UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Scaling psychosocial insights into social systems via computational methodologies

by Rohit Ram

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Dr. Marian-Andrei Rizoiu

University of Technology Sydney
Faculty of Engineering and Information Technology

October 2025

Certificate of Original Authorship

I, Rohit Ram, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctorate of Philosophy, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research was supported by an Australian Government Research Training Program (RTP) Scholarship doi.org/10.82133/C42F-K220.

Signature:

Production Note:
Signature removed prior to publication.

Date: October 27, 2025

Scaling psychosocial insights into social systems via computational methodologies

by

Rohit Ram

A thesis submitted in fulfilment of the requirements for the
degree of Doctor of Philosophy

Abstract

This thesis examines the complexity of human interactions and societal dynamics, focusing on the phenomena associated with modern communication technologies, i.e. social media. It emphasizes the role of computational methodologies, particularly machine learning and deep learning, to analyze social data. Motivated by challenges like misinformation and radicalisation on social media, the thesis advocates for computational solutions and new perspectives on approaching psychosocial research.

The thesis aims to bridge the gap between sociological and computational approaches, asserting that computational methods can advance our understanding of psychological and sociological phenomena. The research covers measurement, modeling, profiling, tooling, and collaboration, exploring the synergy between computational methods and psychosocial practices. It includes measuring social influence, introducing the General Influence Model, profiling political ideologies, developing tools for practitioners, and sharing the author's experiences at the intersection of psychosocial and computational disciplines. The goal is to demonstrate how computational methods offer a promising avenue for addressing enduring societal challenges and improving our comprehension of human behavior.

Acknowledgements

I want to acknowledge the many people who made this thesis possible.

My mum -Althea- and dad -Sairam— have been my unwavering foundation. Though the doctoral path was foreign territory, their belief never wavered. They supported not just my research, but my dream of earning this doctorate. Their love sustained me when the journey seemed endless.

My brother -Rahual- was my first academic rival and set me on this path for a love of knowledge and learning. Our early intellectual explorations ignited the curiosity that eventually led to this doctorate.

My friend -Neeraja- was my study companion and steadfast friend. She made the solitary work of research feel less lonely. Her presence and friendship have been constants in an ever-changing landscape.

My colleagues -Quyu, Frankie, Pio, and Nik- guided my thinking, expanded my intellectual horizons, and made this journey enriching.

My supervisor -Andrei- has given me a gift I can never lose—a way of thinking. Over many years, he went beyond the call of duty, supporting me not just as a scholar but as a person. He fundamentally shaped how I approach problems, organize ideas, and engage with the world. I have immense gratitude for this gift.

Everyone else who has touched this journey is woven into the fabric of this work.

This thesis exists because of all of you.

Contents

List of Figures

List of Tables

List of Publications

Rohit Ram and Marian-Andrei Rizoiu. Empirically measuring online social influence. *EPJ Data Science*, 13(1):53, 2024

Rohit Ram, Emma Thomas, David Kernot, and Marian-Andrei Rizoiu. Practical guidelines for ideology detection pipelines and psychosocial applications. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 1630–1648, 2025

Rohit Ram, Quyu Kong, and Marian-Andrei Rizoiu. Birdspotter: A tool for analyzing and labeling twitter users. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 918–921, 2021

Co-Authorships

Quyu Kong, **Rohit Ram**, and Marian-Andrei Rizoiu. Evently: Modeling and analyzing reshare cascades with hawkes processes. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 1097–1100, 2021

Pio Calderon, **Rohit Ram**, and Marian-Andrei Rizoiu. Opinion market model: Stemming far-right opinion spread using positive interventions. In *Proceedings of the 18th ACM International Conference on Web Search and Data Mining*, 2024

Quyu Kong, Pio Calderon, **Rohit Ram**, Olga Boichak, and Marian-Andrei Rizoiu. Interval-censored transformer hawkes: Detecting information operations using the reaction of social systems. In *Proceedings of the ACM Web Conference 2023*, pages 1813–1821, 2023

Chapter 1

Introduction

The intricacies of social interaction and resulting societal dynamics pose challenges for empirical investigation due to their inherent complexity and vast scale. Human interactions significantly influence critical events such as unrest, war, trade, economics, and social progress, particularly within democratic contexts where decisions are linked to public sentiment.

Advancements in communication technology have historically reshaped societal dynamics. From the telegraph's ability to bridge distances to television's introduction of a visual medium for mass communication, these innovations have transformed human interaction patterns. In contemporary times, the rise of social media platforms further revolutionizes communication, marking a fundamental shift in how people engage and mold society.

Social Media Concerns & Opportunities. Social media, where users simultaneously function as content publishers and consumers, signifies a radical transformation in communication modalities and opinion dynamics. However, its relative novelty raises concerns about its enduring impact. This is particularly evident as phenomena such as misinformation spread at unprecedented speeds, filter bubbles are amplified by recommender systems, and pathways for radicalization are observed.

However, amidst these concerns, the ubiquity of the internet and the widespread adoption of social media present unprecedented opportunities for comprehending society. The digital landscape offers an expansive repository of rich social data, serving as a powerful tool

for deciphering the intricate web of human interactions. The increasing prevalence of social media participation across diverse segments of the population offers an unparalleled opportunity to scrutinize shared opinions, unravel the processes of opinion formation, and probe into the psychosocial characteristics of the broader populace.

Computational Advances. The increasing availability of social data coincides with a substantial expansion of our capabilities to extract insights and understanding from this wealth of information. The advent of deep neural network techniques has revolutionized our approach to data analysis, challenging the conventional notion that certain tasks were exclusively within the domain of human capability. Advances in Natural Language Processing (NLP), Graph Machine Learning (Graph ML), and event-based machine learning have proven particularly effective in providing nuanced insights into human behavior within the realm of social media.

The efficacy of these computational models is intricately linked to the parallel advances in infrastructure, spanning storage, compute, and networking. The ability to process vast volumes of social data in real-time has become possible through the development of sophisticated computational infrastructure. Similarly, the computational requirements for complex analyses have been facilitated by advancements in graphical computing, allowing for the rapid execution of computationally intensive algorithms.

Utilisation Within Psychosocial Literature. Traditionally, the domains of sociology and social psychology have primarily relied on qualitative investigations, focus groups, surveys, and controlled experimentation within laboratory settings. These conventional methods, while valuable, often present limitations in scale, efficiency, and real-world applicability. Notably, the subfields of social network analysis and artificial societies have garnered significant attention, offering insights into the structural aspects of social interactions. However, the deeper computational approaches within sociology and social psychology have remained underutilized.

Motivation

We are compelled to embrace computational approaches due to a collective acknowledgment that our understanding of opinion formation, identity shaping, and societal dynamics is limited. This necessity is particularly pronounced when dealing with large-scale phenomena and the interplay between online spaces and their offline manifestations. The prevailing issues of misinformation, disinformation, and radicalization, exacerbated by the pervasive influence of social media, demand our attention. To address these challenges effectively, there is a compelling need to enhance our comprehension of the underlying mechanisms driving such phenomena.

Furthermore, we bear a responsibility to improve our societal situation by managing and regulating online communities more effectively; building resilient online spaces that are less susceptible to devolving into violence and hatred and fostering environments of social trust. This proactive stance not only mitigates conflicts but also lays the foundation for a more harmonious societal coexistence.

I suggest that problems exacerbated by social media necessitate computational solutions, and computational approaches offer a promising avenue for understanding and addressing enduring issues that have persisted throughout human history.

Scope

While others have explored similar paths, what distinguishes this thesis is its unique synthesis of sociological depth and computational breadth. Sociological approaches often lack computational rigor, and computational methods tend to superficially apply sociological principles. My aim is to bridge this gap, approaching the study of societal dynamics from both sociological and computational ends.

1.1 Thesis Statement & Approach

In this thesis, I assert that applying computational methods will advance our understanding of psychological and sociological phenomena. The research unfolds across five key components within the research pipeline, namely measurement, modeling, profiling, tooling, and (non-traditional) research outputs.

Measurement of psychosocial phenomena can be substantially enhanced in terms of scale, representativity, and significance by employing data recruitment methods such as active learning and leveraging the diverse data sources prevalent on social media. Traditional methods, often confined to controlled laboratory experiments with limited sample sizes, can be expanded through modern computational techniques. This raises the question of how empirical measurement of psychosocial phenomena, like social influence, can be improved.

Modeling computationally offers a means to endogenize psychosocial mechanisms within formal frameworks. This approach enables a more nuanced understanding of the complexities inherent in psychological and sociological dynamics.

Profiling populations of online users becomes crucial for comprehending the composition of online communities, their traits, and the opinions stemming from their discussions. Computational approaches can provide insights into the underlying traits that shape these communities.

Tooling development is essential to empower psychosocial practitioners to readily profile and model relevant phenomena. State-of-the-art computational classification techniques can be integrated into tools, facilitating practical applications for researchers and practitioners alike.

Non-Traditional Research Outputs form the basis of academic ecosystems and are instrumental to not only the sharing of research methods, but also building trust in them. Discourse with stakeholders, presentations, and broad audience articles are crucial, particularly at the interface of disciplines.

Each chapter of this thesis directly aligns with these five components, offering a comprehensive exploration of the synergies between computational methods and psychosocial

practices. They each contain a comprehensive literature review for their respective topics, sufficient to situate the research within the broader research landscape.

1.2 Outline

In Chapter 2, we introduce a methodology for empirical measurement of social influence in social media contexts utilising crowdsource workers in a human-in-the-loop paradigm. The outcome is a framework for measuring psychosocial properties at scale in online contexts, a human-labelled ground truth of social influence, and an application of this empirical measurement to show its correlation with other psychosocial properties (i.e., agency and communion).

In Chapter 3, we introduce the data-driven General Influence Model (GIM), which incorporates two psychosocial inspired mechanisms; conductance and influence-capital distribution. We show that GIM is less arbitrary than computational influence models, and corrects biases introduced through common influence metrics. We apply GIM to a large-scale dataset of social media discussions surrounding COVID-19. Furthermore, by utilising data wrangling techniques, we test the association between inferred social influence and two psychosocial properties; social class (via occupation) and content veracity (via media sharing behaviour). We find that some of the most influential occupations, also spread the most misinformation.

In Chapter 4, we introduce a methodology for profiling the political ideologies of social media users. Ideology reflects views on how society should operate, and intuitively forms the basis for many other opinions. Prior approaches tend to utilise few signals of ideology, which limit the representativity and effectiveness of their approaches. Here we curate a set of ideological signals, characterising both moderate and extreme variations, which forms the basis of our classification technique. We utilised ideological signals that are (somewhat) agnostic to the dataset context and allow for generalisability of the approach. Furthermore, we extract psychological traits from the language of profiled ideological users, to further characterise the association between psychology and ideology. In particular we characterise

their language in terms of morality, grievance, nationalism and cognitive distortions; finding significant relationships with ideology.

In Chapter 5, we introduce a set of tooling aimed at psychosocial practitioners to perform profiling and influence modelling on social media datasets. The tooling comes in the form `birdspotter` – a software tool to analyze and label Twitter users – and `birdspotter.ml` – an exploratory visualisation of the generated metrics. `birdspotter` is an end-to-end analysis tool, which allows interdisciplinary practitioners to assess state-of-the-art computational techniques for social media analysis, with only a few lines of code. Furthermore, the package features tutorials, detailed documentation, and an in-depth case study showing how it can be readily utilised as a state-of-the-art bot detection system.

In Chapter 6, I illustrate a body of experience situating my scholarship in the intersection of psychosocial and computational disciplines. I assert here that for psychosocial practitioners to effectively utilise modern computational approaches, they must observe these approaches championed in their domain. I present a history of event participation, presentations, and collaborations, which showcase non-traditional research outputs, in efforts toward this thesis.

Finally, in Chapter 7 I reassert that applying computational methods will advance our understanding of psychological and sociological phenomena, and demonstrate how the preceding chapters support this assertion.

Chapter 2

Empirical Influence Measurement

Measurement is the foundation of scientific inquiry, providing the basis for accurate modeling and deeper theoretical understanding. In the context of social influence, effective measurement is particularly challenging due to the complexity of human interactions. Traditional experimental approaches in psychology, like controlled lab studies, have been instrumental in isolating social phenomena but often fail to scale beyond small groups, limiting their applicability to real-world contexts. Classic experiments, such as those by Asch on conformity [12], demonstrate how social influence operates but are difficult to replicate at scale due to the constraints of traditional psychometrics, which rely on carefully controlled environments that do not easily extend to broader populations.

This research seeks to overcome these limitations by introducing a human-in-the-loop active learning method to empirically measure social influence. By using crowdsourced pairwise comparisons and integrating them with computational simulations, this method leverages large-scale sociometric data to capture the nuances of social dynamics that traditional lab experiments miss. The research addresses the challenges of scaling in psychometrics, offering a practical way to gather and analyze data across diverse, real-world contexts, such as social media platforms.

A critical aspect of measurement in social sciences is managing systematic noise – consistent biases that might stem from the study design or environment – and differentiating it from random noise, which introduces variability. This research illustrates how to navigate

these complexities, accounting for systematic noise in experimental design and random noise using simulation tools to estimate research budgets based on design parameters and participant accuracy.

To ensure the validity of these measurements, this study employs multiple approaches: phenomenological validation by aligning findings with established psychological constructs and predictive validation through correlation with expected social outcomes. For example, the study deploys the active learning method to rank the influence of 500 X/Twitter users, revealing a strong association between empirically derived influence metrics and the psychological constructs of agency and communion. Agency, or the perception of assertiveness and goal-oriented behavior, emerged as a critical dimension, suggesting that these traits significantly drive social influence.

Precise and scalable measurement techniques are essential to advancing the study of social phenomena. The utilization of the active learning method exemplifies the thesis, demonstrating how computational tools can empirically capture complex social behaviors in a manner that aligns with established psychological frameworks.

By positioning this work within the broader scientific discourse on measurement and validation, the thesis underscores the importance of interdisciplinary collaboration. Combining the precision of computational methods with the depth of psychosocial insights offers a promising pathway to address enduring societal challenges, such as misinformation and polarization. This synthesis aims to provide a more comprehensive understanding of human behavior, bringing new perspectives to the study of social influence and the development of effective interventions.

Author Declaration

The following chapter contains content from the following publication.

Rohit Ram and Marian-Andrei Rizoiu. Empirically measuring online social influence. *EPJ Data Science*, 13(1):53, 2024

Author Contributions: R.R. led the research for this study, managed the data processing and collection, and conducted the experiments and analysis. M.A.R. provided supervision through all stages of the study. R.R. and M.A.R. collaboratively developed the model and experimental design. R.R. and M.A.R. interpreted the results and contributed to manuscript writing and editing.

Production Note:
Signature removed prior to publication.

Rohit Ram

Production Note:
Signature removed prior to publication.

Marian-Andrei Rizoiu

2.1 Introduction

Social influence is a pervasive force that shapes our interactions with others and contributes to the emergence of complex societal behaviors. Understanding social influence mechanisms is crucial for comprehending how individuals and groups decide and act in concert. Empirical measurement of online influence can help us identify the factors that shape online behavior, such as the influence of opinion leaders, the latent topology of social networks, and the role of algorithms in directing the flow of information. This, in turn, assists us in developing strategies for managing and regulating online behavior and ultimately contributes to creating healthier and more sustainable communities.

Social influence is defined as *a change in a person's cognition, attitude, or behavior, which has its origin in another person or group* [13]. Influential people are those capable of effecting this change and here we measure this influence in people. This element of behavioral dynamics makes influence surprisingly challenging to quantify. There are two main approaches to achieving this, each with merits and shortcomings; psychosocial and quantitative. Psychosocial experiments are conducted in controlled laboratory environments [14–16], capable of investigating the minute nuances of the influence phenomenon but are consequently limited in size and cannot assess emergent behavior. On the other hand, quantitative methods for measuring influence based on online social media data [17] are often ad hoc, somewhat arbitrary, and do not relate strongly to social phenomena [18]. The insights from one domain remain largely orthogonal to the other [19]. Furthermore, the capability to apply psychometrics at an online scale is lacking. The question is, therefore, **how can the influence phenomenon be measured at scale, while maintaining a high-quality experimental environment?**

This paper explores how to measure social influence empirically and at scale by ranking a set of online social media users (hereafter called *targets*). We base our quantification on the perception of social influence by peers – we ask people to compare two targets and choose the more influential one. Note that, the peers are Amazon Mechanical Turk (MTurk) workers and targets are online users, who represent disjoint sets. We assume that people can judge the relative social influence between individuals. While this may seem like a strong assumption, sociometric research often utilizes peer-perceived measures of social attributes

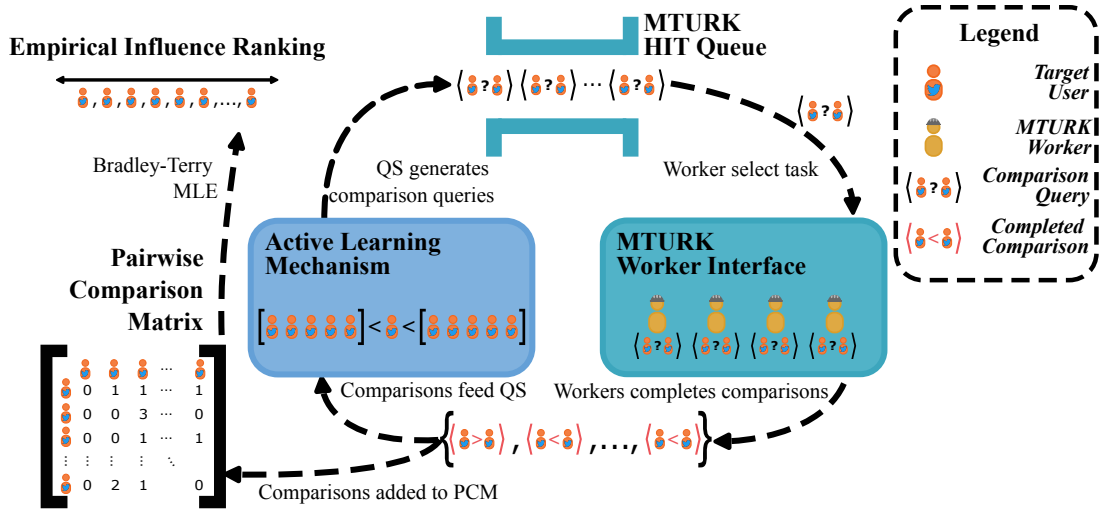


FIGURE 2.1: The schema illustrates the human-in-the-loop system for generating comparisons.

– such as popularity and reputation – and finds these measures are robust and reliable [20–22]; however, the methodology is rarely applied to the scale of our cohort. Influence measurement has two main challenges: scaling up the experiments to hundreds or thousands of targets and constructing the ranking from pairwise comparisons. The naive approach would require performing all possible comparisons – which scales quadratically with the number of targets. This quickly becomes prohibitive in a crowdsourcing environment where one pays a fixed price per comparison.

We solve both challenges by proposing a system with three components. First, we address the scaling challenge. As the *first component*, we use a human-in-the-loop active learning technique – human crowdsource workers are allocated tasks dynamically by an algorithm based on the previous decisions of other workers. Fig. 2.1 illustrates our human-in-the-loop system. The active learning mechanism generates comparison queries and adds them to a queue. Multiple crowdsource workers concurrently complete batches of queries – each query compares two target users and asks who is the more influential. The completed pairs are returned to the active learning mechanism, which uses them to generate more comparison queries, thus closing the loop. Section 2.3.1 further details our empirical influence measurement approach and shows it reduces the required pairwise comparisons and improves each decision's utility. For example, for 300 targets the naive approach would cost $\sim\$2,500$, but our approach would only cost $\sim\$67$. Our framework scales loglinearly

and, consequently, gains improve with scale. For 600 targets the naive approach would cost $\sim\$10,100$, but our approach would only cost $\sim\$145$.

Next, we address the ranking challenge. The *second component* is an augmented Bradley-Terry model [23] that we leverage to build an influence ranking from pairwise comparisons while accounting for systematic noise in worker decisions. We show that the systematic noise is theoretically linked to the expected worker decision accuracy, allowing us to recover, via simulation, the relationship between the required budget, the systematic noise, and the quality of the influence ranking. This allows answering questions such as "What is the expected ranking quality for a given number of targets and my maximum budget?" Conversely, it can answer the question, "What budget do I require to achieve a certain ranking quality for my given number of targets?" With systematic noise approximated through experiments with real workers, we infer the ranking quality achievable.

The *third component* is an Amazon Mechanical Turk (MTurk) survey instrument that we develop and optimize to estimate the impact of design features on worker decision accuracy and reduce the required comparison budget. We account for several design features – such as user metrics, proxies, and qualifications – and show through a series of ablative pilot studies that improving the MTurk worker interface improves the accuracy of worker decisions. We find that each feature individually improves worker accuracy, and jointly using all features provides the best results.

We validate our empirical influence measure by linking it to the Big Two of social cognition (i.e., agency and communion). The Big Two are known as the fundamental dimensions of social comparison, and they are strong predictors of social factors, such as gender, class, and power [24, 25]. Furthermore, prior works [26–28] have linked the Big Two and social influence. It follows that this link should also be measurable between the Big Two and our proposed empirical influence measure. We build regression models for the empirical influence using the intensity of the Big Two as features. We find that both agency and communion are significantly linked to empirical influence. Agency (the drive *to get ahead*) appears more important than communion (the drive *to get along*) for influence formation in our X/Twitter users, given the transient and non-ongoing nature of relationships on the

platform. Interestingly, the user follower count – a widely used proxy for social influence – does not have any detectable connection to social cognition.

The main contributions of this work include:

- A **crowdsource empirical influence measurement framework** to build the influence ranking of large cohorts that leverages human-in-the-loop active learning.
- A set of **simulation and fitting tools** for the measurement framework letting practitioners assess the required budget, design quality of annotation environment, and annotation fidelity.
- A **pilot study of MTurk design features** that improves the expected worker decision accuracy and minimizes the required budget.
- A showcase of the **link between empirical influence and the Big Two of social cognition**.

Challenges Of Measuring Social Influence. Social influence is a complex behavioral phenomenon, which is inherently subjective. Our measure is based on influence as perceived by MTurk workers; while we enforce interventions to manage worker subjectivity, we do not remove the inherent societal biases around influence. For example, would a profound impact on a small group or a minor impact dispersed over a vast group display more influence? Furthermore, influence is often described with a social context and a goal. One might choose to measure influence by whether an influencee performs an influenced action (i.e., compliance). In this work, we limit our scope to the facet of social influence related to online opinion formation.

We assume that people have a latent influence factor, which describes their influence with respect to the population. Additionally, we assume that a total ordering of the social influence of online users exists. In our protocol, we do not necessarily compare every pair of target users, however, we infer a relation between every pair using this latent influence factor. Note, that our monadic interpretation of influence has differing applicability compared to a dyadic one. This interpretation allows for a macro-scale analysis of the attributes associated with influence in general. However, it does not allow for analysis of micro-scale interactions and the attributes of relationships that reinforce influence (e.g., homophily-induced influence).

2.2 Preliminaries and Related Work

We structure the discussion of related work and preliminaries into two parts. Section 2.2.1 presents previous influence estimation approaches, accenting crowdsourced quantification and psychometrics. Section 2.2.2 introduces the required concepts of pairwise comparisons, the Bradley-Terry model and how to construct a ranking.

2.2.1 Influence Estimation and Crowdsourcing

Empirical Influence Measurements. There are two popular avenues in literature to investigate the dynamics of influence. The first avenue within the sociology literature uses agent-based simulations, which can demonstrate sufficient conditions for particular emergent outcomes [18, 29–31]. However, this avenue has rarely correlated outcomes with empirical data, and is predominantly validated at a phenomenological level [30, 32]. The second (and more practiced) avenue emanates from psychology and tests for particular behaviors in a social context, such as social distancing [15]. Researchers aim to identify the characteristics that allow individuals to exert influence – such as authority, likeability, attractiveness, and expertise [33]. Neither avenue deals with measuring the relative influence between individuals. Our work uses human workers to perform pairwise comparisons and creates a ranking of a set of targets based on their human-perceived influence. The approach is inspired by sociometry – used to measure social quantities, such as status, popularity, and reputation [34]. Generally, sociometric methods elicit (positive and negative) nominations about the quantity of interest (e.g., status) from members within a group and derive a measure from these nominations. Our approach differs from traditional sociometry in its application in the online space and, consequently, to peers in a large population who do not necessarily know each other.

Crowdsourcing Psychometrics. Extant literature has used crowdsourcing for psychometrics to measure context relevance [35], factors of explainability [36], and interest [37]. Unlike our work, these studies treat workers as the subject and fail to scale using modern active learning methods.

Influence Co-Variates. Influence has been related to several co-variates, such as likeability and authority. Given its fundamental role in human interactions there are likely many such co-variates. Bhattacharya et al. [38] show that *embeddedness* within a community is crucial to the formation of social movements. Furthermore, Shaik et al. [39] map the types of *engagement*, as a measure of cognitive mobilization, throughout a social movement. Embeddedness and engagement are likely covariates with influence, however their measurement remains elusive in setups like ours. In this work, we choose to investigate the big two of social cognition, given their recent utility in understand social cognition.

Crowdsourcing Influence. The investigation of influence through crowdsourcing is not unique to our work. There are two main themes in the literature: recruitment to artificial social platforms, and identification of micro-influencers. In the first theme, the extant literature constructs artificial social platforms and recruits crowdsource workers to interact on these platforms. This allows researchers access to metadata, which is usually proprietary and inaccessible for the typical social media platforms. Furthermore, this allows them to construct traditional controlled experiments with interventions. For example, Liu et al. [40] construct an artificial social platform with recruited MTurk workers and construct an influence ground truth (notably not a full ranking). Guilbeault et al. [41] use a similar approach to investigate opinion polarization. These approaches suffer from the artificial nature of the experiment, and their limitations are typically acknowledged within [40]. Our work does not leverage a synthetic world, and it uses human workers to compare pairs of targets – therefore, the conclusions suffer less from artificially introduced biases. In the second theme, works enlist the knowledge of crowdsource workers and the communities they are situated to identify micro-influencers (often as practical solutions to influence maximization problems). For example, Arous et al. [42] build a framework to ask workers open-ended questions for identifying micro-influencers. However, these approaches address a distinct task (classification rather than ranking), leverage complex worker responses (i.e., open-ended questions), and are constructed predominantly for marketing applications. Our work leverages an easily quantifiable human decision (pairwise comparison) and an algorithmic method to create a ranking of large populations of targets.

Crowdsource Design. Several papers point to the importance of properly designing the interface of crowdsource annotation tools and providing the proper context [43, 44].

The design choices directly link to the crowdsourced annotation quality, and prior work suggests keeping tasks simple and clear [45]. Our work defines a measure of performance for the MTurk interface design – the average accuracy of workers – and uses it to optimize the interface design by ablating over several features and procedures.

Crowdsourcing Platforms, such as Amazon Mechanical Turk (MTurk), allow large pools of human workers to complete tasks that computers cannot do. The research community uses them for labeling tasks (say, identifying objects in pictures) and psychometric studies. They are significantly cheaper, quicker, and induce a more diverse participant pool than traditional surveys [46]. MTurk provides a programmatic interface to serve tasks and process worker responses, making it amenable to active learning setups [47]—i.e., construct the next task based on previous answers. Additionally, researchers have meticulous control of the worker interface and can filter workers by their characteristics.

Pairwise Comparisons. There are two main psychometric approaches to measurement. The first approach asks people to rate items according to the measured trait (e.g., using a likert scale [48]). However, this direct ranking has scale calibration issues when performed by non-experts [49]. The second approach is the pairwise comparison, which features several advantages; it leads to lower measurement errors than the direct ranking of a set of targets [50]. It induces a quicker and more straightforward experimental task [51], suitable for non-experts in crowdsourcing setups [52] and is amenable to the active learning paradigm [47], which reduces the number of comparisons required for high measurement fidelity. Particularly in noisy and uncertain setups, pairwise comparisons outperform rating scales [53]. Methods for recovering a scoring from pairwise comparisons, such as Bradley-Terry and Thurstonian models, have a strong precedence in psychometric analysis [54, 55].

2.2.2 Pairwise Decisions and Ranking

The **Pairwise Comparison Matrix** (PCM) $M \in \mathbb{R}^{n \times n}$ represents the outcome of n items being compared where $M[i, j]$ is the number of times item i is favoured to item j—denoted hereafter as $j \prec i$.

Stochastic Pairwise Decisions and Ranking Items. An influence ranking is an ordered list $i \prec j \prec k \prec \ldots \prec z$, meaning that i is the least influential, j is more influential than i but less than k, and so on. Going from pairwise comparisons to ranking depends on the difficulty of the pairwise decision task. For simple tasks, human decisions can be considered deterministic – i.e., the same decision is made at multiple repetitions. For such deterministic pairwise decisions, the optimal ranking complexity requires $O(n\log(n))$ comparisons [56]. However, estimating influence is a difficult problem, even for humans; when presented with the same task, two humans might make different choices – e.g., one would say that $i \prec j$ and the other that $j \prec i$. We denote this as a *stochastic pairwise comparison*. When pairwise comparisons are stochastic, all pairs must be compared t times to overcome intransitivity. This requires a dense PCM, which takes $O(tn^2)$ comparisons and is prohibitively expensive when worker remuneration is per comparison.

Bradley-Terry Model (BT) [57] proposes a method for ranking individuals with sparse PCM – when only incomplete pairwise comparisons are available. It is commonly used in sports analysis [58, 59] (e.g., ranking chess players from matches) and psychometric studies [54]. The BT model is intimately linked to the work of the Weber-Fechner laws [60], and Thurstone's Law of Comparative Judgement [61]. Each individual i (i.e., a *target*) is ranked by its latent intensity $\theta_i \in \mathbb{R}$. The probability that target j is preferred to i is $\mathbb{P}(i \prec j) = \frac{1}{1+e^{-(\theta_j-\theta_i)}}$. The maximum-likelihood estimates (MLE) $\hat{\boldsymbol{\theta}}$ are computable even from incomplete sets of pairwise comparisons containing circular comparison results (e.g. $i \prec j \prec k \prec i$) [23, 55]; particularly relevant for crowdsourcing experiments where workers make difficult choices differently. Furthermore, adaptive methods for choosing the pairs to compare were proposed [62] to obtain high fidelity measurements with minimal comparisons.

2.3 The Empirical Influence Ranking Model

This section introduces the empirical influence methodology – our cost-effective method to construct empirical influence rankings using peer perceptions by MTurk workers. First, we describe the active learning approach that leverages an augmented BT model [23] and the ranking inference procedure (Section 2.3.1). Next, we propose a set of simulation

and fitting tools to estimate the required annotation budget (Section 2.3.2). Furthermore, we show the connection between parameters (for modeling systematic noise) and MTurk worker accuracy (Section 2.3.2).

2.3.1 Empirical influence measurements

Building the dense pairwise comparisons matrix M requires $O(tn^2)$ comparisons for n targets (because decisions are stochastic, see Section 2.2.2). As n grows, the process becomes prohibitive using crowdsourcing platforms, where costs are directly proportional to the number of comparisons. The Bradley-Terry model has been successfully applied on sparse versions of M (i.e., not all pairs are compared) to build approximate rankings [63]. The question is selecting which pairs to compare to maximize the ranking quality with the minimum number of comparisons. Passive techniques choose pairs before running the experiment; however, they do not use the information learned during the experiment. Here, we employ a solution that exploits active learning to choose comparisons on the fly. Past comparisons inform future choices, which, in turn, are more informative than random choices.

Sparse Pairwise Comparisons Matrix Via Human-In-the-Loop Active Learning. Our empirical influence quantification method builds on the active learning approach introduced by Maystre and Grossglauser [23]. We use the Quicksort (QS) algorithm to select pairs in the sparse pairwise comparison matrix M. We implement a human-in-the-loop system, in which human judges make the pairwise comparisons (using the MTurk platform), while the algorithm chooses which pairs to compare and builds the final ranking. The QS algorithm chooses a pivot point (target) and compares every other target in the set to this pivot. Crowdsourcing workers perform these comparisons. Two partitions are then formed based on these comparisons. Furthermore, we implement QS recursively; at each iteration, the sorting of the left ($<$) and right ($\geq$) subpartitions are performed in parallel, taking full advantage of MTurk's massive worker pool. In its design, QS exploits information from past comparisons to reduce the number of future comparisons required to complete the task, minimizing the total experiment cost.

We compare a pair of targets at most once during each QS execution (denoted as a *run*); usually, multiple runs are required. Maystre and Grossglauser [23] show that Kendall's Tau – a ranking quality metric – improves with the number of comparisons made, and the estimated ranking approaches asymptotically the true ranking.

The Augmented BT Model. Response fidelity in psychometric experiments suffers from two types of noise. The first type is *systematic noise*, associated with worker subjectivity, worker inauthenticity, and perception biases. This type of noise can be minimized via experimental design interventions (see Section 2.5). The second type of noise is *stochastic noise* that we average out using repeated trials. To account for response fidelity, we use an augmented BT model [23] that introduces the noise λ into the probability of preferring the target j over i as

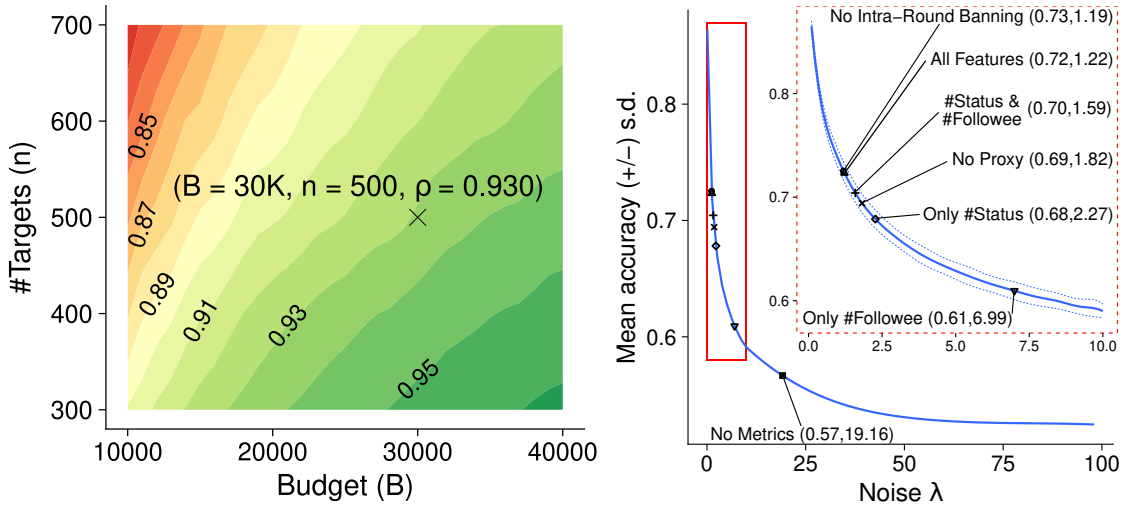
$$\mathbb{P}_{aug}(i \prec j) = \frac{1}{1 + e^{\frac{-(\theta_j - \theta_i)}{\lambda}}} \ . \tag{2.1}$$

Maystre and Grossglauser [23] show that Kendall's Tau deteriorates as noise increases.

Target Ranking Inference. Finally, given an observed set of comparisons $\{i \prec j\}$, we infer the influence scores $\hat{\boldsymbol{\theta}}$ by maximizing the log-likelihood:

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\text{argmax}} \sum_{\{i \prec j\}} log(\mathbb{P}_{aug}(i \prec j)) \\
&= \underset{\boldsymbol{\theta}}{\text{argmax}} \sum_{\{i \prec j\}} log\left(\frac{1}{1 + e^{\frac{-(\theta_j - \theta_i)}{\lambda}}}\right) \\
&= \underset{\boldsymbol{\theta}}{\text{argmax}} - \sum_{\{i \prec j\}} log\left(1 + e^{\frac{-(\theta_j - \theta_i)}{\lambda}}\right) \\
&= \underset{\boldsymbol{\theta}}{\text{argmin}} \sum_{\{i \prec j\}} log\left(1 + e^{\frac{-(\theta_j - \theta_i)}{\lambda}}\right) \ .
\end{aligned} \tag{2.2}$$

In this work, we solve Eq. (2.2) using maximum likelihood estimation (MLE) via iterative Luce Spectral Ranking [64]. The following section shows how to estimate the required budget needed to run a real-world experiment.



(A) (B)

FIGURE 2.2: (a) *Estimate required MTurk budget.* Here we infer the budget, given a required quality of ranking and number of targets. We vary the number of targets (n, y-axis) and maximum budget (B, x-axis). The color map and contour annotations show the Spearman correlation between the BT-estimated influence ranking ($\hat{\boldsymbol{\theta}}$) and a synthetic ground truth ($\boldsymbol{\theta}$). The cross denotes the chosen setup and estimated budget for our real-world experiments. (b) *Select worker interface design features.* Here we determine design features associated with higher worker decision accuracies, and find the relationship between noise and accuracy. The blue line shows the relation between the average accuracy (y-axis) and the noise (λ) (averaged over 100 simulations, $n = 500$ targets, $B = 30,000$). The points show ablations of design features (see Table 2.1) and their MLE fitted noise (relevant area zoomed in the inset). As more features are shown, worker accuracy increases.

2.3.2 Estimate required budget: noise, accuracy and simulations

This section introduces tools to estimate the required budget to run an influence estimation experiment at scale. First, we discuss the intertwining of noise, budget, and ranking quality. Next, we introduce a tool to simulate worker decisions with a given accuracy, and we show how to compute the required budget given the expected worker accuracy and the desired ranking quality. Finally, we show the theoretical link between the noise parameter λ and the average worker decision accuracy.

Intertwining Of Noise, Budget, and Ranking Quality. Deploying the empirical influence measurement introduced in Section 2.3.1 requires making a tradeoff between three intertwined factors: the noise parameter λ, budget (number of comparisons which translate into dollars), and ranking quality. For example, a higher noise level would require a higher budget to obtain a ranking of a given quality. Conversely, reducing the systematic

noise reduces the required budget. We measure the ranking quality as the Spearman rank correlation between the targets' inferred and real ranking.

Estimate Budget Requirements. It is often desirable to be able to estimate the required budget prior to launching the MTurk experiment. We achieve this in two steps.

In the *first step*, we simulate the worker's decision process with a given noise parameter, infer the synthetic influence ranking, and compute the ranking quality. The simulation requires three parameters: the maximum number of comparisons B (*budget*), the number of targets n (*#targets*), and the noise parameter λ (*noise*). We sample the synthetic latent influence intensities θ_i from a power-law distribution. We chose power-law as the literature observes that social metrics tend to follow a rich-get-richer paradigm [65]. For example, Twitter follower count is power-law distributed with exponent 2.016 [66], which we use for sampling the synthetic θ_i. For a pair of targets (i, j), our simulated workers produce correct decisions with probability $\mathbb{P}_{aug}(i \prec j)$ (see Eq. (2.1)), which is completely defined by θ_i, θ_j, and λ. We use the QS procedure to select B comparisons and compute the BT estimates $\hat{\boldsymbol{\theta}}$ from the recorded responses. Finally, we measure ranking quality ρ as the Spearman correlation between $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$. Therefore, the first step expresses the ranking quality as $\rho = \text{function}(B, n, \lambda)$.

In the *second step*, we perform a grid search over the budget B and the number of targets n. We therefore obtain $B = \text{function}(\rho, n|\lambda)$. We note that λ is not a parameter; it is linked to the worker accuracy (see later in this section) and depends on the MTurk interface design (see Section 2.5). Fig. 2.2a shows as a colormap ρ as a function of B (x-axis) and n (y-axis). Visibly, for a target quality (the colored area in Fig. 2.2a), the required budget increases with the number of targets. Here, $\lambda = 1.22$ based on our pilot study in Section 2.5. The labeled crossmark in Fig. 2.2a shows the configuration that we use in our experiments in Section 2.5; we chose a correlation level of 0.93 and 500 targets. Therefore, we estimate we require $30,000$ comparisons (approximately US$120).

Link Between the Noise Parameter and Worker Decision Accuracy. The systematic noise λ is intuitively linked to the accuracy of worker decisions: a higher noise is linked to a lower accuracy and vice-versa. Here, we show the theoretical connection between these two quantities. Let the decision of a worker comparing targets i and j be described by a

Bernoulli random variable

$$\begin{aligned} X_{ij} &\sim Bern(\mathbb{P}_{aug}(i \prec j)) \\ \text{with } \mathbb{E}[X_{ij}] &= 1 \times \mathbb{P}_{aug}(i \prec j) + 0 \times (1 - \mathbb{P}_{aug}(i \prec j)) \\ &= \mathbb{P}_{aug}(i \prec j) \ . \end{aligned}$$

The MTurk experiment is characterized by a series of Bernoulli trials, one trial per pair (i, j). Consequently, the accuracy of the human choices in the MTurk experiment is $\frac{\sum_{(i,j)} X_{ij}}{N}$. Note that X_{ij} are independent but not identically distributed since they depend on the choice of (i, j) for a given λ. The expected accuracy of the MTurk experiment over all worker choices is

$$\mathbb{E}[accuracy] = \mathbb{E}\left[\frac{\sum_{(i,j)} X_{ij}}{N}\right] = \frac{\sum_{(i,j)} \mathbb{E}[X_{ij}]}{N} = \frac{\sum_{(i,j)} \mathbb{P}_{aug}(i \prec j)}{N} \tag{2.3}$$

where N is the total number of compared (i, j) pairs. Eq. (2.3) links the mean worker accuracy and the λ noise parameter (via Eq. (2.1)). Visibly,

$$\begin{aligned} \lim_{\lambda \to \inf} \mathbb{E}[accuracy] &= \frac{\sum_{(i,j)} \lim_{\lambda \to \inf} \mathbb{P}_{aug}(i \prec j)}{N} \\ \text{cf. Sec. 2.3.1} \quad &= \frac{\sum_{(i,j)} \frac{1}{2}}{N} = \frac{1}{2} \ , \end{aligned} \tag{2.4}$$

e.g., the accuracy of unbiased random choice is a binary classification problem.

We use the synthetic worker decision generator described above to compute the relation between the noise λ and the mean accuracy. Fig. 2.2b plots this relationship and the standard deviation determined from 1000 process simuations. We make two observations. First, the mean accuracy converges asymptotically to 0.5 as the λ increases, as indicated by Eq. (2.4). Second, we notice that the standard deviations (dotted lines) are minimal, implying the relationship between accuracy and noise is fairly robust. In Section 2.5 we perform a series of pilot experiments to optimize the MTurk worker interface. We observe that, as we refine the worker interface design, the slider on the noise-accuracy line moves towards higher accuracy and lower systematic noise.

2.4 Dataset, Implementation, and Setup

This section introduces the foundational implementation details for running our QS ranking using real-life crowdsourcing workers. Firstly, we introduce the base experimental setup. Next, we describe the X/Twitter dataset that we use in this work. Finally, we describe how we sample the target and proxy users from the dataset.

MTurk experimental setup. The base experimental setup detailed here is used consistently across most variants explored in Sections 2.5 and 2.6. We use the ubiquitous MTurk crowdsourcing platform to implement the QS active learning procedure. The implementation runs QS partitions concurrently, so comparison pairs enter a First-In-First-Out (FIFO) queue and are served to MTurk workers in batches of 10. The workers were presented with two target users and a proxy user (see Table 2.1 and Section 2.5). Through a pool of differently worded questions, workers were asked to determine which target user was more influential to the proxy. These questions are *"Which user is the proxy user most likely to retweet?"*, *"Who will the proxy user be more socially influenced by?"*, and *"Which user would sway the proxy user's opinion more?"*. Fig. 2.3 shows the MTurk user interface as it would be presented to a worker.

#ArsonEmergency Dataset was collected by Graham and Keller [67] from X/Twitter in the context of the Australian 'Black Summer' Bushfires. It contains discussions around claims that arsonists caused the bushfires – now debunked as misinformation. The dataset was collected between 22 November 2019 and 9 January 2020—using keywords like *arsonemergency*, *bushfireaustralia*, *bushfirecrisis*, and other—, and contains $197,475$ tweets emitted by $129,778$ users.

Target and Proxy Sample Selection. We selected from the #ArsonEmergency dataset a sample of 500 targets and 500 proxies, controlled for availability, language, and Hawkes-modeled influence [1]. The users' availability (suspended or protected status at the time of the experiment) was queried through the Twitter API before the experiment. We selected solely English-speaking users, so the majoritively English-speaking MTurk workers could appropriately judge them. We use a triple-agreement approach between three language detection systems `langid` [68], `cld3` [69], and `whatthelang`. There may be

a significant interaction between language and influence, which could affect our sampling. To remedy this, we verified that the (55.8% remaining) filtered users were uniformly distributed with respect to Hawkes-modeled influence, via a chi-square test at a 95% significance level. From this set of valid users, we used an inverse CDF sampling method, with nearest-matching, to sample users with respect to Hawkes-modeled influence.

2.5 Pilot Study: Optimize MTurk Interface Design

In this section, we show how to design the MTurk worker interface to increase the worker decision accuracy and reduce the systematic noise. Intuitively, a proper experimental design and a worker interface presenting the appropriate information increase the accuracy of the decision. We begin by generally describing our pilot study methodology. Next, we detail each design feature and its corresponding impact on decision accuracy. Finally, we apply our design learnings and generate our final empirical measurement for influence.

Methodology. An ablative pilot study has a few ingredients: a set of design features, a procedure for running pilots, and a method to compare them. To begin, the set of experimental design features includes the *user component*, *proxies*, and *qualifications*. Table 2.1 shows the complete list and the feature description. Next, each pilot is a variation of the basic setup with design features added or removed. We performed three QS runs for each pilot and cleared the worker blocklist (see 'qualification system' below) between pilots. For clarity, workers who have participated in a prior pilot could participate in future pilots. In effect, workers may be exposed to different information associated with the same targets between pilots. Given the number of targets and potential comparisons, we assume this *memory bias* is negligible. Furthermore, we apply this protocol because while MTurk worker pools are vast, we do not wish to disenfranchise high-quality workers. Finally, to compare pilots we infer the decision accuracy they induce in workers. We fit the noise hyper-parameter λ by minimizing $\sum_i^N (\hat{\theta}_i - \theta_i)^2$, where $\hat{\boldsymbol{\theta}}$ is the influence determined using our QS active learning procedure and $\boldsymbol{\theta}$ is the ground truth. Note that λ is a hyper-parameter of our empirical influence measurement model, as it depends on the quality of workers' decisions and not the evaluated targets (see Sections 2.3.1 and 2.3.2). Consequently, λ and $\boldsymbol{\theta}$ cannot be jointly fit from pairwise comparisons (see Appendix A.1

TABLE 2.1: Design features for the MTurk user study.

Feature name	**Feature description**
Username, picture & link	Twitter user profile information
Description	User-reported description
Tweet samples	Sample of 5 tweets emitted by the user
Follower Count	The number of people who follow the user
Followee Count	The number of people the user follows
Status Count	The number of posts the user has authored
Proxy User	A third user on whom the influence of each of the targets is projected
Qualifications	A mechanism to blocklist low-quality workers

for a formal proof). Our pilot study uses follower count as a proxy for $\boldsymbol{\theta}$ because it is a widely adopted metric of influence (although it has been shown to be sub-optimal [65, 70, 71]). Fig. 2.2b plots the obtained noise λ and corresponding worker accuracy for each ablated MTurk design. In the rest of this section, we detail the impact of each design feature.

The User Component allows workers to quickly glean the relevant information about users, including the users' names, pictures, descriptions, hyperlinks to their Twitter profile, and a small sample of their tweets presented. We found that the most important user metrics are the *follower count*, *followee count*, and *statuses count*. Fig. 2.2b shows that removing these metrics significantly reduces worker accuracy – when we remove all metrics (the `No Metrics` annotated point on Fig. 2.2b), we observe the worst decision accuracy of 0.57. Showing the followee count or the status count significantly increases the accuracy to 0.61 (`Only #Followee`) and 0.68 (`Only #Status`), respectively; when showing both metrics above, the accuracy increases to 0.70 (`#Status & #Followee`). Adding the follower count (`All Features`) further boosts the accuracy to 0.72. While this incremental improvement from adding the follower count might seem counterintuitive—given that we use it as a proxy for the ground-truth influence parameter $\boldsymbol{\theta}$—two factors likely explain this effect. First, *follower count* is correlated with both *followee count* and *statuses count* in how workers perceive influence, providing limited unique information beyond these related cues. Second, workers were not informed that follower count serves as the influence ground truth in our model; when other cues are available, they may not treat follower count as a

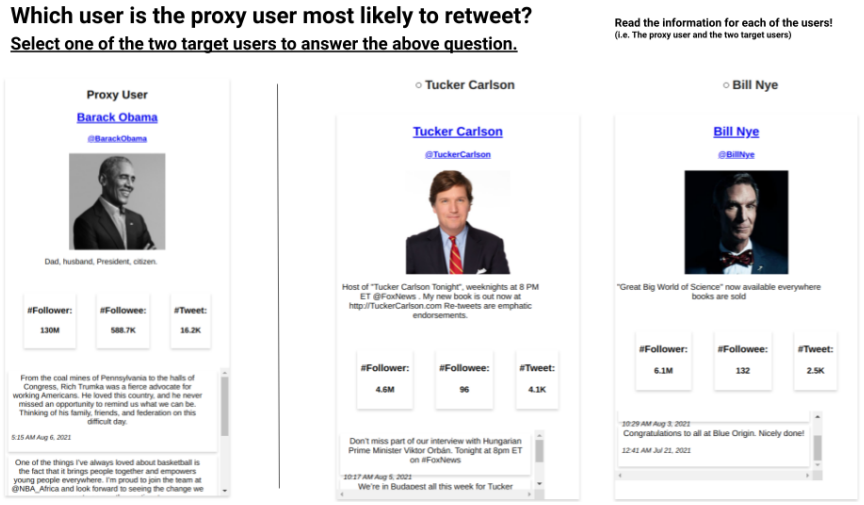


FIGURE 2.3: An example of the user panel within the MTurk interface, showing profile features (name, picture, and handle), metrics (follower, followee, and status counts), and a scrollable sample of authored tweets.

definitive indicator of influence. The above results suggest that all metrics are independently important signals of influence.

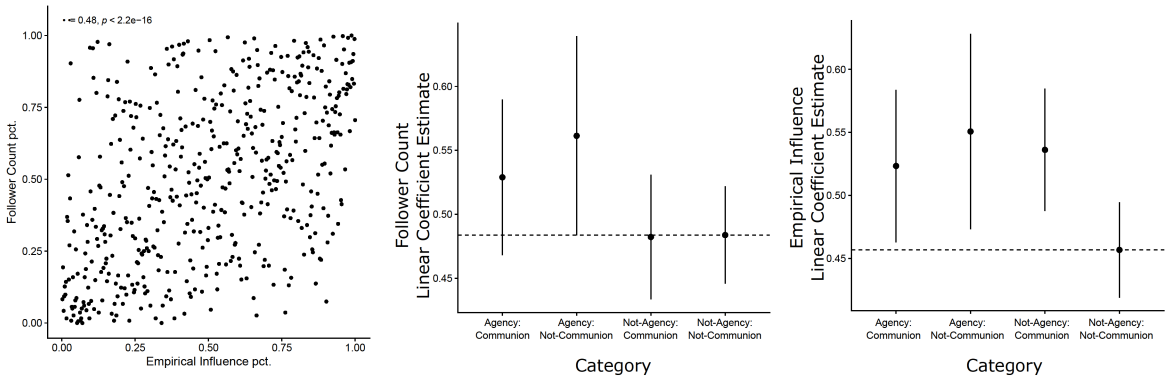
Proxy Users are used to reduce the effect of a worker's opinion on their influence judgment. In judging between two targets, we ask workers "who the proxy would find more influential". Proxy users do not eliminate the workers' subjectivity; however, they increase the worker decision accuracy by 3% – `No Proxy` has an accuracy of 0.69 compared to `All Features`.

The Qualification System restricts designated workers from completing tasks. The incentive structure of MTurk encourages workers to do as many tasks as quickly as possible. This leads to workers performing low-quality work; increasing the payment for each HIT (individual piece of work) does not alleviate the problem [72]. We broadly label such workers as *low-quality workers* and implement a blocklist mechanism to stop them from doing additional work and further reduce the decision accuracy.

Within a run, we measure the quality of a worker as the accuracy of the workers' responses concerning the target follower count percentiles. To clarify this procedure: workers provide binary pairwise comparisons (selecting which of two users is more influential), and we evaluate their accuracy by checking whether their selections align with follower count rankings on unambiguous cases. As the task is inherently subjective and difficult, we add some leniency to this banning scheme. Firstly, only comparisons where the difference of

follower count percentiles is more significant than 20% are included in determining accuracy. The intuition is that targets with a clear difference in the number of followers should be easier to judge. Secondly, banning is only implemented after completing 100 comparisons that satisfy the prior condition. Workers with accuracy below 50% on these clear-cut comparisons are blocked from future tasks. Lastly, the banning decision is only made once per run (but a banned worker stays banned for all remaining runs). This banning scheme is lenient enough to complete work quickly while restrictive enough that response quality remains high. Importantly, this filtering procedure is unlikely to train workers to focus specifically on follower count: workers must perform particularly poorly (worse than random chance) to be banned, and banned workers cannot return to learn from the experience, preventing any feedback-based learning about our evaluation criteria. Counterintuitively, Fig. 2.2b shows that removing the *intra-banning* mechanism (`No Intra-Round Banning`) leads to slightly better performance (accuracy 0.73) than `All Features` (accuracy 0.72). We believe the 0.73 score is too optimistic because low-quality workers were consistently discouraged from our task by the prior implementation of banning; when we removed it to measure impact, they did not return. This also corroborates with previous findings in the literature [72, 73].

Optimal MTurk interface and required budget. The above pilot studies lead us to the final MTurk worker interface with `All Features`, shown in Fig. 2.3, with a worker accuracy of 0.72 (corresponding a systematic noise $\lambda = 1.22$). Given the 500 targets to rank (see Section 2.4) and a desired ranking quality $\rho = 0.930$, we use the budget estimation procedure in Section 2.3.2. We determined that we require $36,252$ comparisons. According to MTurk ethics and US Federal minimum wage, we pay MTurk workers US$0.04 per HIT (a HIT contains 10 pairwise comparisons). As a result, we estimate the total cost of building the influence ranking using our proposed QS ranking for 500 targets to US$145. By comparison, building the dense comparison matrix for 500 targets would cost more than US$7,000 (see Section 2.2.2).



(A) (B) (C)

FIGURE 2.4: Plot (a) illustrates the insufficiency of follower count (i.e., the canonical metric) in representing empirical influence. It shows a moderate (spearman) correlation of 0.48 between empirical influence (x-axis) and follower count (y-axis). Plot (b) and (c) illustrate the nessecity of an empirical influence measure, in uncovering relationships with psychosocial attributes, over ad hoc metrics such as follower count. They show the coefficients of a linear regression predicting follower count (b) and empirical influence (c), respectively. The predictors are the use (or lack of use) of agency and communion in user tweets, and there is no intercept. The dashed line shows the estimate for the reference group (using neither agency nor communion), and the range shows a 95% confidence interval. Accordingly, coefficient ranges encompassing the dashed line are not significantly different from the reference group.

2.6 Empirical Social Influence and Social Cognition

Social influence is difficult to quantify directly; however, several works [26, 27] have shown it is conceptually linked to social cognition. In this section, we illustrate how our empirical influence measure captures the relationship between social influence and the *Big Two* of social cognition [24]; however, the more widely used follower count fails to do so.

Empirical Influence and Follower Count. Many prior works use ad hoc proxies for social influence that are simply correlated with it. One example of a widely used measure correlated with social influence is follower count [3, 74, 75]; however, it has been shown to be a biased measure of influence [65, 70, 71]. The QS ranking proposed in this work moves towards measuring social influence at scale through peer perception measurements. For example, Fig. 2.4a shows only a moderate Spearman correlation between the follower count and our empirical social influence measure, suggesting a significant variation is not explained by the former. Both follower count and empirical influence are converted to rank-percentiles within our sample for this analysis. The remainder of this section shows

why our empirical social influence estimation is a superior quantification of true social influence by connecting it with social cognition.

Linking the Big Two Of Social Cognition To Influence. Many factors impact the social influence phenomenon [33], some of which relate to the context of interaction and others to the individuals. Factors related to individuals include likeability and authority (i.e., expertise, confidence, social class, and status). The *Big Two* of social cognition – i.e., agency and communion – are categories of social motives that have been linked to gender [76], and social class [77]. Agency – the drive *to get ahead*, acquire skills, status, and power – is associated with masculinity and independence. Communion – the drive *to get along*, build trust, generate goodwill, and stoke mutual interest – is associated with femininity and dependence. The *Big Two* are intimately related to one's perception of self, others, and groups [25]. Intuitively, social cognition is linked to influence formation (for example, through agentic authority and communal likeability), and prior literature draws links between these social cognition concepts and social influence. For example, Abele and Wojciszke [25] show that liking is related to communion and respect is related to agency (and both liking and respect are features of influence [33]). Furthermore, Tveleneva et al. [26] link communion to higher susceptibility to influence. Frimer et al. [28] find that influential figures (such as public figures) use agentic and communic motives in specific ways. Finally, Abele and Wojciszke [24], Abele and Bruckmüller [78] suggest the Big Two are the fundamental ways we judge others; accordingly, if any personal traits affect influence, it would be these.

Linking Empirical Influence and the Big Two. Here, we analyze the relationship between the *Big Two* and social influence and illustrate the advantage of our empirical measure over ad hoc metrics like the follower count. We quantify the Big Two using a dictionary-based method [79], which matches the n-grams associated with agency and communion with the n-grams in the posts of our 500 target users (see details in Section 2.4). For example, posts containing the n-grams *authoritative* and *persistent* indicate higher agency, while those containing *benevolent* and *kindness* indicate higher communion[1].

[1]See `https://osf.io/jfct2` for the complete list of Big Two n-grams.

Next, we fit two linear regression models on the same dataset of 500 users: one for the follower count percentile and another for the empirical influence score, using the presence or absence of agentic and/or communic n-grams as predictors. The unit of analysis is individual users (not individual posts), where each user is coded as binary presence/absence of agentic and communic n-grams across their entire tweet corpus. Given the relatively small corpus of posts per user, we use binary indicators rather than continuous counts to avoid sparsity issues. This creates mutually exclusive categories: users with only agentic language, only communic language, both types, or neither (reference group). The regression uses standard linear regression with binary indicators, where each coefficient represents the average score for members of that group. Figs. 2.4b and 2.4c show the coefficients for each fitted model. For the follower count (Fig. 2.4b), we observe that no predictor accounting for Big Two features is significantly different from the reference group, as all confidence intervals intersect the dashed line of the reference group (see caption of Fig. 2.4). However, in Fig. 2.4c, we observe that all predictors are significantly different from the reference group. This means using agentic or communal language is associated with higher empirical social influence but not with a higher follower count. In other words, the follower count seems disconnected from social cognition measures, whereas our empirical influence measure is tightly dependent. Furthermore, we observe that agency alone is most highly associated with social influence, followed by communion alone. We speculate this is because X/Twitter is a debating environment [10] where relationships are transient and non-ongoing. In such environments, the agency is more important to project competence and influence quickly. We might expect communion to have a higher role in influence formation on platforms more conducive to forming social groups (such as Facebook). This is an avenue for future work. Interestingly, having both agency and communion leads only to a modest increase in social influence. This is because Abele and Wojciszke [25] show that agency and communion are negatively correlated – showing one dimension makes people assume the opposite about the other. As a result, Gebauer et al. [80] suggest that congruence (showing both dimensions together) leads to ambivalence in an observer. Testing the above hypotheses requires a larger sample, which we leave to future work.

2.7 Discussion

This work's contribution is four-fold; a human-in-the-loop empirical influence measurement framework, simulation and fitting tools for the framework, an empirical study of experimental design context features, and an analysis linking the Big Two of social cognition to social influence. The empirical influence measurement methodology is a novel contribution; robust to noise (see Fig. 2.2b) and highly correlated with true influence in a broad range of flexible simulation studies (see Fig. 2.2a). Furthermore, we find that upon applying the framework to online users, the empirical influence scores are better correlated with factors intuitively related to social influence than the baseline. It is important to note that the empirical method is limited to measuring the peer-perceived social influence of target individuals, which might be epistemologically distinct from latent social influence. We assume that workers are capable of distinguishing the relative social influence between individuals, and this is an accepted sociometric approach [20, 21]. The method could be appropriated to measure other psychosocial attributes of online individuals (such as reputation and trustworthiness) inexpensively, reliably and simply.

Chapter 3

Generalized Influence Measurement Model

Social influence plays a vital role in shaping human behavior, guiding how information is adopted, norms are followed, and collective decisions are made. During crises, such as the COVID-19 pandemic, social influence can determine whether life-saving information is effectively communicated, public health measures are observed, and vaccination campaigns reach their intended targets. However, the study of online social influence faces methodological challenges. Traditional psychosocial approaches, which emphasize understanding the social dynamics that drive influence, often rely on techniques that are slow, non-scalable, and limited in their ability to generalize to emerging issues. Meanwhile, computational approaches, which excel in processing large-scale data, frequently overlook the complex mechanisms underlying influence, instead relying on oversimplified metrics like network structures or event counts.

This chapter introduces research to address these challenges by developing a more integrative approach to modeling social influence. It proposes the Generalized Influence Model, a data-driven framework that bridges psychosocial theory and computational analysis. The model incorporates two novel mechanisms inspired by psychosocial concepts: conductance of social influence through ties and the distribution of influence capital throughout a

network path. These mechanisms are designed to capture the underlying dynamics of influence and to correct the biases introduced by simplistic metrics like follower counts.

Models play a critical role in the scientific understanding of social phenomena by capturing the mechanisms, informed by theory, that drive processes. In this research, the Generalized Influence Model serves not only as a descriptive tool but also as a theoretical framework that reflects the complex interplay of social factors. By incorporating psychosocial insights, the model provides a more nuanced understanding of influence, addressing limitations in existing computational methods and offering a clearer picture of how influence functions in online settings.

This research exemplifies the use of a model in a social science context. The model has been rigorously tested and shown to outperform state-of-the-art approaches, providing more accurate predictions of influence by correcting biases inherent in conventional models. By applying the model to discussions surrounding COVID-19, we were able to quantify users' influence and compare it against the veracity of the content they shared. Findings revealed that professions like executives, media figures, and military personnel were often more influential than experts such as life scientists and healthcare professionals. Troublingly, the analysis showed that some of these influential non-experts were also among the most active spreaders of misinformation, raising critical concerns about the effectiveness of information dissemination during crises.

Beyond individual case studies, models must be scalable to have a broad impact, and the Generalized Influence Model was designed with this in mind. Its ability to process large-scale data allows for the analysis of influence patterns across diverse contexts, providing a tool that can be applied widely in studying social phenomena. This scalability is essential for making reliable, data-driven predictions that inform policy and intervention strategies.

This research emphasizes that effective models must do more than process data; they must capture the mechanisms behind social behaviors, inform those mechanisms with robust theoretical frameworks, and be adaptable enough to apply at scale. By combining the depth of psychosocial theory with the precision of computational analysis, this approach highlights the transformative potential of models as predictive, scalable tools that can provide deeper insights into complex social dynamics.

This chapter demonstrates how interdisciplinary approaches can address significant societal challenges, such as misinformation and the spread of harmful narratives during crises. It exemplifies how computational models, informed by theory, can serve as robust, predictive, and scalable tools for understanding and managing social influence, supporting better decision-making and crisis response in a digital age.

Author Declaration

The following chapter contains content from the following publication.

Rohit Ram and Marian-Andrei Rizoiu. Conductance and influence-capital: Modeling online social influence. *Preparing Submission*, 2022

Author Contributions: R.R. led the research for this study, managed the data processing and collection, and conducted the experiments and analysis. M.A.R. provided supervision through all stages of the study. R.R. and M.A.R. collaboratively developed the model and experimental design. R.R. and M.A.R. interpreted the results and contributed to manuscript writing and editing.

Production Note:
Signature removed prior to publication.

Rohit Ram

Production Note:
Signature removed prior to publication.

Marian-Andrei Rizoiu

3.1 Introduction

Understanding public opinion formation remains an enigma with ample impact on our increasingly digitized society. The polarization of public opinion signals breakdowns in social trust [81], extreme opinions form pathways to radicalization [82], and in recent years, opinions opposing expert medical advice have led to loss of life by undermining immunization campaigns and stoking vaccine hesitancy. While we are beginning to understand the factors that drive opinion formation, we lack scalable modeling and quantification tools. These tools are essential when trying to mobilize our peers to enact change—whether to accept the vaccine or alter their consumption patterns to avert catastrophic climate change. One of the foundational factors in opinion formation and in enacting change is *influence* [83]. This force governs interpersonal relationships and contributes to establishing societal institutions and reforms.

Social media ubiquity has expanded the role of influence by providing a fertile ground for influence mechanisms to unfold and for a minority of users to exert disproportionate control. There exists ample evidence that online dynamics have offline repercussions in terms of collective movement [84] and radicalization [85]. Although there is a rich trove of literature on social influence, stretching from the psychosocial to computational domains, a quantitative approach to measuring online influence remains elusive [17, 18]. The computational approaches focus heavily on measuring message propagation propensities and exposure to messages through one's social network [66, 86, 87]; however, social influence is more accurately defined as the ability to change the behaviors, opinions, or beliefs of others [14]. Many psychosocial complexities of influence (like the social attributes of the source of influence) are lost when influence is narrowly interpreted as merely a reaction to exposure. Therefore, we ask **how can online influence be quantified, such that it relates to the theoretical psychosocial influence phenomenon?**

Furthermore, sociologists have long espoused that influence is most directly related to concepts such as social class and expertise [88]. Social class, a multifaceted construct encompassing education, wealth, occupation, and subculture, strongly aligns with established determinants of influence such as authority and likability [33]. While historically linked to political power, in contemporary egalitarian societies with social mobility, class

is predominantly indicated by occupation. Expertise, on the other hand, is related to a specific context and typically derives influence through demonstrated outcomes. Despite the compelling appeal of these sociological theories of influence, they remain largely untested on a large scale in online domains. Evaluating these theories is crucial, as adherence to expert advice reflects optimal societal behavior, and the extent to which social class either hinders or reinforces such advice can have critical implications. For instance, in an ideal scenario, populations would follow expert advice (e.g., from epidemiologists, biomedical scientists, healthcare professionals) during pandemics. However, the emergence of the anti-vaxxer movement and the proliferation of vaccine misinformation illustrates a contrasting reality. Therefore, we ask **how does influence correlate with its proposed determinants in the online domain?**

In this paper, we model online influence on Twitter/X. We achieve this in a sequence of two parts. In *the first part*, we build the Generalized Influence Model (GIM) – a novel contribution that bridges the divide between psychology and the quantitative approaches [12]. In Section 3.2, we review an essentially quantitative model [1] for online influence based on exposure. In Section 3.3, we build GIM by augmenting this model with two factors known in psychosocial literature to modulate the exertion of social influence: *conductance* [33, 89] and *social attribution* [33, 90–92]. We endogenize these factors by incorporating a social network conductance and a influence-capital distribution mechanism and propose several flexible implementations that maintain scalability. In Section 3.4, we show GIM to outperform several baselines, including the current state-of-the-art influence quantification [1], and to correct the biases introduced by the widely used follower count metric. In *the second part*, in Section 3.5, we empirically test the hypothesis that social class and expertise are the primary influence determinants. Following prior work, we assume that social class is primarily associated with one's occupation [93] and expertise is associated with the quality of the sources they share. On a Twitter/X dataset containing discussions about the COVID-19 pandemic, we utilize GIM to identify the occupational groups that yield the highest influence. We determine user occupation using the O*NET taxonomy [94], and numerically quantify the veracity of information users spread using a dataset of COVID-19 related misinformation [95]. The analysis reveals experts (i.e., epidemiologists, biomedical

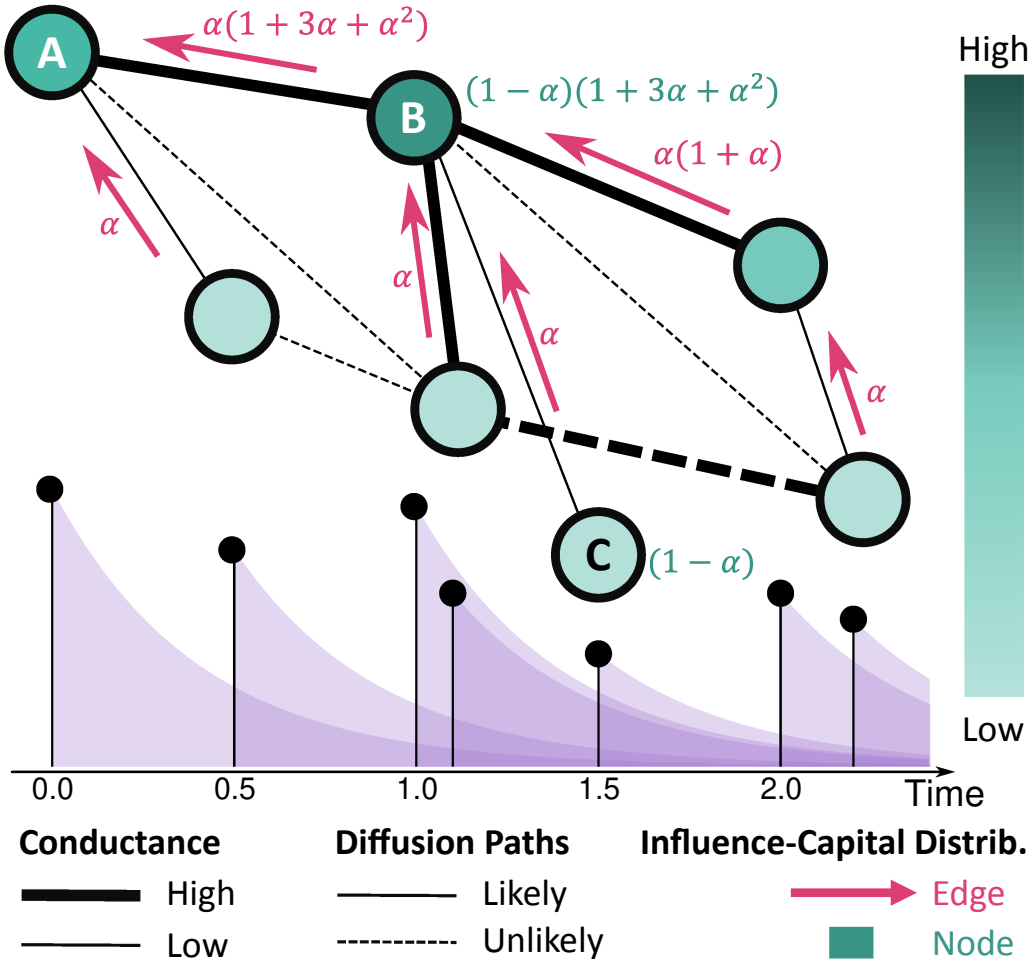


FIGURE 3.1: Schema of the Generalized Influence Model (GIM). (bottom) An example cascade is modeled using Hawkes processes. Each event (timestamp on x-axis) has a mark (y-axis) and spawns new events following a time-decaying intensity (magenta areas). (top) The latent branching structure is shown with solid lines, and other valid pathways are shown with dotted lines. GIM has two psychosocial-inspired components. *Conductance*: Edge thickness represents conductance, which modulates the likelihood of observing diffusions along that edge. *Influence-capital distribution*: A percentage α of a node's capital (green shades) is transferred along diffusion edges (red arrows), from target to source. Influence is proportional to the accumulated capital.

scientists, and healthcare professionals) are not always the ones that shape the discussions. So who does exert the greatest influence?

Successfully modeling online influence identifies the actors who shape societal views. It serves as the first step toward addressing societal issues—such as spreading misinformation amid a pandemic. **The main contributions of this work include:**

- The **General Influence Model** which introduces two psychosocial inspired mechanisms: *conductance* and *influence-capital distribution*, to move toward psychosocial influence.
- An **evaluation analysis** of GIM, where we find the optimal hyperparameters, show that GIM outperforms the state-of-the-art, and illustrate how GIM is an unbiased estimator.
- An **analysis of sociological determinants of influence**; investigating the relationship between social class, expertise, and GIM social influence.

3.2 Preliminaries & Related Works

Quantitative influence approaches exploit readily available metadata, such as timing and network features, which has historically been underutilized in psychosocial literature. These approaches, and the axioms from which they are built, form a strong foundation for an inquiry into influence. Influence is traditionally difficult to observe, as we cannot easily observe the minds of the influenced. We build upon prior works that take a longitudinal view and hypothesize that the endorsement patterns over extended periods reveal the influence structures in an observed cohort. On Twitter/X, this endorsement is signaled via retweeting, which is widely accepted as a form of endorsement of an emitter (the person being retweeted) by a receiver (the retweeter) [96]. However, platforms typically do not expose the structure of these endorsements. Nearly all social media platforms have a 'resharing' behavior (with varying degrees of API accessibility). The technique of reconstructing influence structures relies on acquiring the timing of reshare events. Rizoiu et al. [1] infer these structures by assuming retweets arrive following a Hawkes point process. In this section, we review approaches to influence quantification and articulate the Hawkes-modeled influence [1] from which GIM is built.

Prior Influence Quantification. Common approaches to quantitative models of influence include: using centrality on social graphs as an influence proxy [4, 97], approximating solutions to the influence maximization task [86, 87], and explicitly modeling popularity through approaches such as Hawkes Processes [98]. However, these approaches suffer from the conflation of social influence with other concepts such as popularity or activity. Several

studies have highlighted the discordance between measures of popularity and influence [65, 70, 71]. Recent work has proposed several novel approaches to influence measurement. Nickel and Le [99] efficiently model entity interactions via a Multivariate Hawkes Process (though this approach is entirely based on propagation propensity rather than grounded influence), and Smith et al. [70] estimate the impact by removing confounders through a causal inference network framework (though this approach must effectively reconstruct the network, makes restrictive diffusion assumptions, and focuses on influence operator classification). Unlike these prior methods, our work directly accounts for psychosocial factors that mediate influence propagation: influence-capital and network conductance.

Retweet Cascades consist of an original tweet and subsequent retweets. We denote a marked cascade up to time T as $\mathcal{H}_\zeta(T) = \{v_1, v_2, \ldots\}$, where $v_i = (t_i, \zeta_i)$ denotes the ith tweet, t_i is the event time relative to the original tweet ($t_1 = 0$), and $\zeta_i \in \mathbb{R}$—dubbed the *mark*—is the event meta-data. We denote a possible realization of the endorsement structure as $\mathcal{G}$, and the set of all potential endorsement structures as Υ.

The Branching Structure of retweet cascades is a latent graph (G, E) where G are the tweets, and E contains direct retweet relations. Given a cascade of n events, we define the set of all valid branching structures as $\Upsilon = \{G | (v_i, v_j) \in G \text{ so that } t_i < t_j\}$. Rizoiu et al. [1] show that $|\Upsilon| = (n-1)!$. Fig. 3.1 shows an example retweet cascade, its most likely branching structure in solid lines and other potential branching structures in dashed lines. Next, we compute the probability mass function over the branching structure set Υ.

Hawkes Processes [100] have emerged as a powerful tool for modeling social media events and information diffusion, providing a mathematical framework for capturing the self-exciting nature of online user activities [101**?**]. The self-exciting property—the arrival of an event increases the likelihood of future events— is applicable here due to the property of social affirmation (i.e., past social actions encourage incoming actions).

In Hawkes processes, events arrive following the conditional intensity

$$\lambda(t | \mathcal{H}_\zeta(t)) = \mu(t) + \sum_{v_i \in \mathcal{H}_\zeta(t) : t_i < t} \zeta_i^b \phi(t - t_i) \ ,$$

where $\mu(t)$ is the baseline intensity; each event v_i increases the overall intensity by $\zeta_i^b \phi(t-t_i)$; ζ_i^b is one way to model marks where b mediates the marks' effect, and the kernel $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ controls the event intensity decay. The exponential $\phi_{exp}(t) = e^{-rt}$ and the power-law $\phi_{pow}(t) = (t+c)^{-(1+r)}$ are common parametric forms.

Hawkes and Oakes [102] propose the branching representation of Hawkes processes, where each event v_i generates offspring following a non-homogenous Poisson process of intensity $\zeta_i^b \phi(t-t_i)$—illustrated by the magenta areas in Fig. 3.1. Lewis and Mohler [103] use the branching representation to estimate the probability that v_j is a direct offspring of v_i as

$$p_{ij} = \frac{\zeta_i^b \phi(t_j - t_i)}{\mu(t) + \sum_{t_k < t_j} \zeta_k^b \phi(t_j - t_k)}, t_i < t_j. \tag{3.1}$$

Intuitively, p_{ij} is the proportion of intensity that v_i contributed to the total intensity at time t_j. Note that for retweet cascades $\mu(t) = 0$, and $p_{ij} = 0,$ when $t_i > t_j$. Finally, the probability of a valid branching structure is $\prod_{(v_i,v_j)\in E(G)} p_{ij}$.

Hawkes-Modeled Influence. Rizoiu et al. [1] measure influence as the expected number of offspring of a tweet across all valid branching structures. They formalize influence as $\varphi(v_i) = \mathbb{E}_{\mathcal{G}\in\Upsilon}\Big[\sum_{t_j > t_i} \mathbb{1}(\mathcal{G}_{i\to j})\Big]$, where $\mathbb{1}(\mathcal{G}_{i\to j})$ indicates a path between v_i and v_j in the branching structure $\mathcal{G}$. Using the *independent cascades* assumption [86] (that generating a tweet at t_j is independent of the diffusion structure up to t_j), Rizoiu et al. [1] devise an efficient iterative procedure for computing $\varphi(v_i)$. They introduce m_{ik} the pairwise influence exerted by v_i on v_k, either directly when v_k is a direct offspring of v_i or indirectly when v_k lies on the same diffusion path as v_i. Formally, $m_{ik} = \sum_{j=i}^{k-1} m_{ij} p_{jk}, i < k$, $m_{ik} = 1$ when $i = k$, and $m_{ik} = 0$ when $i > k$. Intuitively, m_{ik} is the sum of the probabilities of all valid paths between i and k. Consider an example with three tweets, v_1, v_2, v_3, the influence m_{13} is computed as $m_{13} = m_{11}p_{13} + m_{12}p_{23} = p_{13} + p_{12}p_{23}$ representing all possible paths between v_1 and v_3. A tweet's influence is the total influence it exerts, i.e. $\varphi(v_i) = \sum_{k=i}^{n} m_{ik}$. The $\{m_{ik}\}$ matrix is computed in n matrix multiplications, and the total time complexity is $O(n^3)$. Finally, a user's influence is the average influence of all their tweets.

Towards Psychosocial. Our work is inspired by improved performance, via theory-based social features, in tasks related to social influence such as the influence prediction task [104, 105] (classifying whether users perform a behavior after exposure), social graph

embedding [106, 107], and recommender systems [108]. This suggests that incorporating psychosocial theory into computational models [109] can yield more accurate and nuanced representations of social influence processes. Our work aims to bridge the gap between computational efficiency and psychosocial theory by introducing mechanisms that capture the relational qualities of influence while maintaining computational scalability.

3.3 Methodology

Prior work estimates influence through endorsement timing and accessible user metadata; however, these approaches fail to consider the nature of relationships between people and the structure of these relationships. To address these limitations we propose the Generalized Influence Model (GIM) that quantifies influence based on observed information cascades. Our proposed GIM generalizes the Hawkes-modeled influence (see Section 3.2) by incorporating two mechanisms that model crucial social information, depicted in Fig. 3.1. These mechanisms – *conductance* ([33, 89], see Section 3.3.1) and *social attribution* ([33, 90–92], see Section 3.3.2) – are factors that psychosocial literature has identified as modulating the exertion of social influence.

We first formally introduce GIM, before detailing its mechanisms. Given a retweet cascade $\mathcal{H}_\zeta(T) = \{v_1, v_2, \ldots\}$, GIM quantifies the social influence of a tweet v_i as:

$$\varphi_\gamma(v_i) = \sum_{\mathcal{G}\in\Upsilon} \sum_{t_j > t_i} \underbrace{\mathbb{P}_\gamma(\mathcal{G}_{i\to j})}_{\text{Conductance}} \quad \underbrace{\Psi(\mathcal{G}_{i\to j})}_{\text{Capital Distrib.}} \quad , \tag{3.2}$$

where $\Psi(\mathcal{G}_{i\to j})$ is the influence-capital allocated to v_i along the endorsement pathway from v_j, and $\mathbb{P}_\gamma(\mathcal{G}_{i\to j})$ is the conductance-mediated probability that a path exists between v_i and v_j in $\mathcal{G}$ (defined in Section 3.3.1). Ψ is defined at the level of edges, preserving the efficiency of the Hawkes-modeled influence computation (see Section 3.3.3). The influence of a user is the average influence of their tweets.

3.3.1 Conductance

Quantitative models often assume that social ties are the primary influence channels, while other conductive channels (like homophily) are underexplored. Cognitive science [110] and social psychology [33, 89**?**] literatures suggest that some relationships are more influential than others. Notably, people in the same community (or who share similarities) are more influential to each other. Conductance assumes that different types of relations between users propagate influence more effectively (e.g., one might be more influenced by close family than by distant work colleagues); this makes some people more likely sources of influence than others. Intuitively, the social system propagates influence similarly to physical materials conducting electricity or heat. The conductance of a social connection modulates the likelihood of adopting an opinion. The higher the conductance, the more likely the receiver will adopt the opinion of the emitter. The influence conductance encapsulates user relationships; by measuring which links are more conducive to influence, we can infer more realistic endorsement pathways.

The conductance mechanism modulates the likelihood of cascade pathways using user lexical and following similarity, and relationship ties. We denote as $\gamma_{i,j}$ the conductance of an edge (v_i, v_j), and define the updated probability that v_j is a direct offspring of v_i as

$$p'_{ij} = \frac{\zeta_i^b \phi(t_j - t_i)\gamma_{i,j}}{\mu(t) + \sum_{t_k<t_j} \zeta_i^b \phi(t_j - t_k)\gamma_{k,j}}, t_i < t_j. \tag{3.3}$$

We consider two choices for conductance: *topological* (users' social network) and *homophilic* (users' similarity with others). We further operationalize homophilic conductance, using two lenses: *following* and *lexical*.

Topological Conductance assumes influence flows between users who are connected in the social graph (i.e., the follower relationship). Each valid edge (v_i, v_j), with $t_i < t_j$, has a baseline conductance β_{top}, regardless of whether u_i is connected to u_j, accounting for alternative influence conduits (such as news feeds or users following topics and hashtags). Formally, we define the topological conductance $\gamma_{ij}^{top} \in [0, 1]$ as $\gamma_{ij}^{top} := \beta_{top} + (1 - \beta_{top})a_{ij}$, where $a_{ij} = 1$ if u_j follows u_i and 0 otherwise (note that this creates a directed influence relationship: $u_i \to u_j$).

Homophilic Conductance of an edge, models the connection between similarity and influence, i.e., people similar to us influence us more [111, 112]. For each user u_i we first build a user representation $h_i \in \mathbb{R}^n$. Next, we quantify $\gamma_{i,j}^{hom}$ the homophilic conductance between two users using the cosine similarity between their user representations plus the baseline conductance β_{hom}. Formally, $\gamma_{i,j}^{hom} := \beta_{hom} + (1 - \beta_{hom})cosine(h_i, h_j)$.

The *following lens* leverages the observation that similar people consume similar content. On social media, following popular users is akin to consuming content. Accordingly, we measure the similarity between two users based on whether they follow the same people. For the following lens, we identify the 1000 most followed users, and collect the followees of all the users in our dataset (i.e., users they follow). We represent a user u_i as $h_i \in \mathbb{R}^{1000}$, where $h_i[j] = 1$ if u_i follows the jth most followed user ($h_i[j] = 0$ otherwise).

The *lexical lens* exploits the insight that similar people use similar language. The vocabulary and language style of users can be a strong indicator of their community, and we measure the similarity of users based on their choice of language. For the lexical lens, we construct user documents by concatenating the user tweets; we represent them using TF-IDF (Term Frequency-Inverse Document Frequency); and a feature hashing dimensionality reduction technique [113]. Finally, we represent each user u_i as $h_i \in \mathbb{R}^{1,048,576}$. Note that the homophilic conductance with the lexical lens does not require knowledge of the user following graph, which can be prohibitively expensive to obtain for Twitter. Related to our work, **?**] characterize the relationship between homophilic ties and the influence phenomena, however their network is artificial and they illustrate only one homophilic lense.

3.3.2 Influence-capital Distribution

The endorsement pattern is not a sufficient explanation of the ability of individuals to exert influence. Sociologists have long pointed to the importance of weak-ties [114] and the accumulation of social capital by *bonders* [115]. Consider the example in Fig. 3.1. Alice influences Bobbie, who influences several people; should Alice (the initiator) or Bobbie (the connector) be considered more influential?

Psychologists have recognized that particular characteristics in individuals are correlated with influence, namely authority [90], attractiveness [91], likeability and others [33]. We define *influence-capital* as the congruence of characteristics that enable the exertion of influence. We propose that users possess latent influence-capital, such that an emitter with higher influence-capital exerts more influence. We posit that a distribution mechanism that explains endorsement pathways, must measure this influence-capital. The question remains, how do we allocate influence-capital to explain the exerted influence along an inferred endorsement pathway?

We propose a *influence-capital distribution* that leads to the accumulation of influence-capital and offers a post-hoc explanation for the inferred endorsement patterns. The mechanism transfers a proportion of a node's (tweet's) influence-capital to its parent in the information cascade, allowing upstream and highly connected nodes to accumulate influence-capital (which translates to influence). The distribution mechanism that we propose here aims to explain the endorsement pathways inferred from the data, and it is related to the concept of value-allocation schemes [92, 116]. Several studies [92, 117, 118] have utilized allocation schemes to recognize the role of *bonders* by assuming benefits, generated by a node, decay with distance within a social graph. We migrate this intuition to a directed diffusion scenario. Intuitively, the allocation rewards users responsible for bonding (i.e. connectors), and those who reach distant communities (i.e. initiators).

We construct the influence-capital distribution as follows. Whenever the user u_i directly influences u_j (i.e., v_j is a direct offspring of v_i in an endorsement pathway), u_j transfers a portion of their influence-capital to u_i (denoted as π_{ij}). Each tweet is endowed with 1 influence-capital for participation; it pays a proportion $\alpha \in (0, 1)$ of all its capital to its parent (if they exist) and keeps $(1 - \alpha)$. Formally:

$$\pi_{ij} = \begin{cases} \alpha, & 1 \leq i < j, \quad j \text{ transfers } \alpha\% \text{ capital to } i, \\ (1-\alpha), & i = j \neq 1, \quad j \text{ keeps } (1-\alpha)\% \text{ capital}, \\ 1, & i = j = 1, \quad \text{The initiator does not transfer.} \end{cases}$$

Fig. 3.1 illustrates the mechanism. Charlie passes Bobbie α of her endowed capital (keeping $1-\alpha$), and Bobbie receives capital from her endowment (1), her three direct influencees (3α),

and one indirect influencee (α^2). She keeps $(1-\alpha)$ of this sum. The capital distribution of a path $\mathcal{G}_{i\to j}$ is $\Psi(\mathcal{G}_{i\to j}) = \prod_{(k,l)\in E(\mathcal{G}_{i\to j})} \pi_{kl}$, and a user's social influence is proportional to the total influence-capital they accumulate via the capital distribution mechanism. The scheme naturally conserves total value, which is equal to the number of participants.

The current framework operates under the assumption that reshares constitute signals of endorsement within information cascades. However, social media interactions encompass a broader spectrum of engagement mechanisms, including replies, mentions, and tags, each carrying distinct semantic implications. While replies may indicate disagreement or constructive debate rather than endorsement, mentions and tags can facilitate influence propagation across non-follower networks. Future research could explore incorporating these alternative interaction modalities to capture more nuanced forms of social engagement, potentially enabling analysis of conflicting discourse patterns and discussion tree dynamics beyond simple endorsement cascades.

3.3.3 Iterative Computation

GIM can be computed efficiently by extending the pairwise influence m_{ik} (introduced in Section 3.2) to incorporate the concepts of conductance and influence-capital distribution. Formally, $m_{ik} = \sum_{j=i}^{k-1} m_{ij} p'_{jk} \pi_{jk}, i < k$, where $m_{ik} = p'_{ik}\pi_{ik}$ when $i = k$, and $m_{ik} = 0$ when $i > k$. Consequently (and similar to the Hawkes-modeled influence), we obtain $\varphi_\gamma(v_i) = \sum_{k=i}^{n} m_{ik}$. The full derivation of the latter, from Eq. (3.2) via Eq. (3.3) and Section 3.3.3, is shown in the online appendix [119]. GIM recursively generates all possible paths in the same way as the Hawkes-modeled Influence [1], which allows an efficient iterative algorithm of temporal complexity $O(n^3)$. Visibly, the Hawkes-modeled influence [1] is a special case of GIM, with $\pi_{jk} = \gamma_{jk} = 1, \forall j, k$.

3.4 GIM Evaluation

Having developed the theoretical foundation for our Generalized Influence Model, we now turn to its empirical evaluation. This section addresses three key questions: (1) what combination of conductance and distribution mechanisms yields optimal performance?

(2) how does GIM compare to existing influence estimation methods? and, (3) can GIM overcome the known biases of traditional influence metrics?

Ground Truth Influence. Social influence is inherently difficult to quantify, particularly in online settings where direct behavioral change is challenging to observe. We utilize our previously developed empirical influence ground truth [6] which employs a human-in-the-loop active learning approach to measure peer-perceived influence. The method leverages crowdsourcing workers who perform pairwise comparisons between users, determining which is more influential. To optimize the required number of comparisons, the method implements a Quicksort-based active learning algorithm that selects the most informative pairs to compare, reducing the quadratic complexity of naive approaches to a loglinear one. The comparisons are then used to fit a Bradley-Terry model, which produces a complete ranking of social influence scores. The resulting dataset consists of influence rankings for 500 Twitter users selected from the #ArsonEmergency dataset [67]— containing Twitter discussions about the Australian bushfires collected between November 2019 and January 2020. This empirical influence measure has been validated by showing strong correlation with the Big Two of social cognition (agency and communion), theoretical determinants of social influence. We utilize this empirical influence ranking to calibrate and evaluate GIM.

Evaluation Metrics. We evaluate GIM ranking against the ground truth (described above) using two measures: NDCG-AUC (see next) and MAPE. The information retrieval literature uses the *Normalised Discounted Cumulative Gain* (NDCG) to measure the overlap between two rankings. It privileges the correct ranking of the top-ranked positions and discounts errors in the lower rankings. Applied to influence, NDCG@k aims to order the k most influential users correctly. We compute the Area Under Curve for NDCG@k (AUC-NDCG) by varying k and producing a single metric value. We also compute the Mean Absolute Percentage Error (MAPE) of the difference in the ranking percentiles for each target. Having established our evaluation methodology, we next introduce the baseline methods against which we compare GIM.

Influence Ranking Baselines. We compare GIM to four baselines. Two baselines are widely used heuristics; *PageRank* [2, 120] assumes influence flows via random-walks on constructed social graphs (here the follower network) and *retweet influence* [3], counts the

retweets of a user's authored tweets. They are centrality- and feature-based approaches, respectively. The other two baselines are purpose-built state-of-the-art influence estimators: *Hawkes-modeled influence* [1] and *ProfileRank* [4] (a PageRank variant). Next, we conduct a comprehensive search to identify the optimal configuration of our proposed model.

GIM search. For each combination of conductance (topological, lexical, and following) and distribution mechanism (influence-capital, and none), we perform a grid search over the hyper-parameters β (conductance) and α (distribution). At each grid point, we compare the influence scores obtained by GIM against the ground truth empirical influence ranking. Fig. 3.2 shows the baselines and GIM with several conductance-distribution combinations in the space of the performance measures: negative MAPE (x-axis) and NDCG-AUC (y-axis) (the top-right corner optimizes both measures).

Analysis of the experimental results reveals three key observations. First, GIM consistently Pareto-dominates (i.e., outperforms) all baselines for almost every hyperparameter combination, showing that our psychosocial-inspired mechanisms render automatic influence quantification closer to the human judgment. Among baselines, the next best performing is the Hawkes-modeled influence, followed by PageRank, retweet influence, and the purpose-built ProfileRank. Second, we observe that the homophilic conductances (i.e., lexical and following) typically outperforms the topological conductance, and between homophilic conductances lexical outperforms the following conductance. Third, only two models are not Pareto-dominated: the topological-none (best NDCG-AUC) and the lexical-influence-capital (best neg. MAPE). Notice, however, that the topological conductance requires recovering the follower network. In practice, this is prohibitive for large datasets (such as the #Covid-19 dataset) due to rate limitations of the Twitter/X API. Based on these findings, in our analysis in Section 3.5 we use the homophilic lexical conductance ($\beta = 0.18$) with the influence-capital distribution ($\alpha = 0.02$). The 18% baseline conductance indicates that the accounted channels do not explain a relatively large proportion of conductance. Note that while passing 2% of the influence-capital to the parent might not seem much, this adds up for nodes with high degrees (particularly given the longtail distribution of follower count [121]). Having established GIM's superior performance against existing models, we now examine its ability to address a key limitation of traditional influence metrics.

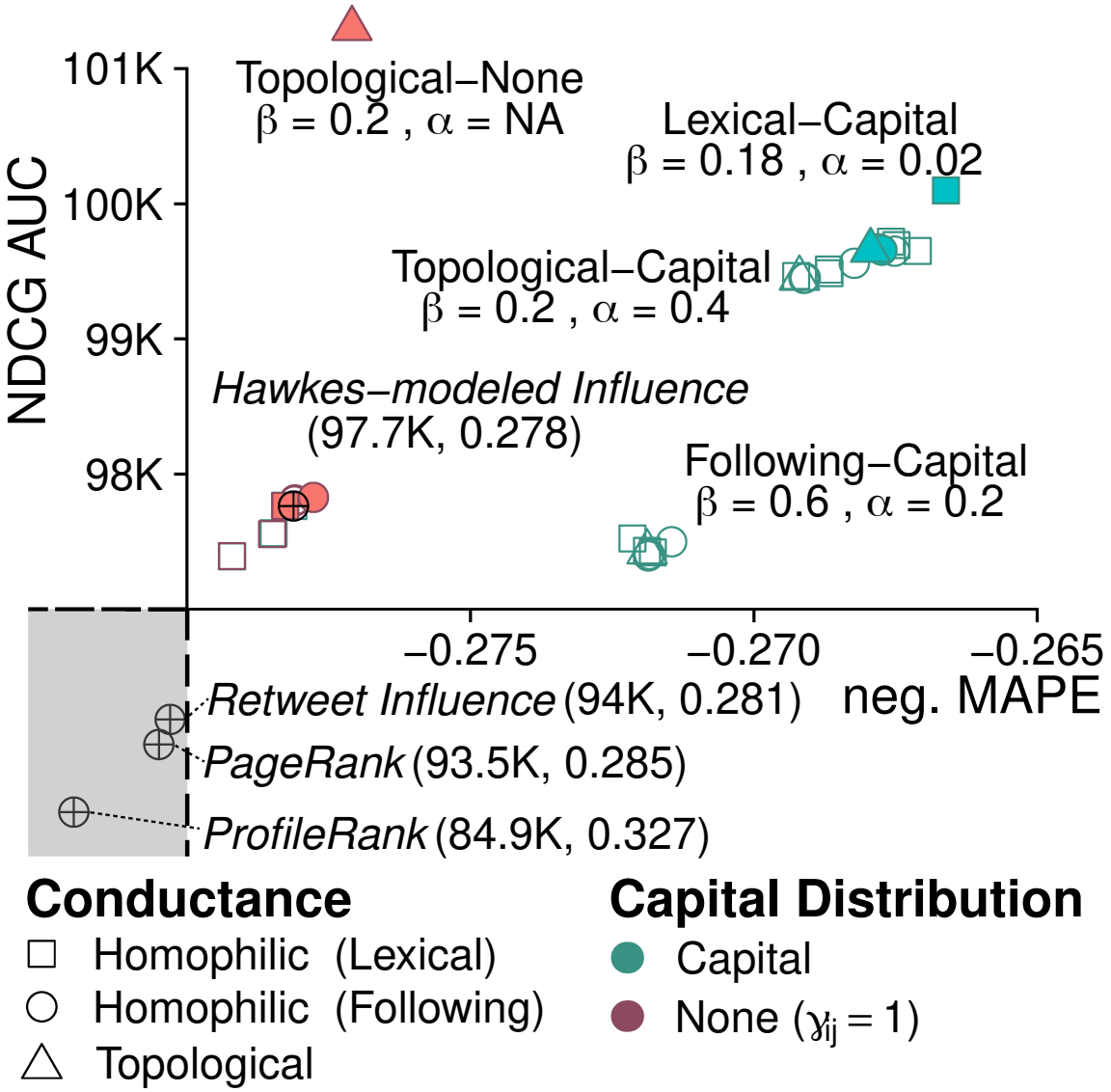


FIGURE 3.2: *Evaluate GIM against the ground-truth in the space of NDCG-AUC (y-axis) and negative MAPE (x-axis).* Higher NDCG-AUC and more negative MAPE indicates a better performing model. The solid shapes are the best models for each combination (conductance–capital distribution). The empty shapes show the Pareto-dominated models in each combination, obtained via grid search in the space (β, α). The circle-crosses denote the baselines: Hawkes-modeled influence baseline [1], PageRank [2], Retweet Influence [3], and ProfileRank [4]. Note, the gray box is not to scale, and the coordinates for baselines are shown in brackets.

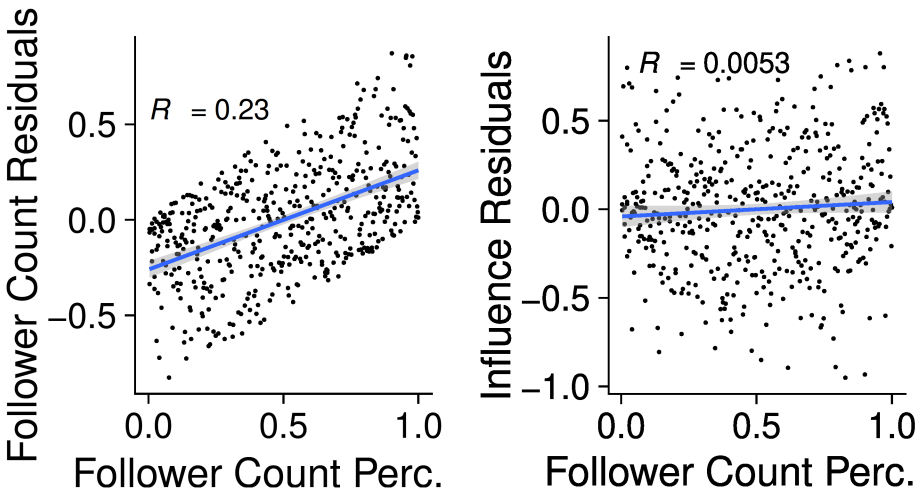


FIGURE 3.3: Residuals (relative to the empirical follower ranking) for the follower count (left) and GIM (right) against the follower count (x-axis).

Debiasing the Follower Count. The follower count is widely used as a proxy for influence [3, 74, 75]; however, it has been repeatedly shown to be biased [65, 70, 71]. We examine whether GIM can overcome this known limitation. Fig. 3.3(left) shows that the follower count residuals—the difference between the follower count percentile and the empirical score percentile—are positively correlated with the follower count percentile ($R^2 = 0.48$). In other words, the follower count overestimates the influence of the highly followed users and underestimates the lowly followed users. In contrast, Fig. 3.3(right) shows that GIM residuals are not correlated with the follower count percentile ($R^2 = 0.073$). That is, GIM is an unbiased influence estimator with respect to the follower count. Having demonstrated GIM's effectiveness at quantifying influence in an unbiased manner, we next proceed to apply it to real-world data to investigate the relationship between social class, expertise, and influence in online discourse.

3.5 Social Class, Expertise, and Influence during COVID

Having established GIM as an effective influence quantification model that outperforms state-of-the-art approaches and corrects inherent biases, we now apply this tool to examine fundamental sociological hypotheses. Sociologists have long espoused that power and influence are primarily related to social class and expertise [88]. The influence of experts, and trust in their advice, is important for the optimal functioning of society. Failure to

adhere to advice can lead to unnecessary death (e.g., vaccine hesitancy) and existential threats (e.g., climate change). Social class is often nebulously defined but is related to wealth, occupation, and subculture. In our increasingly egalitarian societies, occupation provides a suitable proxy. The COVID-19 pandemic presents an ideal context for investigating these relationships, as it constitutes a crisis where expert communication should be paramount for public safety. This global health emergency provides several methodological advantages: clear expert consensus on key issues, readily identifiable subject matter experts (epidemiologists, healthcare professionals, biomedical scientists), widespread public discourse across occupational groups, and established datasets for measuring information veracity. Additionally, the COVID-19 dataset is large and well-studied, with the contextual factors underlying influence behaviors being well understood in the literature. Unlike more ambiguous political issues, COVID-19 offers an objective standard against which to evaluate the alignment between expertise, social class, and actual influence. Ideally, influence, expertise, and social class are aligned; however, this does not seem to have been the case as Bailo et al. [122] observed that far-right accounts, which had been peripheral during previous crisis events, managed to assume more central positions in online COVID-19 discussions.

In this section, we leverage our validated GIM framework to empirically examine the interplay between social influence, expertise, and occupational class during the COVID-19 pandemic. We first introduce the COVID dataset and use GIM to compute the influence of all users in the #Covid-19 dataset. Next, we develop methods to quantify two key variables: users' occupations as proxies for social class, and the veracity of information they spread as an indicator of expertise. We extract users' occupations and the veracity of the information they spread, and we tabulate their influence and veracity against their occupation. Finally, we analyze the relationships between these variables to assess whether influence patterns align with theoretical expectations about expertise and social class.

Dataset. The dataset was collected from Twitter/X in the context of the COVID-19 pandemic. The #Covid-19 dataset was constructed using the keyword *covid19* during August 2020 and contains $143,356,591$ tweets by $21,527,913$ users. Note, this dataset is different from the #ArsonEmergency dataset, we used to tune GIM.

Determine the Occupations Of Twitter/X Users. To establish occupational classifications as proxies for social class, we implement a systematic approach for identifying users' professions [123]. We match user occupation against the Minor Group Occupational Classes of the O*NET occupational taxonomy [94] using textual fuzzy matching [124]. O*NET described users better than other taxonomies investigated. We search each user's Twitter/X description and select the first matched occupation (following Sloan et al. [93])—assuming people list their actual occupation first, before hobbies and other information. We validated the classifier on 100 labeled users, where ground-truth labels were derived by two annotators, with disagreement resolved by discussion. The classifier has a mean macro-F1 of 0.54, comparable to the classification performance reported by literature [125].

Quantifying Expertise. We assume that expertise is correlated with the veracity (i.e., quality) of the content users share. To operationalize expertise in the online domain, we develop a method for measuring the quality of information shared by users. We use the links and tweets that users share to quantify the veracity of the information they spread. We follow prior research [126] and compute a veracity score for the domain names of the URLs. First, we extract from the CoAID dataset [95] all the links and tweets associated with *true* and *fake* information related to COVID-19. The dataset contains full URLs, and some URL domains appear many times. For each URL domain within CoAID, we count the true ($\#R$) and fake ($\#F$) entries recorded in the dataset. Finally, we generate for each domain a normalized score as $\frac{\#R-\#F}{\#R+\#F}$. The score is between -1 (domain is fully unreliable and spreads misinformation) and 1 (fully reliable). Furthermore and similar to [126], we curate a set of *high-quality health sources* (HQHS) from prominent health websites (e.g., CDC, WHO, and Mayo Clinic), and medical journal websites (e.g., The Lancet, and Nature). We compute the veracity score of a post with a link as follows: (1) -1 or 1 if the full link appears in CoAID as fake or true, respectively; (2) 1 if the link domain appears in the HQHS set; (3) the domain's veracity $\in [-1, 1]$, if neither (1) nor (2) applies. A user's veracity score is the mean veracity of the posts they share.

The Influence and Veracity Of Occupations. Having quantified influence, occupation, and information veracity, we now analyze how these variables interact across different professional groups during the pandemic. Fig. 3.4a shows the distribution of user influence

and user veracity scores, for occupations with more than a thousand users in the #COVID-19 dataset. *Executives*, the *Media*, *Entertainers* and the *Military* yield among the highest online influences. This result is hardly surprising for the former three. Notably, *Media* and *Entertainers* are not only influential but have prominent online presences, perhaps on account of their attention-related business models. The latter (*Military*) is an occupation with high bipartisan support and respect in the US – home of most English-speaking Twitter/X users. Examining the veracity distributions reveals important patterns in information quality across occupational groups. For all occupations, Fig. 3.4a shows that the veracity score has a bimodal distribution. One mode is around -1 (the users who spread mainly misinformation) and another at 1 made of users who spread high-quality information. Users typically do not engage with both types of information, probably due to homophily and online polarization. As all occupational subpopulations contain both types of users, the mean occupation veracity (colored horizontal bars in the right panel of Fig. 3.4a) represents the ratio between misinformation and high-quality information spreaders within an occupation. Contrary to theoretical expectations about expertise and information quality, we find that several occupations traditionally associated with authority spread substantial misinformation. Surprisingly, we observe that a significant proportion of *Military*, *Firefighters* and *Police* users spread misinformation. Closer investigation shows they frequently share from controversial publishers, such as *foxnews.com*, *zerohedge.com*, and *breitbart.com*. To better understand the relationship between influence and information quality, we examine their correlation across occupational groups. Fig. 3.4b shows the scatterplot of the occupations, in the space of mean influence (x-axis) and mean veracity (y-axis). First, we observe that the two quantities are uncorrelated ($R^2 = 0.016$). This finding contradicts theoretical expectations that expertise (measured as information quality) would align with influence. We also see that *Life Scientists* (including epidemiologists) and *Social Scientists* typically spread high-quality information; however, they experience limited influence. *Healthcare* workers are moderately influential and have a good proportion of high-quality information spreaders; however, *Healthcare Technicians* including paramedics and medical technologists, have a significant proportion of misinformation spreaders.

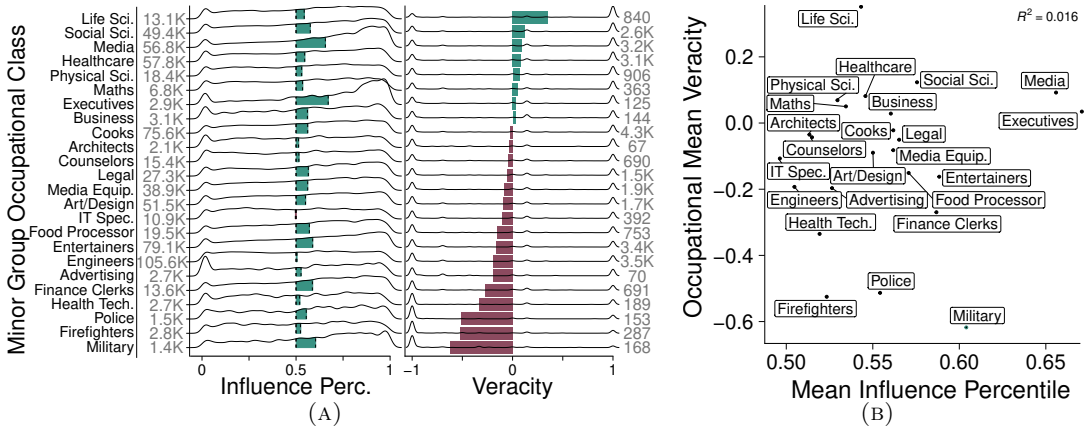


FIGURE 3.4: (a)(left panel) The influence distribution for the O*NET Minor Group occupations with more than 1,000 users in #COVID-19 (number of users shown on the left). (a)(right panel) The veracity distribution of the same occupations (number of spreaders shown on the right). Color bars show the difference between the occupation mean and the distribution's center (0.5 for influence and 0 for veracity). (b) The mean veracity (y-axis) and mean influence percentile (x-axis) of occupations.

3.6 Discussion

Existing methods for online influence measurement rely on heuristic definitions and fail to model complex phenomena. In this work, we have developed an online social influence framework. Our conductance mechanism departs from traditional approaches in the literature which frequently considers the social network (i.e. the topological lens) as the primary channel of influence [86, 87]. As online interactions are increasingly mediated by recommender systems, web search, and cross-platform engagement, modeling the similarity between users (i.e. the homophilic lens) provides a simple and effective method for modeling influence channels. Our findings demonstrate that lexical features are readily attainable and perform influence modeling well (see Fig. 3.2). The conductance mechanism we have introduced extends beyond the current application and could be used to mediate diffusion models or as a more general conception of the social graph, for social tasks (i.e. recommendation) or social network analysis.

The influence-capital distribution mechanism in our GIM framework provides an intuitive approach for allocating influence-capital to explain influence (as perceived by peers). This approach significantly improves upon the arbitrary allocations of prior influence models and integrates naturally with the scalable procedure for estimating influence over the stochastic diffusion graph.

The literature constructed the theoretical relationship between social influence and social class, with occupation serving as a key indicator of social class via educational attainment, income, and prestige. Through our analysis, we have measured occupational influence in discussions around the COVID-19 pandemic (see Fig. 3.4b). Our results reveal that certain aspects of social class can explain our results; for example, the influence of *Executives* might be explained by their income and occupational prestige. Interestingly, we find that individuals with high educational attainment are not influential in this context. Additionally, our findings suggest a significant determinant of influence, with both *Media* and *Entertainers*, having high influence. More broadly, the modeling of social influence can have broader implications for understanding how extreme opinions infiltrate mainstream discussions [127].

An intriguing direction for future analysis involves examining the veracity-influence relationship within occupational categories rather than solely comparing between occupations. While our cross-occupational analysis reveals important patterns, investigating whether higher-veracity individuals are more influential within their respective professions could illuminate different dynamics of expertise and credibility across domains. For instance, within *Life Sciences*, higher veracity might correlate with greater influence as peers recognize and amplify authoritative voices. Conversely, within *Firefighters* or *Military* populations, different patterns might emerge where influence operates through alternative mechanisms such as institutional authority or community solidarity rather than information quality.

Such within-occupation analyses could reveal that people seek different targets for different purposes—turning to *Life Scientists* for evidence-based information while following *Entertainers* or *Military* accounts for other social or political reasons. This could explain why the aggregate cross-occupational analysis shows weak correlations between influence and veracity: the relationship may vary substantially across professions, with opposing or null correlations within different occupational groups canceling each other out in aggregate analyses. However, conducting such fine-grained analyses would require larger sample sizes within each occupation and more granular veracity assessments that can capture content quality variations within professional domains.

In conclusion, we have presented an online influence model that could be used to; provide

insights into downstream modeling such as opinion dynamics, understand the spread of misinformation, or identify influential individuals in political/social campaigns. Our findings reveal that *Media*, *Executives*, *Entertainers*, and the *Military* are the most influential; however the latter contain a large proportion of misinformation spreaders. In contrast, pandemic experts (i.e. *Life Scientists*) have only limited influence. These results highlight the critical need for the amplification of our experts' content.

Chapter 4

Ideology Detection

Profiling is valuable to researchers as it enables systematic analysis of complex human behaviors and social dynamics. Mining large sets of data uncovers hidden patterns, relationships, and trends, allowing researchers to gain deeper insights into group behaviors, belief systems, and identity formation. This further facilitates the development of targeted interventions, informs policy decisions, and enhances the capacity to predict and respond to social phenomena effectively. Scalable profiling ensures findings are applicable across varied contexts and makes research more robust and impactful.

In particular, profiling is crucial for analyzing ideological behavior in online environments, helping to counter violent extremism and understand broader opinion dynamics. However, practitioners face significant obstacles, including the high cost and labor-intensive nature of gathering gold-standard training data, and reliance on ideological signals (e.g., hashtags) with uncertain annotation requirements, limited context transferability, and potential biases.

This chapter addresses these challenges by presenting a comprehensive framework for large-scale, real-time profiling of left, right, and extreme ideologies in online settings. Unlike traditional methods reliant on small, controlled samples, this study emphasizes population-level profiling, enabling scalable, deterministic, and repeatable assessments of complex sociographics. By categorizing ideological signals according to labor and context transferability, the framework offers practical guidelines for constructing effective profiling

pipelines. Our evaluation quantifies biases in various signals and produces one pipeline that surpasses current state-of-the-art methods, achieving a 0.95 AUC ROC across five datasets with over 1.12 million users.

The research also examines whether established psychosocial theories, traditionally applied in offline settings, hold in digital contexts. It evaluates hypotheses around how ideologies are expressed through language, particularly in themes of morality, grievance, nationalism, and dichotomous thinking. Findings indicate that right-wing ideologies often exhibit patterns of vice-moral language, grievance-laden rhetoric, black-and-white thinking, and associations with national symbols, such as flags. These insights not only validate offline theories in an online setting but also help refine the methodologies for detecting and understanding ideological behavior at scale.

Profiling at the population level offers significant advantages over traditional methods, as it enables broader insights into sociographic patterns. The integration of computational techniques requires less labor and resources. The deterministic and repeatable nature of the methods developed here ensures that the results are reliable across various contexts, offering a more consistent understanding of ideological behaviors.

However, profiling practices also raise important ethical considerations. This study addresses these concerns by emphasizing transparency in methodological design, advocating for ethical guidelines in the application of profiling technologies, and encouraging responsible use by practitioners. Ensuring that profiling does not lead to harmful stereotyping or invasive surveillance is critical for maintaining ethical standards in digital research.

This interdisciplinary synthesis offers new perspectives and scalable solutions for analyzing complex sociographic patterns, ultimately aiming to enhance the understanding of human behavior in digital environments. Through ethical, scalable, and repeatable practices, the following chapter seeks to empower practitioners to navigate the complexities of online ideological profiling responsibly, fostering a safer and more comprehensively understood digital landscape.

Author Declaration

The following chapter contains content from the following publication.

Rohit Ram, Emma Thomas, David Kernot, and Marian-Andrei Rizoiu. Detecting extreme ideologies in shifting landscapes: an automatic & context-agnostic approach. *arXiv preprint arXiv:2208.04097*, 2022

Author Contributions: R.R. led the research for this study, managed the data processing and collection, and conducted the experiments and analysis. M.A.R. provided supervision through all stages of the study. E.T. provided direction and validation of psychological approaches. D.K. provided direction for the application of methods. R.R. and M.A.R. collaboratively developed the model and experimental design. R.R., M.A.R. interpreted the results and contributed to manuscript writing and editing. E.T. and D.K. contributed to manuscript editing.

Production Note:
Signature removed prior to publication.

Rohit Ram

Production Note:
Signature removed prior to publication.

Emma Thomas

Production Note:
Signature removed prior to publication.

David Kernot

Production Note:
Signature removed prior to publication.

Marian-Andrei Rizoiu

4.1 Introduction

Ideologies are the collection of beliefs and opinions about the ideal arrangement of society [128]. Tracking extreme ideologies is particularly important in detecting extreme voices that can spread harmful and false information, leading to dangerous and even deadly outcomes. Ideology is canonically (and inexactly) projected onto a left-right spectrum, where the left is associated with equality and reform, and the right is associated with authority and tradition. There has been a recent increase in fringe and extreme-leaning worldviews, including the far-right – a prominent archetype of extreme ideologies associated with ultranationalism and opposition to multiculturalism. Worryingly, this has increased Ideologically Motivated Violent Extremism (IMVE) [129] – a term coined to encompass religious, political and nationalist extremism. Far-right ideologies are disproportionately associated with violent extremism compared to far-left movements, making them a priority for detection and monitoring efforts. Ideology detection is a lead indicator for IMVE, fortifying individual and collective security. It facilitates understanding these ideological groups' values and beliefs, which helps design interventions, build political bridges and tackle radicalization.

Radicalization can occur in a matter of weeks [130, 131], both offline (face-to-face) and online (forums and social media platforms). To combat this, practitioners – such as law enforcement and national security agencies – need practical, real-time ideology detection tools that minimize human effort and can be applied across diverse contexts. Despite the significant existing literature, practical and effective detection guidelines remain scarce. This study establishes a framework for ideology detection pipelines, examining diverse constructions and demonstrating practical implementations using off-the-shelf components. Our first aim is to identify practical pipelines that reduce annotation efforts while maintaining transferability across different contexts. Our second aim is to validate insights into the psychosocial asymmetries of ideologies. We leverage five large datasets, totaling 1.12 million profiles, and test several hypotheses from the psychosocial literature at scale, mainly developed in offline laboratory setups. We answer two specific research questions.

The first question involves *ideological proxies* – measurable user behavior signals correlating with "true" ideology – that minimize annotation labor and are transferable across contexts.

We define a *context* as the tuple (topic, time, geography, platform). Prior works rely on various sources of ideological knowledge, including manually labeling users, labeling ideological proxies, and detecting group behavior differences. However, these approaches have limitations: the former two require extensive expert labeling – an expensive resource – and often fail to transfer across contexts. The latter often lacks robustness. Of the three, ideological proxies are the most common approach to reducing labor; however, they vary in reusability. See Section 4.2 for a complete discussion.

Furthermore, few users partake in direct ideological activity, and some actively avoid disclosure. Consequently, many proxies reveal only the vocal subset of users, biasing downstream analyses [132, 133]. Despite this, prior works commonly use proxies as ground truth [134–136] without quantifying the bias this entices. Our first research question is: **Which ideological proxies minimize annotation labor, maximize context transferability, and reduce bias?**

The second question involves the psychosocial asymmetries of ideologies. Understanding the values and beliefs of ideological groups is instrumental in modeling their polarization and user radicalization. Ideological asymmetry studies are abundant in relevant disciplines [137, 138]; often shown via offline surveys. For example, moral values delineate left-from-right ideologies [139]; and grievance/grudge language delineate moderate-from-extreme ideologies [140, 141]. Many of these hypotheses were developed with offline populations, and there is limited evidence for online populations. We know online and offline populations differ demographically [142], but we do not understand their psychosocial differences. We ask **can we build psychosocial profiles of ideological groups and employ them to evaluate hypotheses related to the psychosocial traits of these groups?**

Solution Outline. First, we evaluate ideological proxies. We make the widely adopted assumption that homophily – the tendency of similar individuals to associate – propagates ideology. We build a framework to construct ideology detection pipelines. We qualitatively evaluate proxies, by their minimization of labelling efforts and how readily they transfer to new contexts, and quantitatively evaluate proxies on their prediction of human-annotated ground truth. We show that a pipeline constructed through a proxy based on media consumption and a lens based on text, is both qualitatively advantageous and quantitatively

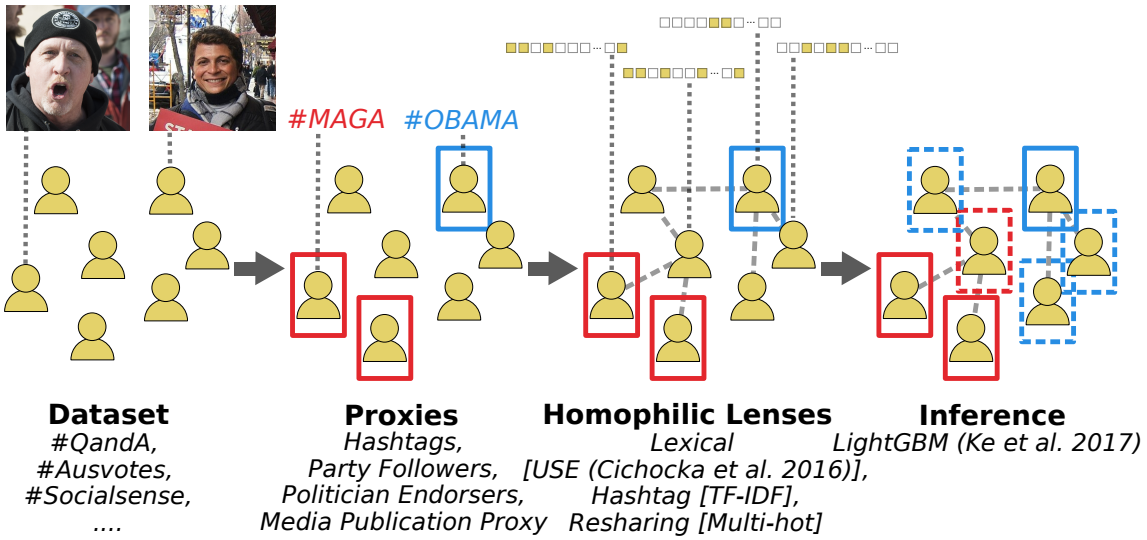


FIGURE 4.1: **The schema conceptualizes the four components of the pipeline**; (1) the datasets contain information about users (two examples are shown), (2) the ideological proxies assign labels on some of the users based on external information (here #MAGA indicates right-leaning, while #OBAMA indicates left-leaning), (3) the homophilic lenses build numeric descriptions for user and a way to measure their similarity, and (4) inference architecture predicts the likely labels of all other users in the dataset.

performant. Second, we use a pipeline to test hypotheses of the psychosocial asymmetries of ideology at scale.

We address the first question in Section 4.4. We introduce our pipeline framework, consisting of four components: dataset, ideological proxy, homophilic lens, and inference architecture. We use five *social media datasets*, collected from three platforms (Parler, Facebook and Twitter), containing 1.12 million users, and spanning social domains such as TV shows, elections, climate change, antivaccination and the January 6th US Capitol Insurrection. We frame the problem as user classification: the left-right detection as ternary classification (left, right, and neutral), and the far-right detection as binary classification. We limit our scope to Anglo-centric, English-speaking contexts with a dominant uniaxial political spectrum[1]. We explore four left-right and two far-right *ideology proxies*, leveraging behaviors such as posting politically-charged hashtags, following political parties, endorsing politicians, and sharing media websites. Our focus on far-right rather than far-left proxies reflects both the greater prevalence of far-right movements in contemporary discourse and the practical

[1]This is not to discount the need for ideology detection in other regions, like the Global South. Nor to suggest that a uniaxial spectrum is sufficient to encompass the complex politics of global communities. See Section 4.8 for a discussion.

challenge of identifying reliable far-left proxies in the Australian political landscape, where such movements have less distinct organizational structures. We build three *homophilic lenses* based on language, endorsements, and topics. We use the ideology proxies and homophily lenses to build ideology pipelines with an *off-the-shelf classifier*. See Fig. 4.1.

In Section 4.6 we evaluate the performance of ideology detection pipelines. We construct gold-standard benchmarks for left-right and far-right classification via human annotation and use them to evaluate bias introduced by ideological proxies. Furthermore, we assess various combinations of ideology proxy and homophilic lens to observe interaction effects and find the best performing combination. Finally, we compare this pipeline to state-of-the-art baselines: TIMME [136], UUS [134], and UUS+ [143] and achieve the best area-under-the-receiver-operating-curve (AUC ROC) of 0.95, an improvement of 6.7% over the next best, TIMME.

We address the second question in Section 4.7. We evaluate psychosocial hypotheses relating to morality, grievance, nationalism, and dichotomous thinking. For *morality*, we evaluate the seminal Moral Foundations Theory [139] hypotheses, operationalized via FrameAxis [144] (see Section 4.3). In its two subsets of hypotheses, individualizing and binding; we find relatively more support for the prior. However, only 46% of hypotheses are supported overall. As alternative hypotheses, we find that the right uses the language of vice more than the left, with statistical significance. For *grievance*, following literature that theorized that grudges and grievances are requirements for radicalization [140], we find large-scale proof that the far-right uses grievance language more than moderates. We operationalize via the Grievance Dictionary threat-assessment tool [141] (see Section 4.3). For *nationalism*, we show that the right exhibit nationalism via flag emojis, adding validity to our inferred grouping. Finally, for *dichotomous thinking*, we apply a dictionary-based approach, showing that the far-right, followed by the right, exhibits more black-and-white thinking (supporting prior work).

The main contributions of this work are as follows:

- An ideology detection pipeline applicable in large-scale online setups, that minimizes labor requirements and improves transferability to multiple contexts.

- The most comprehensive discussion and analysis of ideological proxies (to our knowledge); quantifying their bias independently and jointly with homophilic lenses. One construction outperforms state-of-the-art methods.
- Evaluation of psychosocial hypotheses concerning ideologies in a large-scale online setting.

Glossary. For readability, we collocate and define terms here. *Ideological Proxy*: measurable user behaviors correlating with ideology (e.g., emitting hashtags, following ideological users, sharing ideological media, etc.). *Homophilic Lens*: a representation of users highlighting specific behaviors under the homophilic assumption (users who act similarly are likely to share similar ideological beliefs). *Inference Architecture*: a classifier used to infer user connections in a latent space.

4.2 Related Work

Two corpora relate to our study; ideology detection and psychosocial asymmetries. Our primary concern, for the prior, is pipeline delineation criteria and, for the latter, is evidence bases for hypotheses.

4.2.1 Ideology Detection Delineation

Ideology detection is becoming popular and relevant for researchers and practitioners across the computer, social, and political sciences. We delineate prior work by population scope, homophilic lenses, and ideological proxies.

Population Scope describes *who the technique applies to?* Many works limit their scope to a population subset: legislators, elites [136], the politically active [134], or everyone [143]. Subsets offer clearer ground truth and easier inference, but lack representativeness of the population, leading to biases when applied broadly [132, 133] and constraining the representativeness of correlational analyses [145]. This work applies to all users, providing representative downstream analysis.

Homophilic Lenses describe *which features are utilized to infer ideology?* Underlying detection is the homophilic assumption – people who act similarly are likely to share similar ideological beliefs. Prior works operationalize this via several lenses: content (including metadata, images [146], and text [147]), network (such as followership and resharing [136]), or a combination [148]. In political science, the modus operandi is Ideal Point Estimation [149] using homophily via legislator voting behavior. Ideal Point Estimation techniques are largely unsupervised and rely on distinct behavioral patterns but are used in most political science ideology measurement work [150, 151]. In particular, Barberá [152] utilize the *following of politicians* on Twitter to estimate user ideal points, and their work is employed in correlation analysis [153]. Given the host of behaviors that portray ideology, novel lenses continue to emerge, including media sharing [154, 155], and community participation [156]. Prior works commonly engineer salient lenses and seek their optimal combination [134, 157]; however the complexity of data context, inference architecture, and ideological proxy choices often make the conclusions unclear. For example, Darwish et al. [134] recommend a retweet lens, while Aldayel and Magdy [157] recommend a network and lexical lens combination. The ideological salience of lenses and their combinations is not our work's focus. We implement three homophilic lenses previously shown to be ideologically salient, to limit interaction effects with ideological proxies concerns.

Ideological Proxy describes *what is the source of ideological knowledge?* Prior work utilizes three paradigms for detection: supervised, unsupervised, and weak supervision. Each employs distinct ideological knowledge sources– dubbed *ideological proxies*. In this study, we focus on both the proxies' performance and their expert annotation labor requirement when used across multiple contexts. We delineate proxies by (1) the extent to which they require expert annotation, (2) are transferable to different contexts, and (3) how well they represent *true* ideology. These criteria describe how well proxies generalize to arbitrary datasets and how much manual effort is required for switching contexts.

Direct user annotation for supervised learning [136, 158] is simple, the most representative, and accommodates fine-grained distinctions between ideologies [159]; however, it is also the most restrictive, requiring laborious expert evaluation of users, across every new context. Conversely, unsupervised approaches need little annotation and, in theory, are applicable in any context. Some apply embedding and clustering techniques [134, 135, 143]. Others

utilize matrix factorization to jointly learn representations of users and their behaviors [160, 161]. These methods are not robust in practice, require highly polarized contexts, fail on homogenous user sets, and depend heavily on lenses. Furthermore, they require expert knowledge in post-analysis (e.g., identifying clusters) [134], and clusters do not always align with ideology. Weak supervision trades-off between the high labor of supervised and the instability of unsupervised methods. It employs an ideological proxy, a user behavior strongly correlated with ideology. Prior work utilizes a range of ideological proxies, including; politically-charged hashtags [1], political party relationships [155], politician relationships, community participation [160], and news media sharing [122, 162, 163]. We assess proxies' labor minimization and context transferability qualitatively in Section 4.4.2 and assess their representativity quantitatively in Section 4.6.2.

Related Ideology Detection. Jiang et al. [162] use text and retweet features, and a combined media-hashtag proxy which they validate. However, they limit scope to active users who retweet and require hashtag proxy labeling.

4.2.2 Psychosocial Profiling of Ideological Groups.

Many social science works detail the nuanced profiles of fine-grained ideological groups and highlight the asymmetries between ideologies [137, 138, 164, 165], often requiring painstaking surveys and ethnographic inquiry. In this work, we supply large-scale online evidence for hypotheses surrounding psychosocial asymmetries of ideologies, relating to morality, grievance, nationalism, and dichotomous thinking.

Morality. Moral Foundations Theory [139] is an explanation of moral values variations between liberals and conservatives (see Section 4.3). Despite its support in psychological survey data [139], and a handful of online studies [166, 167], online social data inconsistently supports this explanation [145, 168].

Grievance and Grudge are linked to extreme ideologies in psychological theory;Van der Vegt et al. [141] link grievance to extremism, and Stankov [140] link grudge to the far-right.

Nationalism is definitionally associated with right-wing politicians. Prior work has shown that flags are associated with nationalism [169], emojis hold identity and semantics information [170], and that flag emojis are significant in right-leaning political communication [171]. However, this research is limited to politicians in a US context.

Dichotomous Thinking is a cognitive distortion in people with internalizing disorders, is tied to language [172], and is associated with the right [173].

Our concern is evaluating hypotheses in large-scale online populations in various contexts. Accordingly, we limit our scope to automated techniques using online metadata alone. Prior work, online, analyses left-right [166] or extremist asymmetries, but rarely both [145]. Additionally, they analyze small and non-representative samples. This work analyzes left, right, and far-right ideologies in several large-scale online contexts.

4.3 Preliminaries

Our study relies on several techniques from prior work.

Encoding Techniques are employed to implement homophilic lenses; the Universal Sentence Encoder (USE) [174] for our lexical lens (a mature, off-the-shelf, transformer-based model), Term-Frequency Inverse Document Frequency (TF-IDF) for our hashtag lens, and a multi-hot encoding for our resharing lens. We utilize simple encoding techniques, as they are not our work's main focus.

Inference Architecture Implementation. We use LightGBM [175] – an efficient tree-based classifier – and FlaML [176], a system that infers hyperparameters based on dataset characteristics in pipelines.

Moral Foundations Theory (MFT) [139] explains variations in moral reasoning through five modular foundations.It espouses that liberals express individualizing foundations (care and fairness) while conservatives express binding foundations (loyalty, authority, and sanctity) relatively more. We characterize users' language with FrameAxis [144], a dictionary embedding technique, to identify a user's value for each foundation. It supplies measures, *bias* and *intensity*. Importantly, dictionary-embeddings are generally a

refinement over dictionaries alone, particularly for smaller documents, however they do not capture the complexities of human language. For example, such approaches will not handle negations (for example, "I do not care") and do not consider the context around word usage. Large-language model (LLM) approaches may improve these deficits; however, LLMs introduce their own complexities [177] and the dictionary/embeddings approaches are better validated.

Grievance Dictionary [141] is curated for threat assessment, including categories such as fixation, violence, and paranoia. It is validated on social media data, and provides features for distinguishing extremist texts.

State-Of-The-Art Baselines. In Section 4.6, we compare our approach to three state-of-the-art detection approaches. *UUS* [134] encodes the k most active users, applies dimensionality reduction, clusters these embeddings, and assigns clusters stances via expert annotation. The authors tune parameters including; k, features (based on retweets, retweeted accounts, and hashtags), dimensionality reduction schemes, and clustering schemes. They recommend encoding 1000 users via retweets, then applying UMAP and Mean-Shift. *UUS+* [143] extends *UUS* by finetuning BERT with *UUS*-labels; applying it to remaining users. Finally, *TIMME* [136] is a supervised multi-task multi-relation deep graph method using five user relationships to embed and classify users.

4.4 Ideology Framework and Implementation

In this section, we describe our ideology pipeline framework in two parts; Section 4.4.1 partitions pipelines into four components and Sections 4.4.2 and 4.4.3 provides component implementation details.

4.4.1 Pipeline Constructions Framework

In this section, we abstract four components of ideology detection, shown in Fig. 4.1: the dataset, ideological proxy, homophilic lens, and inference architecture.

The Dataset is a set of unlabelled users and their activity metadata within a context. It has an underappreciated effect on observed pipeline performance. Section 4.5 discusses classification *difficulty* and introduces our evaluation datasets.

The Ideological Proxy infuses ideological knowledge via weak supervision. A user subset is labeled (left, right, or far-right) via ideology-correlated behaviors, such as sharing hashtags, following political parties, endorsing politicians, or sharing news media. See Section 4.4.2 for details.

The Homophilic Lens characterises ideologically salient user similarity. Section 4.4.3 describes three homophilic lenses: the lexical lens, the hashtag, and the resharing lens.

The Inference Architecture propagates labels from a user subset to the remaining unlabelled users. We train a classifier on the ideology-proxy-labeled users represented via homophilic lenses. We use LightGBM with FlaML as our classifier[2]. The remainder of this section enumerates the ideological proxies (Section 4.4.2) and homophilic lenses (Section 4.4.3) evaluated, and their implementations.

4.4.2 Implementating Ideological Proxies

Here, we qualitatively compare proxies and describe the implementations of the proxies evaluated in our study.

Proxy Qualitative Comparison. We conduct an assessment of proxies, based on their utility for practitioners. Based on our reading of the thematic review presented in Section 4.2, we qualitatively build three criteria to assess each proxy. The criteria are designed to partially order proxies as a guide to practitioners. Therefore, we apply a four-star rating (one star is lower) for each criterion, as shown in Section 4.4.2.

The first criterion we construct is *labor minimization* (AL) defined as the extent to which expert labor is required to generate the proxy. Proxies which require human experts to perform the entire contruction will score one star, whereas an approach with no human

[2]The hyperparameter `n_estimators` is inferred for the far-right detection due to the sparsity of labeled users; it is fixed to 200 for left-right detection to prevent overfitting. We set the `is_unbalance` flag due to label imbalance.

Proxy	AL	CT	AV
Hashtags	*	*	*
Community Participation	**	**	*
Politician Endorsers	**	**	**
Party Followers	***	**	***
Media	***	****	****

Table 4.1: **Ideology Proxy Qualitative Comparison** for application by practitioners based on three-part criteria; annotation labor minimization (AL), context transferability (CT), and Availability (AV). Criteria are rated out of four-stars.

intervention scores four stars. The second criterion is *context transferability*[3] (CT), defined as the number and diversity of contexts in which a proxy can be applied. If a proxy is only available in a given context it will score one star, whereas if the proxy is available with no restrictions, it will score four stars. The third criterion is *availability to practitioners*[4] (AV), defined as the extent to which a proxy or its ingredients are openly available, either for ideology detection or independent tasks.

Hashtags shared is commonly used as a proxy, but requires domain knowledge and is time-consuming to generate (one star on AL), and generally requires reannotation for every dataset (* for CT). Furthermore, not all social media platforms use hashtags therefore it has a low availability (* for AV). Community Participation uses user activity in ideological communities (e.g., subreddit posting). The communities tend to be fewer and more persistent (** on AL) and there is some detectable overlap of the communities between platforms (** on CT). However, they are unavailable on some platforms (e.g., Twitter/X) and are inconsistent across countries (* on AV). Furthermore, it requires experts for annotation, and datasets linking communities to ideologies are few. Party Followers and Politician Endorsers leverage databases of political parties and politicians with their online profiles, which are intermittently available (** on AV). Such databases are usually country- and period-specific – political parties emerge, change, and become relegated in time. The advantage of these proxies is their stability and non-ambivalent nature during the studied context (** on CT). Furthermore, databases do not encode all ideologically relevant information, such as the lean of parties or specific politicians, requiring an expert

[3]Note that context transferability has a multiplier effect on annotation labor since a failure to transfer requires reannotation.

[4]We do not discount prior work labor. However, we recognize that availability differs, independently of ideology tasks, and proxies' maintenance should be considered in practitioner guidelines.

instead (** and *** on AL, respectively). MEDIA proxies utilize users sharing news media, which often have known political slants. They leverage available and well-maintained data on media slants (*** on AL), which have intrinsic value in communication studies, the news ideology detection task, and general consumer value. There is strong evidence linking news readership [178, 179] and sharing [180] to ideology. Media slants are fairly consistent across time, media-sharing behaviors occur on most platforms (**** on AV), and media tend to be ideologically consistent across topics (**** on CT). There are limitations to the media proxy (see Section 4.8), but it outperforms its alternatives in terms of annotation labor, context transferability, and general availability.

Left-Right ideological proxies. We build four proxies.

Hashtags proxy requires experts to code hashtags. We qualitatively inspect the $1,000$ most common hashtags in our datasets and label their political lean; -1 if left-leaning, 0 if non-partisan, and 1 if right-leaning. We quantify a user's political lean as the mean of the labeled hashtags they emit and their ideology label as the sign of this lean.

Party Followers proxy requires collecting the followers of the major political parties' online accounts for each target country (i.e., Australia and USA). We code the political parties by their ideology. The users in the dataset who follow a single party receive the party's ideology label.

Politician Endorsers proxy requires a dataset of politicians, their political affiliations, and social media handles. We use the Twitter Parliamentarian Database [181]. We code the politicians using their party's ideology (where independents are excluded). Note that independents' exclusion reduces the proxy representativity, but this is preferable to manually labeling all independents. We label users who retweet politicians using the majority vote of the politicians' ideologies.

Left-Right MPP (Media Publication Proxy) requires a dataset of media websites with their political slants. We utilize an extensive survey [182, 183] of news consumption behavior within English-speaking countries (Australia, New Zealand, UK and the USA), collected in 2020 and 2021 by Reuters. Participants indicated the news media they read and self-reported their political leaning ranging from -3 (extreme left) to 3 (extreme right).

We compute a publication's slant as the weighted mean political lean of the participants who consume that publication, where each participant is weighted by the inverse number of publications they consume. Since countries' perspectives on what constitutes left- and right-leaning differ, we calibrate scores across countries with the AllSides Media Bias Ratings [184]. We encode the ratings' five-point scale onto a numerical scale from -1 to 1. We align each country's scores, minimizing the sum of squared differences between a country's scores and AllSides scores for overlapping publications. Finally, we generate slant scores for each publication as the average slant over all countries and years. We associate publications (and their slants) with their website domains, averaging where a domain is shared. We present the media organizations and their constructed slant scores in online appendix [119]. We compute a user's political lean as the average lean of the media domains they share and their ideology label as the sign of this lean.

Far-Right ideology proxies. We build two proxies.

Far-Right MPP is constructed from the media slant scores of mainstream media built for Left-Right MPP. Next, we label users 'far-right' if their political lean exceeds 0.5 or as 'moderate' otherwise.

MBFC MPP is constructed from the Media Bias Fact Check [185] dataset, including both media slant and veracity, and containing conspiratorial and fake news sources. We label users sharing the right-most media category as 'far-right'.

4.4.3 Homophilic Lenses

Homophily is the tendency of similar users to be similar [186] and is commonly assumed in ideology detection. A *homophilic lens* is a user embedding that encodes ideologically relevant information. Here we convert content about user behavior into numerical vectors. This section details three lenses.

Lexical Lens (USE). Language is a strong indicator of one's political ideology [187]; since a sociolect is formed through associations with others.

Hashtag Lens (HT). Hashtags signal users' interests and the discussion topics they participate in [188].

Resharing Lens (RT). Resharing is a signal of endorsement [96]. We assume users endorsing the same people likely share similar ideologies [189].

Implementation. For the *lexical lens*, we preprocess text, to prevent potential data leaks, by removing URLs, hashtags, and mentions. We concatenate each user's tweets and encode them as 512 dimensional vectors via the universal sentence encoder (USE) [174]. The encoder choice is arbitrary and based on its prior user in literature for social media-originating text [135]. For the *hashtag lens*, we use the Term-Frequency Inverse Document Frequency (TF-IDF) of users (i.e., documents) via hashtags (i.e., words) they use if used at least 10 times. TF-IDF is a refinement over the bag-of-words model that weights terms used by their occurrence within a corpus, providing a simple but salient vector representation. Finally, for the *resharing lens*, we generate a multi-hot encoding for users based on the 1000 most reshared posts. We represent a user u_i as $h_i \in \mathbb{R}^{1000}$, where $h_i[j] = 1$ if u_i reshares the jth most reshared post ($h_i[j] = 0$ otherwise). In summary, there are three representations of users; lexical $\mathbb{R}^{512}$, hashtag $\mathbb{R}^{|\#\text{hashtags}|}$, reshare $\mathbb{R}^{1000}$.

4.5 Contexts, Datasets, and Ideology Labels

This section introduces datasets and their contexts. Section 4.5.1 describes the five datasets, and Section 4.5.2 shows how we qualitatively construct ideology ground truth, used to evaluate proxies and pipelines' performance.

4.5.1 Contexts and Datasets

Section 4.5.1 summarizes the datasets; there are three Australian and two American datasets; one originates from Parler, another is a mixture of Facebook and Twitter, and the remainder are Twitter-based. In prior work datasets, ideology correlates with explicit user behavior (e.g., discussion topics); this simplifies detection but rarely holds in practice. Here, we use data where detection is difficult, as one would likely encounter in the wild. We

Dataset	#Users	#Posts	Country	Hopkins
#QandA	103,074	768,808	AUS	0.2624
#Ausvotes	273,874	5,033,982	AUS	0.2445
#Socialsense	49,442	358,292	AUS	0.2591
Riot	574,281	1,067,794	US	0.1490
Parler	120,048	603,820	US	0.3016

TABLE 4.2: **The datasets used in this work:** source, profiling, and country of origin (AUS and US refer to Australia and USA, respectively). The last column represents the Hopkins statistics [5] for the lexical lens.

quantify the *detection difficulty* using Hopkin's statistics [5] of the lexical lens, indicating the clustering tendency of data, ranging from 1 (highly clustered, easy detection) to 0 (uniformly distributed, difficult detection). Hopkin's statistic is common measure of clustering tendency, effectively characterizing the probability that embeddings are drawn from a uniform distribution. We assume that embeddings with a high clustering tendency are easier to classify. Note clusters do not necessarily align with classes, however they often do in real-world data; baselines, like UUS and UUS+, directly employ this axiom to infer labels (relying heavily on the underlying clustering tendency of the data). Quantifying detection difficulty of datasets is uncommon in literature and prior work often vary dataset difficulty be construction [190] or require class labels to infer it [191]. We employ Hopkin's Statistic as a simple quantification of difficulty (which is not the focus of our work). It is likely related to the decision boundary aspect of classification complexity [191]. Section 4.5.1 shows values $\in [0.14, 0.3]$ indicating no clustering tendency.

Briefly, the datasets are: *#QandA* [Twitter/X] surrounding a political panel show with audience questions; *#Ausvotes* [Twitter/X] surrounding the 2022 Australian Federal Election; *#Socialsense* [Twitter/X and Facebook] [10] surrounding the Australian Black Summer Bushfires; *Riot* [Twitter] [192] and *Parler* [Parler] [193] both surrounding the US capitol insurrection. See [119] for details.

4.5.2 Build a Ground Truth

We qualitatively annotate a subset of #QandA users to generate both a left-right and far-right ground truth.

Left-Right Ground Truth. Due to the imbalance and sparsity of some ideological classes[5], we employ the proposed pipeline to construct a candidate set of users for manual annotation. Platforms such as X/Twitter have been shown to lean-left, and the imbalance in datasets (such as Q+A which attracts a left-leaning audience) can be substantial. While it can be argued that using the pipeline to generate a ground truth to train future pipelines may skew the data selection, it has advantages over the alternatives. For example, (1) conducting a manual search through a random candidate set and generating a proportionately low-volume of right-leaning users is prohibitively expensive, and (2) employing a proxy directly as our ground truth (following the baselines we compare against) defeats the purpose of evaluating the proxies and introduces significant biases.

We generate the candidate set using the following four components; (1) we select each of the four proxies (Hashtags, Party Followers, Politician Endorsers, Left-Right MPP), (2) using labels derived from the selected proxy, we train the classifier to predict user labels (since even proxy do not necessarily produce sufficient volumes of right-leaning users), (3) we apply to proxy-trained classifier to the entire #QandA dataset (including those already labelled), (4) finally, we extract the 100 left- and 100 right-leaning users with the highest classifier confidence (estimated through the classifier sigmoid scores). We collect the pool of 800 users in one set, deduplicate, shuffle it, and remove users who are unavailable (either private or suspended). This results in 695 users; we sample 200 users, inspect their profiles and categorize them as left-leaning, right-leaning, far-right, or indeterminable.

Next, two experts manually labeled each profile. The experts both had extensive knowledge of the Australian political context, and were native English speakers. They were given examples of left, right, far-right, and indeterminable user profiles for context. They were instructed to use any signals of ideological-alignment they observed to make their assessments (see [119] for details). Finally, they were given links to each user profile and instructed to categorize them. They achieved moderate inter-annotator agreement i.e., Cohen's κ of 0.515 As a result, our left-right ground truth contains 103 left- and 74 right-leaning users.

[5]Predicted label counts show this imbalance [119].

Far-Right Ground Truth. Bailo et al. [122] snowball sample Australian far-right users, starting with a 'seed' user and recovering 'lists' (a Twitter feature documenting similar users) they belong to. They intersected the sample with their dataset, manually validated their far-right status, crawled this validated set's followers, and manually coded these too. They obtained $1,496$ users, of which 686 are in #QandA, and serve as our far-right ground truth.

4.6 Proxy Bias, Baselines, and Validation

In this section, we first quantify proxy bias (i.e., representativity) and homophilic lens interaction effects, by enumerating all pipeline constructions, in Section 4.6.1. Next, we present a pipeline construction that outperforms three state-of-the-art methods in Section 4.6.2. Finally, we evaluate transfer learning across contexts, illustrating 'in-context' training superiority, and test cross-proxy performance in Section 4.6.3.

To avoid confusion, Section 4.6.1 employs both ground-truth and Section 4.6.2 uses the left-right ground-truth constructed in Section 4.5.2 for the #QandA dataset. Section 4.6.3 does not utilize the constructed ground truth. In its first segment it trains on labels derived from one proxy and tests on labels derived from another, with fixed dataset #QandA. In its second segment it trains on users from one dataset and tests on users from another, with fixed proxy Left-Right MPP.

4.6.1 Quantifying Proxy Bias

Here we jointly assess ideological proxy and homophilic lens combinations and their performance against our ground truths, to infer proxy representativity. The top section of Section 4.6 shows all combinations. The columns represent proxies, and the rows show the seven possible concatenations of our lens implementations. We use the respective ground truth for validation and testing in a $50\% : 50\%$ split, employing the validation set for threshold calibration (for converting continuous scores to discrete predictions), and removing neutral ideologies from training, as they do not appear in testing. Cells show AUC ROC scores for pipelines trained with respective proxy and lens combinations. A

	Left-Right				Far-right	
	Hashtags	Party Follow.	Pol. Endors.	L.R. MPP	F.R. MPP	MBFC MPP
USE	0.881	0.868	0.788	0.946	0.691	0.773
HT	0.873	0.876	0.812	0.849	0.559	0.633
RT	0.840	0.844	0.752	0.879	0.538	0.668
USE+HT	0.949	0.879	<u>0.870</u>	0.939	<u>0.715</u>	<u>0.785</u>
USE+RT	0.880	0.821	0.785	<u>0.953</u>	0.666	0.762
HT+RT	0.904	<u>0.914</u>	0.799	0.937	0.570	0.632
all	<u>0.950</u>	0.875	0.854	0.929	0.713	0.785
Prec.	0.889	0.873	0.797	**0.892**	0.516	**0.530**
Recall	0.857	0.820	0.794	**0.902**	0.540	**0.557**
F1	0.855	0.821	0.766	**0.893**	0.636	**0.720**

TABLE 4.3: **Determine the optimal proxy and lens combination.** *(top)* AUC ROC for each combination of lenses (rows) and proxy (columns). The underlines show the best lens for a given proxy. *(bottom)* The precision, recall and macro-F1 for each proxy averaged over all lens combinations. The bold show the best-performing proxy.

higher AUC ROC score is better with a maximum score of 1 and a random baseline of 0.5. The bottom section of Section 4.6 shows the precision, recall and F1, averaged over all lens combinations. The purpose is to quantify how well proxies represent 'true' ideology, approximated via our ground truth.

Results. There are two main conclusions. First, Section 4.6 (bottom) shows that MPP consistently outperforms other proxies for left-right detection. In order of representativity, we have LEFT-RIGHT MPP, HASHTAGS, PARTY FOLLOWERS, and POLITICIAN ENDORSERS. The MBFC MPP is the most performant for far-right ideology. This is significant, as we have shown that media-based proxies are both qualitatively advantageous and optimal for representativity; providing clear guidelines for practitioners. Second, Section 4.6 (top) shows that no homophilic lens dominates all others and the best-performing lens combination changes for each proxy. This may explain unclear conclusions within the literature, where lens optimization is performed in isolation of other pipeline components (e.g., proxies). Despite the lack of a dominating lens, we observe that pipelines containing the lexical lens generally outperform their peers, and USE by itself (first row) has competitive performances. In addition, USE is the only platform-independent lens.

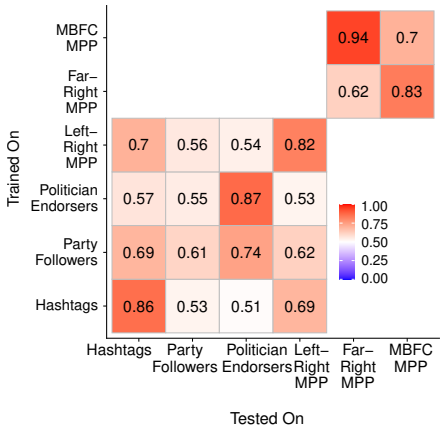


FIGURE 4.2: **Self- and cross-proxy generalization.** The AUC ROC of ideology detection on #QandA when trained on one proxy (y-axis) and tested on another (x-axis) for left-right far-right proxies.

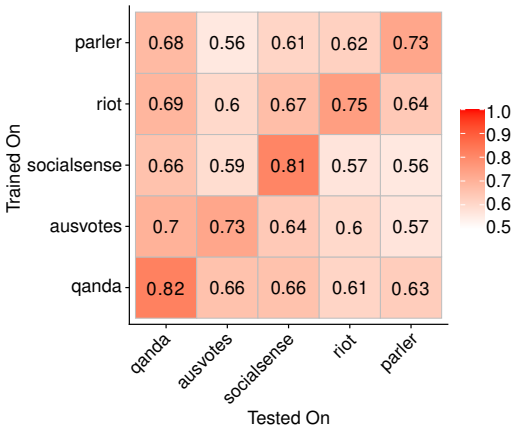


FIGURE 4.3: **Context generalization.** AUC ROC of LEFT-RIGHT MPP trained on one dataset (y-axis) and tested on another (x-axis).

Method	*UUS*	*UUS+*	*TIMME*	Ours
Macro-F1	$0.60_{\pm 0.23}$	$0.61_{\pm\ 0.26}$	0.88	0.92
AUC ROC	–	$0.76_{\pm\ 0.15}$	0.89	0.95

TABLE 4.4: **Baselines.** Left-right classification performance of baselines vs. our pipeline on the ground truth. We report the mean and standard deviation over all setup combinations for *UUS* and *UUS+*. Note that *UUS* does not produce a score, only labels; therefore, AUC ROC cannot be computed for it.

4.6.2 Prediction Performance Against Baselines

Baselines. We evaluate a pipeline construction against three state-of-the-art stance detection techniques: *UUS* [134], *UUS+* [143], and *TIMME* [136] – detailed in Section 4.2. For *UUS*, the authors' recommended setup (UMAP+Mean-Shift, retweet features, and

1000 active users) does not produce any clusters on #QandA. To render UUS competitive, we enumerate the setups similarly to their work. We fix the dimensionality reduction to `UMAP` and clustering to `Mean-Shift` following their recommended setup. We use the default *scikit-learn* settings (*n_neighbors=15*, *min_dist=0.1*, *n_components=2*, *metric=cosine*), and do not enumerate different hyperparameters to (1) faithfully replicate their work, and (2) simulate the experience of a time-poor practitioner. Furthermore, we implement setups for every combination of features (retweets, retweeted accounts, and hashtags) and number of active users (500, 1000, and 5000). In addition, *UUS* only reports the most active users' labels; however, our ground truth users are not the most active. Instead, we use UMAP and Mean-Shift inference methods to acquire labels for these users. For *UUS+*, we use the same set of *UUS* setups. Following the authors, we utilize $\text{BERT}_{\text{base multilingual}}$, using the HuggingFace implementations with PyTorch. We fine-tune BERT by adding a fully-connected dense layer followed by a softmax output layer. We minimize the cross-entropy loss over the training data. As it is not specified by the authors, we choose to fine-tune for 10 epochs (a sufficient quantity for our data volume). Finally, for *TIMME*, we use all relations except the followership network, which is prohibitive to acquire.

Predicting Human-Annotated Ideology. We evaluate performance using the left-right ground truth (see Section 4.5.2) with a 5-fold cross-validation (where applicable). Our ground truth construction leverages multiple proxy types to ensure adequate representation across ideological groups. The baseline methods utilize subsets of these same proxy types: *UUS* and *UUS+* rely primarily on retweets and hashtags, while *TIMME* employs network relations. Our more balanced mix of proxies provides broader coverage than these narrower approaches. Additionally, our strong performance in within-proxy evaluations (see Section 4.6.3) demonstrates robust generalization across different proxy types, validating our approach beyond the specific ground truth sampling procedure. For this task, we use the pipeline constructed from the Left-Right MPP and the USE+RT homophilic lens (the best-performing combination from Section 4.6). Section 4.6.2 shows the F1-macro and AUC ROC scores for each technique. We make several observations. First, our approach consistently outperforms all baselines, with the next best being *TIMME*. Second, *UUS* and *UUS+* show low mean performance and high standard deviation. Most setups failed to cluster users and were removed before computing the mean and standard deviation.

Furthermore, the clusters required an expert for labeling. Our pipeline construction has practical advantages over these baselines and outperforms them.

4.6.3 Cross Proxy and Context Generalization

Cross Proxy Generalization. Here, we characterize the robustness of ideological proxies through their self- and cross-consistency. *Self-consistency* indicates how well the pipeline predictions trained with a given proxy align with the same proxy on a test set. We evaluate self-consistency using a 5-fold cross-validation. *Cross-consistency* indicates that two proxies capture similar ideological signals. We evaluate the directed cross-consistency of a source $\longrightarrow$ target proxy by deploying a pipeline with the source proxy to predict the ideology of every user in the #QandA dataset and testing against the ideology labels set by the target proxy. We report the performance over users whom the target proxy labels, and use a one-vs-one scheme to adjust to the multiclass setting. For a given proxy, we deploy the pipeline with the best lens combination as per Section 4.6.

Section 4.6.2 shows the AUC ROC performance for every pair of source $\longrightarrow$ target proxy for both left-right and far-right ideology detection. *The self-consistency* (main diagonal) is high for all left-right pipelines, except Party Followers. It is worth noting, Politician Endorsers has high self-consistency but a low prediction performance against the ground truth (see Section 4.6). This suggests that politician endorsement behavior is distinct from prototypical ideological behavior. Note, far-right proxies have relatively low self-consistency, perhaps due to the sparsity of far-right users.

Section 4.6.2 shows *cross-consistency* of left-right pipelines is relatively low, except for Left-Right MPP and Hashtags. This supports prior work [132, 133] arguing that different proxies confer diverse ideology prototypes. The Left-Right MPP and Hashtags proxies generalize well to each other and the ground truth (see Section 4.6), suggesting they accurately represent *true ideology*. Both far-right proxies generalize well on each other, but their performance on the ground truth is relatively weak. This indicates they represent similar behaviors not fully aligned with ideology.

In-Context Dominance. Researchers often implicitly suggest political signals from one context transfer to others. Here we demonstrate the importance of 'in-context' training by comparing performance when models are trained and tested within the same dataset versus across different datasets. Each dataset is typically associated with a distinct context (see Section 4.4.1). We evaluate transfer-learning across datasets by training a pipeline on one dataset and testing on another dataset. For this analysis, we use the LEFT-RIGHT MPP proxy to generate ideology labels across all datasets, serving as the ground truth for evaluation. This allows consistent comparison across datasets that lack human-annotated ground truth. Furthermore, we utilise the USE+RT lense for this pipeline. We define 'in-context dominance' as the superior performance achieved when training and testing occur within the same dataset, compared to cross-dataset transfer learning performance. Section 4.6.2 shows the 5-fold cross-validation AUC ROC performance of left-right ideology detection for every pair of datasets. Intuitively, models perform best when trained and tested on the same dataset (i.e., in-context). However, we observe a significant performance drop-off with transfer learning (off-diagonal). Despite this, we see relatively better transfer learning between contexts that share traits. Models trained in Australian contexts perform better when tested within the Australian context, and noticeably underperform when tested in US contexts. Moreover, a further reduction is observed when training or testing with the Parler dataset (i.e., a different social platform context). These observations indicate that signals of ideology differ between contexts. While transfer learning performs better in similar contexts, 'in-context' training is significantly more effective.

4.7 Psychosocial Analysis of Ideology Cohorts

In this section, we test four hypothesis sets for psychosocial asymmetries of ideologies, relating to morality, grievance, nationalism, and dichotomous thinking. This serves two purposes: an application case study for practitioners and to supply online evidence bases for conclusions of prior work. We use pipelines constructed from the MBFC MPP and LEFT-RIGHT MPP proxies, alongside the USE lens for its applicability across all datasets. The first pipeline labels users as 'far-right'. If users are not labeled 'far-right', the second pipeline assigns them as 'left', 'neutral', or 'right'. For most analysis below, we highlight

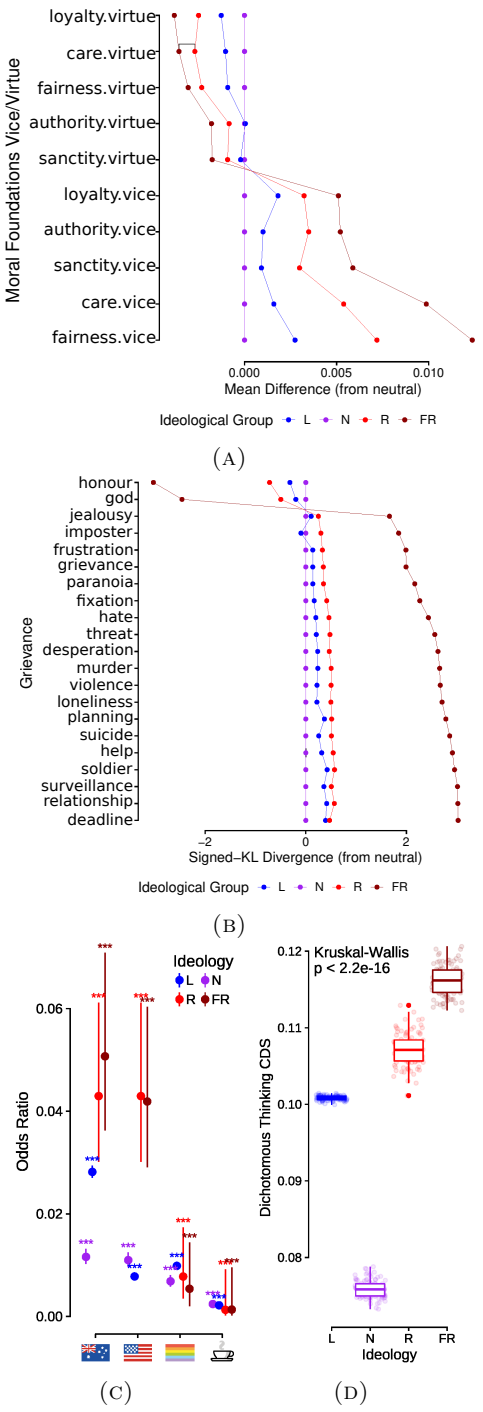


FIGURE 4.4: **(a)(b) Distribution of psychosocial properties** for ideological groups for #QandA and #Ausvotes, respectively. Line color represents ideological groups, and the y-axis shows psychosocial categories. *(a) Vices-Virtues.* The x-axis is the mean difference for each ideological group from neutral, for Moral Foundations vice and virtue categories. *(b) Grievances.* The x-axis is the signed-KL divergence of each group ideological group from neutral for grievance categories. (c) **Emoji Nationalism.** The odds (y-axis) of observing an emoji (x-axis) for a user given their ideological group (color), for #QandA. The odds are determined via logistic regression with no reference group. (d) **Dichotomous Thinking.** The bootstrapped prevalence distribution of dichotomous thinking CDS (y-axis) in tweets by users from ideological groups (x-axis), for #QandA.

results on a single dataset, however we produce the relevant plots for all datasets and label distributions in the supplementary material [119].

Testing Moral Foundations. We begin by evaluating MFT hypotheses. The five MFT hypotheses predict that left-leaning users exhibit higher usage of individualizing foundations (care/harm and fairness/cheating) while right-leaning users exhibit higher usage of binding foundations (loyalty/betrayal, authority/subversion, and sanctity/degradation). For each foundation, FrameAxis extracts two variables from user text: *bias* (indicating direction on the foundation axis, e.g., on care/harm axis, positive bias indicates care-oriented language, negative bias indicates harm-oriented language) and *intensity* (indicating magnitude of foundation usage regardless of direction). We test specific directional hypotheses: for individualizing foundations, we predict left-leaning users show higher positive bias and intensity than right-leaning users; for binding foundations, we predict right-leaning users show higher positive bias and intensity than left-leaning users. We use a Wilcox Rank Sign Test (95%), with Holm adjustment for family-wise error, to evaluate whether group differences are in the predicted direction. We test these hypotheses with both *bias* and *intensity* measures across "left vs. right" and "left vs. far-right" comparisons (i.e., 5 foundations $\times$ 2 measures $\times$ 2 comparisons = 20 hypotheses per dataset). Section 4.7 shows the number of statistically significant tests for each foundation across all datasets. Overall, 46% of hypotheses are supported, with individualizing foundations showing more consistent support than binding foundations. This inconsistency, seen in prior work [158, 168], suggests MFT applies differently online than it does offline.

Next, given the lack of support for MFT, we test an alternative hypothesis, that *right-leaning users, relative to left-leaning users, exhibit vice over virtue foundations.* For each moral foundation, we assign each user a virtue/vice score equal to their intensity, if their bias is positive/negative, respectively. This segregates the population into vice or virtue users. In Fig. 4.4a, we plot each foundation's mean vice and virtue scores for each ideological group in the #QandA dataset. We observe that a significant proportion of right-leaning users partake in the language of vice rather than virtue compared to left-leaning users. We apply the Wilcox Rank Sign Test (95%) between the means of ideological groups, for each category, and find all are significantly different[6]. We show that this is relatively consistent

[6]Except between the right and far-right in the *care-virtue* category, which is irrelevant to our conclusions

	#QandA	#Ausvotes	#Socialsense	Riot	Parler	Total
Fairness	2	2	2	2	2	$^{10}/_{20}$
Care	2	4	3	1	3	$^{13}/_{20}$
Loyalty	2	0	1	1	2	$^{6}/_{20}$
Authority	2	1	2	2	2	$^{9}/_{20}$
Sanctity	2	0	1	2	3	$^{8}/_{20}$
Total	$^{10}/_{20}$	$^{7}/_{20}$	$^{9}/_{20}$	$^{8}/_{20}$	$^{12}/_{20}$	$^{46}/_{100}$

TABLE 4.5: **Moral Foundations Hypotheses testing.** The number of times the MFT hypotheses tests are significant for each foundation (rows) and dataset (columns).

across all datasets in the online appendix [119], This provides a consistent moral asymmetry in the online context.

Testing Extremists' Association With Grievance. Early signals of extremism are of particular concern to national security and law enforcement practitioners. Prior work suggests that *extreme ideologies hold more grievance and grudge beliefs than moderates.* We use the Grievance dictionary [141] to quantify users' grievance and grudge language. In Fig. 4.4b, we plot the Kuller-Leibach divergence (signed by mean difference) between the distribution of each ideological group from the neutral group for each category with the #Ausvotes dataset. We apply the Kruskal-Wallis Test (95%) between ideological groups, for each category, and find all are significantly different. We observe that the far-right users differ significantly from the other ideologies in all categories and generally use more grievance language. Notably, in the #Ausvotes dataset, the far-right users use *honor* and *god* type language less than other groups. In the online appendix [119], we show that this hypothesis holds for most datasets. A takeaway for practitioners is that far-right language and threat assessment indicators overlap, suggesting a method to build effective public safety tools.

Testing Nationalism Via Emoji. Here we add online evidence that *the right-wing are associated with nationalism* via emojis. This hypothesis is widely accepted (and definitional), and supporting it validates our inferred ideologies.

Fig. 4.4c shows the odds of observing an emoji, given a user's ideological group in #QandA. Point ranges indicate the 95% confidence interval. The significance of the emoji in predicting ideological groups, via the Wald Test, is indicated with stars. We make several observations. First, 🇦🇺 is used more by ideological groups than neutral users. Second, the right (and far-right) use 🇦🇺 and 🇺🇸 significantly more than the left. Third, 🏳️‍🌈 is used marginally more by the left than other groups. Finally, we include ☕ as a control (showing no associations with any ideology). We conclude that nationalism, via national flags, is associated with our inferred right-leaning ideologies. The use of 🇺🇸 could be evidence of imported ideology from America to Australia. 🏳️‍🌈 is only marginally associated with the left. While emoji semantics can drift over time and vary across communities—as seen with symbols like the watermelon emoji being adopted for Palestinian solidarity—national flag emojis represent relatively stable semantic anchors, particularly within finite datasets where temporal and cultural context is well-defined. The stability of flag symbolism for nationalism makes them reliable indicators for ideology detection within the scope of this study.

Testing Dichotomous Thinking. Recent work suggests *the right-ideologies applying black-and-white thinking relatively more than left-wing ideologies.* Following [172], we match n-grams relating to cognitive distortions schema (CDS) in user tweets in #QandA. We measure the prevalence – the empirical probability of observing a CDS n-gram in a tweet given an ideological group. Additionally, we utilize 100 bootstrap samples (i.e., repeated sampling of tweets) to estimate the prevalence distributions. Fig. 4.4d shows that all non-neutral ideologies exhibit a significantly higher prevalence of dichotomous thinking, with right-leaning higher than left-leaning and far-right higher than right-leaning. We perform T-Tests (95%) to compare group means and find all differ significantly from each other. These findings support prior literature [173], and extend it by showing that the far-right might engender an even greater extent of dichotomous thinking. Other cognitive distortions' prevalences are summarized in the online appendix [119].

4.8 Conclusion

This work proposes a framework for ideology detection pipelines and quantifies biases introduced by ideological proxies. It tests hypotheses of the psychosocial asymmetries of

ideological groups, in the online space. We present an evaluation of ideological proxies; qualitatively, indicating proxies that minimize labor, are transferable across multiple contexts, and are available; and, quantitatively, indicating the representativity and robustness of proxies. We find the media proxy advantageous, and a pipeline constructed from it and the lexical lens to be optimal, outperforming state-of-the-art approaches. Such research is essential for furnishing practitioners with actionable guidelines for ideology detection and its practical applications.

Limitations. The media proxy has several limitations. Firstly, it relies on the availability of up-to-date media slant data. Publication slant can shift over time, and publication emergence, acquisition and closure can hold significance (particularly on ideological fringes). Secondly, some users share media to refute it. Thirdly, article slants may differ from publication slants. Finally, it will not produce a perfectly representative user subset, although media sharing ubiquity makes it relatively competitive. Furthermore, our conceptualization of ideology is simplistic, and some political systems are complex requiring complex ideological proxies (which are largely unavailable).

Future Work. We limit our scope to English-speaking Anglo-centric countries due to the expertise and language proficiency of the author team. However, the study could be applied broadly. Newman et al. [183] provides data annually for 46 diverse countries, including segments of the Global South. Our study could be extended to any other uniaxial political setting with little amendment.

The Ethics Of Ideology Detection and Broader Perspectives. This work introduces a powerful tool for inferring user ideology based on covert cues such as language patterns. We demonstrate our tool for detecting far-right ideologies; however, it could, in theory, be used by oppressive regimes to infer the true ideologies of their citizens and expose their opponents [194]. The Countering Violent Extremism (CVE) literature [195] explores the ethical concerns of developing tools that can be used for oppressive ends and proposes mitigation strategies. There are also privacy concerns, as one's ideology can be viewed as an intimate and private trait that our tool can expose. Additionally, we show how to use our pipeline to profile entire online populations based on their psychosocial characteristics. We argue that the pipeline predictions are not prescriptive; they should be treated as an

early warning system, requiring human expert investigation. We further note that we only use expert-inferred political affiliation as our ground truth, not private self-reported political indicator data.

Chapter 5

Birdspotter

Tooling is essential for advancing scientific and psychosocial inquiry by making complex analyses more accessible and efficient. Well-designed tools enable researchers to focus on core questions without being hindered by technical barriers, allowing them to extract insights from large datasets and analyze social behavior effectively. For social scientists and practioners, user-friendly tools democratize access to advanced computational methods, fostering wider participation and interdisciplinary collaboration. This accessibility accelerates research, enhances reproducibility, and ensures that findings are robust and impactful. In short, effective tools are key to unlocking data-driven research and driving innovation in understanding complex social phenomena.

The rise of social media has amplified the need for such tools, as it has profoundly impacted societal events and institutions, influencing everything from public discourse to political movements. With user engagement on platforms like Twitter continuing to grow, researchers are just beginning to understand the full scope of its effects. Social scientists and practitioners often analyze online discourse as a proxy for real-world behavior, curating large datasets to study these interactions. However, the lack of accessible tools tailored for non-data science experts means that much of this data remains underutilized, limiting the potential insights that can be drawn.

This chapter addresses this gap by introducing `birdspotter`, an end-to-end tool designed for analyzing and labeling Twitter users, alongside `birdspotter.ml`, an exploratory visualizer

for computed metrics. `birdspotter` provides a streamlined analysis pipeline, enabling users to process pre-collected Twitter data, apply general-purpose labeling, and estimate social influence—all within a few lines of code. By offering comprehensive tutorials and detailed documentation, the tool empowers social scientists, policy analysts, and other non-technical users to engage with advanced computational methods without needing in-depth programming skills.

The development of `birdspotter` highlights the importance of accessible tools in applying computational methods to nuanced domains like social media analysis. Effective tools lower the barriers to complex data analysis, allowing experts from various fields to leverage sophisticated techniques for their specific needs. This approach broadens participation, enabling a deeper exploration of social phenomena and encouraging cross-disciplinary collaboration. `birdspotter` exemplifies this by offering versatile features that can be adapted for different research objectives, from user profiling and influence estimation to bot detection. The tool's ability to function as a bot detector, achieving state-of-the-art performance without real-time Twitter API calls, demonstrates its practical value. This was showcased through an exploratory analysis of COVID-19 discourse, illustrating how `birdspotter` can be used for topical investigations and providing a model for future studies requiring robust, scalable social media analysis.

Within the broader framework of this thesis, `birdspotter` represents the critical role of tooling in integrating computational methods into psychosocial research. The availability of effective tools is essential for applying nuanced analyses across diverse research domains, making it possible to scale sophisticated methods to larger datasets and varied contexts. By providing a user-friendly interface and seamless integration of complex processes, `birdspotter` enables researchers to focus on generating insights rather than managing technical challenges.

Tools like `birdspotter` make it possible to apply sophisticated models at scale, enabling broader collaboration across fields. The tool's design emphasizes usability, ensuring that users can easily harness its capabilities and adapt them to their research needs.

The development and implementation of `birdspotter` illustrate how practical, accessible tools can bridge the gap between technical complexity and real-world application. This

research highlights the transformative potential of well-designed tooling, making it easier for researchers from various fields to engage with, analyze, and interpret social behavior in digital environments. This chapter supports the thesis, underscoring the need for continued development of tools that simplify complex analyses, broadening participation and enabling new, nuanced explorations of social phenomena.

Author Declaration

The following chapter contains content from the following publication.

Rohit Ram, Quyu Kong, and Marian-Andrei Rizoiu. Birdspotter: A tool for analyzing and labeling twitter users. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 918–921, 2021

Author Contributions: R.R. led the research for this study, managed the data processing and collection, and conducted the experiments and analysis. M.A.R. provided supervision through all stages of the study. Q.K. contributed to the case studies. R.R. and M.A.R. collaboratively developed the model and experimental design. R.R., M.A.R., and Q.K. interpreted the results and contributed to manuscript writing and editing.

Production Note:
Signature removed prior to publication.

Rohit Ram

Production Note:
Signature removed prior to publication.

Quyu Kong

Production Note:
Signature removed prior to publication.

Marian-Andrei Rizoiu

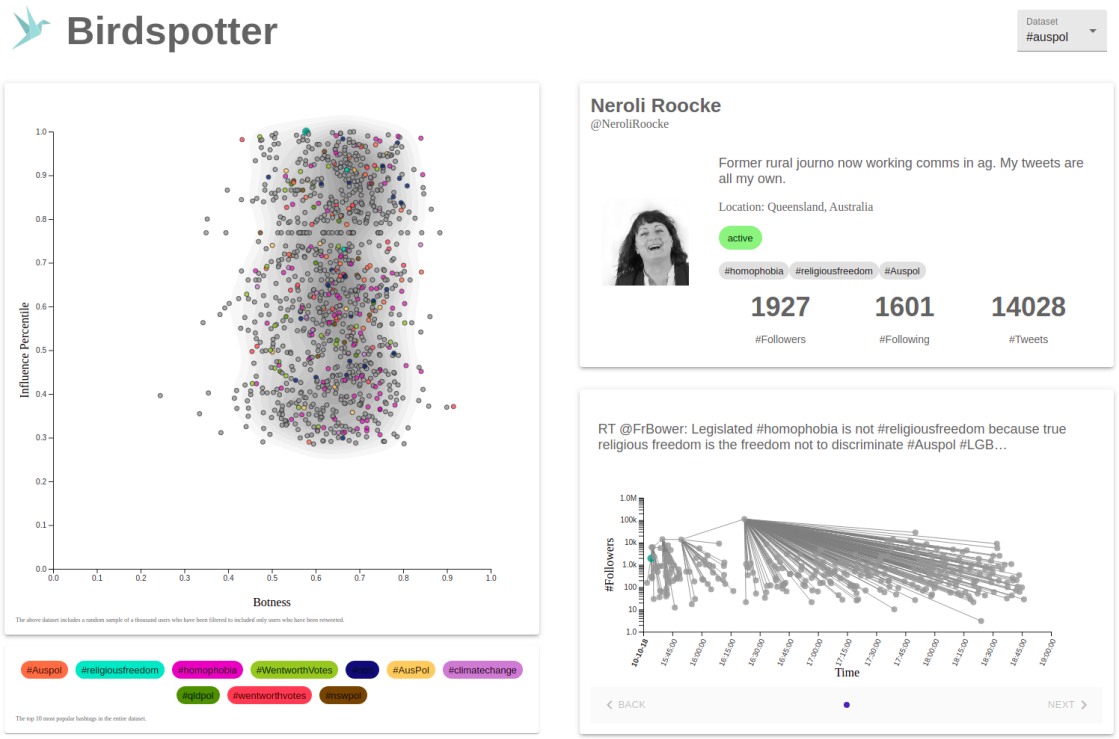


FIGURE 5.1: The `birdspotter.ml` visualization system: Twitter users are plotted based on their user influence and botness (left panel), and we show a selected user's profile (top-right) and cascade history (bottom-right).

5.1 Introduction

Barely a decade old, social media in general — and Twitter in particular — are becoming increasingly important in shaping societal events. They serve as novel fora for a wide array of users to express themselves, discuss, promote agendas and attempt to influence the said societal events. As a result, social and political scientists, journalists, and communication scientists increasingly turn to social media as a proxy to study society. They carefully curate and label large social media datasets, and here a gap emerges. There is a limited offer of tools aimed at non-machine learning experts to analyze users in already existing datasets without making additional web API calls that limits the amount of retrieved data. This paper fills this gap by proposing `birdspotter`, a package aimed at non-computing practitioners with quantitative expertise (basic `R` or `Python`), to analyze, describe, and

automatically label users in Twitter datasets.

This work addresses three specific open questions concerning analyzing Twitter users. The first question relates to the availability of user analysis tools. Existing tools are typically designed for Twitter branding and management, i.e. to either analyze a user's or organization's account (Twitter Analytics[6], or Brandwatch Consumer Research[7]), or one given user account (Account Analysis Tool[8]). The question is **whether a tool exists to retrospectively analyze and label all the users in Twitter dumps, aimed at non-data science experts with computational expertise?** We address this question by proposing `birdspotter`[1], an integrated Twitter user analysis tool, that can achieve three types of analysis in only a couple of lines of code. First, it processes existing Twitter datasets (e.g. `jsonl` data dumps collected through the Streaming API). Second, it describes users using three types of features (relating to the user, content semantics, and hashtag usage). Last, it allows training a classifier against a labeled user subset, which turns `birdspotter` into a general-purpose inferential user analysis tool.

The second open question relates to profiling user botness and influence on previously collected data. The state-of-the-art bot detector, *botometer* [196], can only be accessed through its web APIs and cannot produce predictions for users that are no longer accessible, such as suspended accounts. Since bots have a high tendency of being suspended by Twitter, measuring botness a while after collecting data risks missing a large proportion of the bots involved in discussions. Similarly, existing influence estimation tools require knowledge of the social graph, which often is impossible to capture retrospectively. The question is: **can we design a tool that quantifies users' botness and influence on existing curated datasets, without the need of online API calls or supplementary information?** We address this question two-fold. First, using four existing Twitter bot datasets, we train `birdspotter` to detect bots without requiring additional API calls. We

[1] `birdspotter` source code, tutorial, and feature list: `https://github.com/behavioral-ds/BirdSpotter`
[2] `birdspotter.ml` public installation: `https://www.birdspotter.ml`
[3] `birdspotter` documentation: `https://birdspotter.readthedocs.io`
[4] COVID-19 tutorial: `https://github.com/behavioral-ds/user-analysis`
[5] Supplementary Material: `https://arxiv.org/pdf/2012.02370.pdf#page=5`
[6] Twitter Analytics: `https://analytics.twitter.com/`
[7] Brandwatch Consumer Research: `https://www.brandwatch.com`
[8] Account Analysis Tool: `https://accountanalysis.app`

show that `birdspotter` achieves a higher performance than the current state-of-the-art *botometer* [196]; `birdspotter` ships the bot detector by default, with the package. Second, we implement a diffusion-based influence estimation [101], which is as accurate as using the social graph.

The third open question is **can we visualize and explore both broad and specific views of Twitter users and their activity?** We address this question by proposing `birdspotter.ml`[2], a tool that provides both broad views of the user population and detailed inspections of user activity (see Fig. 5.1 for the main interface).

The main contributions of this paper are as follows:

- `birdspotter`[1] — a software package designed for inferential analysis of online users in pre-collected data, and to estimate online user influence based on the reshare cascades.
- `birdspotter.ml`[2] — an online visualizer designed to perform exploratory analysis of Twitter users.
- an offline bot detector, built using four public labeled datasets; we show that it achieves better than state-of-the-art performance and we showcase it on an example analysis of users discussing COVID-19[4].

Related work. Here, we present the prior work most relevant to `birdspotter`. For a complete related work discussion, please refer to the online appendix[5].

Tree-based ensemble methods dominate social bot detection (over deep learning) due to the heterogeneity of bots and the relative sparsity of training data. The de-facto bot detection tool is `botometer` (formerly `BotOrNot`) [197], which uses more than 1000 user- and recent activity-related features to train a Random Forest classifier. The main limitations of `botometer` are 1) usage of online APIs which are rate-limited by Twitter, 2) lack of reproducibility, since deactivated, protected, and suspended users can no longer be retrieved, and 3) `botometer` scores are likely to vary with user activity and `botometer` versioning. `Birdspotter` addresses the above by predicting bots on pre-collected Twitter dumps.

User influence is typically measured using static user attributes [198], analyzing the online social graph [199], and modeling information diffusion [200]. Closest to our work is `ConTinEst` [201], which requires knowledge of the social graph (often prohibitively expensive to obtain) on which it performs random walks (very slow on large social graphs). `Birdspotter` estimates user influence from resharing dynamics in the absence of knowledge about the social graph.

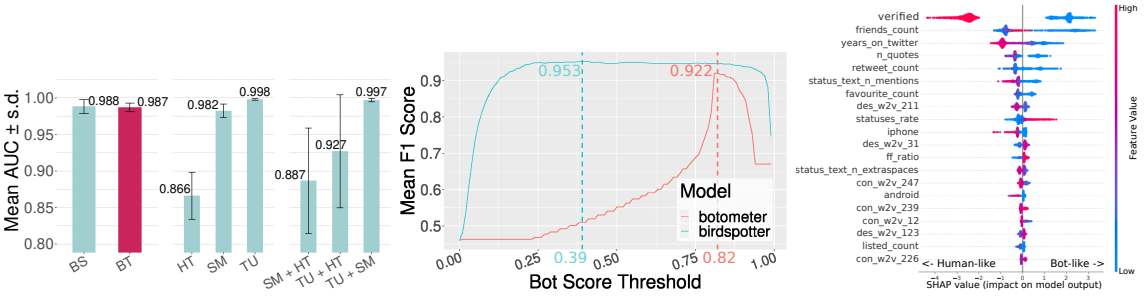
5.2 Preliminaries

In this section, we briefly outline prerequisites concerning influence estimation using point-process models. For a thorough construction of the influence estimation, please refer to the online appendix[5].

User influence estimation. `birdspotter` implements the algorithm in [1], estimating online influence as the mean number of retweets generated, directly and indirectly, by a user's (re)tweet. Rizoiu et al. [1] estimate user influence, absent of the retweet branching structure, by assuming that retweets arrive following a Hawkes point process [101]. They estimate the probability that the tweet $v_j = (m_j, t_j)$ is a direct retweet of v_i as $p_{ij} = \frac{\phi(m_i, t_j - t_i)}{\sum_{k=1}^{j-1} \phi(m_k, t_j - t_k)}$, where m_j is the associated user's follower count, t_j is the time of the event, and $\phi(m, \Delta t) = \kappa\theta m^{\beta} e^{-\theta \Delta t}$ is the marked Hawkes exponential kernel of parameters κ, β and θ. The *pairwise influence* represents the probability that v_i indirectly generates v_j, and is computed as $r_{ij} = \sum_{k=i}^{j-1} r_{ik} p_{kj}$ when $i < j$, $r_{ii} = 1$, and is 0 otherwise. Furthermore, a tweet's influence is the sum of its pairwise influences, and a user's influence is its tweets' influences averaged.

5.3 Package Overview

In this section, we give an overview of `birdspotter` and `birdspotter.ml`, and describe their usage, functionalities, and design.



(A) (B) (C)

FIGURE 5.2: (a) Mean AUC +/- standard deviation, varying ablated models and botometer. Models/Features are indicated by BS (`birdspotter`), BT (`botometer`), HT (Hashtags), SM (Semantic),and TU (Twitter User). (b) Mean F_1 score versus bot threshold for `birdspotter` and `botometer`. (c) SHAP summary plot where points indicate classifier decisions, y-axis shows features in decreasing importance, x-axis shows SHAP impact value, and color indicates feature value. Positive SHAP indicates bots.

5.3.1 `birdspotter`

`birdspotter` labels users and measures influence on previously collected tweets in the standard `jsonl` or `json` format.

Measuring influence. `birdspotter` measures user influence as outlined in Section 5.2, using by default a marked Hawkes exponential kernel with parameters $\beta = 1$, $\kappa = \frac{1}{\theta}$ and $\theta = 6.8 \times 10^{-4}$. These were tuned on a large collection of real cascades [1], and can be customized using the function `getInfluenceScores()`.

Usage and functionalities. Given a dataset of tweets collected externally (e.g. leveraging the Twitter Filter API), `birdspotter`'s core functionality revolves around two steps. In the first step, `birdspotter` loads the Twitter dataset, extracts retweet cascades, and compiles the user-level information. In the second step, it performs the *influence* analysis and user *labeling*. The former is achieved by simply invoking the `BirdSpotter` constructor, while the latter is achieved by calling the function `getLabeledUsers()`, which returns a table with the user features detailed above. For every observed cascade, `birdspotter` also computes the most likely branching structure (see p_{ij} in Section 5.2). This can be achieved using the function `getCascadesDataFrame()`, which returns the reshare cascades (i.e. original tweet and all its retweets) with the additional column `expected_parent` indicating a retweet's most likely parent tweet.

For power users, `birdspotter` provides a number of robust configurations — such as changing the parameters of the Hawkes kernel or using user-defined word embeddings — documented using its `readthedocs`[3] documentation. A usage tutorial is available on `birdspotter`'s repository[1]. For users who prefer to analyze the results outside `python`, `birdspotter` can dump the user table and the reshare cascades in Comma Separated Values (CSV) files, that can be loaded in outside tools. All `birdspotter` functionalities can be accessed in `R` via `reticulate` (`https://github.com/rstudio/reticulate`).

Feature Construction. `birdspotter` constructs user features[1] in three categories: Twitter user, semantic, and topic-based features. **Twitter user features** are engineered directly from twitter user attributes and capture heuristics of common bot behavior. **Semantic features** are constructed (by default) from FastText 300d word2vec embeddings [202] of users' tweets content and descriptions. Content embeddings are averages of tweet embeddings, which are averages of word embeddings. **Topic-based features** are the vectors of the 1,000 most frequent hashtags, scored for each user using the term frequency-inverse document frequency scheme. `birdspotter` is designed to be easily extended with any arbitrary (numerical) features to allow for rapidly evolving bot strategies [203].

User labeling. `birdspotter` implements a supervised labeler. It engineers a large selection of features, and it uses a Gradient Boosting Machine model (XGBoost [204] implementation), with hyperparameters tuned via Random Search and 5-fold cross-validation.

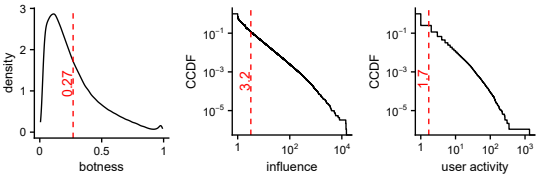
Beyond bot prediction. `Birdspotter` ships by default a pre-trained bot classifier (see Section 5.4), however `birdspotter` can be customized to a particular application or dataset through labeling and re-training. The function `getBotAnnotationTemplate()` outputs a CSV that can be annotated by the user, and `trainClassifierModel()` re-trains the classifier with this annotated data. An option controls whether the model is further tuned starting from the existing model (useful for adapting bot detection to a particular dataset) or retrained from scratch. We exemplify this in Section 5.4.

Data Structures. `birdspotter`'s main class, called `BirdSpotter`, is used to access methods and attributes.

```
from birdspotter import BirdSpotter
bs = BirdSpotter('./corona_2020_01_31.jsonl')
cascades = bs.getCascadesDataFrame()
# user information can be obtained by
labeled_user_labels = bs.getLabels()
labeled_user_influence = bs.getInfluenceScores()

bs.getBotAnnotationTemplate('user_to_label.csv')
bs.trainClassifierModel('users_to_label.csv', update=True,
    iterations=300)
```

(A)



(B)

FIGURE 5.3: Quantifying user *botness* and *influence* analysis on COVID-19 dataset. (a) Code required to load a Twitter dump, generate cascade and user information, annotate and fine-tune the bot classifier. (b) A density plot of user *botness* scores, and complementary cumulative density plots (CCDF) of user *influence* and user *activity*. The red lines show the mean values.

`birdspotter` makes accessible three `pandas` dataframes through the main object after processing: `featuresDataframe` (users and their extracted features), `cascadeDataframe` (tweets and cascade information), and `hashtagDataframe` (TF-IDF of hashtags).

Performance. `birdspotter` performed the extraction, processing, and profiling of a dataset of 196,269 tweets and 129,778 users, in just 5.7 ms per tweet, with an Intel Xeon W-2145 CPU.

Installation. `birdspotter` installs in the canonical `Python` way: `pip install birdspotter`.

5.3.2 birdspotter.ml visualiser

`birdspotter.ml`[2] is a visualizer built on top of `birdspotter`, and designed to analyze Twitter users engaged in online discussions. The visualisation provides both broad and specific views of the data, via the three components shown in Fig. 5.1: a scatter plot component, a user information component, and a cascade view component.

The Scatter Plot. The left panel contains the scatter-plot showing the influence percentile (on the y-axis) and botness (on the x-axis) of a random sample of the users from the dataset, and the underlying 2-D density over the entire data set. Users are colored based on the hashtag they use most and, when selected, the user and cascade views are populated. The plot is pan-able and zoom-able. The view starts with a random sample of 1,000 users and is dynamically populated as practitioners explore cascades.

The User View. The top-right panel provides information about a selected user, including their Twitter image (hyperlink to the user's profile), screen name, location, the hashtags they used, and basic Twitter metrics (such as the number of followers or tweets).

The Cascade View. The bottom-right panel shows the cascades the selected user participates in, which are select-able via a carousel. The component shows the text of the original tweet, the retweets' timing, and the most-likely branching structure inferred using `birdspotter`. The points on this component are select-able and hover-able in the same way as the scatter plot. The component also is pan-able and zoom-able.

5.4 Building a bot detector

In this section, we train `birdspotter` as a bot classifier with better performances than the state of the art `botometer`. We showcase `birdspotter` to profile a topical COVID-19 Twitter dataset.

Training data. `birdspotter` provides the functionality to retrain and update the current model, which we leverage to build a bot detector. We train on four public bot datasets, including {`botometer-feedback-2019`, `political-bots-2019`} [203], and {`verified-2019`, `botwiki-2019`} [205], sourced from *Bot Repository*[9].

Training. The *Bot Repository* only provides account-level data, whereas `birdspotter` is designed to utilize tweet `jsonl`. We use the tool `twarc` to acquire the timeline of each available user's first 200 tweets, to construct `jsonl` training data. We extract and preprocess the data with `BirdSpotter()`, label the resulting dataframe with users' ground truth values, and run `trainClassifierModel()` on this training data to acquire our final model. We ship this model as the default at `birdspotter`'s installation.

Botometer comparison. We compare the derived model against `botometer`, by acquiring their bot scores (universal CAP [203]) for available users through their API. Fig. 5.2a shows that `birdspotter` out-performs `botometer` in terms of mean AUC, despite using less information to make predictions – `botometer` uses more user features extracted from

[9]available from `https://botometer.osome.iu.edu/bot-repository/datasets.html`

the online API. Fig. 5.2b shows that `birdspotter` consistently out-performs `botometer` with respect to mean F_1 scores, over all bot score thresholds.

Ablation study. We test the importance of each set of features through various ablations of our main model. Fig. 5.2a shows the mean AUC obtained for subsets of features. It shows that Twitter user features and semantic features are both informative of bot-like behavior, while hashtag features show more variation. The hashtag model performance may be an artifact of training on the mixture of bot datasets (containing hashtags relating to different topics). We retain hashtag features in `birdspotter`, for better generalizability when users train and test on their own domain datasets. The best performing model uses Twitter user features and semantic features.

SHAP analysis. We use `shap` [206] for explaining the impact of features in our tree ensemble model. Fig. 5.2c shows that the Twitter user features form the majority, and semantic features a minority of the impactful features, in line with the ablation study.

COVID-19 Application Dataset. We apply `birdspotter` to a COVID-19 dataset [207], supplied as tweet IDs which were *re-hydrated* with `twarc` to a `jsonl` format, recovering 68.8% . We limit our analysis to the $\sim$ 1.5M unique tweets relating to posts on January the 31st, resulting in $\sim$ 0.28M users and $\sim$ 0.42M cascades.

Dataset profiling. Fig. 5.3b shows the empirical distributions of botness, influence, and activity (i.e., the number of cascades a user participates in). The distribution of botness indicates two maxima; the larger indicating the humans and the smaller indicating the bots . Conforming with the literature [1], influence and activity are long-tailed (following a "rich-get-richer" paradigm).

(Re-)Labeling Users. Exploring `birdspotter.ml` we observe humans — `@DumplingSays`, `@eddfuentess`, and `@marat_dospolov` — with bot scores of 0.873, 0.83, and 0.925 respectively. Using `getAnnotationTemplate` (see Fig. 5.3a, line 8) we label each user as *human*, and update the classifier with `trainClassifier` (Fig. 5.3a, line 10). The new bot scores are 0.375, 0.296, and 0.559, respectively. Practitioners can use `birdspotter` to classify any latent user attribute.

5.5 Conclusion

We presented `birdspotter`, a Twitter user analysis tool aimed at non-data science experts who analyze discourse and user activity on social media. It provides an end-to-end analysis of users' online characteristics, and populates a visualizer facilitating both broad views of a user population and individual exploration. As with many open-source classifiers, we know that `birdspotter` could be leveraged to infer sensitive features. However, we are currently not aware of any protections that we could implement to prevent this.

Tools like `birdspotter` are integral to the timely, performant, and reproducible analysis of social media users for understanding discourse and society. The framework's reliance on reshare capabilities for influence measurement and user profile information for characterization makes it broadly adaptable to most social media platforms. For instance, it could be readily adapted to work with the Bluesky API without significant amendments, requiring only platform-specific data collection adjustments while preserving the core analytical components.

While `birdspotter` demonstrates effectiveness against traditional automated bot accounts, it is important to acknowledge the evolving landscape of social media manipulation. State-sponsored information operations increasingly employ sophisticated techniques including phone farms—networks of human operators using physical devices to conduct coordinated inauthentic behavior. Unlike traditional bots that exhibit algorithmic posting patterns detectable through behavioral analysis, phone farms present fundamentally different detection challenges. These operations combine human judgment with coordinated messaging, making them appear more authentic in individual-level analysis while maintaining strategic coordination at the network level. Future iterations of tools like `birdspotter` will need to incorporate network-level coordination detection and cross-temporal behavioral analysis to address these more sophisticated manipulation techniques. The detection paradigm must evolve from identifying individual inauthentic accounts to recognizing coordinated inauthentic networks, regardless of whether they are automated or human-operated.

An interesting direction for future research involves examining how influence conductance varies across different subtopics within broader discussions. While topic modeling and

subtopic classification fall outside the scope of `birdspotter` and are well-served by existing specialized tools, the intersection of influence dynamics with topical granularity presents compelling research questions. Different subtopics may exhibit varying influence conductance—the ease with which influence propagates through user networks—suggesting that users may wield domain-specific influence that does not necessarily transfer across related topics. Investigating these patterns could provide deeper insights into how expertise and credibility operate within nuanced discussion landscapes, though such analysis would require integration with dedicated topic modeling frameworks.

Chapter 6

Non-traditional Research Outputs

In this section, I shift focus from the formal research outputs presented throughout this thesis to highlight the impact of non-traditional research outputs (NTROs) as integral elements in advancing the field of computational methods in psychosocial research. While research articles provide foundational contributions, NTROs offer unique avenues to disseminate knowledge, build networks, and foster a culture of innovation among practitioners. By engaging in NTROs, I extend the reach and relevance of my work beyond academic publications, enabling practical, community-driven adoption of computational approaches within psychosocial fields.

The strategic importance of NTROs becomes evident when considering the fundamental challenge of computational social science: bridging the methodological sophistication of computer science with the theoretical depth and practical concerns of psychology and sociology. Traditional academic publishing, while essential for establishing scientific rigor, often fails to reach the practitioners who could most benefit from these advances. Defense analysts need accessible tools for detecting information operations; public health officials require frameworks for understanding misinformation spread; educators seek compelling examples to motivate interdisciplinary learning. These diverse stakeholder needs demand translation efforts that extend far beyond conventional peer review.

Moreover, the computational methods developed in this thesis—from influence modeling to ideology detection—derive their ultimate value from real-world application. Without

sustained engagement with practitioner communities, even the most sophisticated research risks remaining confined to academic circles, limiting its societal impact. NTROs provide essential feedback loops, where practical implementation challenges inform future research directions and where theoretical insights gain validation through applied contexts.

NTROs play a pivotal role in bridging the gap between rigorous research and practical application. These outputs—ranging from public talks and popular articles to collaborations with industry stakeholders—are essential for translating complex findings into accessible formats, enhancing visibility and understanding among diverse audiences. Importantly, NTROs provide opportunities for researchers to engage with sectors like defense, healthcare, and policy, where interdisciplinary knowledge exchange is particularly valuable. Through NTROs, I contribute to a broader, more inclusive research culture that invites non-specialists into the conversation, helping to demystify computational methods and reduce the barriers to adoption.

Through collaborations, public articles, and practical workshops, I have engaged diverse audiences to demonstrate and demystify computational methods, fostering broader adoption and understanding within psychosocial research.

I have engaged in discourse with Australian Defence scientists, including those from the ASD and DSTG, as well as with other stakeholders invested in the dissemination and application of complex psychosocial insights. Additionally, I have authored public-facing articles, such as one in The Conversation, which garnered over 4,700 reads and more than 150 shares, helping to extend the reach and impact of my work.

In highlighting these NTRO activities, I underscore the value of such non-traditional outputs in facilitating a culture of knowledge sharing, collaboration, and interdisciplinary engagement that complements and extends the impact of formal research publications.

Here I present a history of event participation, presentations, collaborations, and broader non-traditional scholarly activities.

6.1 Events & Presentations

This section summarizes key venues where I presented research that bridges computational and psychosocial fields.

Australian Social Network Analysis Conference 2019 (ASNAC'19) The Australian Social Network Analysis Conference brings together social scientists who utilize social network analysis (SNA) to examine human interactions in various contexts. At ASNAC'19, research primarily focused on smaller, highly contextualized networks. I presented an early form of our work on large-scale network analysis, which examined social influence via social media data, contributing a novel perspective to the venue's traditional scope of inquiry.

Data Science Institute Seminar Hosted by the Faculty of Engineering and Information Technology (FEIT) at the University of Technology Sydney (UTS), this seminar series attracts both students and academics. My presentation introduced two of my key research tools: the General Influence Model (GIM) and *birdspotter*, as detailed in Chapters 3 and 5, both of which exemplify the fusion of computational methods with psychosocial research principles.

International Conference On Web Search and Data Mining 2021 (ICWSM'21) The International Conference on Web Search and Data Mining (ICWSM) is an interdisciplinary venue that attracts researchers from both computational and human sciences. At ICWSM'21, I presented my work on *birdspotter*, focusing on the application of computational tools to model social influence on social media platforms. The presentation involved a detailed case study demonstrating the utility of these tools for psychosocial research.

Defence Human Sciences Symposium The Defence Human Sciences Symposium brings together defense sector researchers interested in various aspects of human sciences, including psychology, health, and cyber domains. My presentation focused on the computational detection of online ideological expression, detailed in Chapter 4, underscoring the application of computational techniques to high-stakes psychosocial phenomena.

PSYBER'23 PSYBER'23, co-hosted by the Australian Signals Directorate and the U.S. Department of Homeland Security, is an international conference focused on advancements

in psychological and behavioral sciences relevant to national security. My presentation on empirical measurement frameworks, described in Chapter 2, showcased how computational methods can inform real-world applications, providing cross-domain value.

6.2 Collaborations

This section highlights key interdisciplinary collaborations that illustrate the integration of computational methods into psychosocial research.

Ideology Detection As discussed in Chapter 4, this project represents a collaboration between myself, Professor Emma Thomas (Flinders University), and Dr. David Kernot (DSTG). Our research applies computational psycholinguistics to detect ideological signals in online spaces, requiring both computational techniques and psychological insights into ideological formation.

Wiki Workshop'23 I co-authored a paper with Dr. Francesco Bailo and Dr. Marian-Andrei Rizoiu on the detection of ideology among Wikipedia contributors, focusing on how contributors' activities reflect ideological biases. This work adds to the literature on bias in collective knowledge platforms, with implications for public trust in information sources.

Evently: Modeling Social Media Cascades The *Evently* project, as described in [9], leverages Hawkes processes to model information diffusion on social media platforms. My contribution focused on applying the tool to disentangle the roles of influential users and bots in resharing dynamics, a key issue in understanding the spread of misinformation.

Interval-Censored Transformer Hawkes: Detecting Information Operations In this collaboration, we developed a novel deep learning architecture based on Transformer Hawkes models [208] to detect disinformation operations. My role was to design and conduct experiments that validated the model's robustness, which significantly enhances its practical applicability in detecting incomplete data within social networks.

6.3 Industry Experience

Throughout 2023, I worked as a research data scientist at Thaum, a company providing research services across a range of public and private sector domains. My role involved applying computational and data science methods to interdisciplinary research questions, including scientometrics and digital twinning for risk assessment of marine drones. This position honed my ability to apply computational techniques in diverse real-world scenarios, reinforcing the practical value of interdisciplinary research.

6.4 Teaching

In the first semester of 2021, I taught an introductory course in computer science (CS0), focusing on programming fundamentals using Java. My teaching experience extends beyond this course, having taught multiple subjects across logic, software design, functional programming, and data science at the Australian National University. My pedagogical philosophy emphasizes the importance of showing students the interdisciplinary applications of computational techniques, especially in fields like psychosocial research where they might seem initially disconnected.

Teaching CS0 is particularly challenging in motivating students to recognize the broader value of computational thinking. Drawing from my interdisciplinary experience, I help students see how computational methods can unlock new avenues in fields as diverse as psychology, sociology, and even defense studies.

6.5 Platform Considerations and Broader Applications

While this thesis primarily utilizes Twitter/X data, the computational frameworks developed demonstrate broad applicability across social media platforms. Core mechanisms such as resharing behaviors and textual content analysis are ubiquitous across platforms,

with Chapter 4 explicitly incorporating cross-platform data sources. However, platform-specific characteristics—including communication styles, community structures, and user demographics—may influence observed patterns, warranting future investigation.

The practical applications extend to e-safety and national security domains, while data access limitations across platforms highlight the need for improved research access to serve the common good. A comprehensive discussion of platform generalizability and research implications is provided in Chapter 7.

6.6 Ethical Framework and Societal Considerations

The computational frameworks developed in this thesis operate within established ethical research protocols and utilize publicly available data sources combined with standard machine learning classifiers. This research has been conducted under ethics approval from UTS, ensuring adherence to appropriate research standards and participant protection measures. Each chapter addresses the specific ethical considerations relevant to its particular methodological approaches.

The use of publicly available social media data and off-the-shelf classification techniques represents standard practice in computational social science research. The methodologies developed focus on aggregate patterns and population-level insights rather than individual surveillance or targeting. The research aims to advance scientific understanding of social phenomena and provide defensive tools for detecting harmful manipulation, extremist recruitment, and misinformation campaigns.

A critical consideration in this work is balancing legitimate privacy expectations with the need to protect vulnerable communities from harm. While individual privacy remains paramount, the public nature of social media platforms creates a context where understanding collective patterns of influence, misinformation spread, and extremist recruitment serves essential protective functions. The techniques developed here enable researchers and safety practitioners to identify concerning trends and intervention points while focusing on behavioral patterns rather than individual targeting.

This balance becomes particularly important when considering threats to democratic institutions, public health, and community safety. The ability to detect coordinated inauthentic behavior, monitor extremist recruitment strategies, and understand misinformation propagation provides essential capabilities for protecting vulnerable populations—including children, marginalized communities, and those susceptible to radicalization. These protective applications must be weighed against privacy concerns, with transparent governance ensuring that research serves the public good.

Looking toward future applications, the growing sophistication of computational social science tools necessitates ongoing attention to responsible development and deployment. The frameworks developed in this thesis contribute to creating a safer, better-understood digital landscape while respecting privacy principles and supporting legitimate research needs that serve both individual rights and collective security.

6.7 Summary

The non-traditional research outputs documented in this chapter represent a systematic effort to transform computational social science from an academic exercise into a practical toolkit for addressing real-world challenges. Through strategic engagement across defense, education, industry, and public discourse, these activities have established pathways for computational methods to inform policy decisions, enhance security practices, and improve public understanding of digital influence phenomena.

The defense collaborations, particularly through venues like PSYBER and the Defence Human Sciences Symposium, have enabled direct translation of research findings into operational contexts where understanding social influence and extremist recruitment can protect national security interests. These engagements demonstrate how academic research on ideology detection and misinformation spread can inform counter-intelligence efforts and public safety initiatives.

Educational activities, from university teaching to public articles, have cultivated the next generation of interdisciplinary researchers while simultaneously raising public awareness about computational approaches to social problems. The Conversation article's significant

reach (4,700+ reads, 150+ shares) exemplifies how effective science communication can democratize access to complex research findings, enabling informed public discourse about digital manipulation and social media influence.

Industry collaborations have validated the practical applicability of these methods while revealing implementation challenges that drive future research directions. The experience at Thaum demonstrates how computational social science techniques can adapt to diverse problem domains, from marine risk assessment to scientometrics, illustrating the broad transferability of these approaches.

Collectively, these NTROs establish computational methods as essential tools for understanding and navigating our increasingly digital social landscape. They demonstrate that advancing psychosocial research requires not only methodological innovation but also sustained commitment to knowledge translation, community engagement, and interdisciplinary collaboration. By bridging academic research with practical application, these activities ensure that computational advances serve broader societal needs while fostering continued innovation at the intersection of technology and human behavior.

Overall, NTROs play a crucial role in equipping psychosocial practitioners and stakeholders with advanced computational techniques, bridging the gap between research and practice. By making these methods accessible through public articles, collaborations, and industry engagement, NTROs support the practical adoption of innovative approaches in real-world settings, expanding the utility and impact of my work beyond academic circles.

Chapter 7

Conclusion

In this thesis, I have asserted that applying computational methods advances our understanding of psychological and sociological phenomena. From a high-level perspective, this work has demonstrated the use of a wide variety of computational techniques, including active learning, modeling with point processes, tree-based classification systems, and data wrangling techniques. Furthermore, this research has directly addressed issues of psychological and sociological concern. I have conducted large-scale psychometry, psycholinguistics, characterized political ideologies, and examined the correlations between psychosocial traits. I have investigated misinformation spreaders, looked into the traits of radical ideologies, and generated tools for bot detection. In this section, I summarize the contributions described in each chapter and illustrate how this supports the thesis.

In Chapter 2, we were able to push the boundaries on the scale and representativeness of the empirical measurement of a psychosocial property. We proposed a framework for measurement in an online context that utilized crowdsourced workers completing tasks orchestrated by an active learning mechanism. The framework could be applied broadly to psychometrics; however, in this instance, we chose to measure social influence, a trait intimately tied to the formation of opinions and the trajectory of social dynamics. We conducted extensive simulation studies, pilot studies, and ablative design studies to ensure the robustness and fidelity of the measurement. Finally, we utilized psycholinguistic methods to further validate the measure and advance psychosocial insights.

In Chapter 3, we were able to endogenize psychosocial theory into a computational model, based on Hawkes point processes, and conduct research to evaluate psychosocial hypotheses. We introduced GIM, which has two flexible psychosocial-inspired mechanisms: conductance and influence-capital distribution, that can be utilized with phenomena such as homophily and provide more degrees of freedom to model the complexities of human interaction. We supplied three implementations of homophily and a novel distribution mechanism. Furthermore, after determining an optimal model, we utilized the inferred influence metric to test psychosocial hypotheses relating to social class and content veracity. The analysis was conducted on a large dataset of COVID-19 discussions, requiring the influence model to scale well. Furthermore, data wrangling techniques were required to assess content veracity and extract social class via occupation. We found that *Media*, *Executives*, *Entertainers*, and the *Military* are some of the most influential social classes. Furthermore, the *Military* contain a large proportion who share misinformative sources.

In Chapter 4, we constructed an end-to-end ideology profiler that can generalize across online contexts. In this work, we utilized a diverse range of ideological signals to ensure the robustness of the profiler and to learn robust feature signals of ideology. A robust ideology profiling is the first step toward a complete opinion dynamics modeling system. Furthermore, we evaluated not only ideological direction (i.e., left-right) but also the presence of extremism (i.e., the far-right). We further utilized psycholinguistic dictionary methods, combined with computational word embedding techniques, to characterize users with respect to morality, grievance language, nationalism, and cognitive distortion schemata. This allowed us to evaluate associations between these psychosocial traits and ideology. We found significant associations, namely; the right use more vice-moral and grievance-filled language, exhibit patterns of dichotomous thinking, and are more associated with nationalism (via flag emojis). As such, this work presents novel psychosocial findings that are only available as a result of computational methodologies.

In Chapter 5, we generated tools that can be utilized by psychosocial practitioners for generalized labeling and influence estimation in online contexts. The software tool was designed for a low barrier to entry, and there is a tutorial and extensive documentation. The tool automatically generates rich feature sets, useful in characterizing online users. Furthermore, the tool comes with a visualizer to quickly draw insights utilizing the generated

metrics. In this work, we illustrated how `birdspotter` can be used as a state-of-the-art bot detection system, which is of particular importance to psychosocial practitioners concerned about external manipulation. With enough labeled data, `birdspotter` can be used as a powerful tool for general sociography in online spaces.

Finally, in Chapter 6, I illustrated my activities, via non-traditional research outputs, in the use of computational methods for psychosocial research and my scholarly activities more broadly. I have engaged with psychology and sociology practitioners, including through direct collaborations, having presented at several venues with psychosocial themes and writing broad audience research articles. Furthermore, I have an extensive history of teaching, particularly CS0, and modeling the use of computational methods for interdisciplinary purposes.

From the summary above, it is clear that modern computational methods have been utilized to pursue psychosocial research, but can it be definitively stated that psychosocial research has been advanced? In many instances, computational methods are indispensable for achieving the types of results outlined throughout this thesis. Two pivotal areas stand out where computational methods catalyze progress in psychosocial research: scale and intelligence. Firstly, psychological and sociological investigations have historically been constrained by the challenges of limited sample sizes and restricted access to diverse populations. These limitations undermine the robustness, generalizability, and external validity of findings, contributing to issues such as the replication crisis. Computational methods, by harnessing vast datasets and enabling unprecedented processing speed, allow researchers to transcend these limitations and achieve scale at a level that was previously unattainable without significant and impractical resources. This scalability empowers researchers to conduct large-scale studies, with readily accessible cross-cultural and longitudinal data, that enrich our understanding and ensure more representative results. Secondly, as machine learning technologies continue to evolve, assessments that once required expert human analysis can now be performed by these systems. This includes intricate analyses of natural language, psychological coding, behavioral predictions, and other tasks that demand both precision and scalability. By automating these assessments, computational methods facilitate not only greater efficiency but also a heightened consistency across studies. These advances render research more reproducible, objective, and reliable, breaking down barriers that

limited traditional methodologies. Moreover, beyond the mechanics of scale and intelligence, the adoption of computational tools shifts the paradigm of psychosocial research. It opens new avenues for interdisciplinary collaboration, where insights from computer science, psychology, sociology, and data science converge to produce richer, more nuanced analyses. This multidisciplinary approach expands the horizon of inquiry, embedding psychosocial studies within a broader and more dynamic ecosystem. Therefore, the integration of computational methods represents more than just a technological enhancement; it is a transformative force that reshapes our capacity to understand and investigate psychological and sociological phenomena. It underscores a new era of research—one marked by increased rigor, broader inclusivity, and the potential for deeper insight. As this thesis has illustrated, the future of psychosocial research lies not just in adopting these computational tools, but in applying them to push the boundaries of knowledge and unlock the complexities of human behavior and society.

7.1 Addressing Platform Specificity and Research Generalizability

A critical consideration for this body of work concerns the extent to which findings derived primarily from Twitter/X data can be generalized to the broader digital ecosystem. This question strikes at the heart of contemporary social media research, where platform selection often constrains both methodological approaches and the scope of conclusions that can be drawn.

The methodological foundations established throughout this thesis—encompassing influence measurement, information diffusion modeling, ideological classification, and user behavior analysis—operate on universal principles of human social interaction that transcend specific platform implementations. Whether users share content through retweets, Facebook shares, LinkedIn reposts, or TikTok duets, the underlying mechanisms of information propagation and social endorsement remain conceptually consistent. Similarly, textual expressions of ideology, psychological traits, and social attitudes manifest across platforms wherever users generate written content.

Yet platform ecology significantly shapes user behavior, content moderation practices, and algorithmic mediation in ways that may fundamentally alter the expression and measurement of psychosocial phenomena. Consider how Twitter's brevity constraints encourage different rhetorical strategies compared to Facebook's longer-form posts, or how Instagram's visual-centric design privileges different modes of self-presentation than text-based platforms. Professional networks like LinkedIn cultivate distinct norms around acceptable discourse, while emerging platforms like Bluesky experiment with novel approaches to content curation and community governance.

These variations present both challenges and opportunities for future research. Rather than viewing platform differences as limitations, they offer natural experiments for testing the robustness of computational approaches across varying social contexts. The ideology detection work presented in Chapter 4 already demonstrates this principle through its multi-platform data integration, suggesting that core methodological approaches can indeed transfer across digital environments with appropriate calibration.

Looking forward, the research community faces pressing challenges around data accessibility that extend far beyond methodological considerations. Current trends toward platform closure and API restrictions fundamentally threaten the viability of independent social media research. This development occurs precisely when society most needs robust, independent analysis of digital influence patterns, misinformation dynamics, and platform manipulation tactics.

The emergence of more open platforms like Bluesky signals potential shifts toward research-friendly data policies, but systematic change requires broader recognition that social media research serves essential public interests. Policymakers should consider frameworks that balance platform privacy concerns with legitimate research needs, potentially through mandated research access provisions or data sharing requirements for platforms above certain user thresholds.

Such access would enable the kind of comprehensive, cross-platform comparative research necessary to validate and extend the findings presented in this thesis. More importantly, it would support the development of robust countermeasures against information manipulation,

extremist recruitment, and other threats to democratic discourse that operate across platform boundaries.

The stakes of this research extend well beyond academic inquiry. National security agencies, public health organizations, and democratic institutions increasingly depend on sophisticated understanding of digital influence patterns to fulfill their protective mandates. The computational frameworks developed in this thesis provide foundational tools for such applications, but their full potential can only be realized through sustained, systematic research across the complete spectrum of digital social environments.

Appendix A

Empirically Measuring Online Social Influence

A.1 Bradley-Terry Noise Invariance.

In this section, we show that within the augmented Bradley-Terry model, λ and $\boldsymbol{\theta}$ cannot be jointly fit from pairwise comparisons. We show that any ML estimate of λ and $\boldsymbol{\theta}$ would not be unique.

Suppose there exists unique ML estimates, $\hat{\lambda}$ and $\hat{\boldsymbol{\theta}}$. Then, consider two constructed quantities, with arbitrary scalar k;

$$\lambda^* = k\hat{\lambda},$$
$$\boldsymbol{\theta}^* = k\theta$$

The log-likelihood for the augmented BT model, is given as

$$\ell(\boldsymbol{\theta}, \lambda) = \sum_{\{i \prec j\}} log\left(1 + e^{\frac{-(\theta_j - \theta_i)}{\lambda}}\right).$$

Consider the maximum log-likelihood,

$$\begin{aligned}
\ell(\hat{\boldsymbol{\theta}}, \hat{\lambda}) &= \sum_{\{i \prec j\}} log\left(1 + e^{\frac{-(\hat{\theta}_j - \hat{\theta}_i)}{\hat{\lambda}}}\right) \\
&= \sum_{\{i \prec j\}} log\left(1 + e^{\frac{-(k\theta^*_j - k\theta^*_i)}{k\lambda^*}}\right) \\
&= \sum_{\{i \prec j\}} log\left(1 + e^{\frac{-k(\theta^*_j - \theta^*_i)}{k\lambda^*}}\right) \\
&= \sum_{\{i \prec j\}} log\left(1 + e^{\frac{-(\theta^*_j - \theta^*_i)}{\lambda^*}}\right) \\
&= \ell(\boldsymbol{\theta}^*, \hat{\lambda}^*) \ .
\end{aligned}$$

This implies that $\boldsymbol{\theta}^*$ and $\hat{\lambda}$ are ML estimates, which is a contradiction. Therefore unique ML estimates for both λ and $\boldsymbol{\theta}$, do not exist.

Appendix B

Conductance and Influence-Capital: Modeling Online Social Influence

B.1 Complete derivation of GIM

This section shows the complete derivation from Eq. (3.2) to Section 3.3.3, both shown in the main text.

We define the influence of a tweet given the diffusion scenario $\mathcal{G}$, as:

$$\varphi(v_i|\mathcal{G}) = \sum_{v_j \in V(\mathcal{G})} \theta(i,j|\mathcal{G}) \tag{B.1}$$

where $V(\mathcal{G})$ is the set of nodes in $\mathcal{G}$, and we denote as $\theta(i,j|\mathcal{G}) := \Psi(\mathcal{G}_{i\to j})$ the influence-capital distribution along a path. We start from the definition of GIM given a retweet cascade (main text Eq. (3.2)):

$$\begin{aligned}\varphi_\gamma(v_i) &= \sum_{\mathcal{G}\in\Upsilon} \sum_{t_j > t_i} \underbrace{\mathbb{P}_\gamma(\mathcal{G}_{i\to j})}_{\text{Conductance}} \quad \underbrace{\Psi(\mathcal{G}_{i\to j})}_{\text{Capital Distrib.}} \\ &= \sum_{\mathcal{G}\in\Upsilon} \mathbb{P}_\gamma(\mathcal{G}) \sum_{v_j \in V(\mathcal{G})} \theta(i,j|\mathcal{G}) = \sum_{\mathcal{G}\in\Upsilon} \mathbb{P}_\gamma(\mathcal{G})\varphi(v_i|\mathcal{G})\end{aligned} \tag{B.2}$$

Notably, due to the factorial number of diffusion scenarios in Υ, computing the influence for each graph is intractable.

Incremental construction of diffusion scenarios. We leverage the *independent cascades assumption* (see Section 3.2) to construct an efficient influence computation that overcomes intractablility. The key observation is that each tweet v_k is added simultaneously at time t_k to all diffusion scenarios constructed at time t_{k-1}. v_k contributes only once to the tweet influence of every tweet found on the path to which v_k is attached. The tweet influence is computed incrementally by updating $\varphi(v_i), i < k$ at each time t_k. We denote by $\varphi^k(v_i)$ the value of tweet influence of v_i after adding node v_k. As a result, we only track how the tweet influence increases over time steps, and we do not construct all valid diffusion scenarios.

GIM assumes that a user's tweet is influenced by one of the precedent tweets, chosen stochastically from a discrete distribution over the valid edges. Alternatively, we can interpret that all previous tweets influence the new tweet proportionally to the same discrete distribution (a view inline with recent findings about influence and complex contagion). Let $\Upsilon_{1:k-1}$ be the set of all possible diffusion scenarios at time t_{k-1}, and $\mathcal{G}^- \in \Upsilon_{1:k-1}$ be one such diffusion scenario, with the set of nodes $V^- = \{v_1, v_2, \cdots, v_{k-1}\}$. When v_k arrives, it can attach to any node in V^-, generating $k-1$ new diffusion scenarios $\mathcal{G}_j^+$, with $V_j^+ = V^- \cup v_k$ and $E_j^+ = E^- \cup (v_j, v_k)$. We can write the set of scenarios at time t_k as:

$$\Upsilon_{1:k} = \left\{ \mathcal{G}_j^+ \middle| \forall j < k, \forall \mathcal{G}^- \in \Upsilon_{1:k-1} \right\} \tag{B.3}$$

We write the tweet influence of v_i at time k as:

$$\varphi^k(v_i) = \sum_{\mathcal{G}^+ \in \Upsilon_{1:k}} \mathbb{P}_\gamma(\mathcal{G}^+)\varphi(v_i|\mathcal{G}^+) =^{cf.\ (B.3)} \sum_{\mathcal{G}^- \in \Upsilon_{1:k-1}} \sum_{j=1}^{k-1} \mathbb{P}_\gamma(\mathcal{G}_j^+)\varphi(v_i|\mathcal{G}_j^+) \tag{B.4}$$

Attach a new node v_k. We concentrate on the right-most factor in Eq. (B.4) – the tweet influence in scenario $\mathcal{G}_j^+$. We observe that the terms in Eq. (B.1) divide into two: the paths from v_i to all other nodes except v_k (i.e. the old nodes) and the path from v_i to v_k. We

obtain:

$$\varphi(v_i|\mathcal{G}_j^+) = \sum_{\substack{v_l \in \mathcal{G}_j^+ \\ l>i, l\neq k}} \theta(i,l|\mathcal{G}_j^+) + \theta(i,k|\mathcal{G}_j^+)$$

Note that a path that does not involve v_k has the same influence capital contribution in $\mathcal{G}_j^+$ and in its parent scenario $\mathcal{G}^-$, i.e. $\theta(i,l|\mathcal{G}_j^+) = \theta(i,l|\mathcal{G}^-),$ for $l > i,$ and $l \neq k$. We obtain

$$\varphi(v_i|\mathcal{G}_j^+) = \sum_{\substack{v_l \in \mathcal{G}^- \\ l>i}} \theta(i,l|\mathcal{G}^-) + \theta(i,k|\mathcal{G}_j^+) =^{cf.\ (B.1)} \varphi(v_i|\mathcal{G}^-) + \theta(i,k|\mathcal{G}_j^+) \tag{B.5}$$

Combining Eq. (B.4) and (B.5), we obtain:

$$\begin{aligned}\varphi^k(v_i) &= \sum_{\mathcal{G}^- \in \Upsilon_{1:k-1}} \sum_{j=1}^{k-1} \mathbb{P}_\gamma(\mathcal{G}_j^+) \Big[\varphi(v_i|\mathcal{G}^-) + \theta(i,k|\mathcal{G}_j^+)\Big] \\ &= \underbrace{\sum_{\mathcal{G}^- \in \Upsilon_{1:k-1}} \varphi(v_i|\mathcal{G}^-) \sum_{j=1}^{k-1} \mathbb{P}_\gamma(\mathcal{G}_j^+)}_{A} + \underbrace{\sum_{\mathcal{G}^- \in \Upsilon_{1:k-1}} \sum_{j=1}^{k-1} \mathbb{P}_\gamma(\mathcal{G}_j^+)\theta(i,k|\mathcal{G}_j^+)}_{m_{ik}}\end{aligned} \tag{B.6}$$

Tweet influence at previous time step t_{k-1}. Given the definition of $\mathcal{G}_j^+$ in Eq. (B.3) and the independant cascades assumption, we obtain that $\mathbb{P}_\gamma(\mathcal{G}_j^+) = \mathbb{P}_\gamma(\mathcal{G}^-)p'_{ij}$. Consequently, part A in Eq. (B.6) can be written as:

$$\begin{aligned}A &= \sum_{\mathcal{G}^- \in \Upsilon_{1:k-1}} \varphi(v_i|\mathcal{G}^-) \sum_{j=1}^{k-1} \mathbb{P}_\gamma(\mathcal{G}^-)p'_{jk} \\ &= \sum_{\mathcal{G}^- \in \Upsilon_{1:k-1}} \varphi(v_i|\mathcal{G}^-)\mathbb{P}_\gamma(\mathcal{G}^-) \sum_{j=1}^{k-1} p'_{jk} \overset{cf.\ (B.2)}{=} \varphi^{k-1}(v_i)\end{aligned} \tag{B.7}$$

A is the tweet influence of v_i at the previous time step t_{k-1}. Note that $\sum_{j=1}^{k-1} p'_{jk} = 1$ because v_k is necessarily the direct retweet of a previous nodes $v_j, j < k$ of the retweet cascade.

Contribution of v_k. With A being the influence of v_i at the previous time step, intuitively m_{ik} is the contribution of v_k to the influence of v_i. Knowing that:

$$\mathbb{P}_\gamma(\mathcal{G}_j^+) = \mathbb{P}_\gamma(\mathcal{G}^-)p'_{jk} \text{ and}$$
$$\theta(i,k|\mathcal{G}_j^+) = \Psi(\mathcal{G}_{i\to j})\pi_{jk} = \theta(i,j|\mathcal{G}_j^+)\pi_{jk}^{cf.3.3.2} = \theta(i,j|\mathcal{G}^-)\pi_{jk}$$

we write m_{ik} as:

$$\begin{aligned} m_{ik} &= \sum_{\mathcal{G}^-\in\Upsilon_{1:k-1}} \sum_{j=1}^{k-1} \mathbb{P}_\gamma(\mathcal{G}^-)p'_{jk}\theta(i,j|\mathcal{G}^-)\pi_{jk} \\ &= \sum_{j=1}^{k-1} p'_{jk}\pi_{jk} \underbrace{\sum_{\mathcal{G}^-\in\Upsilon_{1:k-1}} \mathbb{P}_\gamma(\mathcal{G}^-)\theta(i,j|\mathcal{G}^-)}_{m_{ij}} = \sum_{j=1}^{k-1} p'_{jk}\pi_{jk}m_{ij} \ ^{cf.3.3.3} \end{aligned}$$

B.2 Efficient computation of GIM

We define two matrices. First, the transfer matrix $T = [p'_{ij} * \pi_{ij}]$, where the element p'_{ij} is the probability that tweet v_j is a direct retweet of tweet v_i (defined in Eq. (3.3)) and π_{ij} is the proportion of capital transfered from v_j to v_i; Second, the influence accumulation matrix $M = [m_{ij}]$, with m_{ij} defined in Eq. (3.3.3) is the contribution of v_j to the influence of v_i. For each column j of M, we compute the first $j-1$ elements by multiplying the sub-matrix $M_{[1..j-1,1..j-1]}$ with the first $j-1$ elements on the j^{th} column of matrix T, the j-th element is π_{jj}, and the remaining elements are 0. The computation of matrix M finishes after n steps, where n is cascade size.

B.3 Value-Allocation Scheme

Value-allocation schemes over networks are an example of fair division games, concerned with how to allocate the value generated by a network of players among the players; for example, the allocation of advertisement revenue over a chain of marketers, or the allocation of utility revenue across a network of electricity infrastructure providers. Allocation scheme literature generally aims to understand the stability, efficiency, and fairness of network

formation (usually in cooperative undirected networks). Our paper is concerned with the allocation of influence-capital in a non-cooperative setting over an ad-hoc directed-acyclic diffusion graph. The seminal work of Jackson and Wolinsky [117]'s *Connections* game defines the utility of a player as the sum of benefits to all other players decayed by the length of the path between them minus the cost of maintaining direct links for the player.

Appendix C

Practical Guidelines for Ideology Detection Pipelines and Psychosocial Applications

Appendices

This document accompanies the submission. The information in this document complements the submission and is presented here for completeness reasons. It is not required to understand the main paper or reproduce the results.

C.1 Dataset Collection Details

#QandA. We collect discussions related to the Australian panel show Q+A [209], where panelists (public figures, politicians, and experts) answer curated audience questions. Twitter participation is encouraged in airings. We collect #QandA using the filter keyword *qanda* during January-December 2020.

#Ausvotes. We collect discussions about the 2022 Australian Federal election, tracking the lead-up and aftermath. It follows the major parties and their leaders: the left-leaning

Australian Labor Party led by Anthony Albanese and the right-leaning Liberal-National Coalition led by Scott Morrison. We collect #Ausvotes using the keywords *auspol* and *ausvotes*, and for mentions of *@ScottMorrisonMP*, *@AlboMP*, and *@AusElectoralCom*, between 9 May and 15 June 2022 (the elections occurred on 21 May).

#Socialsense [10] features discussions related to the Australian Black Summer bushfires, which gathered discourse concerning climate change, and contains far-right opinions. #Socialsense contains 90 days of Twitter and Facebook discussions, from 1 November to 29 January 2020.

Riot [192] features discussions about the January 6th US Capitol Insurrection, including election fraud and insurrection topics. The dataset spans 6 January to 1 February 2021 and was collected with the filter keywords *TrumpRally*, *Democracy*, *USCapitol*, *Capitol*, *DCProtests*, and *AshliBabbit*.

Parler [193] features discussions about the US Capitol Insurrection from Parler. We collect all posts emitted during the day of 6 January 2021.

C.2 All UUS/UUS+ Metrics

Thi section shows all possible runs for the *UUS* and *UUS+*. We notice that in many instances *UUS* fails to seperate clusters, and even in instances where seperation can be achieved many suffer from poor performance. This shows that these techniques lack robustness for more difficult datasets.

C.3 Left-Right Annotation Procedure

Ideology is the subject of considerable subjectivity, not only because experts have their own ideology, but because annotators are often unclear as to what evidence is permissible for use. For this task we issued the following guidelines to annotators:

> It is not always clear what should count as an ideological signal. For our purposes, we will include the following as signals of ideology:

TABLE C.1: **All Baseline Performances.** The table shows to performances for all combinations of the *UUS* and *UUS+* baselines.

Representation	**Active Users**	**F1-Macro**	**AUC ROC**	**UUS F1-Macro**
H	500	0.37	0.68	0.37
H	1000	-	-	-
H	5000	0.37	0.54	0.37
HR	500	-	-	-
HR	1000	-	-	-
HR	5000	0.89	0.92	0.85
R	500	-	-	-
R	1000	-	-	-
R	5000	0.93	0.93	0.87
T	500	-	-	-
T	1000	-	-	-
T	5000	-	-	-
TH	500	0.4	0.58	0.54
TH	1000	-	-	-
TH	5000	0.92	0.91	0.87
TR	500	-	-	-
TR	1000	-	-	-
TR	5000	-	-	-
TRH	500	-	-	-
TRH	1000	-	-	-
TRH	5000	0.41	0.75	0.36

- If a target user promotes/retweets someone or an organisation with a known ideological affiliation, you may assume that the target endorse them. For example, if a target user retweets a labor MP then you can label the user as 'left'.
- If a target user, has a stance against someone with a known ideological affiliation, then you might infer that the target user's ideology is the opposing ideology. For example, if a target user calls a labor MP an insult, then you can label the user as 'right'.
- If a target user expresses a view about a issue related to an ideology, you can infer the user's ideology. For example, if a user supports LGBTQ or environmental issues, then (if there is enough evidence) you may label them as 'left'.

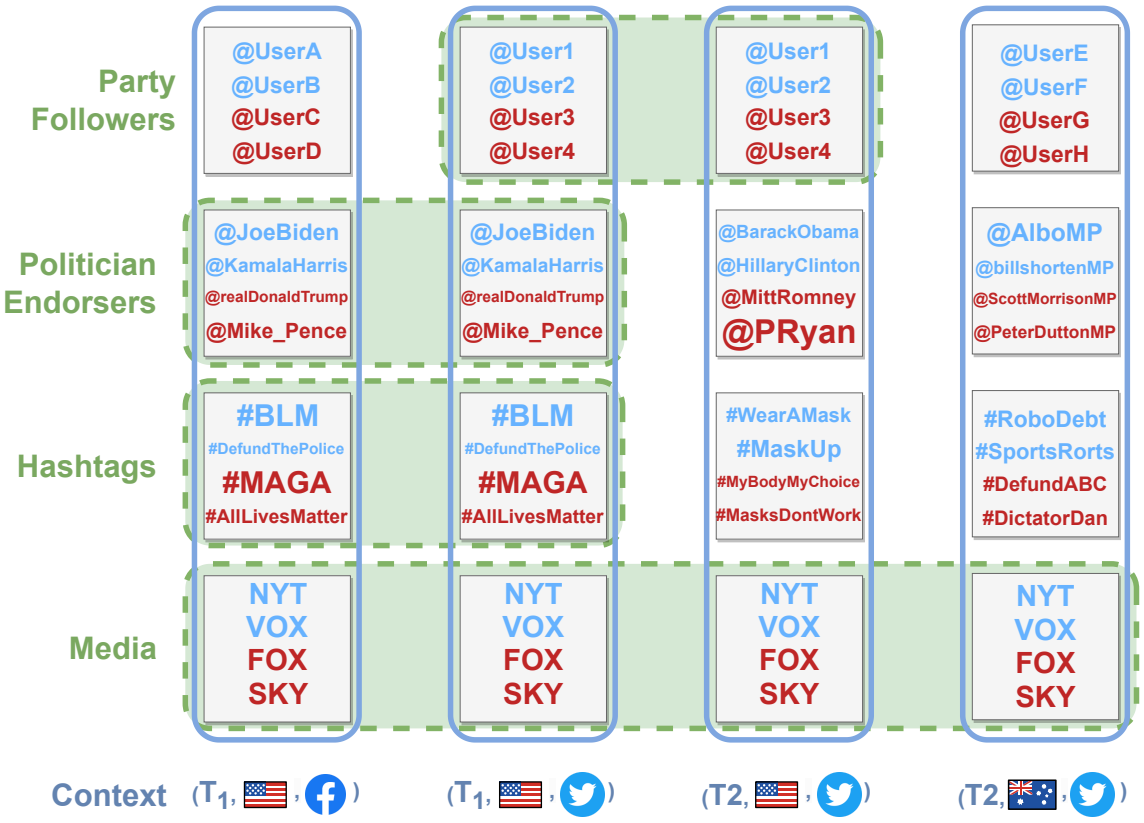


FIGURE C.1: **Most ideology proxies do not generalize across contexts.** The x-axis shows four contexts that vary in time (T_1 and T_2), country (Australia and USA), and platform (Twitter and Facebook). The y-axis show four proxies: endorsing political parties or political figures, using politically charged hashtags and the consumed media slant. The green dashed boxes indicate whether a proxy is applicable across contexts.

These guidelines aim to increase the clarity of the annotation task. In countries where political affiliation is obvert (e.g. the united states), this labelling task is often unambiguous; however, in Australia ideological signals are often implicit. The full annotation briefing material is available in the code repository [`https://github.com/behavioral-ds/ideology_prediction`].

Context-Transfer Illustration

Fig. C.1 further illustrates the difficulty with utilizing particular proxies as ground truth. We observe that some ideological proxies are consistent across only some contexts (represented by the dashed green boxes). For example, `#RoboDebt` (in relation to an Australian incident)

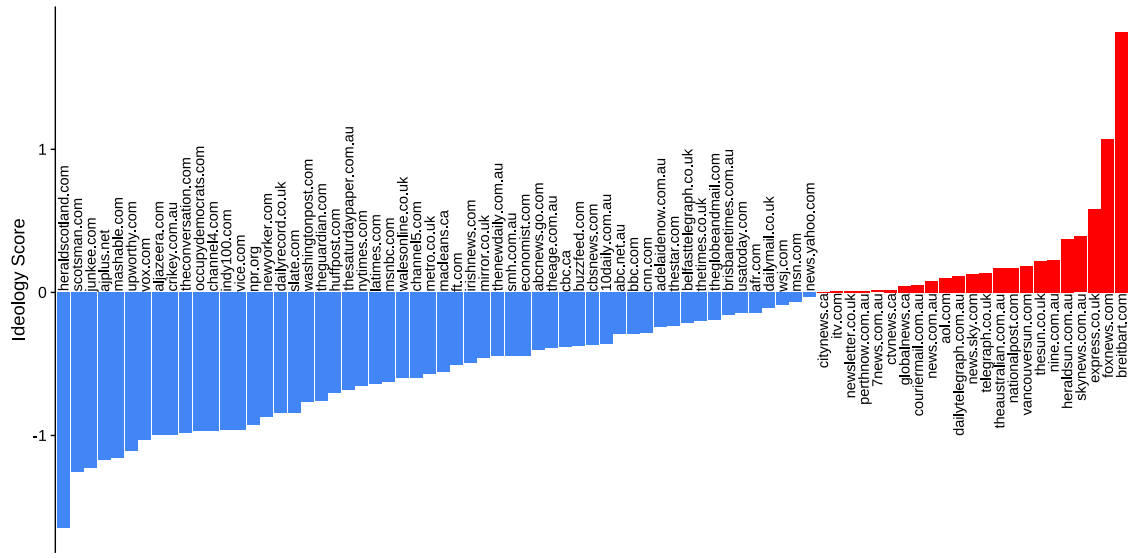


FIGURE C.2: **Media Publication Slants.** The plot shows the slants of Media Publications, as averaged over the year, country, and source point estimates.

is not relevant to the USA and did not exist before 2016; and, although `@MittRomney` signaled right-wing ideology in 2012, the right has shifted since Trump's election.

Prior ideology detection techniques fail to easily *context-switch* and cannot be readily applied to multiple distinct domains.

Media Publication Slants

The media slant scores are shown in Fig. C.2, where we observed publications like *Breitbart* and *Fox News* are extremely right-leaning, and *Vox* and *NYTimes* are left-leaning.

Cognitive Distortions Schemata Prevalence

Fig. C.3 shows the prevalence of all twelve cognitive distortions in each of the ideological groups, for #QandA. Note that many CDS n-grams are extremely rare (or do not appear), namely; *emotional reasoning* and *mental filtering*. In several CDS the left exhibit higher prevalence, such as *catastrophizing*, *fortune-telling*, *disqualifying the positive*, and *should statements*.

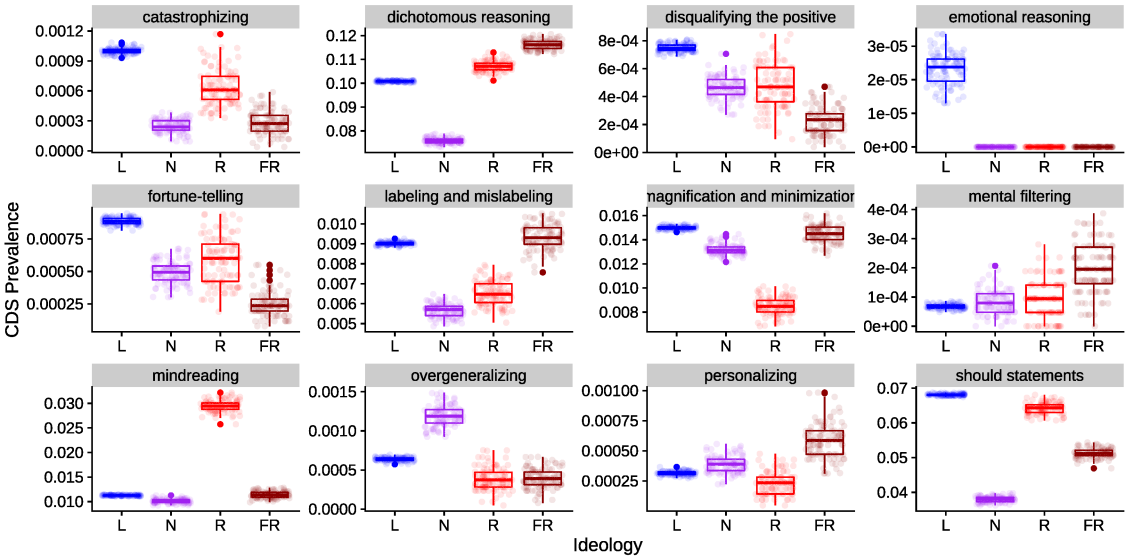


FIGURE C.3: CDS Prevalence

Flag Emoji Hurdle Model

For completeness, we present the results of the hurdle model (used to model zero-inflated count data, such a tokens in a corpus). The hurdle model is a mixed model, comprised of a logistic regression to model the presense of no emoji, and a truncated poisson with log linkage, to model the count of the emoji. Fig. C.4 shows the coefficients for each model, including the reference groups. 🇦🇺 is observed more for far-right users in both the zero and the count models. The count models for the other flags show mixed results and not significant.

Precision-Recall of Pipelines

Fig. C.5 shows precision and recall for every lens combinations and proxy.

Predicted Label Distribution

Table C.2 shows the distribution of predicted labels, to provide context for the psychosocial analysis.

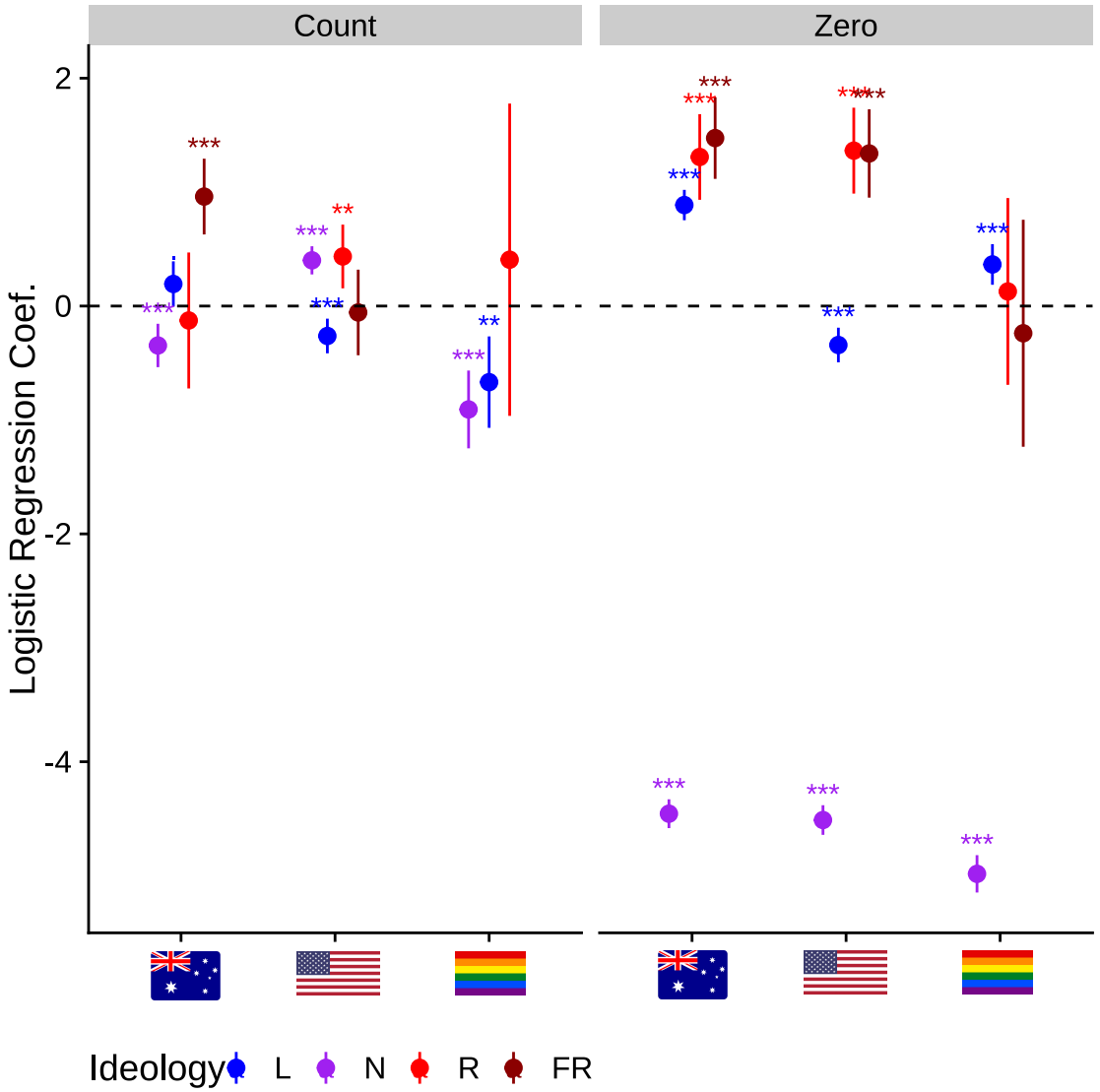


FIGURE C.4: Hurdle Model

Dataset Profiling

Activity levels are often a concern for ideology detection frameworks, given that low-activity users reveal few signals of ideology. Fig. C.6 shows the distribution of activity for users for each dataset. It shows long-tailed activity distributions and the proportion of low-activity users. Riot shows a significant proportion of low-activity users, who're often difficult to classify.

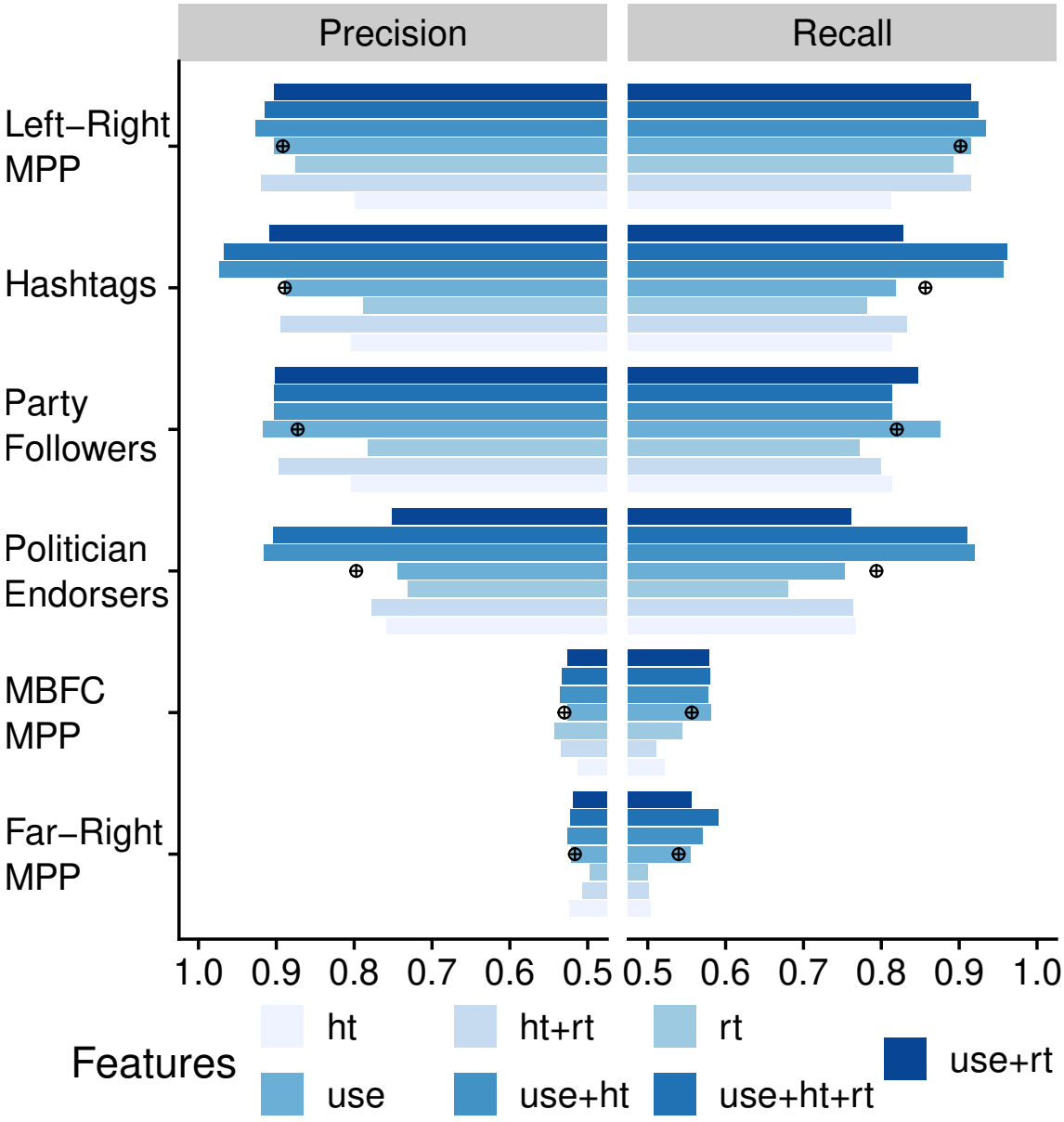


FIGURE C.5: **Precision-Recall.** The plot shows the macro-averaged precision and recall of pipelines, trained with each proxy (y-axis) and each feature set (colors), probability calibrated with the hold-out validation set for F1-macro scores.

TABLE C.2: **Distribution of Predicted Labels.** The number of users predicted to be in each class (rows) for each dataset (columns). Note that for many datasets there is a significant imbalance toward the left (except Parler which is a right-leaning platform).

	#QandA	#Ausvotes	#Socialsense	Riot	Parler
Left	80,375	189,233	48,056	339,095	293
Neutral	21,176	79,221	604	227,839	48,829

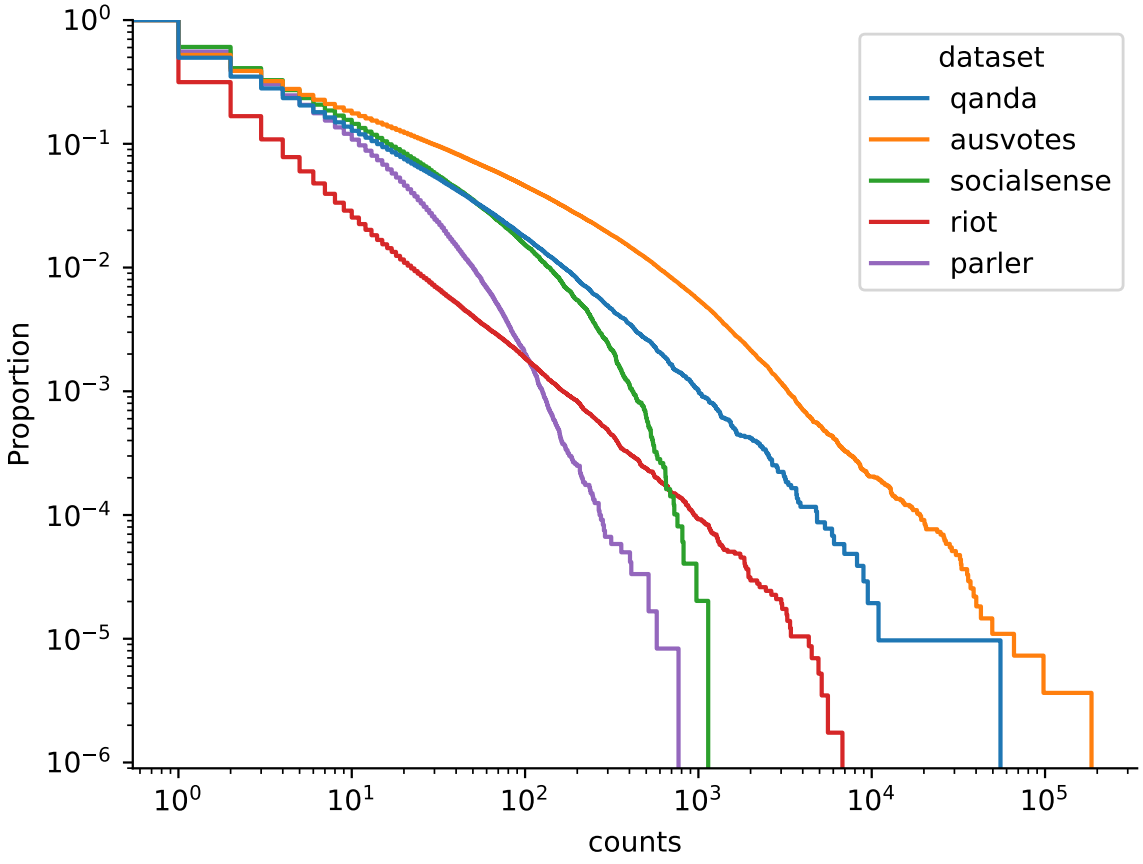


FIGURE C.6: **Activity Distribution.** The log-log ECCDF distribution of activity (number of posts per user) for each dataset.

Exhaustive Psychosocial Analysis

Grievance

This section shows the difference between ideological groups in terms of grievance categories for all available datasets.

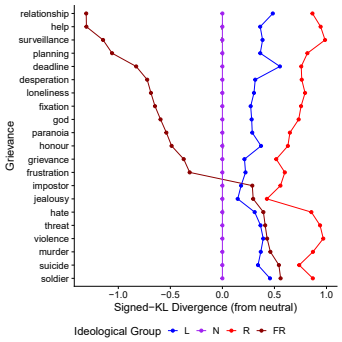


FIGURE C.7: Grievance #QandA

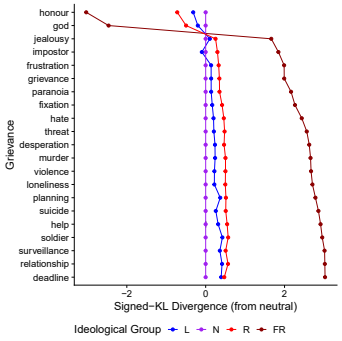


FIGURE C.8: Grievance #Ausvotes

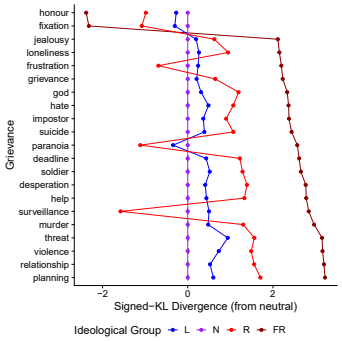


FIGURE C.9: Grievance #Socialsense

MFT

This section shows the difference between ideological groups in terms of moral foundations for all available datasets.

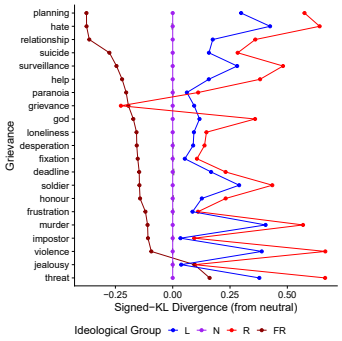


FIGURE C.10: Grievance Riot

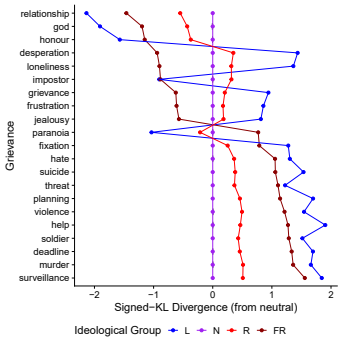


FIGURE C.11: Grievance Parler

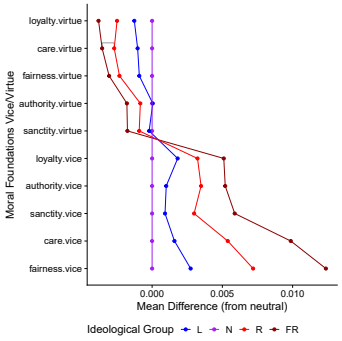


FIGURE C.12: MFT #QandA

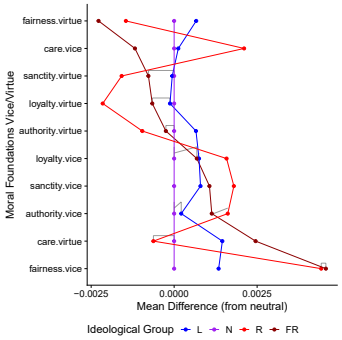


FIGURE C.13: MFT #Ausvotes

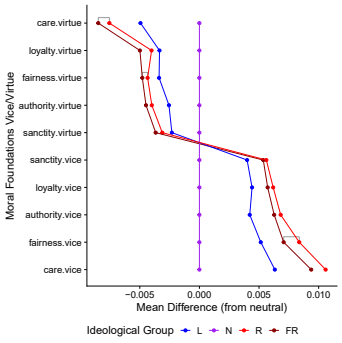


FIGURE C.14: MFT #Socialsense

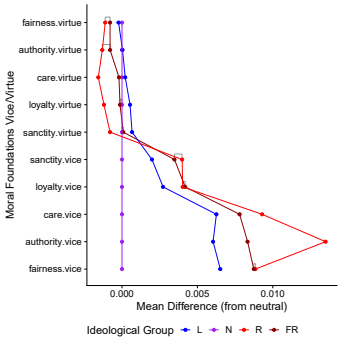


FIGURE C.15: MFT Riot

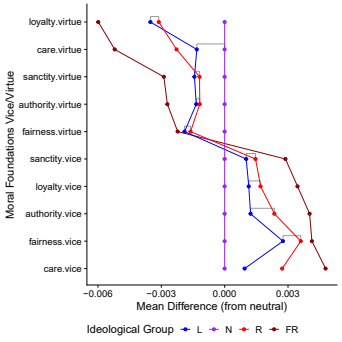


FIGURE C.16: MFT Parler

Appendix D

Birdspotter: A Tool for Analyzing and Labeling Twitter Users

Accompanying the submission *Birdspotter: A Tool for Analyzing and Labeling Twitter Users*.

D.1 Additional Related Work

In this section, we outline other approaches to bot detection and influence measurement in the literature.

Detecting Twitter bots. There have been a myriad of approaches to detect bots on Twitter. There are three motifs within the literature. The first motif are supervised methods used to determine if an individual user is a bot, usually employing feature construction. Such approaches include NLP approaches [210, 211], deep-learning approaches [212], feature-engineering [197, 213, 214] and other methods [215, 216]. The second motif are unsupervised methods used to discover coordinated online behavior/real-time online campaigns; and the third motif are adversarial methods which achieve better bot detection by generating better bots.

`birdspotter` falls in the first category. It uses a supervised approach to retrospectively analyze datasets. It satisfies a different use case than coordinated online behavior tools like `BotSlayer` [217]. Adversarial approaches are fairly novel, however it is unclear whether they might simply improve bot technology, as they provide recipes to build better bots.

The de-facto bot detection tool in the social science community is `Botometer` (formerly `BotOrNot`) [197], which uses more than 1000 user- and recent activity-related features to train a Random Forest classifier. `Botometer` is currently at version 4, at the time of writing, and serves half a million queries a day [196].

The main limitation of `botometer` for practitioners is its dependence on an online API. It cannot be used to profile the users in offline Twitter datasets which have been collected in the past (like used in many works [218–220]). Furthermore, the API is rate-limited by Twitter, and requires registration through both Twitter and *RapidAPI* service. For scientific purposes, `botometer` makes local reproducibility difficult to achieve, since deactivated, protected, and suspended users can no longer be retrieved, and `botometer` scores are likely to vary with user activity and `botometer` versioning.

`Birdspotter` addresses the above-stated shortcomings by producing bot predictions on already collected Twitter dumps, and exposing a simple interface to allows researchers to annotate their own Twitter user collection.

Tools for quantifying online influence. There are many features used to score the influence, reputation or popularity of online users. We delineate these into three areas: those using static user attributes (including lexical features and information on a user's profile) [198], those that analyze the online social graph (e.g. degree, PageRank, HITs, etc.) [3, 199], and those modeling information diffusion [200]. However, few of these have translated into accessible tools for the non-experts in the field. For instance, Cossu et al. [198] provide a set of scripts to perform their influence measurement method. Other tools, like `ConTinEst` [87, 201], require knowledge of the social graph (which is often prohibitively expensive to obtain) on which it performs random walks (which are very slow on large social graphs). `Birdspotter` estimates user influence from reshare dynamics, in the absence of knowledge about the social graph, and provides an end-to-end tool to analyze Twitter users.

D.2 Influence measure

We review the theoretical prerequisites concerning modeling reshare cascades using point processes, and estimating reshare influence.

Reshare cascades. `birdspotter` analyzes the spread of online information in the form of online *reshare cascades*. A reshare cascade consists of an initial user post and some reshare events of the post by other users. On Twitter, for example, this can happen when users use the retweet functionality. We denote a cascade observed up to time T as $\mathcal{H}(T) = \{t_0, t_1, \dots\}$ where $t_i \in \mathcal{H}(T)$ are the event times relative to the first event ($t_0 = 0$). We denote cascades with additional information about events — dubbed here as *event marks* — as marked cascades. We use the notation $\mathcal{H}_m(T) = \{(t_0, m_0), (t_1, m_1), \dots\}$, where each event is a tuple of the event time and the event mark. For example, for retweet cascades, the numbers of followers of a Twitter user are commonly adopted as event marks [66, 98, 221].

The Hawkes processes. `birdspotter` models reshare cascades using Hawkes processes [100] — a type of point processes with the self-exciting property, i.e., the occurrence of past events increases the likelihood of future events. The occurrence of events in a Hawkes process is controlled by the event intensity function:

$$\lambda(t \mid \mathcal{H}(T)) = \mu(t) + \sum_{t_i < t} \phi(t - t_i) \tag{D.1}$$

where $\mu(t)$ is the background intensity function and $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ is a kernel function capturing the decaying influence from a historical event. We note that, for reshare cascades, all events are considered to be offspring of the initial event, i.e. there is no background event rate $\mu(t) = 0$. Two widely adopted parametric forms for the kernel function ϕ include the exponential function $\phi_{EXP}(t) = \kappa\theta e^{-\theta t}$ and the power-law function $\phi_{PL}(t) = \kappa(t+c)^{-(1+\theta)}$.

Marked Models. `birdspotter` implements marked versions of the point processes, where the mark is the number of followers that the user emitting the tweet has. This is because the mark of each event governs the number of future events, e.g., a tweet from a largely followed user is likely to attract more retweets. The marked versions of Hawkes processes [66] are

then derived by rescaling the kernel functions with the marks, i.e., $\phi(m,t) = m^{\beta}\phi(t)$; β controls the warping effect of the mark.

User influence estimation. `birdspotter` adopts the following definition for user influence, widely used in literature [1, 87, 222]:

Definition 1. *Online user influence $\varphi(u)$ is defined as the mean number of reshares generated directly and indirectly by a message posted by u, irrespective if it is an original message or a reshare.*

Estimating influence from retweet cascades has the additional difficulty of not observing the branching structure of the diffusion — i.e., the Twitter API attributes all retweets to the original tweet. `birdspotter` estimates Twitter user influence using only the observed retweet cascade $\mathcal{H}_m(T) = \{v_0 = (t_0, m_0), v_1 = (t_1, m_1), \dots\}$, where marks correspond to users' number of followers.

Rizoiu et al. [1] propose a method to estimate user influence in the absence of the branching structure by assuming that retweets arrive following a Hawkes point process [101]. We can quantify the probability that an event v_j is generated by a previous event v_i as the ratio of the event intensity generated by v_i and the total intensity at time t_j. Formally, the probability v_j retweets v_i is

$$p_{ij} = \frac{\phi(t_j - t_i)}{\sum_{k=1}^{j-1} \phi(t_j - t_k)} \tag{D.2}$$

Rizoiu et al. [1] also introduce the pairwise influence score m_{ij}, intuitively defined as the amount of *influence* that v_i exerts over v_j either directly (when v_j is a direct retweet of v_i) or indirectly (when v_j is a retweet of a descendant of v_i):

$$m_{ij} = \begin{cases} \sum_{k=i}^{j-1} m_{ik} p_{kj} & , i \leq k < j \\ 1 & , i = j \\ 0 & , i > j \end{cases}, \tag{D.3}$$

Finally, the influence of v_i is $\varphi(v_i) = \sum_{k=i}^{n} m_{ik}$, and the influence of a user u is the average of the influences of all of their tweets:

$$\varphi(u) = \frac{\sum_{v \in \mathcal{T}(u)} \varphi(v)}{|\mathcal{T}(u)|} \tag{D.4}$$

where $\mathcal{T}(u)$ is the set of all the tweets emitted by user u.

Bibliography

[1] Marian-Andrei Rizoiu, Timothy Graham, Rui Zhang, Yifei Zhang, Robert Ackland, and Lexing Xie. # DebateNight: The role and influence of socialbots on twitter during the 1st 2016 US presidential debate. In *ICWSM*, 2018.

[2] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford, 1999.

[3] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna Gummadi. Measuring user influence in twitter: The million follower fallacy. In *ICWSM*, 2010.

[4] Arlei Silva, Sara Guimarães, Wagner Meira Jr, and Mohammed Zaki. ProfileRank: Finding relevant content and influential users based on information diffusion. In *SNAKDD*, 2013.

[5] Brian Hopkins and John Gordon Skellam. A new method for determining the type of distribution of plant individuals. *Annals of Botany*, 18(2):213–227, April 1954. ISSN 0305-7364. doi: 10.1093/oxfordjournals.aob.a083391.

[6] **Rohit Ram** and Marian-Andrei Rizoiu. Empirically measuring online social influence. *EPJ Data Science*, 13(1):53, 2024.

[7] **Rohit Ram**, Emma Thomas, David Kernot, and Marian-Andrei Rizoiu. Practical guidelines for ideology detection pipelines and psychosocial applications. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 1630–1648, 2025.

[8] **Rohit Ram**, Quyu Kong, and Marian-Andrei Rizoiu. Birdspotter: A tool for analyzing and labeling twitter users. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 918–921, 2021.

[9] Quyu Kong, **Rohit Ram**, and Marian-Andrei Rizoiu. Evently: Modeling and analyzing reshare cascades with hawkes processes. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 1097–1100, 2021.

[10] Pio Calderon, **Rohit Ram**, and Marian-Andrei Rizoiu. Opinion market model: Stemming far-right opinion spread using positive interventions. In *Proceedings of the 18th ACM International Conference on Web Search and Data Mining*, 2024.

[11] Quyu Kong, Pio Calderon, **Rohit Ram**, Olga Boichak, and Marian-Andrei Rizoiu. Interval-censored transformer hawkes: Detecting information operations using the reaction of social systems. In *Proceedings of the ACM Web Conference 2023*, pages 1813–1821, 2023.

[12] Solomon E Asch. Effects of group pressure upon the modification and distortion of judgments. In *Documents of Gestalt Psychology*. 1961.

[13] Bertram Herbert Raven. *Social influence and power*. University of California, Department of Psychology California (US), 1964.

[14] Mehdi Moussaïd, Juliane E Kämmer, Pantelis P Analytis, and Hansjörg Neth. Social influence and the collective dynamics of opinion formation. *PloS one*, 8(11):e78433, 2013.

[15] Bahar Tunçgenç, Marwa El Zein, Justin Sulik, Martha Newson, Yi Zhao, Guillaume Dezecache, and Ophelia Deroy. Social influence matters: We follow pandemic guidelines most when our close circle does. *Br. J. Psychol.*, 2021.

[16] Benjamin Schüz, Thalia Papadakis, and Stuart Ferguson. Situation-specific social norms as mediators of social influence on snacking. *Health Psych.*, 2018.

[17] Sancheng Peng, Yongmei Zhou, Lihong Cao, Shui Yu, Jianwei Niu, and Weijia Jia. Influence analysis in social networks: A survey. *Journal of Network and Computer Applications*, 2018.

[18] Winter A Mason, Frederica R Conrey, and Eliot R Smith. Situating social influence processes: Dynamic, multidirectional flows of influence within social networks. *Personality and social psychology review*, 2007.

[19] Alicia Cork, Richard Everson, Mark Levine, and Miriam Koschate. Using computational techniques to study social influence online. *Group Processes & Intergroup Relations*, 2020.

[20] Mitchell J Prinstein. Assessment of adolescents' preference-and reputation-based peer status using sociometric experts. *Merrill-Palmer Quarterly (1982-)*, pages 243–261, 2007.

[21] Katja Košir and Sonja Pečjak. Sociometry as a method for investigating peer relationships: What does it actually measure? *Educational Research*, 47(1):127–144, 2005.

[22] Antonius HN Cillessen and Peter EL Marks. Conceptualizing and measuring popularity. *Popularity in the peer system*, pages 25–56, 2011.

[23] Lucas Maystre and Matthias Grossglauser. Just sort it! A simple and effective approach to active preference learning. In *ICML*, 2017.

[24] Andrea E Abele and Bogdan Wojciszke. The Big Two in social judgment and behavior. 2013.

[25] Andrea E Abele and Bogdan Wojciszke. Communal and agentic content in social cognition: A dual perspective model. In *Advances in Experimental Social Psychology*. 2014.

[26] Arina Tveleneva, Christin Scholz, Emily B Falk, Carolyn Yoon, Matthew D Lieberman, Nicole Cooper, Matthew Brook O'Donnell, and Christopher N Cascio. The relationship between agency, communion, and neural processes associated with conforming to social influence. *Communion, and Neural Processes Associated with Conforming to Social Influence*, 2023.

[27] Magdalena Marszał-Wiśniewska and Magdalena Siembab. Power and the self-ascription of agency and communion. *Current Psychology*, 2012.

[28] Jeremy A Frimer, Lawrence J Walker, Brenda H Lee, Amanda Riches, and William L Dunlop. Hierarchical integration of agency and communion: A study of influential moral figures. *Journal of Personality*, 2012.

[29] Thomas C Schelling. Dynamic models of segregation. *Journal of mathematical sociology*, 1971.

[30] Michael W Macy, Manqing Ma, Daniel R Tabin, Jianxi Gao, and Boleslaw K Szymanski. Polarization and tipping points. *PNAS*, 2021.

[31] Arnout Van de Rijt. Self-correcting dynamics in social influence processes. *American journal of sociology*, 2019.

[32] Martin Gestefeld and Jan Lorenz. Calibrating an opinion dynamics model to empirical opinion distributions and transitions. *JASSS*, 2023.

[33] Robert B Cialdini. Influence: Science and practice, 2001.

[34] Antonius HN Cillessen. Sociometric methods. 2009.

[35] Yinglong Zhang, Jin Zhang, Matthew Lease, and Jacek Gwizdka. Multidimensional relevance modeling via psychometrics and crowdsourcing. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2014.

[36] Catherine Chen and Carsten Eickhoff. Evaluating search explainability with psychometrics and crowdsourcing. *arXiv preprint arXiv:2210.09430*, 2022.

[37] Emily Michelle Wetherell. *The Use of Crowdsourcing in the Development of Measurement Instruments*. PhD thesis, The University of Iowa, 2019.

[38] Sayantan Bhattacharya, Billy Spann, and Nitin Agarwal. Solidarity to storming: Assessing the socio-technical factors behind modern social movements. 2024.

[39] Mainuddin Shaik, Niloofar Yousefi, Nitin Agarwal, and Billy Spann. Evaluating role of instagram's multimedia in connective action leveraging diffusion of innovation and cognitive mobilization theories: Brazilian and peruvian social unrest case studies. In *BESC*, 2023.

[40] Rui Liu, Kevin T Greene, Ruibo Liu, Mihovil Mandic, Benjamin A Valentino, Soroush Vosoughi, and VS Subrahmanian. Using impression data to improve models of online social influence. *Nature Scientific reports*, 2021.

[41] Douglas Guilbeault, Andrea Baronchelli, and Damon Centola. Experimental evidence for scale-induced category convergence across populations. *Nature communications*, 2021.

[42] Ines Arous, Jie Yang, Mourad Khayati, and Philippe Cudré-Mauroux. Opencrowd: A human-ai collaborative approach for finding social influencers via open-ended answers aggregation. In *WWW*, 2020.

[43] Ailbhe Finnerty, Pavel Kucherbaev, Stefano Tranquillini, and Gregorio Convertino. Keep it simple: Reward and task design in crowdsourcing. In *Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI*, 2013.

[44] Bahareh Rahmanian and Joseph G Davis. User interface design for crowdsourcing systems. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces*, 2014.

[45] Turki Alelyani, Paul T Grogan, Yla Tausczik, and Ye Yang. Software crowdsourcing design: An experiment on the relationship between task design and crowdsourcing performance. In *HCI International 2020–Late Breaking Papers: Interaction, Knowledge and Social Media: 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22*, 2020.

[46] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? *Perspect Psychol Sci.*, 2016.

[47] Nir Ailon. Reconciling real scores with binary comparisons: A new logistic based model for ranking. *NIPS*, 2008.

[48] R Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 22 140:55, 1932.

[49] Kristi Tsukida and Maya R Gupta. How to analyze paired comparison data. Technical report, WASHINGTON UNIV SEATTLE DEPT OF ELECTRICAL ENGINEERING, 2011.

[50] Nihar B Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramchandran, and Martin J Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *JMLR*, 2016.

[51] Andrew P Clark, Kate L Howard, Andy T Woods, Ian S Penton-Voak, and Christof Neumann. Why rate when you could compare? using the "elochoice" package to assess pairwise comparisons of perceived physical strength. *PloS one*, 2018.

[52] Maria Perez-Ortiz and Rafal K Mantiuk. A practical guide and software for analysing pairwise comparison experiments. *arXiv preprint arXiv:1712.03686*, 2017.

[53] Alexander Peysakhovich, Virot Chiraphadhanakul, and Michael Bailey. Pairwise choice as a simple and robust method for inferring ranking data. In *WWW 2015 Conference Proceedings*, 2015.

[54] Patrick Mair. *Modern Psychometrics with R.* 2018.

[55] Manuela Cattelan. Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science*, 2012.

[56] Charles AR Hoare. Quicksort. *The Computer Journal*, 1962.

[57] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 1952.

[58] Ernst Zermelo. Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 1929.

[59] Manuela Cattelan, Cristiano Varin, and David Firth. Dynamic Bradley–Terry modelling of sports tournaments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2013.

[60] Gustav Theodor Fechner. Elements of psychophysics. 1948.

[61] Louis Leon Thurstone. The measurement of values. *Psychological review*, 1954.

[62] Burr Settles. Active learning literature survey. 2009.

[63] Richard Lenton. *Using the Method of Paired Comparisons in Non-Designed Experiments*. PhD thesis, Griffith University, 2006.

[64] Lucas Maystre and Matthias Grossglauser. Fast and accurate inference of plackett–luce models. *NIPS*, 2015.

[65] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. Everyone's an influencer: Quantifying influence on twitter. In *WSDM*, 2011.

[66] Swapnil Mishra, Marian-Andrei Rizoiu, and Lexing Xie. Feature driven and point process approaches for popularity prediction. In *CIKM*, 2016.

[67] T Graham and TR Keller. Bushfires, bots and arson claims: Australia flung in the global disinformation spotlight. *The Conversation*, 10, 2020.

[68] Marco Lui and Timothy Baldwin. Langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, 2012.

[69] Google. CLD3, August 2020.

[70] Steven T Smith, Edward K Kao, Danelle C Shah, Olga Simek, and Donald B Rubin. Influence estimation on social media networks using causal inference. In *Ieee Ssp*, 2018.

[71] Daniel M Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A Huberman. Influence and passivity in social media. In *Ecml Pkdd*, 2011.

[72] Lanqin Yuan and Marian-Andrei Rizoiu. Detect Hate Speech in Unseen Domains using Multi-Task Learning: A Case Study of Political Public Figures. aug 2022. URL `http://arxiv.org/abs/2208.10598`.

[73] Andrew Law. *Exposing the Stance of Reddit Users Towards Brexit*. PhD thesis, The Australian National University, 2021.

[74] Terrill L Frantz, Marcelo Cataldo, and Kathleen M Carley. Robustness of centrality measures under uncertainty: Examining the role of network topology. *Computational and Mathematical Organization Theory*, 15(4):303–328, 2009.

[75] Jeff Riddell, Alisha Brown, Ivor Kovic, and Joshua Jauregui. Who are the most influential emergency physicians on Twitter? *Western Journal of Emergency Medicine*, 2017.

[76] Ning Hsu, Katie L Badura, Daniel A Newman, and Mary Eve P Speach. Gender,"masculinity," and "femininity": A meta-analytic review of gender differences in agency and communion. *Psychological Bulletin*, 2021.

[77] Anna Lisa Aydin, Johannes Ullrich, Birte Siem, Kenneth D Locke, and Nurit Shnabel. The effect of social class on agency and communion: Reconciling identity-based and rank-based perspectives. *Social Psychological and Personality Science*, 2019.

[78] Andrea E Abele and Susanne Bruckmüller. The big two of agency and communion in language and communication. *Social cognition and communication*, 2013.

[79] Agnieszka Pietraszkiewicz, Magdalena Formanowicz, Marie Gustafsson Sendén, Ryan L Boyd, Sverker Sikström, and Sabine Sczesny. The big two dictionaries: Capturing agency and communion in natural language. *European journal of social psychology*, 2019.

[80] Jochen E Gebauer, Gregory R Maio, and Ali Pakizeh. Feeling torn when everything seems right: Semantic incongruence causes felt ambivalence. *Personality and Social Psychology Bulletin*, 2013.

[81] Carolin Rapp. Moral opinion polarization and the erosion of trust. *Social science research*, 58:34–45, 2016.

[82] Clark McCauley and Sophia Moskalenko. Mechanisms of political radicalization: Pathways toward terrorism. *Terrorism and political violence*, 20(3):415–433, 2008.

[83] Andrzej Nowak, Robin R Vallacher, Marek Kus, and Jakub Urbaniak. The dynamics of societal transition: Modeling nonlinear change in the Polish economic system. *International Journal of Sociology*, 2005.

[84] Hedy Greijdanus, Carlos A de Matos Fernandes, Felicity Turner-Zwinkels, Ali Honari, Carla A Roos, Hannes Rosenbusch, and Tom Postmes. The psychology of online

activism and social movements: Relations between online and offline collective action. *Current opinion in psychology*, 2020.

[85] Kate Gunton. The impact of the internet and social media platforms on radicalisation to terrorism and violent extremism. In *Privacy, Security And Forensics in The Internet of Things (IoT)*. 2022.

[86] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *KDD*, 2003.

[87] Nan Du, Le Song, Manuel Gomez-Rodriguez, and Hongyuan Zha. Scalable influence estimation in continuous-time diffusion networks. In *NIPS*, 2013.

[88] Michael W Kraus, Bennett Callaghan, and Peter Ondish. Social class as culture. 2019.

[89] Theodore M Newcomb. An approach to the study of communicative acts. *Psychological review*, 1953.

[90] Stanley Milgram and Christian Gudehus. Obedience to authority, 1978.

[91] Joann Horai, Nicholas Naccari, and Elliot Fatoullah. The effects of expertise and physical attractiveness upon opinion agreement and liking. *Sociometry*, 1974.

[92] Faisal Ghaffar and Neil Hurley. Structural hole centrality: Evaluating social capital through strategic network formation. *Computational Social Networks*, 2020.

[93] Luke Sloan, Jeffrey Morgan, Pete Burnap, and Matthew Williams. Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. *PloS one*, 2015.

[94] National Center for O*NET Development. O*NET OnLine,.

[95] Limeng Cui and Dongwon Lee. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*, 2020.

[96] Panagiotis Metaxas, Eni Mustafaraj, Kily Wong, Laura Zeng, Megan O'Keefe, and Samantha Finn. What do retweets indicate? Results from user survey and meta-review of research. In *ICWSM*, 2015.

[97] Qi Liu, Biao Xiang, Nicholas Jing Yuan, Enhong Chen, Hui Xiong, Yi Zheng, and Yu Yang. An influence propagation view of pagerank. *TKDD*, 2017.

[98] Swapnil Mishra, Marian-Andrei Rizoiu, and Lexing Xie. Modeling popularity in asynchronous social media streams with recurrent neural networks. In *ICWSM*, 2018.

[99] Maximilian Nickel and Matthew Le. Modeling sparse information diffusion at scale via lazy multivariate hawkes processes. In *WWW*, 2021.

[100] Alan G Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 1971.

[101] Marian-Andrei Rizoiu, Young Lee, Swapnil Mishra, and Lexing Xie. A tutorial on hawkes processes for events in social media. In *Research Frontiers of Multimedia*. 2017.

[102] Alan G Hawkes and David Oakes. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 1974.

[103] Erik Lewis and George Mohler. A nonparametric EM algorithm for multiscale Hawkes processes. *Journal of Nonparametric Statistics*, 2011.

[104] Luca Luceri, Torsten Braun, and Silvia Giordano. Analyzing and inferring human real-life behavior through online social networks with social influence deep learning. *Applied network science*, 2019.

[105] Jiezhong Qiu, Jian Tang, Hao Ma, Yuxiao Dong, Kuansan Wang, and Jie Tang. Deepinf: Social influence prediction with deep learning. In *KDD*, 2018.

[106] Yupeng Gu, Yizhou Sun, Yanen Li, and Yang Yang. Rare: Social rank regulated large-scale network embedding. In *WWW*, 2018.

[107] Anton Tsitsulin, Davide Mottin, Panagiotis Karras, and Emmanuel Müller. Verse: Versatile graph embeddings from similarity measures. In *WWW*, 2018.

[108] Wenhui Yu and Zheng Qin. Spectrum-enhanced pairwise learning to rank. In *WWW*, 2019.

[109] Marian-Andrei Rizoiu, Swapnil Mishra, Quyu Kong, Mark Carman, and Lexing Xie. Sir-hawkes: Linking epidemic models and hawkes processes to model diffusions in finite populations. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, pages 419–428. ACM Press, 2018. ISBN 9781450356398. doi: 10.1145/3178876.3186108.

[110] Lynda C Lin, Yang Qu, and Eva H Telzer. Intergroup social influence on emotion processing in the brain. *PNAS*, 2018.

[111] David Crandall, Dan Cosley, Daniel Huttenlocher, Jon Kleinberg, and Siddharth Suri. Feedback effects between similarity and social influence in online communities. In *KDD*, 2008.

[112] Rahul Goel, Sandeep Soni, Naman Goyal, John Paparrizos, Hanna Wallach, Fernando Diaz, and Jacob Eisenstein. The social dynamics of language change in online networks. In *International Conference on Social Informatics*, 2016.

[113] John Moody. Fast learning in multi-resolution hierarchies. In *NIPS*, 1989.

[114] Mark Granovetter. The strength of weak ties: A network theory revisited. *Sociological theory*, 1983.

[115] Paul Dekker and Eric M Uslaner. *Social Capital and Participation in Everyday Life*. Routledge, Oxfordshire, 2003.

[116] Lloyd S Shapley. A value for n-person games. *Classics in game theory*, 69, 1997.

[117] Matthew O Jackson and Asher Wolinsky. A strategic model of social and economic networks. In *Networks and Groups*, pages 23–49. Springer, New York, 2003.

[118] Karthik Subbian, Dhruv Sharma, Zhen Wen, and Jaideep Srivastava. Finding influencers in networks using social capital. *Soc. Net. Analysis & Mining*, 2014.

[119] Online Appendix. Appendix:, 2021.

[120] Biao Xiang, Qi Liu, Enhong Chen, Hui Xiong, Yi Zheng, and Yu Yang. Pagerank with priors: An influence propagation perspective. In *IJCAI*, 2013.

[121] Siqi Wu, Marian-Andrei Rizoiu, and Lexing Xie. Estimating attention flow in online video networks. *Proceedings of the ACM on Human-Computer Interaction*, 2019.

[122] Francesco Bailo, Amelia Johns, and Marian-Andrei Rizoiu. Riding information crises: The performance of far-right Twitter users in Australia during the 2019–2020 bushfires and the COVID-19 pandemic. *Information, Communication & Society*, pages 1–19, April 2023. ISSN 1369-118X. doi: 10.1080/1369118X.2023.2205479.

[123] Margaret L Kern, Paul X McCarthy, Deepanjan Chakrabarty, and Marian-Andrei Rizoiu. Social media-predicted personality traits and values can help match people to their ideal jobs. *PNAS*, 2019.

[124] Arthur Turrell, Bradley Speigner, Jyldyz Djumalieva, David Copple, and James Thurgood. Using job vacancies to understand the effects of labour market mismatch on UK output and productivity. 2018.

[125] Sourav Mukherjee, David Widmark, Vince DiMascio, and Tim Oates. Determining standard occupational classification codes from job descriptions in immigration petitions. In *ICDMW*, 2021.

[126] Lisa Singh, Leticia Bode, Ceren Budak, Kornraphop Kawintiranon, Colton Padden, and Emily Vraga. Understanding high-and low-quality URL Sharing on COVID-19 Twitter streams. *Journal of computational social science*, 2020.

[127] Quyu Kong, Emily Booth, Francesco Bailo, Amelia Johns, and Marian-Andrei Rizoiu. Slipping to the extreme: A mixed method to explain how extreme opinions infiltrate online discussions. In *ICWSM*, 2022.

[128] J Christopher Cohrs. Ideological bases of violent conflict., 2012.

[129] Hon. Jim Carr, Raquel Dancho, Kristina Michaud, Paul Chiang, Pam Damoff, Dane Lloyd, Alistair MacGregor, Ron McKinnon, Taleeb Noormohamed, Peter Schiefke, Doug Shipley, Tako Van Popta, and Sameer Zuberi. Rise of ideologically motivated violent extremism in canada. Technical report, Standing Committee on Public Safety and National Security, House of Commons, Canada, 2022.

[130] Clark McCauley and Sophia Moskalenko. Mechanisms of political radicalization: Pathways toward terrorism. *Terrorism and Political Violence*, 2008.

[131] Emily Booth, Jooyoung Lee, Marian-Andrei Rizoiu, and Hany Farid. Conspiracy, misinformation, radicalisation: understanding the online pathway to indoctrination and opportunities for intervention. *Journal of Sociology*, 2024.

[132] Kenan Alkiek, Bohan Zhang, and David Jurgens. Classification without (proper) representation: Political heterogeneity in social media and its implications for classification and behavioral analysis. In *ACL*, 2022.

[133] Raviv Cohen and Derek Ruths. Classifying political orientation on Twitter: It's not easy! In *ICWSM*, 2013.

[134] Kareem Darwish, Peter Stefanov, Michaël Aupetit, and Preslav Nakov. Unsupervised user stance detection on twitter. In *ICWSM*, 2020.

[135] Ammar Rashed, Mucahid Kutlu, Kareem Darwish, Tamer Elsayed, and Cansın Bayrak. Embeddings-based clustering for target specific stances: The case of a polarized turkey. In *ICWSM*, 2021.

[136] Zhiping Xiao, Weiping Song, Haoyan Xu, Zhicheng Ren, and Yizhou Sun. TIMME: Twitter ideology-detection via multi-task multi-relational embedding. In *KDD*, 2020.

[137] Silvan Tomkins. Left and right: A basic dimension of ideology and personality. 1963.

[138] John T Jost. Asymmetries abound: Ideological differences in emotion, partisanship, motivated reasoning, social network structure, and political trust. *Journal of Consumer Psychology*, 2017.

[139] Jesse Graham, Jonathan Haidt, and Brian A Nosek. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 2009.

[140] Lazar Stankov. From social conservatism and authoritarian populism to militant right-wing extremism. *Personality and Individual Differences*, 175:110733, June 2021. ISSN 01918869. doi: 10.1016/j.paid.2021.110733.

[141] Isabelle Van der Vegt, Maximilian Mozes, Bennett Kleinberg, and Paul Gill. The grievance dictionary: Understanding threatening language use. *Behavior research methods*, 2021.

[142] Brooke Auxier and Monica Anderson. Social media use in 2021. *Pew Research Center*, 2021.

[143] Younes Samih and Kareem Darwish. A few topical tweets are enough for effective user stance detection. In *ACL*, 2021.

[144] Haewoon Kwak, Jisun An, Elise Jing, and Yong-Yeol Ahn. FrameAxis: Characterizing microframe bias and intensity with word embedding. *PeerJ Computer Science*, 2021.

[145] Meysam Alizadeh, Ingmar Weber, Claudio Cioffi-Revilla, Santo Fortunato, and Michael Macy. Psychology and morality of political extremists: Evidence from Twitter language analysis of alt-right and Antifa. *EPJ Data Science*, 2019.

[146] Nan Xi, Di Ma, Marcus Liou, Zachary C Steinert-Threlkeld, Jason Anastasopoulos, and Jungseock Joo. Understanding the political ideology of legislators from social media images. In *ICWSM*, 2020.

[147] Daniel Preoţiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. Beyond binary labels: Political ideology prediction of twitter users. In *ACL*, 2017.

[148] Souvic Chakraborty, Pawan Goyal, and Animesh Mukherjee. Fast few shot self-attentive semi-supervised political inclination prediction. In *ICADL*, 2022.

[149] Keith T Poole and Howard Rosenthal. A spatial model for legislative roll call analysis. *American journal of political science*, 1985.

[150] Yupeng Gu, Ting Chen, Yizhou Sun, and Bingyu Wang. Ideology detection for twitter users with heterogeneous types of links. *arXiv preprint arXiv:1612.08207*, 2016.

[151] Sean O'Hagan and Aaron Schein. Measurement in the age of llms: An application to ideological scaling. *arXiv preprint arXiv:2312.09203*, 2023.

[152] Pablo Barberá. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political analysis*, 2015.

[153] Vivienne Badaan, Mark Hoffarth, Caroline Roper, Taurean Parker, and John T Jost. Ideological asymmetries in online hostility, intimidation, obscenity, and prejudice. *Scientific reports*, 2023.

[154] Tristan JB Cann, Iain S Weaver, and Hywel TP Williams. Ideological biases in social sharing of online information about climate change. *Plos one*, 2021.

[155] Gregory Eady, Richard Bonneau, Joshua A Tucker, and Jonathan Nagler. News sharing on social media: Mapping the ideology of news media content, citizens, and politicians. 2020.

[156] Kamalakkannan Ravi, Adan Ernesto Vela, and Rickard Ewetz. Classifying the ideological orientation of user-submitted texts in social media. In *ICMLA*, 2022.

[157] Abeer Aldayel and Walid Magdy. Your stance is exposed! analysing possible factors for stance detection on social media. *CSCW*, 2019.

[158] Emma F Thomas, Nathan Leggett, David Kernot, Lewis Mitchell, Saranzaya Magsarjav, and Nathan Weber. Reclaim the beach: How offline events shape online interactions and networks amongst those who support and oppose right-wing protest. *Studies in Conflict & Terrorism*, 2022.

[159] Songtao Liu, Ziling Luo, Minghua Xu, LiXiao Wei, Ziyao Wei, Han Yu, Wei Xiang, and Bang Wang. Ideology takes multiple looks: A high-quality dataset for multifaceted ideology detection. In *EMNLP*, 2023.

[160] Angela Lai, Megan A Brown, James Bisbee, Joshua A Tucker, Jonathan Nagler, and Richard Bonneau. Estimating the ideology of political youtube videos. *Political Analysis*, 2022.

[161] Preethi Lahoti, Kiran Garimella, and Aristides Gionis. Joint non-negative matrix factorization for learning ideological leaning on twitter. In *WSDM*, 2018.

[162] Julie Jiang, Xiang Ren, and Emilio Ferrara. Retweet-BERT: Political leaning detection using language features and information diffusion on social networks. *ICWSM*, 2023.

[163] Adam Badawy, Kristina Lerman, and Emilio Ferrara. Who falls for online political manipulation? In *WWW*, 2019.

[164] Akshay R Rao. Red, blue and purple states of mind: Segmenting the political marketplace. *Journal of Consumer Psychology*, 2017.

[165] Ashwin Rao, Fred Morstatter, and Kristina Lerman. Partisan asymmetries in exposure to misinformation. *Scientific reports*, 2022.

[166] Markus Reiter-Haas, Simone Kopeinik, and Elisabeth Lex. Studying moral-based differences in the framing of political tweets. In *ICWSM*, 2021.

[167] Negar Mokhberian, Andrés Abeliuk, Patrick Cummings, and Kristina Lerman. Moral framing and ideological bias of news. In *SocInfo*, 2020.

[168] Sze-Yuh Nina Wang and Yoel Inbar. Moral-language use by US political elites. *Psychological Science*, 2021.

[169] Markus Kemmelmeier and David G Winter. Sowing patriotism, but reaping nationalism? Consequences of exposure to the American flag. *Political Psychology*, 2008.

[170] Jinhang Li, Giorgos Longinos, Steven Wilson, and Walid Magdy. Emoji and self-identity in Twitter bios. In *NLP+CSS*, 2020.

[171] Ankit Kariryaa, Simon Rundé, Hendrik Heuer, Andreas Jungherr, and Johannes Schöning. The role of flag emoji in online political communication. *Social Science Computer Review*, 2022.

[172] Krishna C Bathina, Marijn Ten Thij, Lorenzo Lorenzo-Luaces, Lauren A Rutter, and Johan Bollen. Individuals with depression express more distorted thinking on social media. *Nature Human Behaviour*, 2021.

[173] Patrick Herman Meyer. Political ideology and black-and-white thinking. 2020.

[174] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

[175] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *NeurIPS*, 2017.

[176] Chi Wang, Qingyun Wu, Markus Weimer, and Erkang Zhu. FLAML: A fast and lightweight automl library. *MLSys*, 2021.

[177] Enrico Liscio, Oscar Araque, Lorenzo Gatti, Ionut Constantinescu, Catholijn M Jonker, Kyriaki Kalimeri, and Pradeep K Murukannaiah. What does a text classifier learn about morality? an explainable method for cross-domain comparison of moral rhetoric. In *ACL*, 2023.

[178] Kiran Garimella, Tim Smith, Rebecca Weiss, and Robert West. Political polarization in online news consumption. In *ICWSM*, 2021.

[179] Eytan Bakshy, Solomon Messing, and Lada A Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 2015.

[180] Jisun An, Daniele Quercia, and Jon Crowcroft. Partisan sharing: Facebook evidence and societal consequences. In *COSN*, 2014.

[181] Livia van Vliet, Petter Törnberg, and Justus Uitermark. The Twitter parliamentarian database: Analyzing Twitter politics across 26 countries. *PLoS one*, 2020.

[182] Sora Park, Caroline Fisher, Kieran McGuinness, Jee Young Lee, and Kerry McCallum. *Digital News Report: Australia 2021*. News and Media Research Centre, 2021.

[183] Nic Newman, Richard Fletcher, Anne Schulz, Simge Andi, Craig T Robertson, and Rasmus Kleis Nielsen. Reuters Institute digital news report 2021. *Reuters Institute for the Study of Journalism*, 2021.

[184] AllSides. AllSides media bias ratings, 2022.

[185] D Zandt. Media Bias/Fact check, 2022.

[186] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 2001.

[187] Aleksandra Cichocka, Michał Bilewicz, John T Jost, Natasza Marrouch, and Marta Witkowska. On the grammar of politics—or why conservatives prefer nouns. *Political Psychology*, 2016.

[188] Leticia Bode, Alexander Hanna, Ben Sayre, JungHwan Yang, and Dhavan V Shah. Mapping the political Twitterverse: Finding connections between political elites. 2013.

[189] Livia Van Vliet, Petter Törnberg, and Justus Uitermark. Political systems and political networks: The structure of parliamentarians' retweet networks in 19 countries. *International Journal of Communication*, 2021.

[190] Núria Macià, Albert Orriols-Puig, and Ester Bernadó-Mansilla. Genetic-based synthetic data sets for the analysis of classifiers behavior. In *HAIS*, 2008.

[191] Ana C Lorena, Luís PF Garcia, Jens Lehmann, Marcilio CP Souto, and Tin Kam Ho. How complex is your classification problem? a survey on measuring classification complexity. *CSUR*, 2019.

[192] Daniel Kerchner and Laura Wrubel. U.S. capitol riot and #TrumpRally tweet IDs. 2021.

[193] Max Aliapoulios, Emmi Bevensee, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, and Savvas Zannettou. A large open dataset from the parler social network. January 2021. doi: 10.5281/zenodo.4442460.

[194] Courtney Radsch. Media development and countering violent extremism: An uneasy relationship, a need for dialogue. *Center for International Media Assistance*, 2016.

[195] Michelle Betz. Constraints and opportunities: what role for media development in countering violent extremism? 2016.

[196] Mohsen Sayyadiharikandeh, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. Detection of novel social bots by ensembles of specialized classifiers. *CIKM*, 2020.

[197] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Botornot: A system to evaluate social bots. In *WWW*, 2016.

[198] Jean-Valère Cossu, Vincent Labatut, and Nicolas Dugué. A review of features for the discrimination of twitter users: Application to the prediction of offline influence. *SNAM*, 2016.

[199] Fabián Riquelme and Pablo González-Cantergiani. Measuring user influence on Twitter: A survey. *Information processing & management*, 2016.

[200] Zizhu Zhang, Weiliang Zhao, Jian Yang, Cecile Paris, and Surya Nepal. Learning influence probabilities and modelling influence diffusion in twitter. In *WWW*, 2019.

[201] Manuel Gomez-Rodriguez, Le Song, Nan Du, Hongyuan Zha, and Bernhard Schölkopf. Influence estimation and maximization in continuous-time diffusion networks. *ACM Transactions on Information Systems*, 2016.

[202] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. In *LREC 2018*, 2018.

[203] Kai-Cheng Yang, Onur Varol, Clayton A Davis, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, 2019.

[204] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *KDD'16*, 2016.

[205] Kai-Cheng Yang, Onur Varol, Pik-Mai Hui, and Filippo Menczer. Scalable and generalizable social bot detection through data selection. In *AAAI*, 2020.

[206] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NeurIPS'17*. 2017.

[207] Emily Chen, Kristina Lerman, and Emilio Ferrara. Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus twitter data set. *JMIR*, 2020.

[208] Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. Transformer hawkes process. In *ICML*, 2020.

[209] Wikipedia. Wikipedia: Q+a (australian talk show), 2023.

[210] Jürgen Knauth. Language-agnostic twitter-bot detection. In *RANLP 2019*, 2019.

[211] Eric M Clark, Jake Ryland Williams, Chris A Jones, Richard A Galbraith, Christopher M Danforth, and Peter Sheridan Dodds. Sifting robotic from organic text: A natural language approach for detecting automation on Twitter. *Journal of Computational Science*, 2016.

[212] Sneha Kudugunta and Emilio Ferrara. Deep neural networks for bot detection. *Information Sciences*, 2018.

[213] Zi Chu, Steven Gianvecchio, Aaron Koehl, Haining Wang, and Sushil Jajodia. Blog or block: Detecting blog bots through behavioral biometrics. *Computer Networks*, 2013.

[214] Zhi Yang, Christo Wilson, Xiao Wang, Tingting Gao, Ben Y Zhao, and Yafei Dai. Uncovering social network sybils in the wild. *TKDD*, 2014.

[215] Michele Mazza, Stefano Cresci, Marco Avvenuti, Walter Quattrociocchi, and Maurizio Tesconi. RTbust: Exploiting temporal patterns for botnet detection on twitter. In *Proceedings of the 10th ACM Conference on Web Science*, 2019.

[216] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *Comm. ACM*, 2016.

[217] Pik-Mai Hui, Kai-Cheng Yang, Christopher Torres-Lugo, Zachary Monroe, Marc McCarty, Benjamin Serrette, Valentin Pentchev, and Filippo Menczer. BotSlayer: Real-time detection of bot amplification on Twitter. *Journal of Open Source Software*, 2019.

[218] Stefan Wojcik, Solomon Messing, Aaron Smith, Lee Rainie, and Paul Hitlin. Bots in the twittersphere. *Pew Research Center. Retrieved May*, 2018.

[219] Alessandro Bessi and Emilio Ferrara. Social bots distort the 2016 US Presidential election online discussion. *First Monday*, 2016.

[220] Emilio Ferrara. # covid-19 on twitter: Bots, conspiracies, and social media activism. *arXiv*, 2020.

[221] Qingyuan Zhao, Murat A. Erdogdu, Hera Y. He, Anand Rajaraman, and Jure Leskovec. SEISMIC: A self-exciting point process model for predicting tweet popularity. In *KDD*, 2015.

[222] Ali Zarezade, Utkarsh Upadhyay, Hamid R Rabiee, and Manuel Gomez-Rodriguez. Redqueen: An online algorithm for smart broadcasting in social networks. In *WSDM*. ACM, 2017.