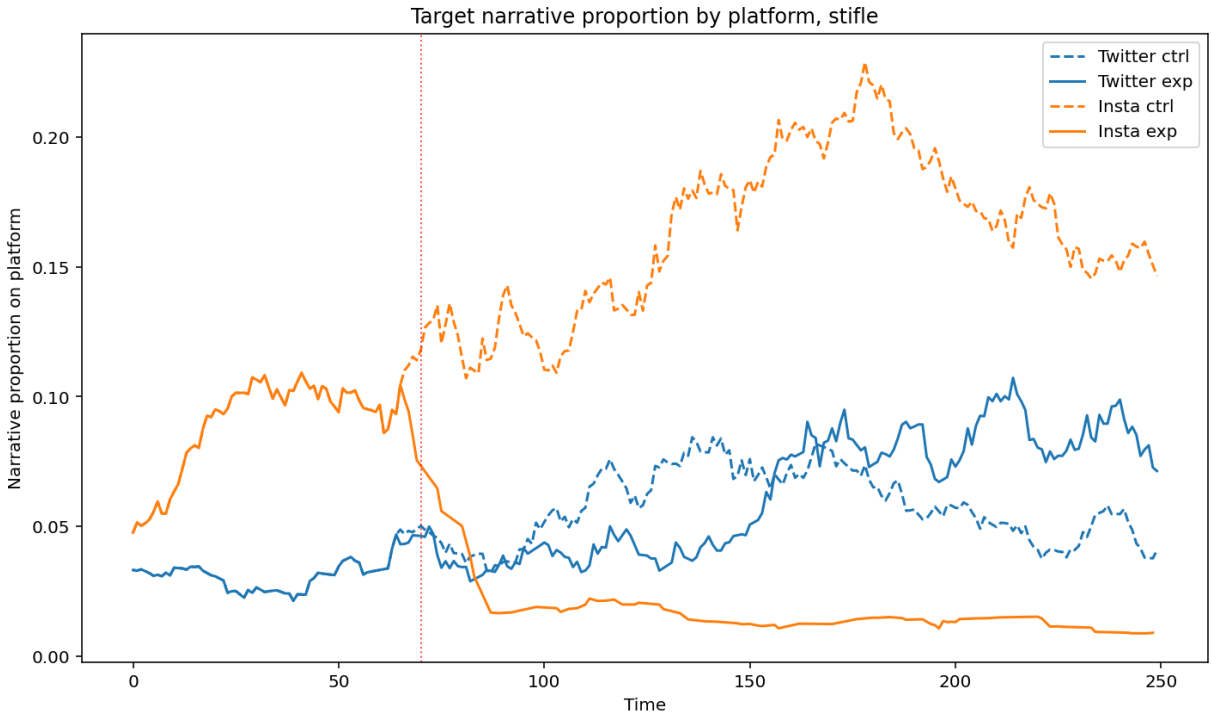

Figure 5.19: Total occurrences of target narrative, top two platforms by subscriber volume.

As with Reveal, the Stifle maneuver is trivially detectable within the OPIEM framework and has significant influence on the IE. In the online ecosystem, propagation of one's own narrative and viewpoint is, in itself, a major influence, even if that narrative does not initially directly alter beliefs or behaviors.

My statistical results are summarized in Table 5.14. I measured two outcomes in the simulation:

- Traffic proportion by narrative, the *change between the control and treatment runs* in the total proportion of corpus messages containing a given narrative at end of simulation. I measured the proportion of the narrative targeted by the Stifle maneuver within each community and on each platform; only a small subset of these groupings saw significant change. This is a measure of the persistence (or absence) of the target narrative, and thus the success of the Stifle maneuver.
- Traffic proportion by platform, the *change between the control and treatment runs* in the total proportion of corpus messages on a given platform. This is an indirect measure of the Stifle maneuver's success, and a measure to demonstrate how the Stifle may drive shifts in user preference and behavior.

As expected, we observe a significant decrease in the stifled narrative occurrence on all communities and on the target platform. Interestingly, we also see a significant decrease when taken over *all* platforms, indicating that the lack of the narrative on a larger platform like Insta makes agents less likely to use it even in environments where it is not blocked. As part of the stifle effort, we observe a significant loss of traffic for Insta, as agents seeking to generate messages

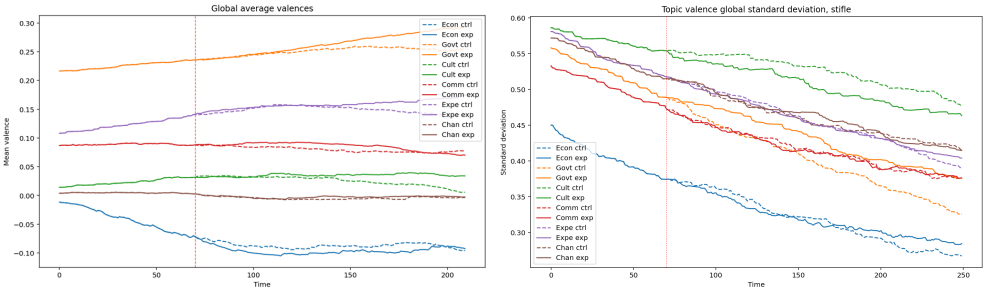

Figure 5.20: Average topic valences (left) and valence standard deviations (right) with (solid) and without (dotted) Stifle (red hash).

on the stifled narrative instead opt to generate elsewhere or not at all.

Table 5.14: Stifle statistical results, N=48

Outcome	**Group**	t **score**	p **value**
Traffic proportion of target narrative	All cmtys	-2.079	0.040
	Target platform (Insta)	-4.071	0.000
	All cmtys, all platforms	-4.554	0.000
Traffic proportion by platform	Target platform (Insta)	-2.352	0.021

Special case: Chinese stifling

As an additional test, I reconfigured the Stifle experiment to mirror Chinese-style "leaky censorship." In this model, *all* platforms participate in Stifling traffic, but only at a 90% rate. That is, any post created with the targeted narrative has a 10% chance of being posted, and a 90% chance of being blocked at creation (and thus never posted or seen). Under these more global conditions, maneuver detection is significantly more clear-cut, as seen in Table 5.15; the global Stifle effort causes the target narrative in-degree to quickly and consistently decrease.

Table 5.15: Network metrics on Topic X Platform network (weighted) (90% censorship)

Measure in $\mathcal{R} \times \mathcal{P}$ *(normalized by platform)*	**Window 1** t=(20,69)	**Window 2** t=(100,149)	**Window 3** t=(200,249)
Total messages	13256.2	12989.6	13612.9
Target narrative in-degree	0.086481	0.002969	0.001593

By construction, the target narrative incidence reduced across the population. As seen in the previous Stifle experiment, the suppression of a single narrative was not sufficient to produce significant valence effects. Of greater interest is the population's behavior in the face of "leaky" censorship. Despite the persistent allowance of some messages, the population's interest declines over time, as the narrative is not visible enough to maintain broad interest and engagement (Figure 5.21).

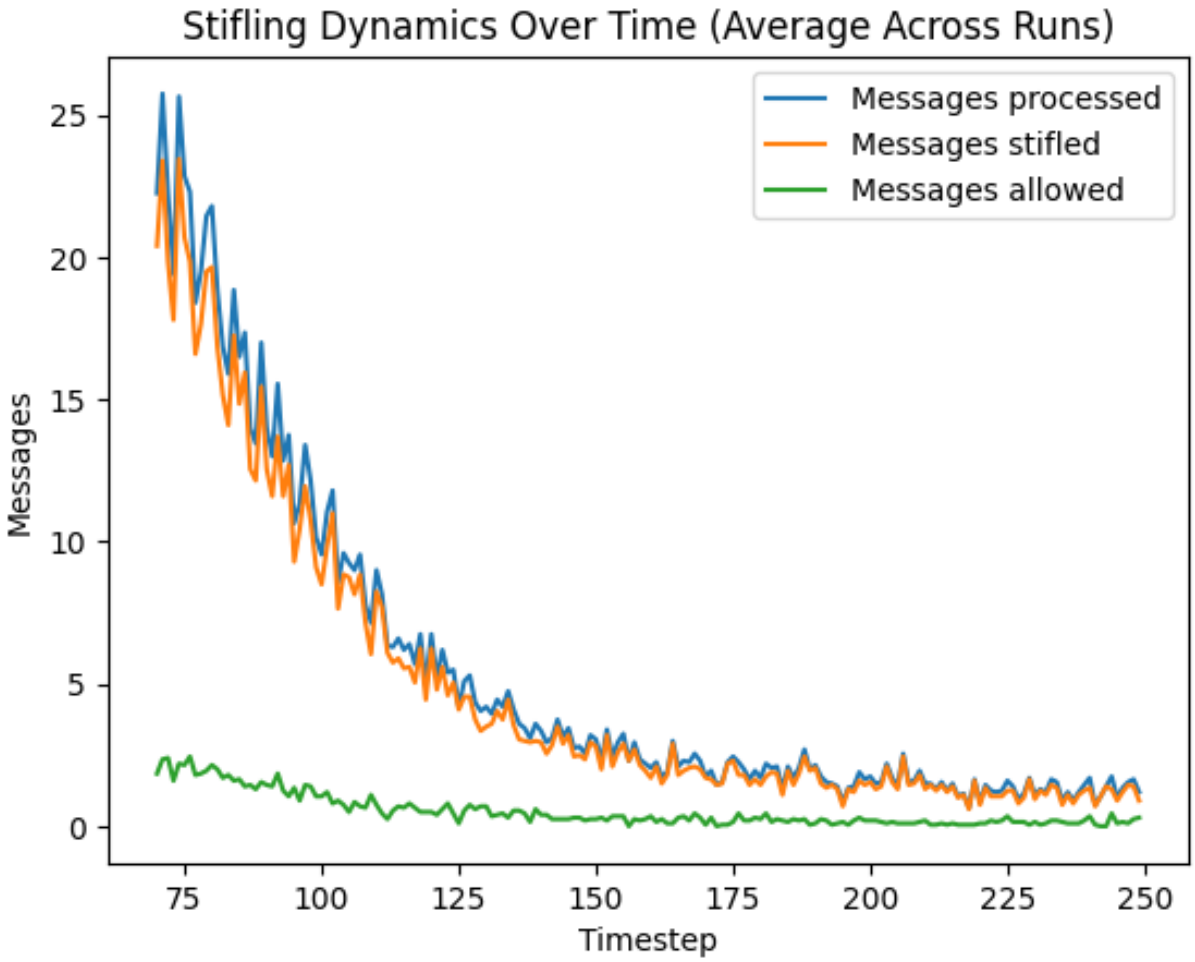

Figure 5.21: Message allowance and blocking over time, 90% censorship regime

This research indicates that imperfect Stifle regimes can still have a strong chilling effect on targeted narratives, while remaining less obviously detectable or offensive to the targeted population.

5.3.7 Repeat

Setup

Table 5.16 summarizes the Repeat experiment: I simulate the action of a bot-farm centered on boosting a particular viewpoint. A marker narrative is added to the narrative set, tied to the Expertise Topic. A set of 15 bot agents activates at $t = 70$, engaging across all platforms to to amplify the target narrative. Importantly bots are limited to reactions (i.e. retweets), meaning they can only amplify actions taken by ”real” agents.

Table 5.16: Simulation configuration, Repeat

Variable	Type	Value
$\mathcal{N}$, narrative set	Independent	$\mathcal{N}_{0R}$, Narrative set with marker narratives added
$\mathcal{P}$, platform set	Independent	$\mathcal{P}_0$: Baseline platform set
$\mathcal{E}$, event set	Independent	Repeat event: Amplifying bot group activates at $t = 70$

Maneuver detection

Reveal manifests in the network as a change in the Occurrence network link value between the target Topics and Platform. The target Occurrence link value shows a significant increase once the Repeat maneuver begins, as shown in Table 5.17.

Table 5.17: Metrics of $\mathcal{R} \times \mathcal{P}$ Occurrence network during Repeat

Network metric	Window 1: $t = [20, 69]$	Window 2: $t = [100, 149]$	Window 3: $t = [200, 249]$
Target topic in-degree	0.025357	1.047392	1.125718

Influence effects

The Repeat maneuver quickly floods the IE with the target narrative, as seen by the sharp spike in narrative proportion across the corpus in Figure 5.22. The target narrative carries a positive Expert valence, and as shown in Figure 5.23, the increased frequency of this narrative induces greater dispersion in opinion. However, the overall impact on actual valence ran *counter* to the target narrative, implying that the bots' relatively low social status failed to effectively persuade anyone and, in fact, may have undercut the position they espoused. Further investigation of this counterintuitive drift is warranted.

The Repeat maneuver is detectable within the OPIEM framework and provides significant influence over the IE. My statistical results are presented in Table 5.18. I measured six outcomes in the simulation:

- Average valence value, the *change between the control and treatment simulation runs* in the average valence on a specified topic within a specified community. I limited measures to the topic of the Repeat target narrative. This outcome detects influence impacts over this topic, which would include impacts from the increased occurrence of the target narrative.
- Valence standard deviation, the *change between the control and treatment runs* in the standard deviation of valence values on a specified topic within a specified community. Again limited to the Expe topic, this measure indicates the extent to which the Repeat maneuver increases dissent or heterogeneity of opinion within various groups.
- Valence value span, the *change between the control and treatment runs* between the maximum and minimum valence values. This is a measure of extremity within the target group.

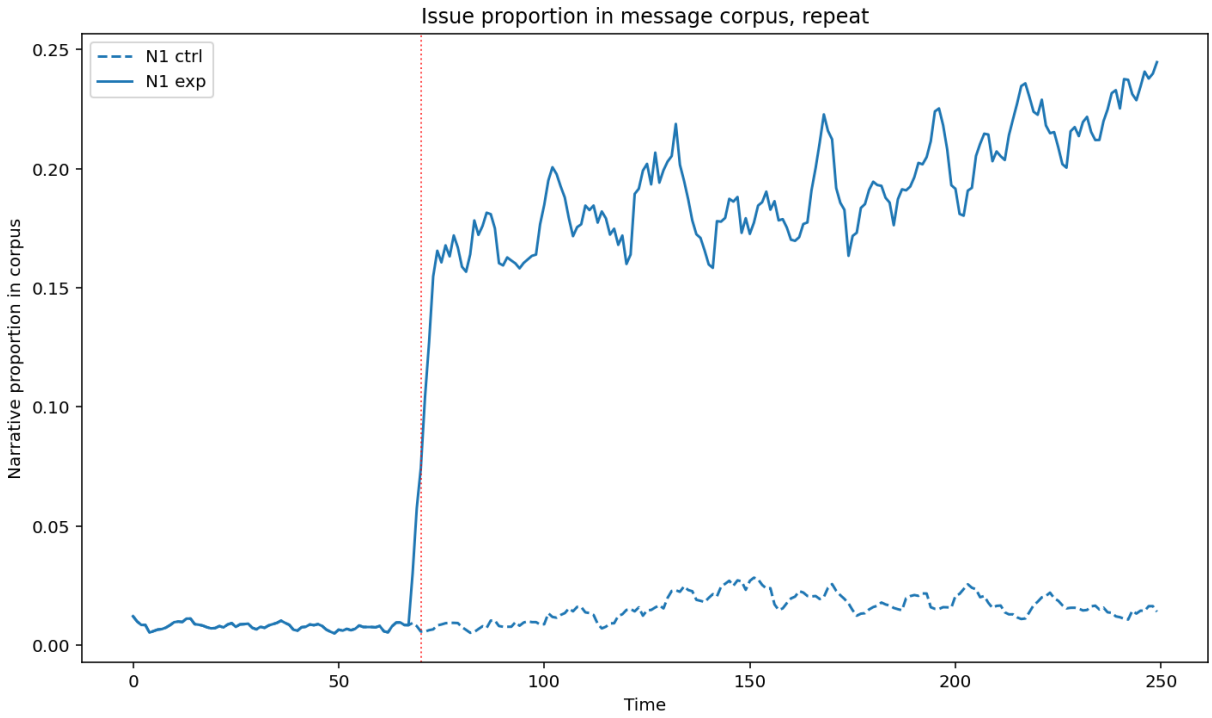

Figure 5.22: Target narrative occurrence with (solid) and without (dotted) Repeat.

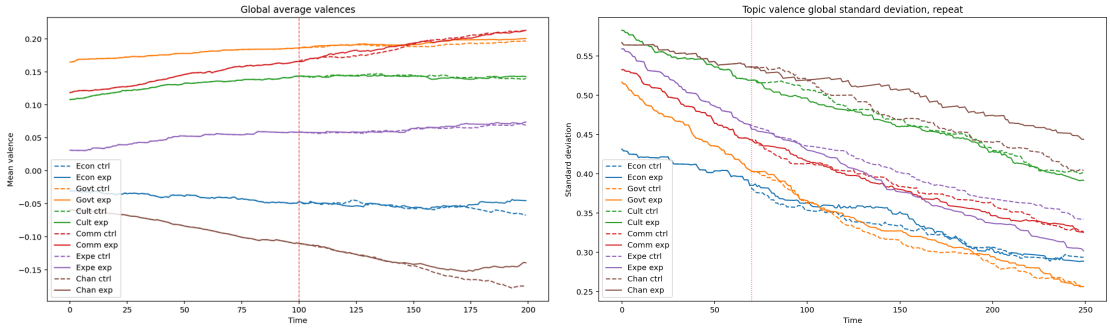

Figure 5.23: Average topic valence values (left) and standard deviations (right) with (solid) and without (dotted) Repeat (red hash).

Again limited to the Expe group, this is a measure of extremity of opinion, and indicates the extent to which a Repeat may push a target group toward more extreme positions.

- Traffic proportion by narrative, the *change between the control and treatment runs* in the total proportion of corpus messages containing a given narrative at end of simulation. I measured the proportion of the narrative targeted by the Repeat maneuver within each community and on each platform, as a means of assessing the maneuver's success and impact.
- Unique engagers, the *change between the control and treatment runs* of the number of unique non-bot users engaging with the target narrative. This measure is a more direct

indicator of the Repeat maneuver's success in driving visibility.

- Number of engagements, the *change between the control and treatment runs* of the number of non-bot engagements with target narrative messages. This measure is a more direct indicator of the Repeat maneuver's success in driving engagement.

The Repeated narrative carries a positive valence on the Expe topic, and we see a significant positive shift in valence value among three of the communities. Across the entire populace, the variance and span of opinions on that topic increases, indicating that increased exposure does not necessarily produce persuasion or agreement at large, but does manage to persuade enough individuals to widen opinions within the populace, increasing discord. As expected, the network of bots produced significant gains for that target narrative across all platforms and within all communities; importantly, however, the number of unique non-bot engagers and the number of non-bot engagements also increased significantly, indicating that the significant proportional gains are not merely bot activity but represent actual agent engagement.

Table 5.18: Repeat statistical results, N=48

Outcome	**Group**	t **score**	p **value**
Average valence value, Expe topic	TradPop cmty	-2.997	0.003
	WorkPrag cmty	-1.909	0.059
	InstMang cmty	2.366	0.020
Valence std deviation, Expe topic	All cmtys	3.741	0.000
Valence value span, Expe topic	All cmtys	1.964	0.053
Traffic proportion of target narrative	On Facebook	10.185	0.000
	On Insta	7.657	0.000
	On LinkedIn	10.301	0.000
	On Reddit	10.637	0.000
	On Twitter	9.559	0.000
	In Margin cmty	6.334	0.000
	In ProgUrb cmty	7.110	0.000
	In TradPop cmty	6.379	0.000
	In WorkPrag cmty	10.081	0.000
	In InstMang cmty	7.228	0.000
	Total corpus (global)	22.116	0.000
Target narrative unique engagers		8.714	0.000
Target narrative number of engagements		18.145	0.000

5.3.8 Smother

Setup

Table 5.19 summarizes the Smother experiment: As with Repeat, I simulate the action of a targeted bot group. For this experiment the narrative set is reduced to four cross-cutting issues (instead of 6), to increase contrast for targeted narratives. A set of 30 bot agents activates at $t =$

80. They join the two largest platforms by subscriber volume, and target the cross-cutting issue with the highest presence in the corpus for Smothering – in other words, they go after the biggest topic on the biggest platforms. The bots limit their actions to reactions (i.e. retweets). They aggressively react to messages about any narrative *except* the narratives of the targeted cross-cutting issue. (For example, from the available cross-cutting issues of abortion, climate change, public health, and immigration, the bots might target immigration, Smothering all narratives within that issue by engaging with all narratives *not* within that issue.)

Table 5.19: Simulation configuration, Smother

Variable	**Type**	**Value**
$\mathcal{N}$, narrative set	Independent	$\mathcal{N}_0^-$, Reduced narrative set (4 issues)
$\mathcal{P}$, platform set	Independent	$\mathcal{P}_{SM}$: No LinkedIn, Telegram replaces Reddit
$\mathcal{E}$, event set	Independent	Smother event: Amplifying bot group activates at $t = 80$

Maneuver detection

Smother maneuvers are detected in the network by the same metrics as Repeat: link values between the targets in the Topic x Platform Occurrence network. Because the Smother target was an issue, I combined all Topics into Issue metanodes in building the $\mathcal{R} \times \mathcal{P}$ network. The link value between the target Platforms and the target Topic metanode shows a significant and sustained decrease once the Smother maneuver begins (Table 5.20).

Table 5.20: Metrics of $\mathcal{R} \times \mathcal{P}$ Occurrence network during Smother

Network metric	**Window 1:** $t = [20, 69]$	**Window 2:** $t = [100, 149]$	**Window 3:** $t = [200, 249]$
Avg target topic in-degree	1.265827	1.028773	0.910698
Avg target topic-platform link weight	0.397414	0.176765	0.139034

Influence effects

The Smother maneuver shifted traffic away from the target narratives, and by the end of the simulation the target issue was significantly less present in the message corpus. In Figure 5.24, the target issue was Immigration, and the Smother produces an immediate decrease in proportionate traffic which remains below control levels. (Note that, even in the control condition, this issue lost ground to others through the internal mechanics of the simulation.) The Smother maneuver was executed on the larger, busier platforms, where fully controlling conversation proved difficult; and, as the already-popular issue became contested, it gained ground on the uncontested platforms. Figure 5.25 demnonstrates this, as smother-induced decreases in target issue traffic on the top platforms (Telegram and Twitter, in this run) coincide with increases in target issue

traffic on the lesser platforms (Facebook and Insta). Because the targeted group of narratives was spread across multiple Topics and represented both sides of the argument, there was no consistent or significant effect on global valence values. The demonstrated influence here is the suppression of conversation about an entire broad social issue, which is effectively achieved and, with sustained effort, could likely be affected further.

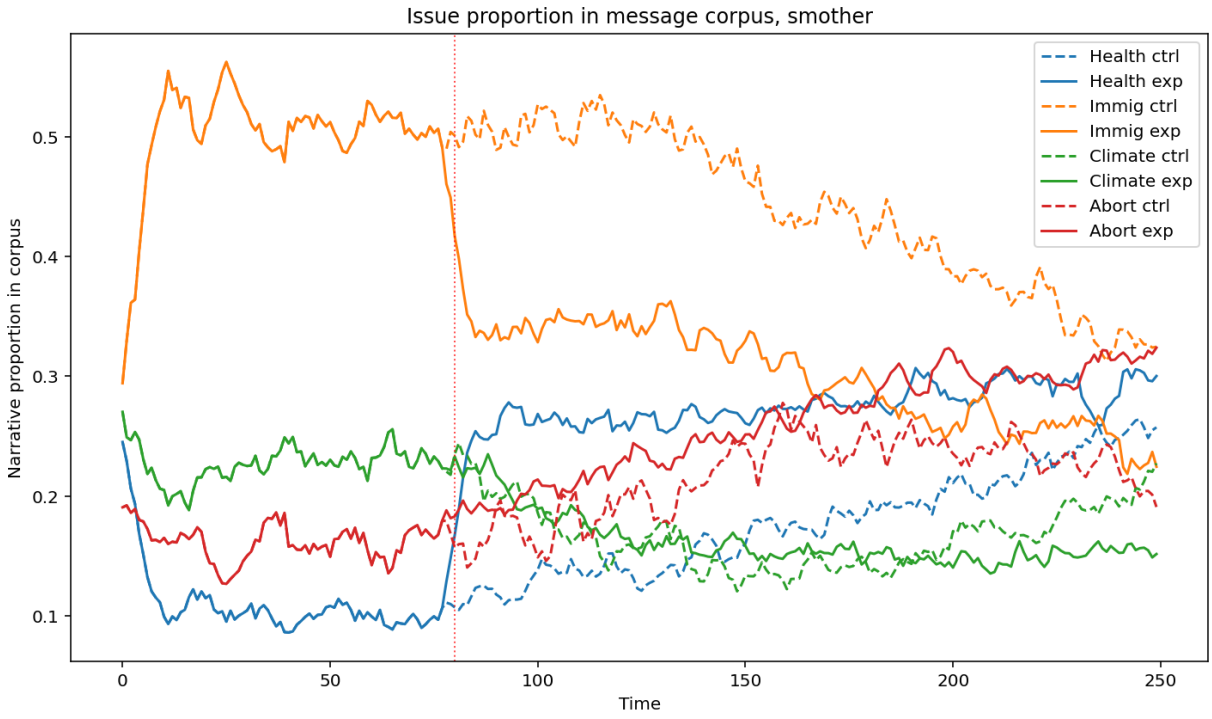

Figure 5.24: Target issue occurrence with (solid) and without (dotted) Repeat.

As demonstrated, the Smother maneuver is detectable within the OPIEM framework and has significant influence on the IE. I present my statistical results in Table 5.21. I measured two outcomes in the simulation:

- Traffic proportion by **issue**, the *change between the control and treatment runs* in the total proportion of corpus messages containing narrative for a given issue at end of simulation. I measured the proportion of the issue targeted by the Smother maneuver one each platform, as a means of assessing the maneuver's success and impact.
- Traffic proportion by platform, the *change between the control and treatment runs* in the total proportion of corpus messages on a given platform. I grouped the platforms into target and non-target because the affected platforms were chosen per-run; I also limited measures to *non-bot* traffic. This is a measure of the Smother maneuver's impact on user behavior relative to platform selection, attempting to detect the extent to which Smothering an issue drives changes in users' social media behavior.

As expected, the issue targeted for Smothering saw significant loss of traffic proportion on the platforms where Smother was implemented; conversely, it saw no significant change on non-targeted platforms. This implies that the Smother effect did not reduce narrative uptake on targets

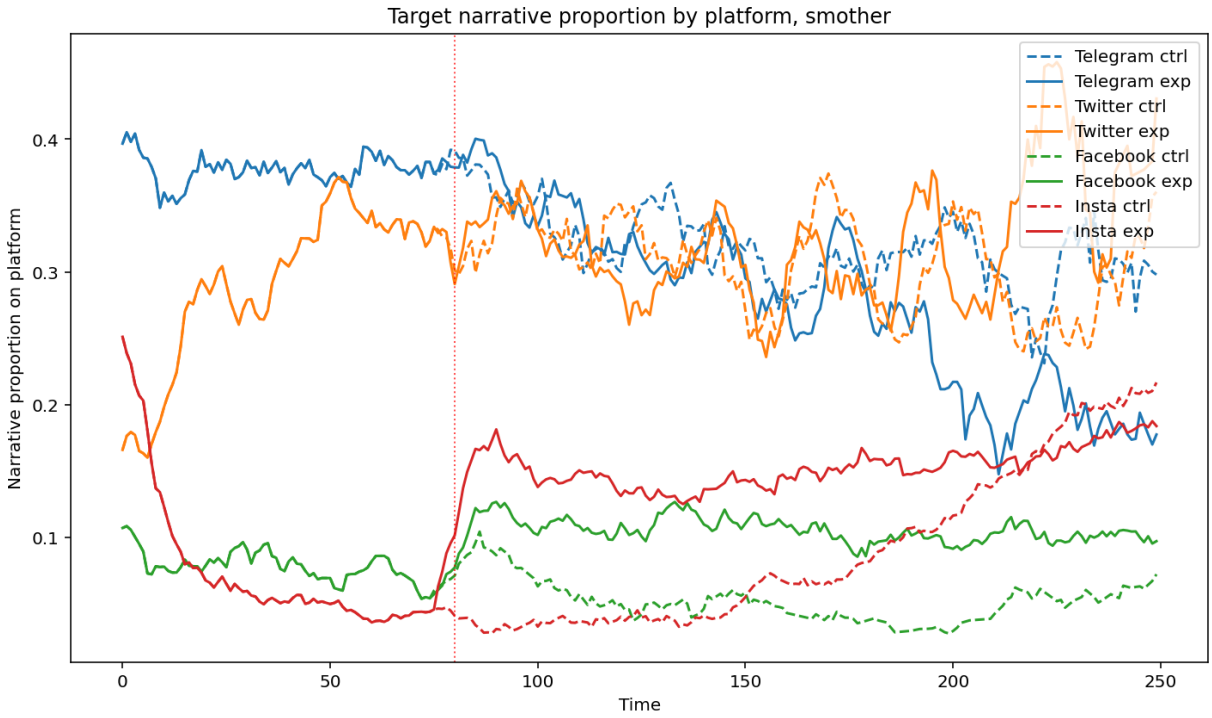

Figure 5.25: Target issue occurrence by platform, with (solid) and without (dotted) Smother (red hash).

not directly affected by bot traffic. In similar fashion, I saw a significant increase in overall non-bot traffic on non-targeted platforms, as agents seeking to discuss the smothered narrative began to favor those platforms; but, importantly, I saw no significant change in traffic proportion for favored platforms. The botnet that smothered the target issue did not otherwise dissuade agents from using the target platforms for other discourse.

Table 5.21: Smother statistical results, N=48

Outcome	**Group**	t **score**	p **value**
Traffic proportion, target issue	Target platforms	-7.058	0.000
	Non-target platforms	0.259	0.796
Traffic proportion, overall (non-bot)	Target platforms	-1.463	0.145
	Non-target platforms	1.983	0.049

5.4 BEND Maneuver Detection

To demonstrate the OPIEM framework's compatibility with the BEND framework, I simulated a selection of BEND maneuvers and measured the corresponding metrics within the constructed OPIEM network.

5.4.1 Back and Negate

To simulate the Back and Negate maneuvers, I created two monster agents, Alpha and Beta. The two are ideologically opposed and members of positionally distant communities. Both monsters had modest followings on social media and were highly active archetypes. At $t = 100$, I increased Alpha's follower count by 50% and activated a group of 20 bots who amplified Alpha's content through reactions. At the same time, I reduced Beta's follower count, and reduced Beta's energy and attention parameters, resulting in the agent engaging less frequently. This simulates exogenous circumstances that "rate limit" Beta. The results are compiled in Table 5.22.

Table 5.22: Node centrality in Agent X Platform X Agent network (weighted)

Agent	**Centrality Measure**	**At t_1**	**At t_2**	**At t_3**
Alpha (Backed)	Authority	0.705	1.0	1.0
	PageRank	0.165	0.155	0.218
	Degree	0.027	0.039	0.030
Beta (Negated)	Authority	0.997	0.524	0.632
	PageRank	0.163	0.109	0.138
	Degree	0.027	0.026	0.019

In every centrality measure, Alpha ends with a higher value, as expected. Similarly, Beta ends with a lower value. Some counter-maneuver "shifts" are observed, such as the decrease in Alpha betweenness centrality from t_1 to t_2. Such shifts are not abnormal for a large network dynamically changing and are likely more visible due to the large time window aggregate used in capturing these measures.

5.4.2 Engage and Dismiss

To simulate the Engage and Dismiss maneuvers, I used a simplified narrative set. Engage and Dismiss are measured through the Salience network, an agent-normalized Agent x Topic network. In real data, the topic nodeset would be a broad array of hashtags and issues constructed post-hoc through natural language processing. In my simulation this nodeset is best approximated by the narratives, which are mapped one-to-one to the underlying agent salience attributes. However, in real data, agent priorities are dominated not by underlying orthogonal principles but by macro-level issues. As such, to better approximate the priority-to-topic decision pathway, I reformulated the narrative array to more closely align to the underlying topic set. In this instance, the topic set is now more a simulation of an "issue" set.

With this simplified narrative set (6 narratives per topic, 3 positive and 3 negative), I ran 300 random agents for 300 time steps. At $t = 100$, I manually adjusted the underlying salience values for agents of two target communities relative to two target topics: I uniformly increased the salience of the Government topic for the TradPop community, while uniformly decreasing the salience of the Culture topic for the Margin community.

The agent salience attribute is unreadable in simulation output; it is an underlying value used by agents when selecting how to engage with social media. These direct value alterations simulate the exact definitions of the Engage and Dismiss maneuvers: any actions taken – within

or external to the social media environment – that increase or decrease the relevance of a topic to the reader. By altering agent salience, I replicate the effect such an action would have, directly increasing or decreasing the agent's concern for that topic.

I then constructed the OPIEM network from message traffic to see if the results of this alteration were visible solely through agent transmission behavior. I constructed the Salience network, an external link Agent x Topic network normalized to the agent. This network captures an agent's proportional engagement with the topics in the IE, indicating the agent's focus and concern. This in turn is a good proxy for an agent's perception of relevance: an agent would not expend limited energy to engage with topics that were irrelevant.

Table 5.23 shows the Salience network link value between the respective target populations and topics, over three time periods. Maneuvers commenced at the end of t_1. The TradPop-Govt link increased from t_1 to t_2 and maintained this increase through t_3. The Margin-Cult link decreased from t_1 to t_2 and continued decreasing into t_3. In both cases, the network indicates the presence of the expected BEND maneuvers.

Table 5.23: Network metrics on Agent X Topic Salience network

Network link (Community-Topic)	$t_1 = (0, 99)$	$t_2 = (100, 199)$	$t_3 = (200, 299)$
TradPop - Govt	0.029877	0.081849	0.07056
Margin - Cult	0.474271	0.411205	0.381191

5.5 Predicting influence effects

The baseline inputs used for these experiments were deliberately simple, sufficient to produce realism without replicating a specific social, ethnic, and/or political context. The demonstrated influence effect of the availability-based maneuvers, and of the OPIEM framework in recognizing both these maneuvers and the larger set of BEND maneuvers, indicates the utility of a tri-modal approach to IE analysis. GhostCell's mechanics provide a baseline representation of the information environment. Specific expansion and adaptation could increase the fidelity of that simulation, as discussed in the following chapter.

Given an accurate set of agents, an appropriately accurate narrative and topic set, and a representative platform set, GhostCell could simulate the case studies referenced in Chapters 2 and 3. I opted not to build such detailed simulations due to the time requirement in creating historically accurate actors and platforms, and in deriving appropriately representative topics and narratives. Simulating real-world incidents requires extensive knowledge of the details of those incidents, to ensure the simulation accurately converges with the historical outcome.

However, given access to the pertinent message traffic surrounding those incidents, any analyst could create a trimodal OPIEM network that, as demonstrated, would provide useful and intuitive understanding of the IE, and would assist in identifying influence momenta within the IE. As social media study expands, multi-platform studies over the same time window and with the same query terms would be greatly simplified using the OPIEM methodology to unify these otherwise disparate traffic sets. The single trimodal network approach to representing the IE

offers a reliable and quantitatively tractable way to analyze and understand socio-cyber hybrid phenomena, availability-based and cognitive-based influence efforts, and the linkages and relationships, within the environment.

5.6 Conclusion

The OPIEM methodology produced a sufficient model of a simulated IE, and, within that model, accurately identified the presence of influence efforts. This success demonstrates the soundness of the model and method, and recommend it for use with more complex simulated environments and with live data.

Chapter 6

OPIEM-enabled Information Maneuver

The demonstrated efficacy of the OPIEM representation of the IE, including the use of the BENDRS lexicon to describe specific actions within the IE, allows me to describe how OPIEM and BENDRS can facilitate more effective maneuver within the information environment.

6.1 Concept of maneuver

The traditional interpretation of military maneuver as the movement of forces on land, sea, and air misses the greater point of those movements: the goal of maneuver is to increase advantage relative to the opponent, to mitigate risk to one's own forces while maximizing the weaknesses of the opponent's [106]. I do not intend to insert myself in the ongoing semantic debate about the precise definition of military maneuver, nor to authoritatively assert how to best intepret the idea in a virtual domain such as the IE. I will adopt for this thesis these tenets of maneuver, gleaned from Hamilton's excellent analysis and the sources he cited in its composition [106].

- The goal of maneuver is relative advantage. Maneuver consists of employing one's assets specifically to capitalize on one's own strengths and/or the opponent's weaknesses.
- Maneuver includes both premeditated and reactive components. Deliberate planning and coordinated action are essential, as is the ability to recognize and exploit emerging opportunity.
- Basic principles of combat apply in discussing maneuver. Ideas such as mass, evasion, and surprise apply in the information domain as in the traditional physical domains of conflict.

Toward these ends, OPIEM and BENDRS are tremendously useful. OPIEM allows leaders to more clearly understand the IE, and therefore to better recognize opportunities within it, both longstanding and newly emergent. BENDRS provides concise labels, a ready-made toolkit to name the efforts, opportunities, and risks that an OPIEM model reveals within the IE.

6.2 Examples of maneuver

I provide several vignettes as examples of OPIEM-enabled information maneuver.

6.2.1 Countering insider threats

Identifying and countering insider threats is a persistent concern for organizations, regardless of size or sector [29]. Research has demonstrated that insider attack precursors can be detected within an IE-derived social network [179]. Analysis of cyber activity has been explored as a means of detecting insider attacks [251].

OPIEM models can bridge these social and cyber approaches, enabling a holistic approach to insider threat handling. Consider the following vignette:

> Vincent is the head of cybersecurity for a major corporation. As part of his duties, he routinely scrapes employees' social media feeds, as well as company-internal traffic (all with employee knowledge and consent, of course). Vincent uses this traffic, along with internal organizational charts, personnel assessments, and other documents, to construct an OPIEM model of his corporate IE. The model contains all of his employees, their respective significant external influences, the topics about which they communicate, and the platforms they access and reference in those communications.
>
> With each update, Vincent runs a full BENDRS analysis on this model. The analysis parses traffic to identify BENDRS maneuver attempts by the various agents, and examines changes in the model over time to identify BENDRS effects within the IE. Vincent can then review his analysis for key insider threat indicators:
>
> - Social isolation within the company, and/or increased social connectivity to hostile external groups. Vincent watches for company agent groups subject to Negate and Narrow, indicating increased isolation of members. He checks for agents conducting Bridge with known hostile groups, indicating increased communication with possible negative influences.
> - Increased engagement with hostile or subversive topics. Vincent checks for Engage and Excite effects between employees and those topics, watching for individuals or groups that seem increasingly interested in counter-company narratives. He monitors for Dismay effects on protected or sensitive company issues, as well as Dismay effects on well-being narratives like finance and future prospects. He further watches for Reveal or Repeat of those problematic topics within company platforms, watching propagation of counter-company dialogue.
> - Expanded access to sensitive material (topics) or systems (platforms). Vincent looks for Reveal of sensitive topics on non-company platforms, clear signs of leaking information. He also watches for the Rollout or Recommend of external platforms relative to employees, vigilant for sudden employee interest in new information systems. He checks for Engage of sensitive topics, looking for individuals or groups with a sudden increased interest in such material.
> - Changes in traffic or cyber behavior (recognized in the agent-platform network). Vincent watches for Repeat of sensitive topics, indicating an increase in access that may be linked to exfiltration attempts. He also watches for Smother and Stifle against those same target topics, as they may represent efforts to

mask threat behavior.

By feeding traffic into an OPIEM-modeling algorithm, and then running an automated BENDRS analysis on the resulting model, Vincent can quickly and quantitatively monitor his company information environment for insider threat precursors and indicators.

6.2.2 Countering malign influence

As previously established, hostile external influence attempts are significant threats to societal function. Multiple frameworks and methods have emerged to help policymakers confront this threat, including DISARM, D-RAIL, ABCDE, and others examined in chapter 1.

OPIEM models can help policymakers identify such threats as they unfold, and offer a useful operational layer in directing counteractions, even in concert with the other frameworks listed. Consider the following vignette:

An NSA information operations cell ingests live social media data at scale to construct an OPIEM model. Cell leaders monitor the shifting topic to designate topics as hostile or problematic, in part by identifying topics that known bad actors are targeting with Engage, Excite, Enhance, Repeat, and/or Reveal maneuvers – all indicators of a desire for greater exposure of these topics.

BENDRS analysis indicates Repeat actions for specific narratives on fringe platforms, with accompanying Boost actions of proponent accounts. These accounts also begin to execute Bridge actions to more mainstream platforms. Analysts determine the likely goal is a Reveal of the problem narratives, to facilitate follow-on Engage and Enhance.

Determined to head the problem off early, cell leaders recommend immediate counteraction. They seek to Boost institutional accounts that will Explain narratives likely to be targeted or undermined by the hostile narrative. They take action to Narrow the proponent community and to Negate specific bridging accounts. They Smother the hostile narrative on the fringe platform, and introduce Distort narratives to fragment discussion. And, they take action to Stifle spillover instances on mainstream platforms in cooperation with platform operators.

The cell then scrapes updated data, rebuilds the OPIEM model, re-runs the BENDRS analysis for updated metrics, and continues to adjust their counteraction plan as necessary.

6.2.3 Public education campaigns

Just as OPIEM can assist in countering malign influence campaigns, it can improve the reach and efficacy of constructive influence campaigns. Consider the following vignette:

Dawn is head of a public health office in a developing nation. Recent outbreaks of polio have produced significant fatalities and taxed local healthcare resources in already-impoverished areas, where vaccination rates have recently declined. Dawn

and her team seek to increase vaccination rates and thereby expand healthcare capabilities in those regions.

Using social media traffic, regional news productions, and internal communciations from local schools and community organizations, Dawn constructs an OPIEM model of the target IE. Agents include the parents, local influencers, healthcare providers, and school officials for the regions in question. Topics include vaccine safety, personal liberty, legal requirements, and conspiracy theories. Platforms include social media channels and groups, local news sites and sources, and local group newsletters and networks (i.e. churches, PTAs, etc).

Dawn's OPIEM-based analysis reveals clusters in the Agent x Topic group. One group is strongly aligned with safety concerns; another with government overreach. Between these clusters is a large but weakly engaged middle populace. The Topic x Platform network shows misinformation concentrated to a few specific social media groups and channels. Neutral and accurate information is present, but has low prominence on widely accessed platforms like the local news. Finally, the Agent x Platform network shows pediatricians and medical sources have high trust, but low reach. PTA leaders and church coordinators emerge as high-centrality bridge nodes between platforms. Dawn assesses that the issue is not an absence of good information, but misaligned availability and weak amplification of trustworthy voices.

Dawn gathers her team of government influencers, a collection of overt messengers (public and civil affairs) and covert messengers (intelligence and cyber operatives). She prescribes a series of maneuvers to rectify the situation:

- Repeat accurate vaccine information across the IE. Increase incidence on all platforms of useful information.
- Boost trustworthy local messengers, such as pediatricians, school nurses, and aligned PTA leaders.
- Bridge high-trust medical actors with socially central but topic-neutral community figures.
- Explain vaccine risks and benefits clearly, directly addressing concerns.
- Recommend reliable information sources, and Sideline known sources of bad information.
- Engage the populace on vaccine issues, encouraging engagement and research.
- Dismay the populace regarding fatalities to induce urgency.

Dawn's team acknowledges her instructions and gets to work. Her public affairs team crafts short, locally branded explainers for wide circulation, and ready-to-share content briefs for influencers. Her civil affairs team reaches out to pediatricians and civic leaders to recruit them as backers, and begins scheduling live Q&A events. Her cyber team sets up moderated comment and question fora online, while also covertly attacking and limiting the most deliberately malignant influencers and platforms.

Using OPIEM and BENDRS, Dawn's public influence campaign is better positioned

to achieve her goals. Crucially, Dawn achieves this success by identifying the necessary BENDRS maneuvers; she does not need to understand or directly specify the way her team goes about effecting the directed impact.

6.3 Implementations and Interventions

As implied in the vignettes above, there are many diverse ways to implement any given BENDRS maneuver. Any of the maneuvers can be attempted through both online and offline means, by both overt and covert methods. Here, I provide examples of various tactics and techniques that might be part of executing various maneuvers. This list is by no means exhaustive, and is intended to serve as an example of the operational-to-tactical abstraction provided by BENDRS. The implementation methods given as examples are domain-specific and rely on the expertise of the executor for efficacy. By specifying not the specific implementation, but only the BENDRS maneuver – and thus the desired impact – operational leaders give greater flexibility to executing experts, and can still lead effectively even if they lack specific expertise themselves.

Where appropriate I link techniques to the MITRE ATT&CK and DISARM frameworks, as they are well-developed taxonomies for actions in cyber and social influence respectively. I stress that not all possible actions are captured under the MITRE frameworks, and that any given MITRE technique can map to multiple BENDRS maneuvers. Further, many DISARM and ATT&CK techniques are too implementation-specific to map to a BENDRS maneuver directly. ATT&CK technique T1484 Domain or Tenant Policy Modification is a tremendously powerful technique, for example, but could be leveraged in service of almost any BENDRS maneuver as a desired outcome. As such, many techniques emerge frequently as possible BENDRS ”nearest match” implementations, while others are too specific or too broadly applicable to bear mention.

ATT&CK techniques are prefixed ”T”, while DISARM techniques are prefixed ”DA”. For the sake of brevity I include only a subset of maneuvers.

6.3.1 Community maneuvers

Back

- Create or expand controlled online personas to visibly support and validate the target (T1585.001 Establish accounts; DA0002 Develop personas)
- Elevate actor prominence via manipulated web presence (e.g. ”featured expert” placement (T1601 Modify system image / content)
- Stand up controlled platforms that cite and reinforce the actor’s authority (T1583.006 Acquire infrastructure)
- Use repeater bots to increase visibility of pro-actor messaging (DA0005 Amplify content)
- Explicitly position actor as authorititative via credentials and endorsements (DA0003 Establish credibility)
- Feature actor in mainstream media (interviews, op-eds); award symbolic recognition (titles, roles, speaking positions)

Bridge

- Enable cross-group communications channels through mailing lists and in-person outreach (T1585.003 Establish accounts (email))
- Identify overlapping members or entry points between groups (T1598 Phishing for information – relationship mapping)
- Direct content specifically across group boundaries (T1608.005 Stage capbilities - link targeting)
- Conduct direct interaction between groups via comments, replies, discussion (DA0006 Engage with target audience)
- Insert messaging into adjacent communities (DA0009 Leverage existing communities)
- Synchronize narrative expression so that they appear shared across separate groups (DA0011 Coordinate messaging across channels)
- Conduct joint events (town halls, panels) featuring representatives from both groups
- Identify and amplify influencers that are members of both groups (DA0005 Amplify content; DA0003 establish credibility)

Narrow

- Create and seed initial ”members” of a desired group (T1585.001 Establish accounts; DA0002 Develop personas)
- Create formal identity hub (e.g. website or forum) for new group (T1583.001 Acquire infrastructure)
- Use bots to simulate coordinated activity, giving the appearance of a cohesive group (T1587.001 Develop capabilities (malware/bots))
- Introduce narratives that define the group's purpose (DA0004 Seed content)
- Advertise a specifically framed version of the population the group claims to represent (DA0001 Identify target audience)
- Create a named coalition
- Launch a social media group with curated membership (T1583 Acquire infrastructure)
- Publish a manifesto or shared statement of identity (DA0004 Seed content)

6.3.2 Narrative maneuvers

Explain

- Deploy structured, explanatory content across multiple platforms (T1608.001 Stage capabilities - upload content; DA0004 Seed content)
- Host authoritative explanatory resources, like FAQs and dashboards (T1583.001 Acquire infrastructure)

- Tailor explanations to audience context (T1591 Gather victim information)
- Answer questions and refine understanding interactively (DA0006 Engage with target audience)
- Reinforce explanations with data, sources, and validation (DA0010 Provide evidence / supporting info)
- Provide visualizations and tools that make complex information accessible

Distort

- Alter source material (documents or datasets) to mislead (T1656.001 Data manipulation - stored data)
- Distribute altered narratives at scale (T1608.003 Stage capabilities - upload modified content)
- Hijack trusted sources to deliver altered or distorted information (T1584.001 Compromise infrastructure)
- Publish material with altered framing, omitted context, or selectively edited information (DA0008 Manipulate content)
- Inject distorted narratives into trusted spaces (DA0009 Leverage existing communities)
- Republish statistics, images, or video with altered or skewed perspective

Distract

- Flood channels with alternative content (T1608.004 Stage capabilities - upload content)
- Support parallel narratives at scale through dedicated sites/channels (T1583.006 Acquire infrastructure - web services; DA0011 Coordinate messaging across channels)
- Generate activity and discussion around competing topics, elevating unrelated or tangential topics (DA0005 Amplify content (competing))
- Keep attention anchored away from topic issue with constant updates, using viral or outrage content (DA0012 Sustain engagement)
- Launch unrelated but emotionally engaging content
- Time major announcements to overshadow target topic developments

6.3.3 Availability maneuvers

Shutdown

- Delete data, rendering the platform unusable (T1485 Data destruction)
- Disrupt platform availability (e.g. DDOS) (T1499 Endpoint denial of service)
- Corrupt system behavior to disable functionality (T1565.003 Data manipulation - runtime)

- Block access to platform (DA0015 Deny access to information channel; T1531 Account access removal)
- Eliminate key nodes from the network (DA0016 Deplatform / remove accounts)
- Enact legal or regulatory bans
- Impose or induce platform self-deactivation through litigation or financial exigency
- Infrastructure-level filtering (e.g. ISP, app store) (DA0015 Deny access to information channel)

Recommend

- Direct users toward preferred content pathways through link placement and recommendations, including onboarding flows or disguides "join here" links (T1598.003 Phishing for information - Spearphishing via service)
- Embed platform-specific links across channels to funnel traffic to target (T1608.005 Stage capabilities - link targeting)
- Create branded entry points, landing pages, or redirects channeling users to target platform (T1583.001 Acquire infrastructure)
- Position platform as trustworthy, authoritative, or aligned with user values (DA0003 Establish credibility)
- Tailor recommendations of the platform to specific communities – "this is where people like you are" (DA0007 Target audience segmentation)
- Encourage established groups to migrate collectively to the preferred platform, offering incentives or advantages (DA0009 Leverage existing communities)
- Secure endorsements from major influencers – "follow me on X instead of Y"
- Offer migration incentives and cross-platform compatibility or import features
- Embed target platform into desirable workflows; bundle as default in other systems or ecosystems

Stifle

- Prevent propagation of certain content (T1562.006 Impair defense - indicator blocking)
- Remove specific instances while leaving broader content intact (T1070.004 Indicator removal - File deletion)
- Replace or obscure content visibility (T1491.001 - Defacement; T1656.001 Data manipulation - stored data)
- Reduce or limit distribution without banning outright, e.g quota implementation (DA0013 Suppress content)
- Impose algorithmic downranking or throttling, i.e. "shadow banning" (DA0014 Limit reach/visibility)

- Rate-limit shares or replies of target, or reduce recommendation eligibility

Repeat

- Sustain repeated messaging through multiple accounts/actors (T1585 Establish accounts)
- Increase posting cadence via additional channels (T1586.002 Compromise - Accounts - social media; DA0002 Establish personas)
- Maintin continuous content injection across channels (T1608.004 Stage capabilities - Upload content)
- Reinforce repeated exposure per user (DA0005 Amplify content)
- Maintain ongoing interaction cycles by varying content without altering underlying topic (DA0012 Sustain engagement)
- Reintroduce core narrative(s) over time (DA0004 Seed content)
- Ensure consistent talking points across speakers (DA0011 Coordinate messaging across channels)

Table 6.1 summarizes the mapping between BENDRS maneuvers and MITRE techniques. I emphasize that not all MITRE techniques map to BENDRS maneuvers; nor are the MITRE techniques an exhaustive catalog of possible methods for executing any given BENDRS maneuver.

Table 6.1: Partial mapping of MITRE techniques to BENDRS maneuvers

BENDRS Maneuver	ATT&CK techniques	DISARM techniques
Back	T1583, T1585, T1601	DA0002, DA0003, DA0005
Bridge	T1585, T1598, T1608	DA0006, DA0009, DA0011
Narrow	T1583, T1585, T1587	DA0001, DA0002, DA0004
Explain	T1583, T1591, T1608	DA0004, DA0006, DA0010
Distort	T1584, T1608, T1656	DA0004, DA0006, DA0010
Distract	T1583, T1608	DA0005, DA0011, DA0012
Shutdown	T1485, T1499, T1531, T1565	DA0015, DA0016
Recommend	T1583, T1598, T1608	DA0003, DA0007, DA0009
Stifle	T1070, T1491, T1562, T1656	DA0013, DA0014
Repeat	T1585, T1586, T1608	DA0002, DA0004, DA0004, DA0011

6.4 Training

OPIEM is, by construction, ready for operational use. This includes use for training purposes. Both operational-level and implementation-level information operators require training to deal with the complex and challenging IE; ideally, this training would be as realistic as possible, but not conducted within or upon actual human beings within the IE.

6.4.1 Live Virtual Construct methodology

In chapter 4 I discussed Project OMEN, an Information Operations training suite that provides a high-fidelity virtual IE suitable for training analysts. OMEN is an example of a Live Virtual Constructive (LVC) training environment, in which Live (real people and real systems), Virtual (real people using simulated systems), and Constructive (simulated people and systems) aspects are combined into a single cohesive environment [114].

LVC training provides the benefits of rigorous effects- and outcomes-based training without the ethical perils of direct real-world testing. It allows analysts to train with actual systems and tools, rather than mock-ups, within realistic (but ultimately synthetic) environments that present the same complexities and challenges anticipated in real-world use.

As described in chapter 4, GhostCell provides reasonably complex social media traffic for OPIEM modeling. Project OMEN produces much more realistic data, in the form of social media traffic and web content (e.g. sites and articles), simulating a large-scale information environment with thousands of actors. Social media data is produced in compliance with the simulated platforms' respective APIs, enabling the use the same analysis tools and methods the trainees would employ for real-world data. Further, OMEN allows exercise controllers to modify or inject new data during an exercise, allowing them to shape the virtual IE in service of specific training objectives. Thus OMEN combines Live aspects (analysts using service-ready systems and tools), Virtual aspects (moderator-guided IE simulations), and Constructive aspects (thousands of simulated agents and platforms) to enable highly effective information operations training.

As the preceding vigenettes demonstrate, both OPIEM and BENDRS are inherently useful in an operational context, and thus can be employed as-are in an LVC training environment. The efficacy of OPIEM, however, is bounded by the fidelity of the data used to construct the network model, which means the utility of OPIEM as a training tool – and as an operational tool beyond – is ultimately driven by the data available to the modeling analyst.

6.4.2 IE simulation fidelity

The quality and availability of real-world operational data are not within the purview of my thesis, as they are issues dominated by political, financial, legal, and ethical concerns. My foray into synthetic social media data production, as realized in GhostCell, offers significant insight into key challenges and nuances that must be addressed in an LVC environment to ensure data – and the resulting OPIEM models – are sufficiently accurate, realistic, and nuanced.

Modeling platforms

As of this publication, Project OMEN generates traffic for two social media platforms – Twitter/X and Telegram – and open platforms as specified by the scenario author. By contrast, GhostCell simulated activity on six social media platforms (Twitter/X, Telegram, Reddit, LinkedIn, Facebook, and Insta), while only simulating five discrete open platforms.

Adding platforms to GhostCell is trivial – I need only specify a handful of parameters (as described in Appendix B). By contrast, adding platforms to OMEN is a significant technical task. The gap in effort is because GhostCell significantly simplifies output format and platform

architecture compared to OMEN. By contrast, OMEN does not incorporate user perception of platforms, and the resulting selection biases, migration behaviors, and attention budgeting.

GhostCell's output format is intended only for use within the simulation, as GhostCell is *not* designed as part of an LVC environment. GhostCell simulated traffic is not compatible with analytical tools designed for use on real data. OMEN, by contrast, produces social media traffic as specified by the respective platforms' API, permitting analysis of the synthetic traffic as if it were gleaned from the actual platform being simulated. Thus the first challenge of adding a new platform to OMEN is mapping the simulation engine's traffic output into the new platform's API. While each platform has specific quirks, this task is generally not challenging, except as complicated by the next consideration: architecture.

Social media platform designers make decisions about architecture that have significant influence over how information and influence travels within their platform; consequently, architectural features must be represented in any simulation purporting to represent information and influence propagation. Simulating a platform to the exact specifics would be extraordinarily resource-intensive, but is likely unnecessary. Social media platforms can be generalized as a single model – a shared, uncurated information space – with variations in interface and features abstracted into parameters like Comfort, as in GhostCell. Within this single model, platform architectural variations can be represented as *channels*, and the different platform architectures can be captured by altering channel membership criteria.

Initially, social media platforms emerged with a specific "channel philosophy." Facebook and MySpace, for instance, were focused on users' pages, where they could post content and receive comments on that content. Each page was a content node, and users accessed new channels – others' pages and personal networks – by forming mutual links ("Friending" one another). Facebook offered a feed of content gleaned from pages users had already joined (via the Friend action), aggregating content from the set of channels in which the user had membership. Initially Facebook constrained searches for other pages, limiting users' ability to find and join new channels (i.e. other users' pages).

By contrast, Twitter was created around the idea of a single, globally shared channel, in which all content was visible to all users and competed for attention. As with Facebook, Twitter offered a news feed, rapidly aggregating content from "followed" or preferred users. However, all users' Tweets were ostensibly visible to any other user through search. This lack of restriction made viral messages much more possible, as seemingly disconnected users could have content quickly recommended to them through relational networks, or thorugh algorithmic promotion; the content was rapidly available in part because all the users were part of the same single, global channel.

Over the intervening decades, every social media platform has ultimately adopted successful features of the others; the result is a fairly homogenous set of features primarily differentiated by their established user bases. X allows private group messaging, eliding the single, global message space so central to its initial design. Similarly, Facebook's news feed now includes content from persons with whom the user has no mutual friendship, subverting the overlapping ego-network design of the original platform in service of a more globally visible, algorithmically-curated space.

This reveals an abstraction mechanic for social media platforms. Alll platforms are ultimately spaces where users can post and ingest content; they are differentiable in which content a user

ingests. The end goal of this endeavor is to model and understand the influence on a given user, which is determined by the content that user sees. This visibility is in turn driven by two broad factors.

The first is **accessibility**, what the user is **permitted** to see, as determined by segmentation, permission, invitation, and other such factors. This is a hard gate on the set of candidate content. I split this factor into three archetypes:

- *Global*: any user's content is, in principle, accessible.
- *Graph-bounded*: content limited by an ego network (friends/follows/invites).
- *Group-gated*: Limited to explicit channels/groups with membership control.

The second is **routing**, what the user is **likely** to see, as determined by algorithmic recommendation and by user curation (following/liking, e.g.). This is the ranking or sampling methodology on the set of candidate content. I posit that user curation is equal among platforms, in that all platforms allow users to actively shape their own feed's stochastic weighting, and therefore focus on differentiating platform recommendation behaviors. As such I split this factor into three archetypes:

- *Recency*: New content is given precedence, regardless of provenance.
- *Social-signal*: Content is weighted by engagement, ties, or votes from the user body.
- *Recommended*: Content is weighted by an interest model or behavioral modeling, often to the individual user.

Table Table 6.2 offers a breakdown of different social media platforms, grouped into these archetypal buckets. As mentioned above, each platform has over time adopted the features of others, and as such no single archetype captures any given platform – or conversely, each platform falls into multiple (if not all) archetypes, in some mode. As such I have listed the specific mode, feed, or feature of each platform where applicable in parens.

This abstraction facilitates modeling any social media platform by varying only the visibility and recommendation parameters, and the API of the output, which simplifies implementation significantly.

Modeling influence

As explored in both chapter 4 and Appendix B, GhostCell relies on a deliberately basic influence model, including holding some key agent factors static across the simulation. Topic salience, emotional state, communication style, and other agent-specific nuances of influence are either omitted or held static for simplicity.

OMEN implements a significantly more detailed model of agent state and influence, though this model currently treats only on agent-agent interactions. Agent influence derived from ingested messages is based solely on the content of the messages and the agent's perception of the author. Research has shown that source credibility plays a significant role in influence, and is partially modulated by platform [147] [54]. As such, a combination of the two approaches is warranted, with OMEN adopting a refined version of the agent-platform perception and preference mechanics explored on GhostCell. Specific possible improvements are further explored in chapter 7 and Appendix B.

Table 6.2: Comparison of platform architectures

		Accessibility		
		Global	*Graph-bounded*	*Group-gated*
Routing	*Recency*	X (following) Mastodon BlueSky 4chan	Instagram (following) Facebook (recent) LinkedIn (following) Snapchat (Friends stories)	Telegram WhatsApp Discord Signal
	Social-signal	Reddit (r/all) Stack Overflow (hot) Tumblr (trending)	Facebook LinkedIn Instagram (home feed) Nextdoor	Discord (threads) Slack Telegram (replies) MS Teams
	Recommended	TikTok (For You) YouTube (Up Next) Instagram (Reels) X (For You)	Facebook (suggested) Instagram (suggested) LinkedIn (rec.) Snapchat (Discover)	Discord (serv rec.) Telegram (chan rec.) Facebook (Grp rec.) Slack (chan sug.)

GhostCell also implements exogenous, non-social-media agent influencers, in the form of both event-driven opinion changes and closed platform influence. GhostCell's design assumption is that, while traditional media outlets (e.g. CNN) may hold a social media presence, they do so primarily to present links to the outlet's actual site in a relatively neutral way. As such their social media accounts serve not to influence, but to direct agents to the outlet's influencing article. Because of this assumption, I did not model organizational entities as agents. In GhostCell, all agents are humans, potential influence targets. I discuss non-human modeling below.

To better simulate the social media environment, OMEN must better simulate the IE *external* to social media. Adding exogenous influence channels for media to directly alter agent internal states is a useful approximation of the off-keyboard world agents encounter, ensuring that in the time between messages, agents do not remain static and inert. A realistic simulation should produce agents that engage with social media from a slightly altered position day-on-day, based on the other factors of the agents' lives.

Modeling migration

While user migration between platforms does occur, it is rarely a binary phenomenon (i.e. I leave one platform completely in favor of the new one). In reality, migration often takes the form of attention splitting, and is rarely permanent. Still, user movement across platforms, including new user adoption and user departure, represents significant churn in the structure of the IE. I discuss improvements to migration mechanics in chapter 7 and Appendix B.

Modeling non-human agents

OMEN offers a more realistic version of social media, in which organizations (both media and non-media) operate accounts and post content to social media platforms, in addition to other con-

tent. Further, OMEN does not assume that organizational accounts are minimally influencing. OMEN does impose some behavioral limitations on such non-human accounts, however: for example, a government account is expected to use more formal language as compared to individual accounts. Organizational accounts are limited in their ability to use hashtags, mentions, and other social mechanics, to ensure that they behave in a manner congruous to the owning organization's goals, whereas individual accounts are driven solely by the simulated individual's opinions.

In GhostCell, I approximated different social media user types using archetypes, as explored in Appendix A. Expansion of these archetypes would easily permit simulation of some non-human agents. At present, these user archetypes are used to generate user behavioral characteristics. With expansion, this generation could include hard gates or requirements that should produce the desired behavior from non-human archetypal accounts, such as corporations, public affairs offices, and news outlets.

These behavioral characteristics must include platform-specific behaviors, including adoption, usage preference, and followerships. Organizations are often "anchors" in an online space, sought out actively for information; such accounts' presence on one platform over another would be a significant draw for users aligned with the organization. This phenomenon is not unlike the network externality dividend discussed in Appendix B, and as noted there, is not currently implemented in GhostCell or OMEN.

The specific platform adoption and migration behaviors of platforms bears further research. In general, we may assume an organization to have significantly more "attention" than individuals, and to desire broad exposure; as such it is likely organizations would join *any* platform that achieved some minimal user base, and was within some tolerance of the organization's value alignment. The specific range of these tolerances and thresholds is grounds for further research.

Modeling preference and recommendation

As explored above, one of the significant differentiators between platforms is the content recommendation algorithm. I represent this mechanism when, for a given agent, I stochastically weight each social media message for potential visibility, as described in Appendix B, specifically section B.5.2. My weighting function is holistic and platform agnostic.

A future improvement would divide the message weight into two stochastic components: user curation and platform curation. User curation would capture user-driven elements that would increase the likelihood of a message's visibility: following or friend-ing the author; having an alert for the message's topic; searching for authorities on the topic in response to some exogenous exciting event; and so forth. Platform curation would capture platform-driven elements that similarly affect visibility: message engagement count; author follower count; trending topics; second- or third-order connections between the author and the reader; and so forth.

By dividing the message weight into these components, the model creates greater possibility to represent both individual users, and separate platforms, via parameterization. Different platforms can provide greater or lesser emphasis to aspects of each message, potentially as determined by the archetypes proposed above. Individual users, meanwhile, can have draw on their own archetypes in determining how they curate their own feeds – and consequently which messages they are more likely to encounter.

This mechanism enables a number of interesting interventions and experiments. Platforms

might alter weights by specific topics, or based on the measured internal polarization of a topic, and seek to minimize polarization by downranking especially polar messages, or by suggesting moderate messages toward users with highly polar positions. The effect of such interventions is unproven, but this methodology would allow reasonably rigorous simulation and thus investigation of their merit.

6.4.3 OMEN exercises

As previously stated, the OPIEM modeling methodology can be utilized in an LVC training environment, as it is apt for modeling both live and synthetic data. The BENDRS lexicon is similarly useful in an LVC context. OPIEM-BENDRS can contribute to OMEN-based LVC training events immediately, and can usefully guide the evolution of those events moving forward.

Immediate uses

Current OMEN exercises focus on information open-source intelligence (OSI) analysts. OPIEM can be utilized to model OMEN data as an agent-topic-platform tri-modal network, which can then be subjected to BENDRS analysis as described in the vignettes above. Current OMEN participants conduct similar modeling and analysis using BEND to understand and describe the IE, and to propose possible interventions in support of their simulated unit's mission. OPIEM would provide additional insight by adding the platform layer, allowing analysts to more directly address sources of information and influence within the IE, and to propose availability-based interventions.

Near-term uses

In the near future, as the OMEN model expands to better simulate multi-platform social media behavior, OPIEM and BENDRS will become increasingly invaluable in tracking and describing high volumes of traffic across multiple social media spaces. OPIEM's platform layer will assist analysts in identifying, creating, or contesting information bridges within the IE.

As OMEN expands its ability to incorporate participant-generated interventions into the simulation, the training audience will necessarily include not just IE analysts, but IE influencers as well. The expanded BENDRS lexicon will make it easier to accommodate different IE-oriented skillsets. In selecting methods of exectuion, public affairs, civil affairs, cyber, and psychological operations experts will be better served by having availability-based effects explicitly separated from narrative and community altering effects.

OMEN's simulation engine, coupled with these interventions, will enable simulated testing of proposed courses of action prior to adoption, something that will prove increasingly valuable as the training audience continues to expand.

Long-term uses

Ultimately, OMEN must seek to expand the training audience to include not only analysts and influencers, but organizational leaders: military and government personnel who hold decision

authority but do not have personal expertise in information operations or social influence. Here, OPIEM provides an excellent method to concisely visualize the IE for intuitive description. OPIEM's mechanic of folding the core network in response to specific queries will allow analysts to quickly and precisely respond to high-level queries from non-experts, as explored in chapter 4. And BENDRS will provide a similarly accessible toolset for those leaders to understand ongoing actions in the IE and to direct changes toward a desired endstate.

Exploring higher-level interventions also becomes possible. At the tactical level, participants might post messages, alter data, or interfere with traffic flow. At the organizational level, leaders might apply pressure to platform leaders through legal, political, or financial means. This pressure can be translated into the simulated environment by altering a platform's behavioral parameters, allowing OMEN to simulate an instance where a platform agreed to moderate content in certain way, or to shift its recommendation algorithm toward a desired outcome.

As an example, we might imagine an instance where a leader seeks to reduce polarization in a target populace, leading to a platform altering its recommendation weights to deprioritize content similar to the user's current feed in favor of "endorsed" or "authoritative" content from the government.

Chapter 7

Conclusion

This dissertation demonstrates the construction of an operationalized model of the information environment, offering greater understanding into socio-cyber influence phenomena. By generalizing the BEND framework to include information availability, through specific representations of information systems within the environment, this work offers policymakers and societal leaders a powerful tool to understand and manage influence, including combating malicious actors. This work also indicates multiple avenues for subsequent academic inquiry and refinement.

7.1 Contributions

7.1.1 Theoretical

Beyond confirming the influencing nature of information availability, this work examines quantitative methods for identifying such influence within the larger information environment. This work also provides a sufficient model for the information environment, incorporating availability-oriented influence with previously established community- and narrative-oriented influence. This model is sufficient for decision-oriented influence modeling within a populace of interest, relative to the decisions of interest, and can scale up to organizational and societal levels.

7.1.2 Methodological

My dissertation presents several methodological contributions in its approach to simulating and analyzing the information environment.

Platform Selection and Preference

In addition to explicitly simulating the function of platforms within the information environment, I propose a platform preference and selection mechanic that models agents' attention allocation within a competitive and crowded information ecosystem, and the influence that results from that allocation. This mechanism is increasingly important in accurately depicting information consumers in the modern environment, where users have ever more options competing for a fixed time/attention budget. My platform preference and selection method enables the model

to represent how user beliefs about platforms impact the influence those platforms can achieve within the populace. This is essential in creating a sufficient model: influencers, especially at the state level, have significant capabilities to influence platforms, including altering public perception of those platforms. The platform preference/selection mechanism allows the model to depict how such perception changes translate into opinion and belief changes regarding broader dialogues.

Mixed channel influence pressures

My work also simulates the varying effects of influence sources within the environment, modeling agents' received influence from peers via social media, and from non-peers via media outlets, additively. Previous work has explored the differences in these sources [193]. My work incorporates both sources as inputs into a single influence model, complete with parameters to alter agent susceptibility to the two respective inputs.

Availability maneuvers

This dissertation extends the BEND framework to include availability-oriented maneuvers, providing a lexicon for such maneuvers to produce the generalized BENDRS framework. As with BEND, BENDRS is a useful tool for non-expert leaders such as military officials and policymakers, who must understand and direct information activities regardless of their own subject matter expertise.

7.1.3 Application

This work includes an algorithm to construct a sufficient OPIEM network of the information environment solely from collected message traffic that could easily be adapted for use with message traffic from any source. Using that derived OPIEM network, this dissertation includes network operations and metrics that identify BENDRS maneuvers within the network. These operations are quantitatively consistent and computationally feasible even as the network scale increases.

7.2 Limitations

Qualitative retrospective analysis: Despite the many case studies available to demonstrate and describe influence-based availability and socio-cyber hybrid influence events, the data required for rigorous quantitative analysis – including constructing an OPIEM network of the information environment around the event – was not available. Thus my ability to generate an OPIEM network for analysis was limited to a simulated dataset.

Simulation assumptions and constraints: My simulation system includes multiple constraints and assumptions, which are further documented in the appendix. The primary limitations include:

- The simulated populace and communities are based on the US population without consideration or inclusion of other groups. Only five archetypal community groupings are used, representing the entire populace as five large blocs.
- The populace is static across the run. All actors are present from $t = 0$ and remain present throughout. For the short timeframe of the simulations presented here this is likely fine, and somehwhat mitigated by my sleep-timer mechanic, but additional insight could be gained by introducing new cohorts and aging out others over the length of the run.
- The population is not inherently anonymized when I take measurements: I have perfect awareness of Agents' cross-platform activities and can track them accordingly. In reality, matching user identities across platforms is a thorny problem and represents additional noise and uncertainty when interpreting network measurements.
- The underlying base topics are assumed to be orthogonal, which is likely not practically true; further, inter-topic influence likely varies at the individual level. Narratives within the simulation are connected to only a single underlying topic which is also unrealistically restrictive, but was appropriate for a first-version simulated environment.
- The Friedkin-Jenkins influence model selected is a convergence-style model, which may be inappropriate to the actual environment being modeled; further, I extensively modified the Friedkin-Jenkins model based on the additional factors and agent behaviors present in the simulation. The resulting model is sufficient to demonstrate influence occurring within the populace, but may not be completely accurate to the environment being simulated.

BENDRS cause and effect separation: As with previous work on the BEND framework, my research does not attempt to establish causal relationships between specific messages and detected maneuvers. Hickman and Blaine were similarly careful around this area [113] [30]. Operational leaders using BEND – and BENDRS by extension – are eager to understand the IE, and often just as eager to assign blame and designate productive targets. My research does not indicate this task is impossible, but it does not address the problem.

Channel vs Device distinction: As mentioned in chapter 1, the Platforms nodeset is broadly defined. It is sufficient in defining an IE build to understand and model decision influence. However, the Platform nodeset as constructed in chapter 4 and chapter 5 would be insufficient to extensively model detailed cyber actions. The Platform nodeset could, if needed, be decomposed into two complementary nodesets, **Channels** and **Devices**. Devices would comprise hardware such as smart phones and computers, while Channels would capture the services accessed through those devices (e.g. web servers, communications apps, video streams, etc.) Within the Devices nodeset, and in the Devices X Devices network, a traditional cyberspace representation could be built to model cyber actions at a more granular level. The Devices X Channels network would, in turn, model the propagation of those cyber actions into user-facing areas. Agents X Channels and Agents X Platforms would both play a role in modeling availability effects, as at that level, some effects would target devices while others would target services.

Cross-platform attribution: As mentioned above, my simulation mechanics allowed me to trace agent actions across multiple platforms. In reality there are a class of users that *desire* to be identified cross platform, which includes major influencers, political figures, and organizations. These are entities that benefit from identity recognition and the commensurate credibility. Rank-and-file users, however, often do not explicitly identify themselves in the same way across platforms, in which case it becomes extremely hard to ascertain with certainty a user's shared identity in separate social media spaces. An OPIEM model can make strong use of known cross-platform linkages, but the resulting network may be limited in its ability to answer specific queries, as the handful of "hub" agents that build cross-platform links may be insufficient to allow productive folds between separate platform-centric user clusters.

7.3 Future work

The potential of the OPIEM framework and the BENDRS paradigm is substantial, both for academic expansion and operational application. I am excited to continue work in this area, including in the efforts listed below.

7.3.1 Quantitative retrospective validation

Although I could not locate a dataset with sufficient Agent, Topic, and Platform detail to build an appropriately robust OPIEM network, historical data may still contain requisite information, albeit in a less accessible form. Opinion surveys from pre-digital times, when information platforms were more monolithic, could permit the construction of a tri-modal IE representation. This would allow a retrospective validation of OPIEM, allowing me to compare the important features found in the network against historical facts. Further, I could verify that the BENDRS metrics accurately detect known historical influence campaigns.

7.3.2 Platform switching and attention budgeting

GhostCell used a fairly simplistic model to drive Agent platform preference, including selection and switching behavior. Further study into how people use multiple platforms, and why, along with a more robust model capturing Platform-Topic and Platform-Target interactions in Agent choice, would be useful in constructing more accurate IE simulations.

7.3.3 Omen integration

GhostCell pioneered the tri-modal apporach to IE simulation, and my dissertation validated the utility and necessity of that approach. Additional work will be required to integrate this tri-modal approach into Omen, which is a significantly more complicated simulation engine.

7.3.4 Channel/Device nodeset division and cyber action modeling

As described in the Limitations section above, extending the tri-modal OPIEM core network to accommodate a higher fidelity representation of the cyber portion of the IE may be useful to practitioners at the implementation edge of an organization. Future work must investigate how to adapt BENDRS to this nodeset split, and whether the split is actually useful in modeling decision making.

Appendix A

Agent generation model details

This appendix documents the Agent creation method used by GhostCell. I refer to this module as AURA-L, for AURORA-Lite, since it is a highly simplified version of Omen's AURORA module. I present both a functional examination of the system, and a more detailed review of the inputs used in my dissertation.

A.1 Agent definition

Table A.1, Table A.2, and Table A.3 list the traits of GhostCell Agents.

Table A.1: Agent demographic traits

Trait	Data type	Values	Influencing factors
Age	Integer	[16,75]	Uniformly random
Community	String	(text)	Age; capped at 25% of populace
Gender	Category	{F,M}	Community
Race	Category	{White, Black, Latin, Other}	Community
Marital status	Category	{0,1,2}	Community
Education level	Float	[0,1]	Community
Wealth level	Float	[0,1]	Community, Education level
Attractiveness	Float	[0,1]	Age, Wealth level
Group Memberships	Set	$\{g_1, g_2, \dots\} \subset \mathcal{G}$	Group sampling algorithm
Core group memberships	Set	$\{g_1, g_2, \dots\} \subset \mathcal{G}$	Group sampling algorithm

A.2 AURA-Inputs

AURA-L takes the following inputs:

Table A.2: Agent positions traits

Trait	Data type	Values	Influencing factors
Initial Econ valence	Float	[-1,1]	Community, Wealth level
Initial Govt valence	Float	[-1,1]	Community, Education level
Initial Cult valence	Float	[-1,1]	Marital status, Age
Initial Comm valence	Float	[-1,1]	Community, Race
Initial Expe valence	Float	[-1,1]	Community, Education level
Initial Chan valence	Float	[-1,1]	Age, Wealth level
Topic saliences	Float	$[0,1]^6$	Community
Topic authority status	Boolean	{T,F}	Topic salience
Initial Platform alignments	Float	[0,1]	Platform positions, Agent saliences
Initial Platform trusts	Float	[0,1]	Platform security, Platform reputability
Platform comforts	Float	[0,1]	Age, Education level, Gender
Platform preferences	Float	[0,1]	Alignment, trust, comfort

A.2.1 Topics

The system takes a list of topics, as strings, to be used in the broader scenario. These are the underlying topics for which Agents maintain valence and salience values. Topics are stated as polars, such that Agents can reasonably be assigned a value in [-1,1].

In my dissertation, i used the following topics:

- Economy (Econ). -1 endorses socialist or redistributive systems, 1 endorses market-driven or capitalist systems.
- Central Government (Govt). -1 endorses libertarian/small government positions, 1 endorses authoritarian/strong government positions.
- Cultural Norms (Cult). -1 endorses pluralist and individualist models, 1 endorses moral regulation and integrative models.
- Community (Comm). -1 holds globalist/inclusive views, 1 holds nationalist/exclusive views.
- Elitism or Technocracy Expe. -1 endorses populism and direct democracy, 1 endorses expert leadership and insulated governance.
- Rate of Change (Chan). -1 favors incrementalism and a conservative approach to change, 1 favors reformism, disruption, and radical change.

Table A.3: Agent user traits

Trait	Data type	Values	Influencing factors
Social Media user level	Category	{Casual, Regular, Programmed, Addict}	Age, Attractiveness, Education level, Gender
Social Media user type	Category	{Average, Influencer, Superfriend, Publisher, Amplifier}	Age, Attractiveness, Wealth level
Attention level	Integer	$[1, \infty)$	Social media level & type
Ego	Float	(0,1)	Social media type
Energy level	Float	(0,1)	Social media level & type
MFT Alignment	Floats	$[0, 1]^6$	Uniform random
Plutchik Alignment	Floats	$[0, 1]^8$	Uniform random
Open platform accounts	Set	$\mathcal{P}_{\text{open}}$	Social media type & level, Preferences
Closed platform subscriptions	Set	$\mathcal{P}_{\text{closed}}$	Preferences

A.2.2 Communities

Unlike AURORA, which derives Agent communities from reference documents, AURA-L requires explicit definitions of communities as inputs. Communities represent mutually exclusive demographic blocs within the simulated populace; Agents are members of a single community.

Communities have significant influence on an Agent's formation. Specifically communities provide parameters used in determining an Agent's:

- Age,
- Gender,
- Race,
- Marital status,
- Education level,
- Wealth level,
- Initial valence and salience values.

Table A.4 shows the communities I used for my experiments. The parameters listed are explained in greater detail below.

Table A.4: Community parameters

Community:	ProgUrb	TradPop	InstMang	WorkPrag	Margin
Description	Progressive, urban	Traditional, populist	Institutional, managerial	Working-class, pragmatist	Marginalized communities
Age weights	[1.4, 1.25, 1.0, 0.75, 0.55, 0.35]	[0.7, 0.85, 1.05, 1.25, 1.4, 1.5]	[0.6, 0.9, 1.3, 1.25, 0.95, 0.8]	[1.05, 1.05, 1.0, 1.0, 0.95, 0.9]	[1.25, 1.15, 1.0, 0.85, 0.7, 0.4]
Gender PDF	[6,4]	[4,6]	[4.5,5.5]	[5,5]	[5.5,4.5]
Race PDF	[0.6, 0.12, 0.15, 0.13]	[0.75, 0.08, 0.1, 0.07]	[0.7, 0.1, 0.1, 0.1]	[0.65, 0.12, 0.15, 0.08]	[0.3, 0.3, 0.25, 0.15]
Marital PDF	[0.5, 0.4, 0.1]	[0.25, 0.65, 0.1]	[0.3, 0.6, 0.1]	[0.35, 0.55, 0.1]	[0.45, 0.4, 0.15]
Education parameter	0.62	0.45	0.72	0.44	0.41
Wealth parameter	0.56	0.53	0.64	0.44	0.46
Econ parameter	-0.55	0.35	0.1	-0.1	-0.3
Govt parameter	0.55	-0.35	0.65	-0.15	0.25
Comm parameter	-0.55	0.45	-0.1	0.05	-0.35
Expe parameter	0.25	-0.45	0.75	-0.2	0.05
Econ hook	0.3	0.1	0.2	0.7	0.4
Govt hook	0.2	0.3	0.8	0	0.2
Cult hook	0.4	0.8	-0.1	0.1	0.6
Comm hook	0.3	0.7	0.1	0.1	0.8
Expe hook	0.2	0.4	0.7	0.1	0
Chan hook	0.4	0.2	0.1	0.3	0.3

A.2.3 Platforms

AURA-L requires a list of the platforms present in the simulation. On Agent creation, AURA-L determines which platforms the Agent uses, and the Agent's status on those platforms as applicable. Platforms have several attributes that are used both for Agent creation and for behavioral simulation. Table A.5 and Table A.6 provide the specific platform parameter values used in my dissertation.

Table A.5: Platforms (open)

Platform	Facebook	Twitter	Reddit	Insta	LinkedIn
Description	Liberal-leaning, facted-checked, moderated	Conservative-leaning, unmoderated	Neutral	Slightly liberal, youth oriented	Slightly conservative
Econ position	-0.25	0.4	0	-0.1	0.35
Govt position	0.2	-0.35	0	0.05	0.3
Cult position	-0.3	0.1	0	-0.45	0.4
Comm position	-0.15	0.25	0	-0.2	-0.05
Expe position	0.35	-0.45	0	0.05	0.6
Chan position	0.05	0.6	0	0.15	0.1
Security	0.5	0.4	0.4	0.5	0.7
Reputability	0.4	0.25	0.5	0.3	0.6
Base parameter	0.7	0.4	0.1	0.9	0.3
Age parameter	-0.9	0.9	0.6	1	-0.2
Education parameter	0.1	0.2	0.7	-0.1	1.1
Gender parameters	0.2	-0.1	-0.25	0.5	0.1
κ	14	12	12	14	13

Table A.6: Platforms (closed)

Platform	LeftNews	RightNews	IntlNews	ExtLeft	ExtRight
Description	Mainstream left-leaning	Mainstream right-leaning	Neutral	Extreme left-leaning	Extreme right-leaning
Econ position	-0.2	0.55	0	-0.7	0.65
Govt position	0.3	0.2	0	0.3	-0.1
Cult position	-0.25	0.4	0	-0.55	0.75
Comm position	-0.2	0.35	0	-0.45	0.7
Expe position	0.4	0.3	0	0.25	-0.4
Chan position	0.15	-0.15	0	0.65	0.25
Security	0.8	0.8	0.7	0.6	0.6
Reputability	0.75	0.75	0.85	0.4	0.4
Base parameter	-0.2	-0.2	-0.4	-0.3	-0.1
Age parameter	-0.1	-0.1	0	0.2	0.3
Education parameter	0.8	0.7	1	0.6	0.4
Gender parameters	0.05	0	0	0.05	-0.05
κ	12	12	13	11	11
Econ beat value	0.16	0.14	0.22	0.26	0.1
Govt beat value	0.18	0.18	0.2	0.14	0.14
Cult beat value	0.19	0.24	0.12	0.16	0.26
Comm beat value	0.1	0.2	0.16	0.08	0.28
Expe beat value	0.22	0.14	0.1	0.22	0.16
Chan beat value	0.15	0.1	0.2	0.14	0.06
ρ_{topic}	0.3	0.35	0.15	0.45	0.5
Responsiveness	0.25	0.3	0.45	0.25	0.35
Prominence	25	25	20	3.5	3.5

A.3 Agent creation

Agents are created in the following sequence.

A.3.1 Core demographic traits

I first determine demographic traits, from which to derive more specific traits. The possible values of these traits, and the parameters by which I defined the sampling distributions, were adjusted to match available US population demographics from the US Census Bureau, Pew Research Center, and Gallup. For Agent a,

1. **Age.** a's age is sampled uniformly from [16,75].
2. **Community.** Community membership is sampled from a piecewise-linear PDF. I block ages at anchor values $\{20, 30, 40, 50, 60, 70\}$. Each community has corresponding raw weights at those points; I derive an Agent's specific raw weight $w_c(a_{age})$ by interpolating. Then, a belongs to community c with probability:

$$P(c|a_{age}) = \frac{w_c(a)}{\sum_h w_h(a)}$$

   I sample from the resulting PDF to select a_c.
3. **Gender.** Gender is sampled from a discrete distribution, parameterized by a_c.
4. **Race.** Same as gender, parameterized by a_c.
5. **Marital status.** Same as gender, parameterized by a_c.

A.3.2 Intercorrelated demographic traits

I next determine secondary traits that are functions of the previously set demographics. Each of these attributes take values in $[0, 1]$ and are drawn from a Beta distribution, paramterized by a mean m and parameter κ such that $\alpha = m\kappa, \beta = (1 - m)\kappa$. In addition, I assign each agent a random "advantage" variable or shock variable, $z \sim \mathcal{N}(0, 1)$, to increase variation within the populace. I used the following function definitions:

$$\sigma(x) = \min\left(\epsilon, \max\left(\frac{1}{1 + e^{-x}}, 1 - \epsilon\right)\right)$$

$$\text{logit}(p) = \min\left(\epsilon, \max\left(\ln\frac{p}{1 - p}, 1 - \epsilon\right)\right)$$

Education Level

Education level a_{edu} is derived as:

$$a_{edu} \sim \text{Beta}(m\kappa_E, (1 - m)\kappa_E)$$

$$m = \sigma\left(\text{logit}(d_c) + \lambda_E \cdot z\right)$$

where d_c is the educational parameter for community c.

A.3.3 Wealth level

Wealth level a_{wth} is derived as:

$$a_{wth} \sim \text{Beta}(m\kappa_S, (1-m)\kappa_S)$$

$$m = \sigma\left(\text{logit}(w_l) + \lambda_S \cdot z + \beta_{SE}(a_{edu} - 0.5)\right)$$

where w_c is the wealth parameter for community c.

A.3.4 Attractiveness

Attractiveness a_{lks} is derived as:

$$a_{wth} \sim \text{Beta}(m\kappa_L, (1-m)\kappa_L)$$

$$m = \sigma\left(\text{logit}(w_l) + \lambda_L \cdot z + \beta_{LS}(a_{wth} - 0.5 - \beta_{LA}\dot{a}_{age})\right)$$

where w_l is the attractiveness parameter for the population and $\dot{a}_{age}$ is a's age, normalized as $[16, 75] \mapsto [0, 1]$.

A.3.5 Parameters

Table A.7 provides specific parameter values for the Beta distributions used. All parameters were tuned to approximate available census and demographic data.

Table A.7: Intercorrelated trait Beta distribution parameters

Trait	Mean weight	Shock weight	Secondary weight	Tertiary weight
Education level	$\kappa_E = 10.0$	$\lambda_E = 0.35$		
Wealth level	$\kappa_S = 10$	$\lambda_S = 0.45$	$\beta_{SE} = 1.2$	
Attractiveness	$\kappa_L = 12$	$\lambda_L = 0.25$	$\beta_{LS} = 1.0$	$\beta_{LA} = 1.4$

A.4 Initial position values

I assign a initial valence values based on a's previously assigned traits, as appropriate to each topic.

A.4.1 Economy

$$a(\text{Econ}) \sim \text{Beta}(m \cdot \kappa_{EC}, (1-m)\kappa_{EC})$$

$$m = \frac{\mu + 1}{2}$$

$$\mu = c(\text{Econ}) + \gamma_{EC}(a_{wth} - 0.5)$$

where $c(\text{Econ})$ is a community parameter, $\kappa_{EC} = 12.0$, and $\gamma_{EC} = 0.9$

A.4.2 Government

$$a(\text{Govt}) \sim \text{Beta}(m \cdot \kappa_{GO}, (1-m)\kappa_{GO})$$
$$m = \frac{\mu + 1}{2}$$
$$\mu = c(\text{Govt}) + \gamma_{GO}(a_{edu} - 0.5)$$

where $c(\text{Govt})$ is a community parameter, $\kappa_{GO} = 14.0$, and $\gamma_{EC} = 0.8$

A.4.3 Culture

$$a(\text{Cult}) = (2x - 1), x \sim \text{Beta}(m\kappa_{CU}, (1-m)\kappa_{CU})$$
$$m = \frac{\mu + 1}{2}$$
$$\mu = \alpha_{CU}(2\dot{a}_{age} - 1) + w_{mar}$$

where $\kappa_{CU} = 10$, $\alpha_{CU} = 0.9$, and w_{mar} is a weight based on a marital status, as:

$$[\text{Single}, \text{Married}, \text{Separated}] \mapsto [-0.25, 0.25, -0.05]$$

A.4.4 Community

$$a(\text{Comm}) = (2x - 1), x \sim \text{Beta}(m\kappa_{CO}, (1-m)\kappa_{CO})$$
$$m = \frac{\mu + 1}{2}$$
$$\mu = c(\text{Comm}) + w_{rac}$$

where $\kappa_{CO} = 10$ and w_{rac} is a weight based on a race, as:

$$[\text{White}, \text{Black}, \text{Latin}, \text{Other}] \mapsto [0.35, -0.25, -0.15, 0.05]$$

A.4.5 Expertise

$$a(\text{Expe}) \sim \text{Beta}(m \cdot \kappa_{EX}, (1-m)\kappa_{EX})$$
$$m = \frac{\mu + 1}{2}$$
$$\mu = c(\text{Expe}) + \gamma_{EX}(a_{edu} - 0.5)$$

where $c(\text{Expe})$ is a community parameter, $\kappa_{EX} = 14$, and $\gamma_{EX} = 1.4$.

A.4.6 Change

$$a(\text{Chan}) = (2x - 1), x \sim \text{Beta}(m\kappa_{CH}, (1-m)\kappa_{CH})$$
$$m = \frac{\mu + 1}{2}$$
$$\mu = \alpha_{CH}(1 - \dot{a}_{age}) - \beta_{CH}(a_{wth} - 0.5)$$

where $\kappa_{CH} = 14$, $\alpha_{CH} = 0.9$, and $\beta_{CH} = 0.6$.

A.4.7 Saliences

Saliences are assigned across topics for each agent, as follows. For topic r, Agent salience $a_S(r)$ is as:

$$a_S(r) \sim \text{Beta}(m\kappa_{sal}, (1-m)\kappa_{sal})$$
$$\kappa_{sal} = k_{min} + w(k_{max} - k_{min})$$
$$w = \sigma(\lambda_g \cdot z)$$
$$m = \sigma(b_r + h_{c,r})$$

where b_r is a baseline value, drawn from a table as:

$$\begin{bmatrix} \text{Econ} \\ \text{Govt} \\ \text{Cult} \\ \text{Comm} \\ \text{Expe} \\ \text{Chan} \end{bmatrix} \mapsto \begin{bmatrix} -0.2 \\ -0.5 \\ -0.1 \\ -0.4 \\ -0.8 \\ -0.6 \end{bmatrix}$$

$h_{c,r}$ is the hook paramter for topic r and community c, $\lambda_g = 1.0$, and k_{min}, k_{max} are population range parameters drawn as:

$$\begin{bmatrix} \text{Econ} \\ \text{Govt} \\ \text{Cult} \\ \text{Comm} \\ \text{Expe} \\ \text{Chan} \end{bmatrix} \mapsto \begin{bmatrix} [4.5, 16.0] \\ [6.0, 20.0] \\ [4.0, 15.0] \\ [4.0, 16.0] \\ [5.5, 20.0] \\ [4.5, 16.0] \end{bmatrix}$$

A.5 User traits

Next I assign traits that dictate a's social media behavior.

A.5.1 Social media level

Social media level archetypes capture different levels of personal focus on social media:

- *Casual:* uses social media when the mood strikes or reactively.
- *Regular:* uses social media as part of a regular routine or habitually.
- *Programmed:* uses social media deliberately as part of a strategy or larger goal, such as a professional influencer.
- *Addict:* uses social media compulsively as a dominant activity in their life.

Social media level is chosen using an ordered-logit model, with Agent age, attractiveness, education level, and gender as model inputs. An agent's latent engagement score s is:

$$s = b_0 + b_A(1 - \dot{a}_{age}) + b_L(a_{lks} - 0.5) + b_E(a_{edu} - 0.5) + b_G 1[\text{male}]$$

The categories are number-coded in order from least to most engaged. Cumulative probabilities are determined using cutpoints, such that:

$$P(Y \leq 0) = \sigma(t_{cas} - s)$$
$$P(Y \leq 1) = \sigma(t_{reg} - s)$$
$$P(Y \leq 2) = \sigma(t_{pro} - s)$$
$$P(Y \leq 3) = 1$$

The model is parameterized as:

$$\begin{bmatrix} b_0 \\ b_A \\ b_L \\ b_E \\ b_G \end{bmatrix} = \begin{bmatrix} 0 \\ 1.4 \\ 1.0 \\ 0.9 \\ 0.25 \end{bmatrix}, \begin{bmatrix} t_{cas} \\ t_{reg} \\ t_{pro} \end{bmatrix} = \begin{bmatrix} 0 \\ 1.9 \\ 2.95 \end{bmatrix}$$

such that the approximate population composition is 35% casual, 40% regular, 10% programmed, and 15% addict.

A.5.2 Social media type

Social media type archetypes capture different styles of social media engagement:

- *Average:* Normal users who follow friends, post personal (non-confrontational) updates, and use social media as a journal or newsletter for life.
- *Superfriend:* Users who 'like' or counter-engage with anyone and everyone, often as part of attention-seeking behavior. In order to keep broad appeal such users do not post as much original content and avoid controversial or strong opinions or positions.
- *Amplifier:* Users who use social media to specifically amplify or boost one or more target accounts, acting as a "hype man" for a specific person, group, or cause, without necessarily contributing strong original content toward the same.
- *Influencer:* Users who seek to maximize both follower count and network centrality, often as part of a monetization strategy. They see social media as an essential means to their broader goals.
- *Publisher:* Users who use social media use social media to announce, reveal, or advertise information, especially novel information. This would include journalism, product reveals, gossip, etc.

As with social media level, type is chosen using an ordered-logit model, with agent age, attractiveness, and wealth level as mdoel inputs. The model is:

$$s = b_0 + b_A(1 - \dot{a}_{age}) + b_L(a_{lks} - 0.5) + b_W(a_{wth} - 0.5)$$

$$\begin{bmatrix} b_0 \\ b_A \\ b_L \\ b_E \end{bmatrix} = \begin{bmatrix} -0.25 \\ 1.8 \\ 1.1 \\ 0.95 \end{bmatrix}, \begin{bmatrix} t_{avg} \\ t_{sup} \\ t_{amp} \\ t_{inf} \end{bmatrix} = \begin{bmatrix} 0 \\ 0.85 \\ 1.5 \\ 2.25 \end{bmatrix}$$

The parameter values are such that the approximate population composition is 40% average, 20% superfriend, 15% amplifier, 15% influencer, and 10% publisher.

A.5.3 Attention span

An agent's attention span a_{att} affects how long messages remain in the agent's awareness for possible engagement.

$$a_{att} \sim \frac{2}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\Phi\left(a\frac{x-\mu}{\sigma}\right)$$

where $a = 3.5$ and $\mu, \sigma = \mu/2$ are determined by the agent's social media level, as below:

$$\begin{bmatrix}\text{Cas}\\ \text{Reg}\\ \text{Pro}\\ \text{Add}\end{bmatrix} = \begin{bmatrix}16\\ 36\\ 48\\ 60\end{bmatrix}$$

A.5.4 Ego

An agent's ego a_{ego} drives how frequently the agent will post original content, as opposed to engaging with others' content.

$$a_{ego} \sim \text{Beta}(m\kappa_{ego}, (1-m)\kappa_{ego})$$

$$m = \sigma(SMT_{ego}), \kappa_{ego} = 14$$

where SMT_{ego} is determined by the social media type as:

$$\begin{bmatrix}\text{Avg}\\ \text{Inf}\\ \text{Sup}\\ \text{Pub}\\ \text{Amp}\end{bmatrix} = \begin{bmatrix}0.0\\ 0.8\\ -0.3\\ 0.3\\ -0.2\end{bmatrix}$$

A.5.5 Energy

An agent's energy a_{egy} affects how many engagements an agent will conduct in a given time slot.

$$a_{egy} \sim \text{Beta}(m\kappa_{egy}, (1-m)\kappa_{egy})$$

$$m = \sigma(SMT_{egy} + SML_{egy}), \kappa_{egy} = 18$$

where SMT_{ego}, SML_{ego} are determined by the social media type and level as:

$$\begin{bmatrix}\text{Avg}\\ \text{Inf}\\ \text{Sup}\\ \text{Pub}\\ \text{Amp}\end{bmatrix} = \begin{bmatrix}0.0\\ 0.2\\ 0.3\\ 0.1\\ 0.15\end{bmatrix}, \begin{bmatrix}\text{Cas}\\ \text{Reg}\\ \text{Pro}\\ \text{Add}\end{bmatrix} = \begin{bmatrix}-1.2\\ -0.5\\ 0.2\\ 1.0\end{bmatrix}$$

A.5.6 Delay

An agent's delay a_{lag} value determines the gap between engagements.

$$a_{lag} = \max(2.0, \mathrm{SML}_{\mathrm{lag}} \cdot \mathrm{SMT}_{\mathrm{lag}} \cdot \omega)$$

where SMT_{lag}, SML_{lag} are determined by the social media type and level as:

$$\begin{bmatrix} \text{Avg} \\ \text{Inf} \\ \text{Sup} \\ \text{Pub} \\ \text{Amp} \end{bmatrix} = \begin{bmatrix} 1.0 \\ 0.8 \\ 0.7 \\ 0.9 \\ 0.85 \end{bmatrix}, \begin{bmatrix} \text{Cas} \\ \text{Reg} \\ \text{Pro} \\ \text{Add} \end{bmatrix} = \begin{bmatrix} 24 \\ 12 \\ 8 \\ 3 \end{bmatrix}$$

and ω is distributed log-normally, $\omega \sim \mathrm{LN}(0, 0.15)$

A.5.7 Communications style

An agent's style is denoted by the agent's MFT alignment $[0, 1]^6$, representing their rhetorical and argumentative style, and the agent's Plutchik alignment $[0, 1]^8$, representing their emotional state. These are both initialized uniformly randomly and are static in the simulation.

A.6 Platform perceptions and preferences

Agents next generate initial perception values for each platform in the scenario, and compute their preference score based on those perceptions. Parameters for these distributions were tuned to produce realistic behavior within the simulation, based on my comparisons with observed and anecdotal evidence online.

A.6.1 Alignment

An Agent's Platform alignment score indicates how closely the agent believes that platform to be with the agent's own opinions. The alignment computed as:

$$\mathrm{align}(a, p) = (1 - \hat{d})^{\alpha}$$

Here, δ is a weighted distance score, and $\alpha = 7$ creates a fall-off that penalizes misaligned platforms in a superlinear fashion. The weighted distance $\hat{d}$ is computed:

$$\hat{d} = \frac{\sqrt{\sum_{\mathcal{R}}^{i} w_i (\phi_{a,i} - \phi_{p,i})^2}}{\sqrt{2 \sum_{\mathcal{R}}^{i} w_i}}$$

where $\phi_{x,r}$ is the position of entity x on topic r, and w_r is a's salience on topic r.

A.6.2 Trust

The Platform trust score indicates how trustworthy the agent believes the platform to be, both as a source and as a steward of information. This encapsulates both reporting veracity and information security. Initial trust score is computed as:

$$\text{trust}(a, p) = \frac{p_{sec} + p_{rep}}{2} + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, 0.05)$, and p_{sec}, p_{rep} are platform parameters.

A.6.3 Comfort

The Platform comfort score captures how easy and comfortable a platform is for the user. It stands in for the user experience, technological requirements, and signup/operation overhead. Comfort is computed as:

$$\text{comfort(a, p)} \sim \text{Beta(m}\kappa_{\text{p}}, \text{(1-m)}\kappa_{\text{p}}$$

$$m = \sigma(w_0 + w_A(1 - \dot{a}_{age}) + w_E(a_{edu} - 0.5) + w_G[1|a_{gen} = \text{male}])$$

where all w values are platform parameters.

A.6.4 Preference

An Agent's preference score for a given platform $\psi_a(p)$ is a function of the Agent's current alignment, trust, and comfort scores for that platform, as below:

$$\psi_a(p) = \text{align}(a, p)^{\alpha_a} \cdot \text{trust}(a, p)^{\alpha_t} \cdot \text{comfort}(a, p)^{\alpha_c}$$

The weight vector $\boldsymbol{\alpha}$ varies between open platforms (social media) and closed platforms (curated or traditional media). Values are below:

$$\begin{bmatrix} \alpha_a^{open} \\ \alpha_t^{open} \\ \alpha_c^{open} \end{bmatrix} = \begin{bmatrix} 1.0 \\ 0.7 \\ 1.3 \end{bmatrix}, \begin{bmatrix} \alpha_a^{closed} \\ \alpha_t^{closed} \\ \alpha_c^{closed} \end{bmatrix} = \begin{bmatrix} 1.2 \\ 1.8 \\ 0.8 \end{bmatrix}$$

A.6.5 Parameters

Table A.6 and Table A.5 list parameter values for the platforms in the simulation.

A.7 Group memberships

In GhostCell, groups simulate social ties between Agents. They represent ties built in formal settings (i.e. shared employment) and ties built informally (i.e. common interest groups). GhostCell creates four separate categories of groups, and then stochastically assigns Agent membership based on group criteria, as follows.

A.7.1 Constructed groups

Constructed groups represent institutional or purpose-built groups within a society, such as professions and employers. These are groups that capture a wide cross-section of society, as they are primarily functional.

Size

Constructed group size is randomly set between 6% and 15% of the total population.

Membership

Once group size is determined, the appropriate number of agents is randomly selected from the populace. Those agents selected become group members; there are no restrictions or additional checks.

Leadership

Group leaders are randomly selected from membership. A first leader is chosen; AURA-L then conducts a Bernoulli trial to determine if another leader will be selected. The Bernoulli trial uses parameter $p_0 = 1.0$, which is reduced as $p_t = p_{t-1}/1.5$ with each successive leader chosen until the result is negative.

Limitations

Ideally constructed groups would have more inherent character, and would thereby impose membership criteria based on Agent demographics (age, education level, etc.). The specifics of the group would also impose appropriate size caps.

A.7.2 Community groups

Community groups represent community-internal structures and associations. Examples include political parties or cultural/regional associations.

Size

A community group is capped at a maximum population of

$$|g|_{max} = \frac{\mathcal{A}}{2|\mathcal{C}^*|}$$

Here, $\mathcal{C}^*$ is the set of input communities. This is important: in other areas, I have used $\mathcal{C}$ to represent any arbitrary community partition of $\mathcal{A}$. Assuming roughly equivalent membership in all communities across the population, the size cap ensures that no more than half of a given community has membership in a single group.

Membership

For each candidate Agent (sampled in random order), AURA-L determines membership via a Bernoulli trial with $p = 0.65$ until all candidates have been considered or group membership reaches the size cap.

Leadership

AURA-L selects a random integer from $[1, 4]$, and then randomly chooses that many group leaders from the membership.

Limitations

Ideally community groups would have additional identifying information, and membership would be conditioned on additional Agent traits (wealth, race, etc.).

A.7.3 Attribute groups

Attribute groups represent homophily ties based on Agent demographic information. Examples include neighborhood associations or hobby groups.

Size

Attribute groups are bounded in the range $[8\%, 20\%]$ of the total population.

Membership

Membership probability is based on the distance between the group's centroid and a candidate agent's attribute coordinate. GhostCell currently uses $\text{centroid}(a) = [a_{age}, a_{lks}, a_{edu}, a_{wth}]$ as the attribute vector.

When created, an attribute group is assigned a random centroid in $[0, 1]^4$. For each Agent (in random sequence), the distance:

$$d_{a,g} = \|\text{centroid}(a) - \text{centroid}(g)\|$$

If $d_{a,g} < \delta_{att}$, a becomes a member of g. If all Agents are considered and g does not yet have the minimum membership, the threshold is adjusted as $\delta_{att}^{t} = 1.25 \cdot \delta_{att}^{t-1}$. This repeats until the group meets minimum size, or until g reaches the size cap. In the current simulation, $\delta_{att} = 0.15$.

Leadership

AURA-L selects a random integer from $[3, 10]$, and then randomly chooses that many group leaders from the membership.

Limitations

Ideally, attribute groups would be defined around specific centroids in correlated positions (e.g. the Young Republicans [age, race, community, wealth], or a recreational sports league [age, attractiveness]).

A.7.4 Position groups

Position groups represent homophily ties based on opinions. Examples include political organizations and online forums or support groups.

Size

Position groups are bounded in the range $[10\%, 25\%]$ of the total population.

Membership

Membership probability is based on the distance between the group's centroid and a candidate agent's valence coordinate $\boldsymbol{\phi}_a$.

When created, an position group is assigned a random centroid in $[-1, 1]^{|\mathcal{R}|}$. For each Agent (in random sequence), the distance:

$$d_{a,g} = \|\boldsymbol{\phi}_a - \boldsymbol{\phi}_g\|$$

Selection algorithm is the same as for Attributte groups. In the current simulation, $\delta_{att} = 0.1$.

Leadership

AURA-L selects a random integer from $[3, 10]$, and then randomly chooses that many group leaders from the membership.

Limitations

The simulation assumes that underlying topics are orthogonal, which is a useful simulation assumption. However, position groups in real life often form around issues, not underlying political/societal principles. Ideally, position groups would be instantiated as issue groups, with one or more centroids defined for both the 'pro' and 'anti' sides of the debate.

A.8 Topic authority

For each topic r, all Agents with salience $a_S(r) > \delta_{sal}$ are candidates for authority status. The number of authorities is randomly selected from the interval $[2, 5]$, and the corresponding number of Agents are randomly selected to be topic authorities.

A.9 Platform accounts

The final step in Agent creation is assigning online accounts and subscriptions.

A.9.1 Number of accounts

Agents are active on a variable number of social media platforms driven by the Agent's social media level and type. This number is drawn from a Gibbs distribution. First, I compute a factor score incorporating the Agent traits:

$$s = 0.05 + w_{SML} + w_{SMT}$$

where weight values are as:

$$\begin{bmatrix} \text{Avg} \\ \text{Inf} \\ \text{Sup} \\ \text{Pub} \\ \text{Amp} \end{bmatrix} = \begin{bmatrix} 0.0 \\ 0.8 \\ 0.4 \\ 0.3 \\ 0.2 \end{bmatrix}, \begin{bmatrix} \text{Cas} \\ \text{Reg} \\ \text{Pro} \\ \text{Add} \end{bmatrix} = \begin{bmatrix} -1 \\ -0.3 \\ 0.7 \\ 1.4 \end{bmatrix}$$

The number of accounts x is then sampled as:

$$P(X = x|s, \tau) = \frac{\exp\left(\frac{x-1}{\tau}s\right)}{\sum_j \exp\left(\frac{j-1}{\tau}s\right)}$$

For this research I set the temperature parameter $\tau = 1.3$. Once the number of accounts has been chosen, the Agent selects the top x accounts by preference score $\psi_a(p)$.

A.9.2 Follower counts

Followership determines an Agent's influence within the simulated ecosystem. I assign followership with a two-tier stochastic system. First, I conduct a Bernoulli trial with $p = 0.2$ to determine if the Agent is an "exceptional" user, as opposed to a standard user.

Standard users

Standard user a draws a follower count $f_a(p)$ on platform p as:

$$f_a(p) \sim \text{Lognormal}(9.1466, 0.5183)$$

where distribution parameters have been tuned against observed average follower counts and spreads on the simulated social media platforms (collected in December 2025).

Exceptional users

Exceptional user a has a power-law derived follower count, drawn using a modified Pareto distribution as:

$$f_a(p) \sim \left(\frac{f_{min}}{1 - u(1-a)} \right)^{1/\alpha}$$
$$a = \left(\frac{x_{min}}{x_{max}} \right)^{\alpha}$$
$$u \sim \mathcal{U}_{[0,1]}$$

where x_{min} =50,000, x_{max} =50,000,000, and $\alpha = 1.9$.

A.9.3 Media subscriptions

Research indicates digital natives maintain, on average, six sources of news [230]. Simplistically, I drive Agents to maintain six platform connections. I prioritize social media platforms (above), counting those accounts as news sources. Any remaining news sources (i.e. $n = 6 - |\mathcal{V}_{a,\mathcal{P}^{\wr\surd}}^{\wr\surd\rceil\setminus}|$), Agents subscribe to closed platforms, choosing the top n choices based on preference score.

A.10 Interpersonal Influence

To simulate Friedkin-Johnsen-type interpersonal influence between Agents [93], I pre-compute $\mathbf{W}$, the interpersonal influence matrix, where w_{ij} captures Agent i's influence from Agent j. For my simulation I opt to use strictly linear relationships. I use five components to compute w:

A.10.1 Common group membership

Letting $\mathcal{G}_{\dashv}$ denote the set of groups in which a is a member, I compute proportional scores:

$$w_1 = \frac{|\mathcal{G}_{\rangle} \cap \mathcal{G}_{|}|}{|\mathcal{G}_{\rangle}|}$$
$$w_2 = \frac{|\mathcal{G}_{\rangle} \cap \mathcal{G}_{|}^{*}|}{|\mathcal{G}_{\rangle}|}$$

where G^* indicates membership as a leader.

A.10.2 Expertise

j exerts some influence on i if i believes j to be an expert or authority. I compute this component score as:

$$w_3 = \frac{|\{r \in \mathcal{R} : \text{auth}(j,r) \cdot \text{sign}(\phi_i(r) \cdot \phi_j(r) > 0\}|}{|\mathcal{R}|}$$

where $\text{auth}(a,r) = 1$ if a is an authority on r and 0 elsewise. In short, I count the number of topics where j's position is aligned to i's, and j' is an authority, divided by all possible topics.

A.10.3 Categorical homophily

I compute trait-based homophily for categorical traits using a Hamming distance, using comparator $eq(a, b) = [a = b]$:

$$w_4 = \frac{1}{4} d_H([i_{community}, i_{gen}, i_{marital}, i_{race}], [j_{community}, j_{gen}, j_{marital}, j_{race}])$$

A.10.4 Continuous homophily

I compute trait-based homophily for continuous-valued traits as a Manhattan difference, as:

$$w_5 = \frac{1}{4} \sum_{\mathbf{x}} |x_i - x_j|$$

where $\mathbf{x_a} = [\dot{a}_{age}, a_{att}, a_{edu}, a_{wth}]$.

A.10.5 Total influence score

With these components, I apply weights: $\boldsymbol{\omega} = [0.2, 0.25, 0.25, 0.15, 0.15]$ and derive:

$$w_{ij} = \boldsymbol{\omega} \cdot \mathbf{w}$$

and stipulate that $w_{ii} = 0 \ \forall i$.

Appendix B

Social media traffic generation model details

This appendix documents the synthetic traffic generation system in GhostCell. I present both a functional examination of the system, and a more detailed review of the inputs used in my dissertation.

B.1 Overview

GhostCell runs a primary time loop. In each timestep, a subset of Agents is chosen to act. On their turn, each of those Agents will *Listen* to non-social media news sources, *Read* social media content, *Think* about what they've heard and read and update their internal state accordingly, and *Engage* with social media by posting and/or reacting to messages.

B.2 GhostCell Inputs

B.2.1 Topics

The Topics used are the same discussed in Appendix A.

B.2.2 Narratives

GhostCell requires a list of the narratives extant in the simulation. As discussed in chapter 4, Narratives are decompositions of larger Issues. Each Narrative is aligned to a single Topic, and has either a positive or negative sign, indicating the stance espoused by the narrative. In most cases, Narratives can be presented as a signed pair, as two directly opposing arguments. Table B.1 lists the base set of Narratives used in my dissertation research.

Table B.1: Base Narratives

Issue	Narrative	Topic	Polarity	Description
Abortion	Govt limits	Govt	-1	Government should not regulate personal medical decisions
	Protect life	Cult	1	Society has moral duty to preserve human life
	Bodily autonomy	Cult	-1	Individuals should have full bodily autonomy
	National values	Comm	1	Protecting life is our moral identity
Immigration	Labor competition	Econ	1	Immigration depresses wages
	Economic engine	Econ	-1	Immigrants drive growth
	Border security	Comm	1	Border enforcement preserves national cohesion and safety
	Nation of immigrants	Comm	-1	Immigration is part of our national identity
Climate change	Trust science	Expe	1	Climate policy should follow science
	Skeptical of experts	Expe	-1	Experts exaggerate climate change for personal gain
	Transform economy	Chan	1	Immediate transformation is needed
	Adapt economy	Chan	-1	Gradual change and innovation best balance risks
	Green investing	Econ	-1	Public investment should drive transition systems
	Regulation harmful	Econ	1	Environmental regulation damages growth
Public health	Federal mandates	Govt	1	Mandates are necessary in crisis
	Local control	Govt	-1	Public health policy should be local
	Trust experts	Expe	1	Public health agencies should lead response
	Do own research	Expe	-1	Health agencies are politicized and unreliable
	Personal risk	Cult	-1	Individuals should decide their own risk
	Herd immunity	Cult	1	Group well-being outweighs personal freedom
	Driving reform	Chan	1	Crisis shows need for healthcare reform

Table B.1: Base Narratives

Issue	Narrative	Topic	Polarity	Description
Crime	Federal intervention	Govt	1	Federal government should intervene when locals fail
	Local control	Govt	-1	Policing is a local authority
	Systemic reform	Chan	1	Police require radical structural overhaul
	Training reform	Chan	-1	Should make incremental change through training and policy
	Moral decay	Cult	1	Crime represents a breakdown in societal norms
	Global context	Comm	-1	Crime is driven by migration and interlopers
Education	National standards	Govt	1	Federal standards ensure equality
	Parental control	Govt	-1	Parents should make educational decisions
	Traditional curriculum	Cult	1	Schools should teach civic and moral values
	Identity inclusion	Cult	-1	Curriculum should reflect diverse viewpoints
	Expert design	Expe	1	Curriculum should be designed by trained educators
	Local accountability	Expe	-1	Educators should answer directly to voters and parents
	Incremental improvement	Chan	-1	Best reform is through pilot programs and data study
	Global competition	Comm	-1	Education prepares students to compete in the global market

Evidence

Narratives include supporting Evidence. In the simulation, Evidence objects represent links, images, quotes, or other externally sourced content that posters use to buttress the position expressed in a post. Evidence is extremely useful in analysis: tracking the movement of Evidence within the ecosystem provides insight into the propagation of different viewpoints. Omen uses actual URLs and images as evidence. GhostCell, by contrast, uses only abstract labels to represent the presence of evidence.

When GhostCell reads in the Narrative list, it identifies the Issue list – the larger Issues the various Narratives address – and generates N_{ev} pieces of evidence per issue. It then randomly assigns each Narrative 'favored' Evidence, meaning this Evidence originated with and/or is uniquely apt for the corresponding Narrative. GhostCell then randomly assigns 2-5 additional non-favored Evidences to each Narrative. Since both Evidence and Narratives are subcomponents of Issues, in all cases, Evidence-Narrative pairings have the same overarching Issue.

B.2.3 Events

GhostCell requires a list of exogenous events that will occur during the simulation. There is no baseline or standard list of Events, since Events are often used to induce experiment conditions during simulation runs. Discussion of the Event system can be found in chapter 4.

B.2.4 Platforms

GhostCell uses the same Platforms discussed in Appendix A.

B.2.5 Agents

GhostCell reads in the Agents created by AURA-L. this input includes the interpersonal influence matrix $\mathbf{W}$ computed by AURA-L. Details are in Appendix A.

B.3 Initialization

Prior to execution, GhostCell prepares the simulation environment as follows.

B.3.1 "Big hitters" list

GhostCell scans the Agent population and identifies the top N influencers, by follower count, on each open Platform. This "big hitter" list is used in later steps to confer additional influence to these top posters, simulating the effects of platform recommendation algorithms. For my research, $N = 5$.

B.3.2 Trend tracker

GhostCell initializes a Trend Tracking system, which captures proportional occurrences of narratives within a time window T_{tr}. This trend value is also used to simulate recommendation algorithm effects. For my work, $T_{tr} = 90$ time steps.

B.4 Main loop: Pre-Agent

The simulation runs for T_{run} time steps. In each time step t, the following occurs.

B.4.1 Event processing

Event activation

The system scans the Event list for Events with a start time t. Any such Events are activated, and immediate effects (the Alteration portion of the event) are applied to the simulation.

Event deactivation

Any Event with an end time t is set to deactivated status.

Event processing

For each active event e, the system executes three steps:

1. Event intensity ξ_e is reduced. The simulation reduces intensity from the initial value I_e linearly over the duration of e.
2. For all Narratives and Topics tied to e, excitement boost values are calculated for each Narrative $q_e(n)$ and Topics $q_e(r)$.
3. Event reflection effects are applied. For each Agent, the event bumps the targeted internal belief. Targetable beliefs include topic valences, topic saliences, and platform perceptions (trust, comfort, and/or alignment).

   An Agent's susceptibility to the bump $\pi_e(a)$ is computed from an Agent trait which is specified in the Event description; for my simulation all Events used $\dot{a}_{age}$. For Agent belief a_x, the bump is applied as:

   $$a_x^t = a_x^{t-1} + e_{sign} * \xi_e * \pi_e(a) * \rho_E$$

   where ρ_E is the global Event effect scale factor.

B.4.2 Activations

GhostCell identifies the in-play factors for time step t.

- Platforms: Platforms flagged as inactive are excluded for use and reference.

- Messages: The visible message corpus $\mathcal{M}^t$ is a reduced set of $\mathcal{M}$. Any messages from platforms that are inactive at t are not included in $\mathcal{M}^t$. Messages with a time stamp $m_t < t - T_{span}$ are excluded to keep run time manageable.
- Media: Closed platforms not flagged as inactive are included as acting media entities in t.
- Agents: Agents with a queue value of 0 are activated for action in t. All other Agents have their queue value decremented by 1.

B.4.3 Media actions

Each active media entity conducts a Bernoulli trial with probability p_{media}. If successful, that entity acts during t; if not, the entity is dormant. Acting entities:

Scan corpus

The media entity reads the available corpus $\mathcal{M}^{\sqcup}$ to measure the occurrence frequency of Topics and Narratives.

Select focus

The media entity (ME) then takes the occurrence frequencies found in its scans, along with any Event boosts, and the Platform's own weights, and determines a Narrative focus for its action in t.

The ME first selects a focus Topic. This choice is sampled from the Topics set with convex probability weights applied. The weight w_r for Topic r is derived as:

$$w_r = \omega_b B_p(r) + \omega_f f_r^t + \omega_e E(t, r)$$

B_p is the set of beat parameters for each closed Platform, shown in Table A.6. These values serve as priors for an outlet's tendency to cover r. f_r^t is the frequency of r in $\mathcal{M}^t$. $E(t, r)$ is the cumulative active event boost on r at t, computed as:

$$E(t, r) = \frac{1 - (\exp(\mathrm{q}_{\mathcal{E}^{\sqcup}}(\mathrm{r})/\lambda_{\mathrm{t}})}{\sum_{\mathcal{R}} 1 - \exp(q_{\mathcal{E}^{\sqcup}}(h)/\lambda_t)}$$

where $\mathcal{E}^t$ is the set of active Events at t, and λ_t is a saturation parameter.

Convex selection weights $\omega_b, \omega_f, \omega_e$ are derived as:

$$\omega_f = \rho_p(r), \quad \omega_e = 1 - \omega_f, \quad \omega_b = 1 - \omega_f - \omega_e$$

where $\rho_p(r)$ is a Platform parameter that balances the Platform's traditional topics with trend influence.

Topics are then weighted with respective w_r and sampled stochastically. The chosen topic $\hat{r}$ drives Narrative selection: candidate Narratives are only those connected to $\hat{r}$. Narrative selection weights w_n are derived as:

$$w_n = (1 + \nu_f \cdot f_n^t) \left(1 + \beta \left(1 - \frac{-q_{\mathcal{E}^{\sqcup}}(n)}{\lambda_e}\right)\right)$$

where ν_f is a global scaling factor, f_n^t is the frequency of n in $\mathcal{M}^t$, β is a bounding factor for Event influence, and λ_e is the saturation parameter for event excitement. $\hat{n}$ is selected using the calculated weights w_n.

Write article

The ME composes an article about $\hat{n}$, which is added to the media stream at t.

B.5 Main loop: Agent

Each active Agent conducts the following actions.

B.5.1 Ingest media ("Listen")

The Agent consumes all articles in the mediastream written by Platforms to which the Agent subscribes. Each artle affects the Agent's beliefs and perceptions, as follows.

Adjust valence

Articles nudge the reader's valence position based on the article Narrative and its underlying Topic. An Agent's updated position at t is:

$$\phi_a^t(r) = \phi_a^{t-1}(r) + stub(\phi_a^t(r)) \cdot \delta \cdot (\overline{\phi^t(r)} - \phi_a^t(r))$$

$stub(\phi_a(r))$ is the stubborness function, simulating Agent resistance to shifting opinion:

$$stub(\phi_a(r)) = \max\left((1 - |\phi_a(r)|)^{\beta_{stub}}, L_{stub}\right)$$

where β_{stub}, L_{stub} are model parameters. δ is the step size of the nudge:

$$\delta = \sum_{r \in \mathcal{R}} \eta_{posn} \cdot \chi(\phi_a(r) - \phi_m(r)) \cdot \text{trust}^*(a, p)$$

η_{posn} is a scaling factor for article influence. $\text{trust}^*(a, p)$ is the capped trust of a for p:

$$\text{trust}^*(a, p) = 0.25 + 0.75\,\text{trust}(a, p)$$

χ is the soft gate function:

$$\chi(x) = \begin{cases} 1.0, & \text{if } x \leq d_{0,A} \\ \exp\left(-\left(\frac{x - d_{0,A}}{\tau_{me}}\right)^2\right) & \text{else} \end{cases}$$

$d_{0,A}$ is a comfort-zone parameter; articles at a further positional distance that d_0 have increasingly reduced influence. τ_{me} is the width parameter determining how quickly that influence decays. $\overline{\phi^t(r)}$ is the average position of all consumed articles on r.

Adjust trust

Agent a's trust in ME p is adjusted based on the read article, as:

$$g_m^+ = \exp\left(-\frac{((\phi_a(r) - \phi_m(r))^2}{2\sigma_{tr}^2}\right)$$

$$g_m^- = 1 - g_m^+$$

$$\delta_m = g_m^+ - (\gamma_{res} g_m^-)$$

$$\delta = \sum_i \eta_{tr} \delta_i$$

$$q = \frac{|\mathcal{M}^t(p)|\eta_{tr}}{|\mathcal{M}^t(p)|\eta_{tr} + K}$$

$$\delta' = \frac{\delta}{|\mathcal{M}^t(p)|\eta_{al}}$$

$$\text{trust}^{\text{t}}(a, p) = \text{trust}^{t-1}(a, p) + q * \delta'$$

Here, $\phi_m(r)$ is the position of the article (or message) on r. In my simulation, $\phi_m = \phi_p \ \forall r \in \mathcal{R}$ if m was produced by p. σ_{tr} is the anti-linear modifer for trust adjustment. γ_{res} is the resilience penalty, making it easier to lose trust than to gain it. η_{tr} is the learning rate for trust changes, and $\mathcal{M}^t(p)$ is all messages (or articles in this case) produced by p currently in the corpus. K is a smoothing factor for learning updates.

Adjust alignment

Agent a's perception of p's alignment is updated based on the read article, as:

$$\delta_m = \exp\left(-\frac{((\phi_a(r) - \phi_m(r))^2}{2\sigma_{al}^2}\right)$$

$$\delta = \sum_i \eta_{al} \delta_i$$

$$q = \frac{|\mathcal{M}^t(p)|\eta_{al}}{|\mathcal{M}^t(p)|\eta_{al} + K}$$

$$\delta' = \frac{\delta}{|\mathcal{M}^t(p)|\eta_{al}}$$

$$\text{align}^{\text{t}}(a, p) = (1 - q)\,\text{align}^{t-1}(a, p) + q * \delta'$$

Notes

At present my trust and alignment update models are very similar; both are driven by the expressed position in the message. This can produce 'lock-in', where Agents never significantly alter their preferences as long as an outlet does not alter its position. The countering force in the simulation is (or would be) exogenous events targeting trust, such as scandals, etc. Future versions will implement a trust model based on social ties to other users, Evidence production on highly salient Topics/Issues, and other non-content factors.

Additional simulation fidelity can be gained by adding noise to article/message position on generation, simulating natural language ambiguity and authorial style effects. I omitted this work in my dissertation as desirable but ultimatley out of scope.

B.5.2 Ingest social media ("Scroll")

Next, the acting Agent considers the available message corpus $\mathcal{M}^t(\neg a)$ (all messages visible at t not written by a). In random order, the Agent considers each message and assigns it a probabilistic weight determining how likely it is that the Agent reads that message. Weights are computed as integer scores:

$$w_m = \beta_{pref}(s_{mem} + s_{ldr} + s_{pop} + s_{eng} + s_{evt})$$

β_{pref} is the scaling platform preference factor, computed as:

$$\beta_{pref} = f_{min} + \left(\frac{\psi_a(p) - \min_j(\psi_a(j))}{\max_j(\psi_a(j)) - \min_j(\psi_a(j))} (f_{max} - f_{min}) \right)$$

for bounding parameters f_{min}, f_{max}. s_{mem} is the shared membership score between a and author a':

$$s_{mem} = 2|\mathcal{G}_{\dashv} \cap \mathcal{G}_{\dashv'}|$$

s_{ldr} is points awarded if a considers author a' to be a leader, based on group membership:

$$s_{ldr} = 5 \cdot 1[\mathcal{G}_a \cap \mathcal{G}^*_{a'} \neq \emptyset]$$

s_{pop} is points awarded based on the author's popularity on the message's platform $m(p)$, measured by author follower count:

$$s_{pop} = \begin{cases} -2 & \text{if } f_{a'}(m(p)) < 5000 \\ 0 & \text{if } 5000 \leq f_{a'}(m(p)) < 15000 \\ 1 & \text{if } 15000 \leq f_{a'}(m(p)) < 25000 \\ 2 & \text{if } 25000 \leq f_{a'}(m(p)) < 40000 \\ 3 & \text{if } 40000 \leq f_{a'}(m(p)) < 65000 \\ 4 & \text{if } 65000 \leq f_{a'}(m(p)) < 150000 \\ 5 & \text{if } 150000 \leq f_{a'}(m(p)) < 500000 \\ 6 & \text{if } 500000 \leq f_{a'}(m(p)) < 1000000 \\ 7 & \text{if } 1000000 \leq f_{a'}(m(p)) \end{cases}$$

s_{eng} is points awarded for the message's engagement, making popular messages more likely to be read.

$$s_{eng} = 0.25 * m_{eng}$$

Finally, s_{evt} is points awarded based on exciting Events. m receives $P_{max} * \xi_e$ points for every active event e tied to m_n, the Narrative of m. All messages in $\mathcal{M}^t$ are scored. The computed w_m are then used to stochastically decide whether the message is visible to the Agent (whether they Agent "saw" that message in time step t):

$$\mathcal{M}_a^t = \{m \in \mathcal{M}^t : U < x\}, U \sim \mathcal{U}_{[0,1]}$$

$$x = \max\left(1 - e^{-\lambda_M w_m}, Q\right)$$

where λ_M is an attentuation parameter. Agents are subject to attention penalties. For each message read, the attention value Q decreases:

$$Q^m = Q^{m-1} - \frac{1}{2a_{att}}$$

Thus each message receives the higher of two scores: Whether the messages is likely to be read based on its merits (w_m), or based on its early presentation to the reader as a function of the reader's attention span (a_{att}).

B.5.3 Update internal state ("Think")

Next the acting Agent updates its internal state, altering Topic positions and Platform perceptions based on the messages chosen in the previous step. The mechanics of update are similar to those described in the "Listen" phase of media consumption.

Update valences

The influence impact of a message is a function of five variables.

1. **Author impact.** The influence of author b on Agent a is stored in the pre-computed interpersonal influence matrix $\mathbf{W}$, as W_{ab}.
2. **Emotional impact.** Emotional impact score w_{EM} is the cosine similarity between the Agent's internal Plutchik vector a_{EM} and the message's Plutchik vector m_{EM}, as:

$$w_{EM} = \frac{a_{EM} \cdot m_{EM}}{\|a_{EM}\|\|m_{EM}\|}$$

3. **Persuasive impact.** Logical impact score w_{MFT} is the cosine similarity between the Agent's internal MFT vector a_{MFT} and the message's MFT vector m_{MFT}.
4. **Platform trust.** Agents are more influenced by trustworthy platforms. Trust impact score is

$$w_{TR} = \text{trust}(a, p)$$

5. **Position difference.** A message's influence on an Agent diminishes as the distance between their positions grows; Agents are more able to ignore content they perceive to be extreme.

$$w_{POSN}(r) = \exp\left(-\left(\frac{|\phi_a(r) - \phi_m(r)| - d_{0,V}}{\tau_V}\right)^2\right)$$

Each message based on Topic r "nudges" a's valence value $\phi_a(r)$. The per-message nudge is:

$$\delta_m(r) = \eta_V \cdot \frac{1}{2} W_{ab} \cdot (\alpha_V w_{EM} + (1 - \alpha_V) w_{MFT}) \cdot w_{TR} \cdot w_{POSN}(r)$$

where α_V parametrizes preference between between emotional and rhetorical impact.

The Agent then computes the cumulative nudges on r over all messages,

$$\delta_{\mathcal{M}_a^t}(r) = \sum_{m \in \mathcal{M}_\dashv^\sqcup} \delta_m(r)$$

and the average nudge size

$$\overline{\delta(r)} = \frac{\delta_{\mathcal{M}_a^t}(r)}{|\mathcal{M}_\dashv^\sqcup|^{\gamma_V}}$$

where γ_V is a sublinear normalization parameter to ensure message quantity increases influence, but not in a linear fashion.

The Agent also computes the cumulative "destination" on r – the advocated valence value on Topic r summed over all applicable messages, with the message's nudge strength scaling its contribution to the final computed position. (The position is translated from [-1,1] to a latent variable space [0,1] via the atanh() function; I denote such latent scores as $\hat{\phi}$..)

$$\phi_{\mathcal{M}_a^t}(r) = \sum_{m \in \mathcal{M}_\dashv^\sqcup} \delta_m(r) \cdot (\hat{\phi}_m(r))$$

The average position of all imbibed messages is then:

$$\overline{\hat{\phi}_{\mathcal{M}_a^t}(r)} = \frac{\hat{\phi}_{\mathcal{M}_a^t}(r)}{\delta_{\mathcal{M}_a^t}(r)}$$

This "destination" position (in latent space) is compared to the Agent's current (latent) position, $\hat{\phi}_a^t(r)$. The Agent then computes a "step" from its current position toward the destination position, as:

$$\Delta_V(r) = 1 - e^{-|\overline{\delta(r)}|}$$

and scales this step by the distance between the current and destination positions:

$$\hat{\phi}_a^t(r) = \hat{\phi}_a^t(r) + \Delta_V(r)\left(\overline{\hat{\phi}_{\mathcal{M}_a^t}(r)} - \hat{\phi}_a^t(r)\right)$$

Finally the variable is translated out of latent space:

$$\phi_a^t(r) = \tanh(\hat{\phi}_{\mathrm{a}}^{\mathrm{t}}(\mathrm{r}))$$

Update perceived platform alignment

Messages influence an Agent's perceived platform alignment based on three factors.

1. **Message position.** Messages espousing positions far from the Agent's belief lose trust, as:

$$w_{POSN} = 1 - \frac{1}{2}|\phi_a(r) - \phi_m(r)|$$

2. **Author prominence.** Prominent authors have greater influence on the reader's perception of a Platform. Agents will judge the space by the loudest voices therein. For author b on Platform p, where $f_b(p)$ denotes b's followership on p,

$$w_{FOL} = 1 - \exp\left(\frac{-f_b(p)}{F_0}\right)$$

   where F_0 is a bending factor that approximates the 50th-percentile follower value.

3. **Message engagement.** Highly engaged messages have greater influence, as readers perceive them to be more representative of the larger Platform:

$$w_{EN} = 1 - \exp\left(\frac{-m_{eng}}{E_0}\right)$$

   where E_0 is again a bending factor.

As with valence, the Agent computes a per-message nudge on perceived alignment. For brevity I denote $A = \text{align(a, p)}$:

$$\delta_m(A) = \eta_p \cdot w_{FOL} \cdot w_{EN}$$

The cumulative nudges on A:

$$\delta_{\mathcal{M}_a^t}(A) = \sum_{m \in \mathcal{M}_\dashv^\sqcup} \delta_m(A)$$

The cumulative destination value, in which nudges are scaled by agreement:

$$A_{\mathcal{M}_a^t} = \sum_{m \in \mathcal{M}_\dashv^\sqcup} \delta_m(A) \cdot w_{POSN}$$

Finally, the Agent adjusts their perceived alignment for p toward the mean destination, with a scaled step size:

$$\overline{A} = \frac{A_{\mathcal{M}_a^t}}{\delta_{\mathcal{M}_a^t}(A)}$$

$$\Delta_A = \frac{\delta_{\mathcal{M}_a^t}(A)}{\delta_{\mathcal{M}_a^t}(A) + K}$$

$$A^t = A^{t-1} + \Delta_A * \overline{A}$$

where K is the learning rate factor.

Update trust in platform

Messages influence an Agent's trust in the delivering Platform based on two factors.

1. **Agreement.** Agents lose trust in Platforms that present content that does not align with the Agent's worldview. Such losses are small but accumulate, as:

$$w_{POSN} = 1 - \frac{1}{2}|\phi_a(r) - \phi_m(r)|$$

   I denote the disagreement factor as

$$\tilde{w}_{POSN} = 1 - w_{POSN}$$

2. **Social proof.** The extent to which a Platform is spreading information the Agent believes to be true or false scales the rate at which the Agent gains or loses trust, as:

$$w_{SP} = w_{FOL} \cdot w_{EN}$$

   using the same scores computed in adjusting alignment above.

I denote $T = \text{trust}(a, p)$. The strength of m's push on T is computed:

$$\eta'_T(m) = \eta_T w_{SP}(m)$$

where η_T is a tuning parameter. The direction and length of that push are computed as:

$$\delta_m(T) = w_{POSN} - \gamma_T \tilde{w}_{POSN}$$

where γ_T is the resilience penalty, making it easier to lose trust than to gain it. The Agent computes the cumulative push strength

$$\eta'_T(\mathcal{M}^{\sqcup}_{\dashv}) = \sum_{m \in \mathcal{M}^t_a} \eta'_T(m)$$

and the cumulative destination

$$T_{\mathcal{M}^{\sqcup}_{\dashv}} = \sum_{m \in \mathcal{M}^t_a} \eta'_T(m)\delta_m(T)$$

The Agent then adjusts their perceived trust of p as:

$$\overline{T} = \frac{T_{\mathcal{M}^{\sqcup}_{\dashv}}}{\eta'_T(\mathcal{M}^{\sqcup}_{\dashv})}$$

$$\Delta_T = \frac{\eta'_T(\mathcal{M}^{\sqcup}_{\dashv})}{\eta'_T(\mathcal{M}^{\sqcup}_{\dashv}) + K}$$

$$T^t = T^{t-1} + \Delta_T * \overline{T}$$

Recompute preferences

The acting Agent next recomputes its internal preferences scores for all Platforms, using the updated trust and alignment values. The preference function is the same described in Appendix A. Note that Agents maintain preference values for *all* platforms, including those to which they are not actively subscribed; only values for used platforms change, since messages and articles from unused platforms are not visible to that Agent. After the Agent updates its preference scores, it compares the worst score in its used platforms $p' = \min_p(\psi_a(\mathcal{P}_a))$ against the best score in its unused platforms $q = \max_p(\psi_a(\mathcal{P}_{\neg a}))$, with a hysterisis threshold h applied to prevent rapid switching:

$$q > hp'$$

If the inequality holds, a adds a subscription to q for a trial period T_p, during which time a slightly favors use of q due to the "novelty factor." After T_p has elapsed, a drops the lowest-ranked platform by preference, whether that is q or some incumbent platform that q has surpassed. During the trial period, a will not add additional platform memberships (the Agent is considered "oversubscribed".)

Crucially, I do *not* model a "loss cost" for platform switching or abandonment. In reality, Agents' willingness to try new platforms would be hindered by their attention budget; a rational Agent would not immediately abandon a platform into which they had sunk considerable energy and on which they had a following. Thus the "handoff" function would be better modeled as a tension between multiple factors. Dissatisfaction and appealing potential social ties would weigh in favor of joining the new platform, while existing social credit and familiarity would weigh against leaving the existing platform. As such the trial period would likely be much longer and multi-stage, with use of the legacy platform tapering off as the Agent came to rely more on the newer platform. This all bears further examination in future work.

Limitations

The adjustment models reviewed in this section, as in the media section ("Listen"), are adapted from Friekin-Johnsen models of influence. They are parametrized on common-sense and anecdotally observed factors, and tuned to create realistic behavior within the system. Each of these models merits further examination and refinement to better replicate the mechanics of internal state change within social media users. I considered this work adequate for an introductory examination of my research questions.

B.5.4 Act on social media ("Engage")

Acting Agents next engage with the social media ecosystem, as follows.

Determine engagement count

Agents will engage multiple times per turn. The number of engagements in an active time slot t is determined stochastically by the Agent's energy level and social media user type:

$$n \sim \text{Poisson}(\lambda_n)$$

$$\lambda_n \sim \Gamma(k_n, \mu/k_n)$$
$$\mu = m(1 + \theta_{SMT} a_{egy})$$
$$m = \alpha_n + \beta_n * a_{egy}$$

The agent's number of engagements is sampled from a Poisson process, which is parametrized by a Gamma distribution. The Gamma distribution is in turn parametrized by:

- α_n, β_n, corner values roughly specifying the engagement rate fo casual users and addict-level users respectively.
- k_n, a dispersion parameter; smaller values of k create heavier tails and thus "burstier" behavior in the population.
- θ_{SMT}, weights determined by a's social media type, as:

$$\begin{bmatrix} \text{Avg} \\ \text{Inf} \\ \text{Sup} \\ \text{Pub} \\ \text{Amp} \end{bmatrix} = \begin{bmatrix} 0.0 \\ 0.6 \\ 1.8 \\ 1.8 \\ 1.2 \end{bmatrix}$$

Select Platform

For each engagement, the Agent first chooses which Platform they will use. The Agent selects $p \in \mathcal{P}_{\dashv}$, using Platform preferences $\psi_a(\mathcal{P})$ as weights.

Select Narrative

The Agent next selects a Narrative to engage. Narrative choice is bounded: from the set of narratives $\mathcal{N}$, a can only choose narratives which were present in their recent message traffic $\mathcal{M}_a^t$, or which are tied to active exogenous events $\mathcal{E}^t$. I denote this set of available narratives $\mathcal{N}_a$. Narrative selection is driven by four factors.

1. **Topic salience.** Agents favor Narratives tied to Topics with higher salience.

   $$w_{SAL} = a_S(\nu(r))^{\alpha_S}$$

   where α_S is a non-linear scaling factor and $\nu(r)$ is the Topic r associated with narrative ν.

2. **Recency.** Agents favor Narratives with higher recurrence in recent reading, the short-term memory effect.

   $$w_{REC} = 1 + \lambda_{REC} \frac{\mathcal{M}_a^t(n)}{\mathcal{M}_a^t(n) + R_0}$$

   where $\mathcal{M}_a^t(n)$ denotes the number of occurrences of n in $\mathcal{M}_a^t$ and λ_{REC}, R_0 are tuning parameters.

3. **Trends.** Agents favor Narratives that the selected Platform would recommend, based on volume.

   $$w_{TRE} = 1 + \lambda_{TRE} \frac{\text{trend}(p, n)}{\text{trend}(p, n) + T_0}$$

$\text{trend}(p, n)$ is computed per platform/narrative pairing at the end of every time step, and is covered below. T_0, λ_{TRE} are tuning parameters.

4. **Events.** Agents are inclined to talk about goings-on in the world.

$$w_{EVT} = \lambda_{EVT} \frac{q_{\mathcal{E}^t}(\nu)}{q_{\mathcal{E}^t}(\nu) + E_0}$$

The stochastic weight for ν is then:

$$w(\nu) = w_{SAL}(w_{REC} w_{TRE} + w_{EVT})$$

All narratives are thus weighted, and the agent stochastically selects a Narrative. There are two special edge cases:

- *Egotism.* For each Agent, the system conducts a Bernoulli test $X < a_{ego}/5, X \sim \mathcal{U}_{[0,1]}$. If the test succeeds (inequality is true), then the Agent uses only $a_S(\nu(r))$ as the weight for *all* $\nu \in \mathcal{N}$. This enables especially egotistic Agents to launch new conversations in spaces where a given Narrative may not be present (i.e. $\mathcal{M}_a^t(\nu) = \emptyset$).
- *Firsts.* In the event that $\mathcal{N}_\dashv = \emptyset$, the Agent similarly uses internal salience as weights and selects from the set of all Narratives $\mathcal{N}$.

Select mode

Having selected Platform p and Narrative ν, the Agent now selects how they will engage. There are four engagement modes:

- **Compose:** The Agent composes a wholly original post.
- **Reply:** The Agent directly replies to another author's post, adding their own content as either reinforcement or rebuttal.
- **Quote:** The Agent references another author's post, adding some minimal content of their own.
- **React:** The Agent reacts to another author's post. (In real life reaction varies by Platform; here, this mode includes X's Retweet, Facebook's Like, Reddit's UpVote, etc., any action that openly avows the reader's support for the post and amplifies the author accordingly.)

Mode selection is driven by Agent energy and ego, and by the size of $\mathcal{M}_a^t$. The mode j is assigned probablistic weights p_j as follows:

$$l_j = (\omega_{0,j} + \omega_{e,j} a_{egy} + \omega_{g,j} a_{ego})/K_M$$

$$l'_{compose} = l_{compose} - \frac{\ln(1 + |\mathcal{M}_a^t|)}{10}$$

$$p_j = \exp(l_j - \max_{\text{i}}(l_i))$$

The probabilities are then normalized across all modes and used as stochastic weights. Some notes:

- The probability for composition is reduced as the message corpus grows, increasing the likelihood that Agents will choose to engage with existing traffic; $l'_{compose}$ is used to compute $p_{compose}$.
- This distribution is parametrized by three sets of weights: a base value $\boldsymbol{\omega_0}$, energy weights $\boldsymbol{\omega_e}$, and ego weights $\boldsymbol{\omega_g}$:

$$\mathbf{j} = \begin{bmatrix} \text{React} \\ \text{Quote} \\ \text{Reply} \\ \text{Compose} \end{bmatrix}$$

$$\boldsymbol{\omega_0} = \begin{bmatrix} 1.9 \\ 0.4 \\ -0.8 \\ -1.3 \end{bmatrix}, \boldsymbol{\omega_e} = \begin{bmatrix} -0.8 \\ -0.6 \\ 0.4 \\ 0.9 \end{bmatrix}, \boldsymbol{\omega_g} = \begin{bmatrix} -0.3 \\ -0.2 \\ 0.2 \\ 1.0 \end{bmatrix}$$

In the special case that $\mathcal{N}_a = \emptyset$, $p_{compose} = 1.0$.

Target messages

A crucial logical branch occurs here. If the Agent has selected to Quote, Reply, or React, they must now choose a target message upon which to engage.

In choosing a message, the Agent first conducts a Bernoulli trial with $p = 0.4$. If the trial is successful, the Agent limits the target message pool as:

$$\mathcal{M}^\dagger = \{m \in \mathcal{M}_a^t : [m(o) \in (\mathcal{G}_o^* \cap \mathcal{G}_a)]\} \cup [m(o) \in p(\mathcal{A}^*)]\}$$

The first clause indicates messages with authors $m(o)$ that are leaders of groups in which a is a member. The second clause indicates messages with authors that are "big hitters" $\mathcal{A}^*$ on platform p. If the chosen engagement mode is React, the pool is further narrowed to messages based on Topics where the author and reader agree (by sign, not magnitude), as:

$$\mathcal{M}^\dagger_{react} = \{m \in \mathcal{M}^\dagger : \phi_m(r) \cdot \phi_a(r) > 0\}$$

If the Bernoulli trial is not successful, the Agent considers all messages on p at time t, $\mathcal{M}^\dagger = \mathcal{M}^t_{a,p}$.

Next, each message in the message pool $\mathcal{M}^\dagger$ is given a score for stochastic selection. Message score is based on four factors and is computed in manner similar to that seen in the "Scroll" stage above:

1. **Shared membership:** As computed previously,

$$s_{mem} = 2|\mathcal{G}_\dashv \cap \mathcal{G}_\wr|$$

2. **Author leadership:** Agents prioritize responding to messages over whom the Agent is a leader (directorship).

$$s_{ldr} = 5 \cdot |\mathcal{G}_a^* \cap \mathcal{G}_o|$$

3. **Authority:** The message receives additional consideration if the author is an authority on the topic. Using the authority function $\text{auth}(a, r)$ which is 1 if a is an authority on r and 0 otherwise,
$$s_{auth} = 3 * 1[\text{auth}(a, r) = 1]$$
4. **Popularity:** The author's followership s_{pop} as computed piece-wise previously in the "Scroll" section.
5. **Engagement:** Message engagements, m_{eng}.

Message m's probability of selection, p_m, is computed as follows:

$$s_m = s_{mem} + s_{auth} + s_{ldr} + s_{pop}$$

with a minimum value of 1. This base score is then scaled to simulate preferential attachment, as:

$$s'_m = s_m \cdot (m_{eng} + 1)^{\gamma_{PA}}$$

Finally, the message is penalized if $m(\nu)$ is not the selected Narrative. This penalty disincentivizes Agents from ignoring their predilections, but also allows author popularity and influence to override the Agents' previous calculations.

$$s''_m = s'_m * \frac{1[m(\nu) = \nu_a^t]}{4}$$

where ν_a^t is a's chosen Narrative for this engagement. The resulting set of scores $s''_{\mathcal{M}^\dagger}$ is used to stochastically select the target message $m^\dagger$.

If the Agent chose to React, the algorithm is complete. A message m_a^t object attributed to a, with follow-on attribution to o, on platform p, is produced and added to $\mathcal{M}$.

Add evidence

If the agent chose to Compose, Quote, or Reply, the Agent must next consider citing evidence to strengthen the expressed position. The probability of an Agent choosing to cite evidence in their message is:

$$p_{ev} = 1 - (c_{min} + (a_{ego} + 0.15 * 1[\text{auth}(a, r) = 1])(c_{max} - c_{min}))$$

where c_{min} denotes the minimum likelihood an Agent will *not* cite evidence, and c_{max} denotes the maximum likelihood. If a opts to cite evidence, the Agent randomly selects one piece of evidence corresponding to ν and attaches it to m_a^t.

If the Agent has chosen to quote or reply to a post, the algorithm is now complete. A message m_a^t object attributed to a, with follow-on attribution to o, on platform p is constructed, along with the chosen narrative $m_a^t(\nu)$ and any evidence $m_a^t(ev)$, is produced and added to $\mathcal{M}$.

Original messages

If the Agent has chosen to compose a post, the Agent will probabilistically decide whether to cite evidence as described above. A message object m_a^t with the same attributes as above will be created, except that the follow-on attribution will list a as the original author (as opposed to some source author o). That message is added to $\mathcal{M}$.

B.5.5 Compute trends

After all designated Agents in t have acted, the system updates the trend tracker per platform. The tracker counts the number of times a given Narrative ν occurs on Platform p within time window $t-T_{tr}$, which is then used to simulate the stochastic effects of Platform recommendation engines.

B.5.6 Reset queue value

As described above, all Agents *not* acting at t have their queue value reduced by 1. Agents that *did* act in t recompute a new queue value, drawing a value x from a geometric distribution parametrized by the actor's delay attribute, as:

$$x \sim \text{Geom}(a_{lag}^{-1})$$

B.6 Model parameters

A full list of control parameters is provided in Table B.2.

Table B.2: Simulation variables

Variable	Type	Range	Value
F_{fav}, recurrence of favored evidence in a Narrative evidence pool	Control	$\mathbb{N}$	6
N_{ev}, number of randomly created Evidence objects per Issue	Control	$\mathbb{N}$	12
p_{ev}, Bernoulli parameter used to determine if non-favored evidence is added to a Narrative	Control	[0,1]	0.6
N, number of big-hitters per platform	Control	$\mathbb{N}$	5
T_{tr}, time window for trend accumulation	Control	$\mathbb{N}$	90
ρ_E, Event effect scale factor	Control	$\mathbb{R}$	0.1
T_{span}, maximum age of messages for consideration	Control	$\mathbb{N}$	90
p_{media}, probability of a media entity acting in a time step	Control	[0,1]	0.75
λ_t, topic saturation parameter for event influence on media	Control	$\mathbb{Z}$	5
β, max percent increase of Event effects on Narrative selection weight,	Control	$\mathbb{R}$	2.0
λ_e, event saturation scale capping event impact on narrative selection,	Control	$\mathbb{R}$	1.0
β_{stub}, semi-linear parameter for stubbornness/polarity,	Control	$\mathbb{R}$	1
L_{stub}, floor value for valence stubbornness,	Control	[0,1]	0.08
η_{posn}, scaling factor for article influence,	Control	$\mathbb{R}$	0.1
$d_{0,A}$, comfort zone parameter for article valence influence,	Control	$\mathbb{R}$	1.0
τ_{me}, tail width parameter for article valence influence,	Control	$\mathbb{R}$	0.6
σ_{tr}, anti linear factor for article trust adjustment,	Control	$\mathbb{R}$	0.5
γ_{res}, resilience penalty for article trust loss,	Control	$\mathbb{R}$	1.02

Table B.2: Simulation variables

Variable	Type	Range	Value
K, learning smoothing factor for article trust updates,	Control	$\mathbb{R}$	1
f_{min}, f_{max}, bounding parameters for platform preference factors,	Control	$\mathbb{R}$	0.6, 1.6
P_{max}, maximum number of points a message can receive from an exciting event,	Control	$\mathbb{N}$	6
λ_M, attenuation parameter for message reading,	Control	$\mathbb{R}$	0.25
$d_{0,V}$, comfort zone parameter for message valence influence,	Control	$\mathbb{R}$	0.8
τ_V, tail width parameter for message valence influence,	Control	$\mathbb{R}$	0.25
α_V, emotion-logic valence influence balance parameter,	Control	$[0, 1]$	0.6
γ_V, sublinear normalization for message valence influence,	Control	$\mathbb{R}$	0.5
F_0, followership bend factor for saturation mapping,	Control	$\mathbb{Z}$	8500
E_0, engagement bend factor for saturation mapping	Control	$\mathbb{R}$	1.35
γ_T, resilience penalty for message trust loss	Control	$\mathbb{R}$	1.05
h, platform comparison hysterisis threshold	Control	$\mathbb{R}$	1.1
α_S, non-linear scaling factor for salience	Control	$\mathbb{R}$	1.0
λ_{REC}, tuning parameter for recency in narrative choice	Control	$\mathbb{R}$	0.5
R_0, tuning parameter for recency in narrative choice	Control	$\mathbb{Z}$	3
λ_{TRE}, tuning parameter for trend in narrative choice	Control	$\mathbb{R}$	1.0
T_0, tuning parameter for trend in narrative choice	Control	$\mathbb{Z}$	10
K_M, mode selection temperature parameter	Control	$\mathbb{R}$	1.0
γ_{PA}, preferential attachment non-linear scaling factor	Control	$\mathbb{R}$	1.2
c_{min}, c_{max}, boundary parameters for citation probability	Control	[0,1]	0.5, 0.95

Bibliography

[1] Information Environment Project. URL https://carnegieendowment.org/projects/information-environment-project. 1.5

[2] Transition 2001, December 2000. URL https://nsarchive.gwu.edu/document/22894-national-security-agency-transition-2001. 2.2

[3] Tallinn tense after deadly riots. *BBC News*, April 2007. URL http://news.bbc.co.uk/2/hi/europe/6602171.stm. 1.7.2

[4] Net Neutrality: Last Week Tonight with John Oliver (HBO), June 2014. URL https://www.youtube.com/watch?v=fpbOEoRrHyU. 2.4.2

[5] Social Psychology Studies Human Interactions, 2014. URL https://www.apa.org/education-career/guide/subfields/social. 1

[6] Market research and competitive analysis, September 2025. URL https://www.sba.gov/business-guide/plan-your-business/market-research-competitive-analysis. 1.3

[7] Countering hybrid threats, January 2026. URL https://www.nato.int/en/what-we-do/deterrence-and-defence/countering-hybrid-threats. 2.2

[8] Gohar Abrahamyan. 'Patriotic hackers' in Armenia and Azerbaijan escalate crisis with cyber attacks, September 2012. URL https://www.atlanticcouncil.org/blogs/natosource/patriotic-hackers-in-armenia-and-azerbaijan-escalate-crisis-with-cyber 2.3.6

[9] Charles Abramson, Imran S Currim, and Rakesh Sarin. An experimental investigation of the impact of information on competitive decision making. *Management Science*, 51(2): 195–207, 2005. 1.3

[10] Hugo González Aguilar. Declarations of international organizations on the right to access to the Internet. In *2023 18th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–5. IEEE, 2023. 1

[11] Alexandre Alaphilippe. Adding a 'D'to the ABC disinformation framework. *The Brookings Institute*, 2020. URL https://coilink.org/20.500.12592/. 1.7.4, 3

[12] David S Alberts and Daniel S Papp. The information age: An anthology on its impact and consequences. 1997. 1

[13] David S Alberts, John J Garstka, Richard E Hayes, and David A Signori. Understanding information age warfare. 2001. 1

[14] Collin Anderson. Dimming the Internet: Detecting throttling as a mechanism of censorship in Iran. Technical report, 2013. URL https://arxiv.org/abs/1306.4361. Publication Title: arXiv preprint. 3.1.2

[15] Laura Andrew. Andrew Tate live streams from a private jet to 90,000 followers on his new social media account after being banned on Instagram and TikTok. *National World*, August 2022. URL https://www.nationalworld.com/news/people/andrew-tate-instagram-tiktok-3823993. 3.1.7, 3.2.2

[16] Ro'Ifatun Anisa, Rafiantika Prihandini, Dinda Jannah, Indi Izzah Makhfudloh, Alvian Agatha, and Yunita Wulandari. Application of Graph Theory in Computer Network Optimization. June 2024. 1.9

[17] Isuru Ariyarathne, Gangani Ariyarathne, Alessandro Flammini, Filippo Menczer, and Alexander C Nwala. Behavior change as a signal for identifying social media manipulation. *arXiv preprint arXiv:2603.03128*, 2026. 2

[18] Miriam Arnold, Mascha Goldschmitt, and Thomas Rigotti. Dealing with information overload: a comprehensive review. *Frontiers in psychology*, 14:1122200, 2023. 1

[19] Esma Aïmeur, Sabrine Amri, and Gilles Brassard. Fake news, disinformation and misinformation in social media: a review. *Social network analysis and mining*, 13(1):30, 2023. ISSN 1869-5450 1869-5469. doi: 10.1007/s13278-023-01028-5. 3

[20] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pages 519–528, 2012. 1.9

[21] Pierre Baldi, Paolo Frasconi, and Padhraic Smyth. *Modeling the Internet and the Web*. Wiley, Chichester, 2003. 3

[22] Jack M Balkin. How rights change: freedom of speech in the digital era. *Sydney L. Rev.*, 26:5, 2004. 3.1.5

[23] N. Balu. Musk's poll results: Elon should step down as Twitter CEO, 2022. URL https://www.reuters.com/technology/musk-poll-shows-575-want-him-step-down-twitter-chief-2022-12-19/. 3.1.4

[24] Zach Beauchamp. Alex Jones, Pizzagate booster and America's most famous conspiracy theorist, explained. *Vox*, December 2016. URL https://www.vox.com/policy-and-politics/2016/10/28/13424848/alex-jones-infowars-prisonplanet. 3.1.8

[25] Bond Benton, Jin-A. Choi, Yi Luo, and Keith Green. Hate Speech Spikes on Twitter After Elon Musk Acquires the Platform. Technical report, Montclair State University, College of Communication and Media, 2022. URL https://digitalcommons.montclair.edu/scom-facpubs/33. 3.1.4, 3.2.3

[26] Johannes Berendt, Sebastian Uhrich, Abhishek Borah, and Gavin J. Kilduff. The ri-

valry reference effect: Referencing rival (vs. nonrival) competitors in public brand messages increases consumer engagement. *Journal of Marketing Research*, 61(6):1058–1078, 2024. doi: 10.1177/00222437241248414. URL https://doi.org/10.1177/00222437241248414. tex.eprint: https://doi.org/10.1177/00222437241248414. 3.1.1

[27] Winnona DeSombre Bernsen. Crash (exploit) and burn: Securing the offensive cyber supply chain to counter China in cyberspace. Technical report, Atlantic Council, 2025. URL http://www.jstor.org/stable/resrep71235.5. 2.2

[28] David M Beskow and Kathleen M Carley. Social cybersecurity: an emerging national security requirement. 2019. 1.7.7, 1.9.5

[29] Matt Bishop and Carrie Gates. Defining the insider threat. In *Proceedings of the 4th annual workshop on cyber security and information intelligence research: Developing strategies to meet the cyber security and information intelligence challenges ahead*, Csiirw '08, Oak Ridge, Tennessee, USA, 2008. Association for Computing Machinery. ISBN 978-1-60558-098-2. doi: 10.1145/1413140.1413158. URL https://doi.org/10.1145/1413140.1413158. Number of pages: 3 tex.address: New York, NY, USA tex.articleno: 15. 6.2.1

[30] Janice T Blane. *Social-cyber maneuvers for analyzing online influence operations*. phd, Carnegie Mellon University, USA, 2023. 2, 7.2

[31] Sam Blazek. SCOTCH: A framework for rapidly assessing influence operations. *Atlantic Council*, 2021. URL https://www.atlanticcouncil.org/blogs/geotech-cues/scotch-a-framework-for-rapidly-assessing-influence-operations/. 1.7.8

[32] Danielle Blunt and Ariel Wolf. Erased: The impact of FOSTA-SESTA and the removal of Backpage on sex workers. *Antitrafficking Review*, 14, 2020. URL https://antitraffickingreview.org/index.php/atrjournal/article/view/448. 3.2.1

[33] Gillian Bolsover. Social media, computational propaganda, and control in China and beyond. In *The World Information War*, pages 122–138. Routledge, 2021. 3.1.2, 3.2.5

[34] John B Bonds. *Bipartisan Strategy: Selling the Marshall Plan*. Praeger, 2002. 1

[35] Michael Bossetta. The weaponization of social media: Spear phishing and cyberattacks on democracy. *Journal of international affairs*, 71(1.5):97–106, 2018. 1.3

[36] Rosie Bradbury. I spent a week on Trump's new social media app Truth Social. I felt like I was exploring a ghost town. *Business Insider*, April 2022. URL https://www.businessinsider.com/trump-truth-social-media-app-review-ghost-town-overrun-bots-2022-3. 3.1.5

[37] R. Brandom. Elon Musk begins reinstating banned Twitter accounts, starting with Jordan Peterson and the Babylon Bee, 2022. URL https://www.theverge.com/2022/11/18/23466625/

elon-musk-twitter-reinstatement-jordan-peterson-kathy-griffin-babylon-
3.1.4

[38] Russell Brandom. Surveillance drives South Koreans to encrypted messaging apps. *The Verge*, October 2014. URL https://www.theverge.com/2014/10/6/6926205/surveillance-drives-south-koreans-to-encrypted-messaging-apps. 3.1.6

[39] Tania Branigan. Accounts invaded, computers infected – human rights activists tell of cyber attacks. *Guardian*, January 2010. URL https://www.theguardian.com/world/2010/jan/14/china-human-rights-activists-cyber-attack. 2.3.1

[40] Haley Britzky. The hints of Weinstein's behavior that went ignored. *Axios*, October 2017. URL https://www.axios.com/2017/12/15/the-hints-of-weinsteins-behavior-that-went-ignored-1513306130. 3.1.9

[41] John T Buchanan, Erez J Henig, and Mordecai I Henig. Objectivity and subjectivity in thedecision making process. *Annals of Operations Research*, 80(0):333–345, 1998. 1.3

[42] C. Orr Bueno. FRACTURES: THE IMPACT OF DISCORD, DISINFORMATION, AND DAMAGED DEMOCRACY. *The Journal of Intelligence, Conflict, and Warfare*, 5(3):183–186, 2023. doi: 10.21810/jicw.v5i3.5194. URL https://doi.org/10.21810/jicw.v5i3.5194. 3

[43] Sanya Burgess. Donald Trump's social media app Truth Social surges in popularity after FBI raid his home. *SkyNews*, August 2022. URL https://news.sky.com/story/donald-trumps-social-media-app-truth-social-surges-in-popularity-after
3.1.5

[44] Richard M Burton. Computational laboratories for organization science: Questions, validity and docking. *Computational & Mathematical Organization Theory*, 9(2):91–108, 2003. 1.8

[45] Sergio Caltagirone, Andrew Pendergast, and Christopher Betz. The diamond model of intrusion analysis. 2013. 1.7.5

[46] Dell Cameron. Parler CEO Says He'll Ban Users for Posting Bad Words, Dicks, Boobs, or Poop. *Gizmodo*, June 2020. URL https://gizmodo.com/parler-ceo-says-hell-ban-users-for-posting-bad-words-d-1844222360. 3.1.5

[47] Kathleen Carley. Knowledge acquisition as a social phenomenon. *Instructional Science*, 14(3):381–438, 1986. 1.9

[48] Kathleen Carley. Formalizing the social expert's knowledge. *Sociological Methods & Research*, 17(2):165–232, 1988. 1.6

[49] Kathleen M Carley. Social cybersecurity: an emerging science. *Computational and mathematical organization theory*, 26(4):365–381, 2020. 1.7.7, 3.2.4, 3.4

[50] Kathleen M Carley and Vanessa Hill. Structural change and learning within organizations. *Dynamics of organizations: Computational modeling and organizational theories*, 2001. 1.6

[51] David F. Carr. Bluesky Sees Greatest Sustained Growth So Far in the US and UK. Summary, Similarweb, November 2024. URL https://www.similarweb.com/blog/insights/social-media-news/bluesky-sustained-growth/. 2.3.2, 3.1.5

[52] John B Casterline. Diffusion processes and fertility transition: Selected perspectives. 2001. 1.6

[53] Giulia Cencetti, Federico Battiston, Bruno Lepri, and Márton Karsai. Temporal properties of higher-order interactions in social networks. *Scientific reports*, 11(1):7028, 2021. 1.3, 1.9

[54] Jiyoung Cha. Predictors of the credibility of social media as a news outlet: An examination of the influences of social media contacts, source perceptions, and media use. *International Journal on Media Management*, 26(1-2):68–93, 2024. doi: 10.1080/14241277.2025.2481826. URL https://doi.org/10.1080/14241277.2025.2481826. tex.eprint: https://doi.org/10.1080/14241277.2025.2481826. 6.4.2

[55] Derek Chadee. *Theories in social psychology*. John Wiley & Sons, 2022. 1

[56] Julia Chan. Top Apps Worldwide for January 2021 by Downloads. Summary, SensorTower, February 2021. URL https://sensortower.com/blog/top-apps-worldwide-january-2021-by-downloads. 3.1.6

[57] Eric S. Charleston, Daniel Girlando, and Hanna Rioseco. No need for access, theft or disclosure: encryption of data is notifiable under PHIPA and CYFSA, September 2025. URL https://www.blg.com/en/insights/2025/09/hospital-for-sick-children-v-ontario-information-and-privacy-commissio 2.3.7

[58] Kyle Chayka. Bluesky's Quest to Build Nontoxic Social Media. *The New Yorker*, (14 Apr 2025), April 2025. URL https://www.newyorker.com/magazine/2025/04/14/blueskys-quest-to-build-nontoxic-social-media. 3.1.5

[59] Carl Clausewitz. *On war*. Penguin UK, 2003. 1.3, 2

[60] James Clayton. Trump's Truth Social app branded a disaster. *BBC*, April 2022. URL https://www.bbc.com/news/technology-60922717. 3.1.5

[61] Gaya Cocca, Paolo Frasca, and Chiara Ravazzi. A coupled friedkin-johnsen model of popularity dynamics in social media. *IEEE Control Systems Letters*, 2025. 1.7.9

[62] Dave Collins. Bankruptcy trustee discloses plan to shut down Alex Jones' Infowars and liquidate assets. *AP*, June 2024. URL https://apnews.com/article/alex-jones-infowars-bankruptcy-sandy-hook-e8bf9a3d11b9506abb18c285d3e7 3.1.8

[63] Kate Conger and Jack Nicas. Twitter Bars Alex Jones and Infowars, Citing Harassing Messages. *New York Times*, September 2018. URL https://www.nytimes.com/

2018/09/06/technology/twitter-alex-jones-infowars.html. 3.1.8, 3.2.3

[64] Sidney A Connor. Military operations in the information age: putting the cognitive domain on top. 2017. 2.1

[65] Terry Crowdy. *Deceiving Hitler: Double-cross and deception in world war II*. Bloomsbury Publishing, 2011. 1

[66] Elizabeth Culliford and Katie Paul. Unhappy with Twitter, thousands of Saudis join pro-Trump social network Parler. *Reuters*, June 2019. URL https://www.reuters.com/article/us-twitter-saudi-politics-idUSKCN1TE32S/. 3.1.5

[67] J. E. Cutting. Perception and information. *Annual Review of Psychology*, 38(1):61–90, 1987. 1.3, 3

[68] Ikran Dahir. Andrew Tate's Hustlers University 2.0 Has Made At Least $11 Million In Just One Month. *Buzzfeed News*, October 2022. URL https://www.buzzfeednews.com/article/ikrd/andrew-tate-hustlers-university. 3.1.7

[69] Shanti Das. Inside the violent, misogynistic world of TikTok's new star, Andrew Tate. *Guardian*, August 2022. URL https://www.theguardian.com/technology/2022/aug/06/andrew-tate-violent-misogynistic-world-of-tiktok-new-star. 3.1.7, 3.2.5

[70] Marco De Falco. Stuxnet facts report: A technical and strategic analysis. *NATO Cooperative Cyber Defense Centre of Excellence*, 118, 2012. URL https://ccdcoe.org/uploads/2018/10/Falco2012_StuxnetFactsReport.pdf. 1.7.2

[71] Stephen Dipple, Michael Kowalchuck, Kathleen M Carley, and Neal Altman. Construct User Guide. 2021. 1.6

[72] Disarm Foundation. A brief history of DISARM. URL https://www.disarm.foundation/brief-history-of-disarm. 1.7.5, 3

[73] Gwyn D'Mello. A Pakistani Group Hacked Into 10 Indian University Websites As Revenge Against Indian Hackers. *India Times*, April 2017. URL https://www.indiatimes.com/technology/news/a-pakistani-group-hacked-defaced-10-indian-university-websites-as-reta
html. 2.3.6

[74] Marie-Claire Dorking. Alyssa Milano's #MeToo hashtag proves shocking number of women have been sexually harassed and assaulted. *Yahoo Lifestyle*, October 2017. URL https://www.yahoo.com/lifestyle/alyssa-milanos-metoo-hashtag-proves-shocking-amount-women-sexually-har
html. 3.1.9

[75] Kimberly Dozier. How and why NSA spies on US allies. *Associated Press*, October 2013. URL https://apnews.com/general-news-88ff930fa0494aa1908e4c2acf1c1bb8. 2.2

[76] Fabio Duarte. Amount of Data Created Daily (2026), February 2026. URL https:

//explodingtopics.com/blog/data-generated-per-day. 1

[77] Brian M Ducote. Challenging the Application of PMESII-PT in a Complex Environment, 2010. URL https://apps.dtic.mil/sti/citations/ADA523040. 1.7.1

[78] K. Duffy. Elon Musk asks Twitter users whether they'd like an "edit" button. Twitter's CEO says results of the poll will be "important.", 2022. URL https://www.businessinsider.com/elon-musk-polls-twitter-edit-button-ceo-agrawal-results-important-2022 3.1.4

[79] Anthony Dukes and Esther Gal-Or. Negotiations and exclusivity contracts for advertising. *Marketing Science*, 22(2):222–245, 2003. ISSN 07322399, 1526548X. URL http://www.jstor.org/stable/4129716. 3.1.1

[80] Matt Egan. Trump Media's stock has plunged by nearly half since the election. Now it's taking action. *CNN Business*, June 2025. URL https://www.cnn.com/2025/06/26/business/trump-media-stock-truth-social. 3.1.5

[81] Millie Elsen, Rik Pieters, and Michel Wedel. Effects of advertising exposure duration and frequency: a theory and initial test. *Journal of Marketing Analytics*, 13(2):386–404, 2025. 3.1.1

[82] Ksenia Ermoshina, Benjamin Loveluck, and Francesca Musiani. A market of black boxes: The political economy of Internet surveillance and censorship in Russia. *Journal of Information Technology & Politics*, 19(1):18–33, 2022. 3.1.2, 3.2.2

[83] Hamid Etemad. The artificial intelligence, digital economy, and global connectivity: Implications and lessons for international entrepreneurship: H. Etemad. *Journal of International Entrepreneurship*, 22(4):409–432, 2024. 1

[84] Ronan Farrow. Harvey Weinstein's Army of Spies. *The New Yorker*, November 2017. URL https://www.newyorker.com/news/news-desk/harvey-weinsteins-army-of-spies. 3.1.9, 3.2.5

[85] Erwin W. Fellows. 'Propaganda:' History of a Word. *American Speech*, 34(3):182–189, 1959. 3.1.2

[86] Joseph Firth, John Torous, José Francisco López-Gil, Jake Linardon, Alyssa Milton, Jeffrey Lambert, Lee Smith, Ivan Jarić, Hannah Fabian, Davy Vancampfort, Henry Onyeaka, Felipe B. Schuch, and Josh A. Firth. From "online brains" to "online lives": understanding the individualized impacts of Internet use across psychological, cognitive and social dimensions. *World psychiatry : official journal of the World Psychiatric Association (WPA)*, 23(2):176–190, June 2024. ISSN 1723-8617 2051-5545. doi: 10.1002/wps.21188. 1

[87] Brian P Fleming. Hybrid threat concept: contemporary war, military planning and the advent of unrestricted operational art. 2011. 2.2

[88] Luciano Floridi. *Information: A very short introduction*, volume 225. Oxford University Press, 2010. 1.5

[89] Camille François. Actors, behaviors, content: A disinformation ABC. *Algorithms*, 2020. URL https://doi.org/10.3390/a13040085. 1.7.4, 3

[90] Sheera Frenkel. The storming of Capitol Hill was organized on social media. *New York Times*, January 2021. URL https://www.nytimes.com/2021/01/06/us/politics/protesters-storm-capitol-hill-building.html. 3.1.5

[91] Noah E Friedkin. *A structural theory of social influence.* Cambridge University Press, 1998. 1.9

[92] Noah E. Friedkin and Eugene C. Johnsen. Social influence and opinions. *The Journal of Mathematical Sociology*, 15(3-4):193–206, 1990. doi: 10.1080/0022250X.1990.9990069. URL https://doi.org/10.1080/0022250X.1990.9990069. tex.eprint: https://doi.org/10.1080/0022250X.1990.9990069. 1.7.9

[93] Noah E Friedkin and Eugene C Johnsen. Influence networks and opinion change. *Advances in Group Processes*, 16(1):1–29, 1999. A.10

[94] gcohen. Throwback Attack: Christmas attack on SickKids hospital prompts rare apology from LockBit. *Control Engineering*, June 2023. URL https://www.controleng.com/throwback-attack-christmas-attack-on-sickkids-hospital-prompts-rare-ap 2.3.7

[95] Lin Ge, Hengrui Cai, Runzhe Wan, Yang Xu, and Rui Song. A review of causal decision making. *arXiv preprint arXiv:2502.16156*, 2025. 1.3

[96] Johannes Gerschewski and Alexander Dukalskis. How the internet can reinforce authoritarian regimes: The case of North Korea. *Georgetown Journal of International Affairs*, 19:12, 2018. 3.1.2

[97] Lea Gerster, Richard Kuchta, Dominik Hammer, and Christian Schwieter. Telegram as a buttress: How far-right extremists and conspiracy theorists are expanding their infrastructures via Telegram. Technical report, Institute for Strategic Dialogue, 2022. URL https://www.isdglobal.org/wp-content/uploads/2022/11/Telegram-as-a-Buttress_How-far-right-extremists-and-conspiracy-theorists-are-expanding-their-pdf. 3.1.6

[98] GETTR Pub. Affairs. GETTR CEO Jason Miller and Former Brexit Party Leader Nigel Farage to Visit Australia as Signups in Country Explode by 10,000 Percent. Press Release, September 2022. URL https://about.gettr.com/press/gettr-ceo-jason-miller-and-former-brexit-party-leader-nigel-farage-to- 3.1.7

[99] Joseph Edward Gleason Jr. Iranian foreign policy shedding the pariah image. 1993. 3.1.2

[100] Col Joshua Glonek. The coming military AI revolution. *Military Review*, 104(3):88–99, 2024. 1.3

[101] Dan Goodin. Parler's amateur coding could come back to haunt Capitol Hill rioters. *Ars Technica*, January 2021. URL https://arstechnica.com/information-technology/2021/01/parlers-amateur-coding-could-come-back-to-haunt-capitol-hill-rioters/.

3.1.5

[102] Jay Graber. How the Open Social Web Will Change Everything, with Bluesky's Jay Graber, October 2024. URL https://dot-social.simplecast.com/episodes/jay-graber/transcript. 3.1.5

[103] Timothy Graham and Mark Andrejevic. A computational analysis of potential algorithmic bias on platform X during the 2024 US election. Technical report, 2024. URL https://eprints.qut.edu.au/253211/1/A_computational_analysis_of_potential_algorithmic_bias_on_platform_X_during_the_2024_US_election-4.pdf. 3.1.4, 3.2.5

[104] Sara Guaglione. News publishers hesitate to commit to investing more into Threads next year despite growing engagement. *Digiday*, December 2023. URL https://web.archive.org/web/20231205135223/https://digiday.com/media/news-publishers-hesitate-to-commit-to-investing-more-into-threads-next 3.1.5

[105] Blessing Guembe, Ambrose Azeta, Sanjay Misra, Victor Chukwudi Osamor, Luis Fernandez-Sanz, and Vera Pospelova. The emerging threat of ai-driven cyber attacks: A review. *Applied Artificial Intelligence*, 36(1):2037254, 2022. 1

[106] Michael A. Hamilton. Words Matter: Demystifying 'Maneuver'. *Infantry*, 113(1):28–34, 2024. ISSN 0019-9532. URL chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.benning.army.mil/infantry/magazine/issues/2024/Spring/pdf/10_Hamilton_txt.pdf. 6.1

[107] Drew Harwell. Trump's Truth Social's disastrous launch raises doubts about its long-term viability. *The Washington Post*, February 2022. URL https://www.washingtonpost.com/technology/2022/02/22/trump-truth-social-disaster/. 3.1.5

[108] Jiahui He, Haris Bin Zia, Ignacio Castro, Aravindh Raman, Nishanth Sastry, and Gareth Tyson. Flocking to mastodon: Tracking the great twitter migration. In *Proceedings of the 2023 ACM on Internet Measurement Conference*, pages 111–123, October 2023. URL https://dl.acm.org/doi/abs/10.1145/3618257.3624819. 2.3.2

[109] Todd C Helmus, Elizabeth Bodine-Baron, Andrew Radin, Madeline Magnuson, Joshua Mendelsohn, William Marcellino, Andriy Bega, and Zev Winkelman. *Russian social media influence: Understanding russian propaganda in eastern europe*. Rand Corporation, 2018. 3.1.2

[110] A Simon Herbert. *Models of Man. Social and Rational*. 1957. 1.3

[111] Lin Herbert and Jaclyn Kerr. On cyber-enabled information warfare and information operations. In *The Oxford Handbook of Cyber Security*, pages 251–272. Oxford University Press, 2021. 2

[112] Alex Hern. Facebook, Apple, YouTube and Spotify ban Infowars' Alex Jones. *Guardian*, August 2018. URL https://www.theguardian.com/technology/2018/

aug/06/apple-removes-podcasts-infowars-alex-jones. 3.1.8, 3.2.3

[113] Matthew Hickman. *AI-Enabled Social Cyber Maneuver Detection and Creation*. Doctoral Dissertation, Carnegie Mellon, Pittsburgh, PA, 2025. 1.7.7, 1.9.5, 3.4.1, 4.5.3, 4.6, 7.2

[114] Douglas D Hodson and Raymond R Hill. The art and science of live, virtual, and constructive simulation for test and analysis. *The Journal of Defense Modeling and Simulation*, 11 (2):77–89, 2014. 6.4.1

[115] Frank G Hoffman. *Conflict in the 21st century: The rise of hybrid wars*. Potomac Institute for Policy Studies Arlington, VA, 2007. 2.2

[116] Amanda Holpuch. John Oliver's cheeky net neutrality plea crashes FCC website. *Guardian*, June 2014. URL https://www.theguardian.com/technology/2014/jun/03/john-oliver-fcc-website-net-neutrality. 2.3.5, 2.4.2

[117] Kalley Huang. Twitter's Rivals Try to Capitalize on Musk-Induced Chaos. *New York Times*, December 2022. URL https://www.nytimes.com/2022/12/07/technology/twitter-rivals-alternative-platforms.html. 3.1.5

[118] Cheyenne Hunt-Majer. Truth Can't Handle the Truth: Censorship on Truth Social. Technical report, Public Citizen, August 2022. URL https://www.citizen.org/article/truth-cant-handle-the-truth/. 3.1.5

[119] Lance Y Hunter. Social media, disinformation, and democracy: how different types of social media usage affect democracy cross-nationally. *Democratization*, 30(6):1040–1072, 2023. 1

[120] Zainatun N Hussin, Christie Pei-Yee Chin, Stephen L Sondoh, and Kim-Kwang Raymond Choo. How do governments leverage the use of social media? A systematic review. *IEEE Transactions on Engineering Management*, 71:7242–7256, 2023. 1

[121] Caitlin Huston. Trump Agrees to Use Truth Social as Primary Social Media Platform. *Hollywood Reporter*, May 2022. URL https://www.hollywoodreporter.com/business/digital/trump-agrees-to-use-truth-social-as-primary-social-media-platform-1235
3.1.5

[122] Marco Iansiti. The value of data and its impact on competition. *Harvard Business School NOM Unit Working Paper*, (22-002), 2021. 1.3

[123] Mike Isaac. Threads Becomes Most Rapidly Downloaded App, Raising Twitter's Ire. *New York Times*, July 2023. URL https://www.nytimes.com/2023/07/06/technology/threads-downloads-twitter.html. 3.1.5

[124] Meera Jacka. Andrew Tate's return stream attracts 430,000 concurrent viewers. *Dexerto*, June 2023. URL https://www.dexerto.com/entertainment/andrew-tates-return-stream-attracts-430000-concurrent-viewers-2177843/
3.1.7, 3.2.2

[125] Margarita Jaitner. Exercising Power in Social Media: A Study of Hard and Soft Power in the Context of Russian Elections 2011–2012. Technical report, 2012. 3.1.2, 3.2.2

[126] A Jayanthi and R R Jagadeeshwari. A Study On Brand Awareness And Its Impact On

Sales. *International Journal of Creative Research Thoughts*, 13(6), June 2025. ISSN 2320-2882. URL https://ijcrt.org/papers/IJCRT2506034.pdf. 3.1.1

[127] Alyssa Kann. State-controlled media experience sudden Twitter gains after unannounced platform policy change, 2023. URL https://dfrlab.org/2023/04/21/state-controlled-media-experience-sudden-twitter-gains-after-unannounc 3.1.4

[128] Matt Kapko. Progress Software shakes off MOVEit's financial consequences, maintains customers. *Cybersecurity Dive*, January 2024. URL https://www.cybersecuritydive.com/news/progress-software-moveit-meltdown/703659/. 2.4.1

[129] Makena Kelly. With Parler down, QAnon moves onto a 'free speech' TikTok clone. *The Verge*, January 2021. URL https://www.theverge.com/2021/1/28/22254411/clapper-parler-gab-qanon-tiktok-free-speech. 3.1.5

[130] Jaclyn A. Kerr. The Russian model of Internet control and its significance. Technical report, Lawrence Livermore National Laboratory, Livermore, CA (United States), 2018. 3.1.2

[131] KeyData Cyber. Lessons Learned From The SickKids Hospital Breach: What Went Wrong and How to Prevent It. URL https://keydatacyber.com/blog/lessons-learned-from-the-sickkids-hospital-breach-what-went-wrong-and- 2.3.7

[132] Elizabeth K Kiessling. Gray zone tactics and the principle of non-intervention: Can" one of the vaguest branches of international law" solve the gray zone problem? *Harv. Nat'l Sec. J.*, 12:116, 2021. 2

[133] Soo-Am Kim. A Study on the Access to Information of the North Korean People. Technical report, 2021. 3.1.2

[134] David Kirichenko. Russia Is Shutting Down Its Own Internet To Stop Ukrainian Drones. *Forbes*, October 2025. URL https://www.forbes.com/sites/davidkirichenko/2025/10/09/russia-is-shutting-down-its-own-internet-to-stop-ukrainian-drones/. 2.3.3

[135] Stephen Kosack and Archon Fung. Does transparency improve governance? *Annual review of political science*, 17(1):65–87, 2014. 1.3

[136] Beth Kowitt. The John Oliver Effect: Why the British comedian's impact is no joke. *Fortune*, September 2015. URL https://fortune.com/2015/09/29/john-oliver-impact/. 2.3.5

[137] Lucio La Cava, Luca Maria Aiello, and Andrea Tagarelli. Drivers of social influence in the Twitter migration to Mastodon. *Scientific Reports*, 13, December 2023. doi: https://doi.org/10.1038/s41598-023-48200-7. URL https://www.nature.com/articles/s41598-023-48200-7. 2.4.4

[138] Frank Landymore. Half a Million Users Flooded to Twitter Competitor After Elon Musk

Handed Creeps the Keys. *Futurism*, October 2024. URL https://futurism.com/half-million-users-join-twitter-competitor. 2.3.2, 3.1.5

[139] Rachel Lerman. The conservative alternative to Twitter wants to be a place for free speech for all. It turns out, rules still apply. *Washington Post*, July 2020. URL https://www.washingtonpost.com/technology/2020/07/15/parler-conservative-twitter-alternative/. 3.1.5

[140] Matthew Levitt. Hezbollah's regional activities in support of Iran's proxy networks. *Middle East Institute*, 26:18, 2021. 3.1.2

[141] Ari Levy. Trump fans are flocking to the social media app Parler — its CEO is begging liberals to join them. *CNBC*, June 2020. URL https://www.cnbc.com/2020/06/27/parler-ceo-wants-liberal-to-join-the-pro-trump-crowd-on-the-app.html. 3.1.5

[142] Helen Lewis. The Weird, Fragmented World of Social Media After Twitter. *The Atlantic*, (Jul 2023), July 2023. URL https://www.theatlantic.com/ideas/archive/2023/07/twitter-alternatives-bluesky-mastodon-threads/674859/. 3.1.5

[143] Qian Li, Qian Liu, Shaoqiang Liu, Xinyue Di, Siyu Chen, and Hongzhong Zhang. Influence of social bots in information warfare: A case study on@ UAWeapons Twitter account in the context of Russia–Ukraine conflict. *Communication and the Public*, 8(2): 54–80, 2023. 2.1

[144] Martin C Libicki and Shari Lawrence Pfleeger. Collecting the dots: problem formulation and solution elements. *Rand Science and Technology*, 103(OP-103-RC), January 2004. 1.6

[145] Zhiang Lin and Kathleen Carley. Proactive or reactive: an analysis of the effect of agent style on organizational decision-making performance. *International Journal of Intelligent Systems in Accounting, Finance and Management*, 2(4):271–287, 1993. 1.9

[146] Joseph Littell and Nicolas Starck. Russian influence operations during the invasion of Ukraine. In *International conference on cyber warfare and security*, pages 209–XIV. Academic Conferences International Limited, 2023. 3.1.2

[147] Xiao Liu and Xiaoyong Zheng. The persuasive power of social media influencers in brand credibility and purchase intention. *Humanities and Social Sciences Communications*, 11(1):15, January 2024. ISSN 2662-9992. doi: 10.1057/s41599-023-02512-1. URL https://doi.org/10.1057/s41599-023-02512-1. 6.4.2

[148] Mallory Locklear. Stitcher removes Alex Jones' podcast from its platform. *Engadget*, August 2018. URL https://www.engadget.com/2018-08-03-stitcher-removes-alex-jones-podcast.html. 3.1.8, 3.2.3

[149] R. Mackey and M. Lee. Left-Wing Voices Are Silenced on Twitter as Far-Right Trolls Advise Elon Musk, 2022. URL https://theintercept.com/2022/11/29/

elon-musk-twitter-andy-ngo-antifascist/. 3.1.4, 3.2.5

[150] Pesha Magid. Iraqi journalists watch Telegram. Telegram is watching them back. *Columbia Journalism Review*, November 2022. URL https://www.cjr.org/the_feature/iraqi-journalists-watch-telegram-telegram-is-watching-them-back.php. 3.1.6

[151] Durairaj Maheswaran and Joan Meyers-Levy. The influence of message framing and issue involvement. *Journal of Marketing research*, 27(3):361–367, 1990. 1.3

[152] Daniel J Mallinson and Peter K Hatemi. The effects of information and social conformity on opinion change. *PloS one*, 13(5):e0196600, 2018. 1.3, 3

[153] Johnnie Manzaria and Jonathon Bruck. Media's Use of Propaganda to Persuade People's Attitude, Beliefs and Behaviors. *Ethics of Development in a Global Environment*, 1999. URL https://web.stanford.edu/class/e297c/war_peace/media/hpropaganda.html. 1.3

[154] Maxwell E McCombs and Donald L Shaw. The agenda-setting function of mass media. *Public opinion quarterly*, 36(2):176–187, 1972. 4.5.5

[155] Ivan Mehta. Telegram founder Pavel Durov says app now has 1B users, calls WhatsApp a 'cheap, watered down imitation'. *TechCrunch*, March 2025. URL https://techcrunch.com/2025/03/19/telegram-founder-pavel-durov-says-app-now-has-1b-users-calls-whatsapp-
3.1.6, 3.2.2

[156] Hannah Metzler and David Garcia. Social drivers and algorithmic mechanisms on digital media. *Perspectives on Psychological Science*, 19(5):735–748, 2024. doi: 10.1177/17456916231185057. URL https://doi.org/10.1177/17456916231185057. tex.eprint: https://doi.org/10.1177/17456916231185057. 3

[157] Marcus Michaelsen. Transforming Threats to Power: The International Politics of Authoritarian Internet Control in Iran. *International Journal of Communication*, 12, 2018. 3.1.2

[158] N. Miguel. What Data Does Grok Train On? Insights & Analysis, 2025. URL https://www.byteplus.com/en/topic/407704?title=what-data-does-grok-train-on. 3.1.4

[159] Carl Miller. D-RAIL: DIRECTING RESPONSES AGAINST ILLICIT INFLUENCE OPERATIONS. *The Journal of Intelligence, Conflict, and Warfare*, 7(3), 2025. 1.7.6

[160] Eduardo Jacobo Miranda Ackerman. Extracting a causal network of news topics. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 33–42. Springer, 2012. 1.9

[161] Emmanuel Mogaji. Market segmentation, targeting, and positioning. In *Strategic marketing management: Principles and practice*, pages 103–133. Springer Nature Switzerland, Cham, 2025. ISBN 978-3-031-99367-1. doi: 10.1007/978-3-031-99367-1_5. URL https://doi.org/10.1007/978-3-031-99367-1_5. 3.1.1

[162] Roger Moorhouse. *The devils' alliance : Hitler's pact with Stalin, 1939-1941*. London : The Bodley Head, 2014. ISBN 978-0-09-957189-6. 1

[163] Jonathan Morgan, Jacob Shaha, Rebecca Marigliano, Matthew Hicks, and Kathleen M Carley. GhostField: Simulating an Integrated Information Training Environment at Scale. In *19th International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction and Behavior Representation in Modeling and Simulation*, Pittsburgh, PA, October 2026. Springer. URL https://sbp-brims.org/2025/papers/working-papers/2025_SBP-BRiMS_paper_8.pdf. 4

[164] M. Murphy. Kathy Griffin kicked off Twitter as "free-speech absolutist" Elon Musk cracks down on parody accounts targeting him, 2022. URL https://www.marketwatch.com/story/free-speech-absolist-elon-musk-cracks-down-on-parody-accounts-targetir 3.1.4

[165] Paul P. Murphy. InfoWars' main YouTube channel is two strikes away from being banned. *CNN*, February 2018. URL https://www.cnn.com/2018/02/23/us/infowars-youtube-videos-trnd/index.html. 3.1.8

[166] José María Díaz Nafría and Francisco Salto Alemany. Towards a transdisciplinary frame: Bridging domains, a multidimensional approach to information. *tripleC: Communication, Capitalism & Critique. Open Access Journal for a Global Sustainable Information Society*, 9(2):286–294, 2011. 1.4

[167] Meghna Manoj Nair, Atharva Deshmukh, and Amit Kumar Tyagi. Artificial intelligence for cyber security: Current trends and future challenges. *Automated secure computing for next-generation systems*, pages 83–114, 2024. 1

[168] Ellen Nakashima, Yeganeh Torbati, and Will Englund. Ransomware attack leads to shutdown of major U.S. pipeline system. *Washington Post*, May 2021. URL https://www.washingtonpost.com/business/2021/05/08/cyber-attack-colonial-pipeline/. 2.3.8

[169] Shanti Naseer. Cable News Fact Sheet, 2023. URL https://www.pewresearch.org/journalism/fact-sheet/cable-news/. 3.1.3

[170] Stefan Neumann, Yinhao Dong, and Pan Peng. Sublinear-time opinion estimation in the friedkin–johnsen model. In *Proceedings of the ACM web conference 2024*, pages 2563–2571, 2024. 1.7.9

[171] Jack Nicas. Alex Jones Said Bans Would Strengthen Him. He Was Wrong. *New York Times*, September 2018. URL https://www.nytimes.com/2018/09/04/technology/alex-jones-infowars-bans-traffic.html. 3.1.8

[172] Ben Nimmo. Anatomy of an info-war: How Russia's propaganda machine works, and how to counter it. *Central European Policy Institute*, 15:1–16, 2015. URL https://www.stopfake.org/en/anatomy-of-an-info-war-how-russia-s-propaganda-machine-works-and-how-t 1.7.7

[173] Erik C. Nisbet, Olga Kamenchuk, and Aysenur Dal. A psychological firewall? Risk perceptions and public support for online censorship in Russia. *Social Science Quarterly*, 98(3):958–975, 2017. 3.1.2

[174] Office of Training. Psychological Warfare: Military Aspects, April 1977. URL https://www.cia.gov/readingroom/docs/CIA-RDP57-00259A000100090010-9.pdf. 2.1

[175] Katherine Ognyanova. In Putin's Russia, information has you: Media control and Internet censorship in the Russian Federation. In *Management and participation in the public sphere*, pages 62–79. IGI Global Scientific Publishing, 2015. 3.1.2

[176] Abby Ohlheiser. The woman behind 'Me Too' knew the power of the phrase when she created it — 10 years ago. *Washington Post*, October 2017. URL https://www.washingtonpost.com/news/the-intersect/wp/2017/10/19/the-woman-behind-me-too-knew-the-power-of-the-phrase-when-she-created- 3.1.9

[177] Jens David Ohlin. A roadmap for fighting election interference. 2021. 2

[178] Barbara Ortutay. Facebook bans 'dangerous individuals' cited for hate speech. *AP*, May 2019. URL https://apnews.com/article/7825d0df3fda4799a78da92b9e969cdc. 3.1.8, 3.2.3

[179] Luke J Osterritter and Kathleen M Carley. Modeling interventions for insider threat. In *International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation*, pages 55–64. Springer, 2020. 6.2.1

[180] John Paczkowski and Ryan Mac. Amazon Will Suspend Hosting For Pro-Trump Social Network Parler. *Buzzfeed*, January 2021. URL https://www.buzzfeednews.com/article/johnpaczkowski/amazon-parler-aws. 3.1.5

[181] Carolina Vendil Pallin. Internet control through ownership: The case of Russia. *Post-Soviet Affairs*, 33(1):16–33, 2017. 3.1.2, 3.2.2

[182] Annie Palmer. Twitter CEO Jack Dorsey has an idealistic vision for the future of social media and is funding a small team to chase it. *CNBC*, December 2019. URL https://www.cnbc.com/2019/12/11/twitter-ceo-jack-dorsey-announces-bluesky-social-media-standards-push.html. 3.1.5

[183] Elliot Panek, Wyatt Harrison, and Jue Hou. Change by default: Exploring the effects of a sudden influx of newcomers on the discourse of r/TwoXChromosomes. *First Monday*, 24(10), October 2019. doi: http://dx.doi.org/10.5210/fm.v24i10.10143. URL https://firstmonday.org/ojs/index.php/fm/article/download/10143/8113. 3.2.1

[184] M. Paris, D. Hull, and M. Adler. Elon Musk's Takeover of Twitter: Everything You Need to Know, 2022. URL https://finance.yahoo.com/news/elon-musk-takeover-twitter-everything-193231867.html. 3.1.4

[185] Aaron Parsley. QAnon-Focused Accounts Banned on Twitter Now Flourish — and Get a Boost from Trump — on Truth Social. *People*, August 2022. URL https://people.com/politics/qanon-focused-accounts-banned-twitter-flourish-boost-from-trump-truth- 3.1.5

[186] Douglas Paton. Risk communication and natural hazard mitigation: How trust influences its effectiveness. *International Journal of Global Environmental Issues*, 8(1-2): 2–16, 2008. 1.3

[187] Kari Paul. From dance videos to global sensation: what you need to know about TikTok's rise. *Guardian*, October 2022. URL https://www.theguardian.com/technology/2022/oct/22/tiktok-history-rise-algorithm-misinformation. 2.4.4

[188] Ryan Paul. Researchers identify command servers behind Google attack. *Ars Technica*, January 2010. URL https://arstechnica.com/information-technology/2010/01/researchers-identify-command-servers-behind-google-attack/. 2.3.1

[189] Alfredo Ribeiro Pereira and César Augusto Silva da Silva. The legality of international espionage based on the nature of the target and the perpetrator. *Expeditions with MCUP*, 2025(1):1–17, 2025. 2

[190] Michael E Porter, Victor E Millar, and others. How information gives you competitive advantage. 1985. 1.3

[191] Olivia Powell. Discord.io exposes personal data of more than 760,000 users. *Cyber Security Hub*, August 2023. URL https://www.cshub.com/attacks/news/discordio-exposes-personal-data-of-more-than-760000-users. 2.3.4

[192] Keith A Preble and Charmaine N Willis. Trading with pariahs: North korean sanctions and the challenge of weaponized interdependence. *Global Studies Quarterly*, 4(2):ksae031, May 2024. ISSN 2634-3797. doi: 10.1093/isagsq/ksae031. URL https://doi.org/10.1093/isagsq/ksae031. tex.eprint: https://academic.oup.com/isagsq/article-pdf/4/2/ksae031/57502267/ksae031.pdf. 3.1.2

[193] Walter Quattrociocchi, Guido Caldarelli, and Antonio Scala. Opinion dynamics on interacting networks: media competition and social influence. *Scientific reports*, 4(1):4938, 2014. 7.1.2

[194] Vignesh Radhakrishnan. Indian hackers bring down Pak websites on Independence day. *Hindustan Times*, August 2015. URL https://www.hindustantimes.com/india/indian-hackers-bring-down-pak-websites-on-independence-day/story-8lJr9kpGMzu5irk62SImyH.html. 2.3.6

[195] Yuqing Ren, Kathleen M Carley, and Linda Argote. The contingent effects of transactive memory: When is it more beneficial to know what others know? *Management Science*,

52(5):671–682, 2006. 1.6

[196] Nicolas Rivero. Hacking collective DarkSide are state-sanctioned pirates. *Quartz*, July 2022. URL https://qz.com/2007399/the-darkside-hackers-are-state-sanctioned-pirates. 2.4.3

[197] Laurie Robinson. Data Everywhere: Making Sense of Our Digital Landscape, March 2023. URL https://ischool.umd.edu/news/data-everywhere-making-sense-of-our-digital-landscape/. 1

[198] Roig-Franzia. How Alex Jones, conspiracy theorist extraordinaire, got Donald Trump's ear. *Washington Post*, November 2016. URL https://www.washingtonpost.com/lifestyle/style/how-alex-jones-conspiracy-theorist-extraordinaire-got-donald-trumps-ea 2016/11/17/583dc190-ab3e-11e6-8b45-f8e493f06fcd_story.html. 3.1.8

[199] Janus Rose. Keep Bluesky Weird. *VICE*, May 2023. URL https://www.vice.com/en/article/keep-bluesky-weird/. 3.1.5

[200] O. R. Royle. Musk's new Twitter CEO Linda Yaccarino issues first rallying cry to employees, 2023. URL https://fortune.com/2023/06/13/elon-musk-twitter-ceo-linda-yaccarino-staff-memo-email/. 3.1.4

[201] John Russell. Oracle Study Reveals Decision Making Paradox: More Data, Greater Uncertainty. *HPCwire*, April 2023. URL https://www.hpcwire.com/bigdatawire/2023/04/21/oracle-study-reveals-decision-making-paradox-more-data-greater-uncerta 1

[202] Maria-Lucia Rusu and Ramona Herman. The implications of propaganda as a social influence strategy. *Scientific Bulletin*, 23(2):46, 2018. 1

[203] U. Samet. The positive influence of large language models on fact-checking practices: A case study of Grok. *World Journal of Advanced Engineering Technology and Sciences*, 15 (3):1727–1738, 2025. 3.1.4

[204] David E. Sanger and Nicole Perlroth. F.B.I. Identifies Group Behind Pipeline Hack. *New York Times*, May 2021. URL https://www.nytimes.com/2021/05/10/us/politics/pipeline-hack-darkside.html. 2.3.8

[205] Jeanine Santucci. Donald Trump announces new social media platform, Truth Social, after being banned from major apps. *USA TODAY*, October 2021. URL https://www.usatoday.com/story/news/nation/2021/10/20/donald-trump-announces-new-media-platform-truth-social-after-twitter-b 6113559001/. 3.1.5, 3.2.1

[206] Shayan Sardarizadeh. Parler 'free speech' app tops charts in wake of Trump defeat. *BBC*, November 2020. URL https://www.bbc.com/news/

technology-54873800. 3.1.5

[207] Isaac Saul. This Twitter Alternative Was Supposed To Be Nicer, But Bigots Love It Already. *Forward*, July 2019. URL https://forward.com/news/427705/parler-news-white-supremacist-islamophobia-laura-loomer/. 3.1.5, 3.2.1

[208] Zoë Schiffer and Casey Newton. Yes, Elon Musk Created a Special System for Showing You All His Tweets First, 2023. URL https://www.theverge.com/2023/2/14/23600358/elon-musk-tweets-algorithm-changes-twitter. 3.1.4, 3.2.3

[209] Olivier Schmitt. When are strategic narratives effective? The shaping of political discourse through the interaction between political myths and strategic narratives. *Contemporary security policy*, 39(4):487–511, 2018. 1.3

[210] Ben Schreckinger. Amid censorship fears, Trump campaign 'checking out' alternative social network. *Politico*, May 2019. URL https://www.politico.com/story/2019/05/28/trump-campaign-twitter-1345357. 3.1.5, 3.2.2

[211] Michael Schwirtz and Nicole Perlroth. DarkSide, Blamed for Gas Pipeline Attack, Says It Is Shutting Down. *New York Times*, May 2021. URL https://www.nytimes.com/2021/05/14/business/darkside-pipeline-hack.html. 2.3.8

[212] Adam Segal. Why China Hacks the World. *Christian Science Monitor*, February 2016. URL https://www.csmonitor.com/World/Asia-Pacific/2016/0131/Why-China-hacks-the-world. 1.7.2

[213] Jeff Seldin. IS Struggles to Regain Social Media Footing After Europe Crackdown. *VOA News*, December 2019. URL https://www.voanews.com/a/europe_struggles-regain-social-media-footing-after-europe-crackdown/6180532.html. 3.1.6

[214] Alfonso Semeraro, Salvatore Vilella, and Giancarlo Ruffo. PyPlutchik: Visualising and comparing emotion-annotated corpora. *Plos one*, 16(9):e0256503, 2021. 4.3

[215] Jody Serrano. Here's a Long List of What Went Wrong When Trump's Truth Social App Launched on the App Store. *Gizmodo*, February 2022. URL https://gizmodo.com/trump-truth-social-problems-launch-presidents-day-app-s-1848574162. 3.1.5

[216] Jonathan Shieber. Parler jumps to No. 1 on App Store after Facebook and Twitter ban Trump. *TechCrunch*, January 2021. URL https://techcrunch.com/2021/01/09/parler-jumps-to-no-1-on-app-store-after-facebook-and-twitter-bans/. 3.1.5

[217] Catherine Shu. Meet Telegram, A Secure Messaging App From The Founders Of VK, Russia's Largest Social Network. *TechCrunch*, October 2013. URL https://techcrunch.com/2013/10/27/

meet-telegram-a-secure-messaging-app-from-the-founders-of-vk-russias-l 3.1.6

[218] Antionette Siu. Trump's Truth Social App Plummets in Traffic, Sees 93% Drop in Signups Since Launch Week (Exclusive). *The Wrap*, March 2022. URL https://www.thewrap.com/trump-truth-social-app-93-drop-signups-traffic/. 3.1.5

[219] Katherine Taken Smith, Lawrence Murphy Smith, Marcus Burger, and Erik S. Boyle. Cyber terrorism cases and stock market valuation effects. *Information & Computer Security*, 31(4):385–403, 2023. 2.3.1

[220] John F Sowa. Semantic networks. *Encyclopedia of artificial intelligence*, 2:1493–1511, 1992. 4.5.4

[221] Todd Spangler. Trump's Truth Social Bans Users Who Post about Jan. 6 Hearings. *Variety*, June 2022. URL variety.com/2022/digital/news/trumps-truth-social-is-banning-users-who-post-about-jan-6-hearings-acc 3.1.5

[222] Christopher St. Aubin and Jacob Liedke. Social Media and News Fact Sheet, September 2025. URL https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/. 1

[223] Adam Staten. Truth Social Users "Losing Interest" in Trump's Social Media App. *Newsweek*, February 2022. URL https://www.newsweek.com/truth-social-users-losing-interest-trumps-social-media-app-1683375. 3.1.5

[224] Galen Stocking, Regina Widjaya, Luxuan Wang, and Naomi Forman-Katz. Bluesky has caught on with many news influencers, but X remains popular. Summary, Pew Research Center, May 2025. URL https://www.pewresearch.org/short-reads/2025/05/29/bluesky-has-caught-on-with-many-news-influencers-but-x-remains-popular 3.1.5

[225] Blake Strom. ATT&CK 101, September 2018. URL https://medium.com/mitre-attack/att-ck-101-17074d3bc62. 1.7.5, 3

[226] Sun Tzu. *The Art of War*. The Internet Classics Archive, tr lionel giles edition, 400 BCE. URL https://classics.mit.edu/Tzu/artwar.html. 1.3

[227] Morgan Sung. Bluesky sends some users personalized apologies after racism controversy. *TechCrunch*, July 2023. URL https://techcrunch.com/2023/07/27/bluesky-racism-slur-apology-feedback/. 3.1.5

[228] Sylwia Szybowska, Krzysztof Chochowski, and Anna Chochowska. Social media as a battlefield for information warfare. In *International conference on optimization and data science in industrial engineering*, pages 322–339. Springer, 2025. 2.1

[229] Chao Tang and et al. In-depth analysis of the Great Firewall of China. Technical report, 2016. URL http://www.cs.tufts.edu/comp/116/archive/fall2016/

ctang.pdf. 3.1.2

[230] The Media Insight Project. The news consumption habits of 16- to 40-year-olds. Technical report, American Press Institute, August 2022. URL https://americanpressinstitute.org/the-news-consumption-habits-of-16-to-40-year-olds/. 3.2.3, 5.1, A.9.3

[231] Michele Theil and Kieran Press-Reynolds. 'One-stop hate shop': Andrew Tate's rise to TikTok superstardom is fueling violent misogyny and men's rights extremism. *Business Insider*, August 2022. URL https://www.businessinsider.com/andrew-tate-tiktok-hustlers-university-misogyny-women-comments-mens-ri 3.1.7

[232] Nway Minn Thila. How Myanmar lobbyists use Telegram to spread propaganda, fake news. *Mizzima*, June 2022. URL https://www.mizzima.com/article/how-myanmar-lobbyists-use-telegram-spread-propaganda-fake-news#google_vignette. 3.1.6

[233] S. A. Thompson, T. M. Terol, K. Conger, and D. Freedman. How Elon Musk Is Remaking Grok in His Image, 2025. URL https://www.nytimes.com/2025/09/02/technology/elon-musk-grok-conservative-chatbot.html. 3.1.4, 3.2.5

[234] Craig Timberg and Isaac Stanley-Becker. QAnon learns to survive — and even thrive — after Silicon Valley's crackdown. *Washington Post*, October 2020. URL https://www.washingtonpost.com/technology/2020/10/28/qanon-crackdown-election/. 3.1.5

[235] V Ya Tsvetkov. Assessment of information advantage. *European Journal of Natural History*, (3):36–39, 2017. 1.4

[236] Megan Twohey, Jodi Kantor, Susan Dominus, Jim Rutenberg, and Steve Eder. Weinstein's Complicity Machine. *New York Times*, December 2017. URL https://www.nytimes.com/interactive/2017/12/05/us/harvey-weinstein-complicity.html. 3.1.9, 3.2.4, 3.2.5

[237] US Department of Defense. US Army Field Manual 3-0, Operations, 2008. 1.7.1

[238] US Department of Defense. US Army Field Manual 3-12, Cyberspace Operations and Electromagnetic Warfare, 2021. 1.7.3

[239] US Department of Defense. Information in Joint Operations (JP 3-04), 2022. 1.5, 1.7.2

[240] Apurva Venkat. LockBit apologizes for ransomware attack on hospital, offers decryptor. *CSO*, January 2023. URL https://www.csoonline.com/article/574271/lockbit-apologizes-for-ransomware-attack-on-hospital-offers-decryptor.html. 2.3.7

[241] Vasilis Ververis, Sophia Marguel, and Benjamin Fabian. Cross-Country comparison of Internet censorship: A literature review. *Policy & Internet*, 12(4):450–473, 2020. 3.1.2

[242] K. Vlamis. Elon Musk vows to change his AI chatbot after it apparently expressed similar left-wing political views as Chat-

GPT, 2023. URL https://www.businessinsider.com/elon-musk-vows-to-make-xai-chatbot-grok-politically-neutral-2023-12. 3.1.4

[243] Claire Wardle and AbdelHalim AbdAllah. The information environment and its influence on misinformation effects. *Managing infodemics in the 21st century: addressing new public health challenges in the information ecosystem*, pages 41–51, 2023. 1.5

[244] R Kent Weaver. Policy leadership and the blame trap: Seven strategies for avoiding policy stalemate. *Governance Studies, Brookings Institution*, 2013. 1.3

[245] Ashley V Whillans, Chelsea D Christie, Sarah Cheung, Alexander H Jordan, and Frances S Chen. From misperception to social connection: Correlates and consequences of overestimating others' social connectedness. *Personality and Social Psychology Bulletin*, 43(12):1696–1711, 2017. 1.3, 3

[246] R Wilkov. Analysis and design of reliable computer networks. *IEEE Transactions on Communications*, 20(3):660–678, 2003. 1.9

[247] Jason Wilson. Rightwingers flock to 'alt tech' networks as mainstream sites ban Trump. *Guardian*, January 2021. URL https://www.theguardian.com/us-news/2021/jan/13/social-media-trump-ban-alt-tech-far-right. 3.1.5

[248] Edward Wong. Hackers Said to Breach Gmail Accounts in China. *New York Times*, January 2010. URL https://www.nytimes.com/2010/01/19/technology/companies/19google.html. 2.3.1

[249] Angela Yang. Andrew Tate sues Meta and TikTok for 'deplatforming' him in 2022. *NBC News*, August 2025. URL https://www.nbcnews.com/tech/tech-news/andrew-tate-tristan-tate-lawsuits-meta-tiktok-rcna225654. 3.1.7

[250] Becky Yerak. Ex-Owner of Parler Social-Media App Files for Chapter 11 Bankruptcy. *Wall Street Journal*, April 2024. URL https://www.wsj.com/articles/ex-owner-of-parler-social-media-app-files-for-chapter-11-bankruptcy-68 3.1.5

[251] Shuhan Yuan and Xintao Wu. Deep learning for insider threat detection: Review, challenges and opportunities. *Computers & Security*, 104:102221, 2021. 6.2.1

[252] Tebany Yune. What is Parler and why won't conservatives shut up about it? *Mic*, June 2020. URL https://www.mic.com/impact/what-is-parler-why-wont-conservatives-shut-up-about-it-27627361. 3.1.5

[253] Michael Zakharin and Timothy C Bates. Moral foundations theory: Validation and replication of the MFQ-2. *Personality and Individual Differences*, 214:112339, 2023. 4.3

[254] Christina Zhao. Vimeo Removes Alex Jones's InfoWars Content: 'Discriminatory and Hateful'. *Newsweek*, August 2018. URL https://www.newsweek.com/vimeo-removes-alex-jones-infowars-profit-discriminatory-hateful-107003 3.1.8, 3.2.3

[255] Marrian Zhou. Alex Jones' Infowars removed from LinkedIn and MailChimp, still up on Instagram and Twitter. *CNet*, August 2018. URL https://www.cnet.com/tech/services-and-software/alex-jones-infowars-removed-linkedin-pinterest-still-up-on-instagram-t 3.1.8, 3.2.3