Ph.D. Thesis 2023
ISBN 978-87-94336-84-0

Laura Jahn

Curbing Amplification Online
Towards Improving the Quality of Information Spread on Social Media
Using Agent-Based Models and Twitter Data

Laura Jahn

Curbing Amplification Online

2023

UNIVERSITY OF COPENHAGEN
FACULTY OF HUMANITIES

SEMPER ARDENS
Sponsored by Carlsberg Foundation
CENTER FOR INFORMATION AND BUBBLE STUDIES
UNIVERSITY OF COPENHAGEN

Laura Jahn
Center for Information and Bubble Studies

Curbing Amplification Online

Towards Improving the Quality of Information Spread on Social Media
Using Agent-Based Models and Twitter Data

Curbing Amplification Online

Curbing Amplification Online

Towards Improving the Quality of Information Spread on Social Media
Using Agent-Based Models and Twitter Data

by Laura Jahn

UNIVERSITY OF
COPENHAGEN

Thesis for the degree of Doctor of Philosophy
Thesis supervisors:
Prof. Vincent F. Hendricks & Assistant Prof. Rasmus K. Rendsvig
Assessment committee:
Associate Prof. Yong-Yeol Ahn & Prof. Olivier Roy
Chair of the committee:
Prof. Irina Shklovski

Curbing Amplification Online

Towards Improving the Quality of Information Spread on Social Media Using Agent-Based Models and Twitter Data

by Laura Jahn

UNIVERSITY OF
COPENHAGEN

Funding information: The Carlsberg Foundation is gratefully acknowledged for funding the work on this doctoral thesis within the project 'Attention Economics and Social Influence' under The Center for Information and Bubble Studies, Faculty of Humanties, University of Copenhagen, [CF15-0212] (Principal Investigator Vincent F. Hendricks).

Cover Illustration: Shadow of a bridge in Ubbedissen, Bielefeld, by Klaus Jahn.

Faculty of Humanities, Department of Communication, Center for Information and Bubble Studies

Contents

List of Papers

This thesis includes the following papers, referred to by their Roman numerals.

I **Detecting Coordinated Inauthentic Behavior in Likes on Social Media: Proof of Concept**
Laura Jahn, Rasmus K. Rendsvig, Jacob Stærk-Østergaard
Submitted to *Social Network Analysis and Mining*, July 15 2022.

II **Towards Detecting Inauthentic Coordination in Twitter Likes Data**
Laura Jahn and Rasmus K. Rendsvig
Draft.

III **Friction Interventions to Curb the Spread of Misinformation on Social Media**
Laura Jahn, Rasmus K. Rendsvig, Alessandro Flammini, Filippo Menczer, Vincent F. Hendricks
Draft.

Abstract

This Ph.D. thesis studies ways to curb the *amplification of low-quality content*, such as misinformation, on social media using agent-based models and data from the social media platform Twitter. The thesis focuses explicitly on the amplification through *one-click user reactions* such as *likes* and *shares*. Liking and sharing are central ways by which information spreads in a social network while informing platforms' content-sorting algorithms, further increasing reach.

Amplification through likes and shares may be driven by coordinated and/or inauthentic actors such as social bots. Yet, also authentic human users may spread low-quality content. In light of social influence and cognitive biases, authentic users may engage with high-engagement posts allocating little to no attention to assess accuracy or quality. Both inauthentic and authentic dynamics amplify misinformation online and undermine the *wisdom of crowds*: High engagement does not reliably point to high quality. While the inflation of engagement metrics is a readily available manipulation strategy undermining the wisdom-of-crowds effect, research has yet to extensively study the amplification of low-quality content through likes and shares. A major reason is that data on one-click user reactions is non-trivial to collect.

The main part of this thesis consists of three research articles. From different angles, these articles address threats to the wisdom of crowds. They share the goal of improving the (epistemic) quality of the information that gets amplified on social media. Articles I and II study computational methods to *detect* inauthentic, coordinated metric inflation and suspicious correlations in reactions data. This part of the thesis is based on computer-simulated data from an agent-based model (Article I) and novel empirical data live-collected through Twitter with a scripted algorithm written with the purpose of overcoming the data shortage on one-click user reactions (Article II). Article III studies behavioral interventions based on *friction* to *prevent* the amplification of low-quality content analyzed with an agent-based model.

An introductory chapter precedes the articles. The introduction describes the interdisciplinary basis of the thesis and presents key concepts and theories, subsequently serving as a scaffold in the presentation of each article. The introduction discusses key concepts such as social media, information diffusion, attention economy, notions of quality (e.g., misinformation and disinformation), and amplification. These concepts are considered theoretically from a voting perspective. This perspective is used to assess how today's social media landscape challenges the wisdom of crowds—mathematically based on the *Condorcet Jury Theorem*—through cognitive biases, social influence, and the presence of inauthentic actors, coordination, and influence operations.

A methodology section explains and justifies modeling choices and assumptions concerning agent-based modeling and empirical data collection from Twitter. The methodology section additionally elaborates on data processing and analysis and concludes with a description of the data and a discussion of ethical considerations surrounding data collection and censorship. Subsequently, the introduction chapter reviews related work on coordination, bot detection, and behavioral interventions. Finally, implementation possibilities of the proposed procedures and future research avenues are described.

Resumé

Denne Ph.d.-afhandling undersøger måder at bremse *forstærket spredning* af *lavkvalitetsindhold*, såsom misinformation, på sociale medier. Dette undersøges gennem brug af agentbaserede modeller og data fra den sociale medieplatform Twitter. Afhandlingen fokuserer eksplicit på forstærkningen gennem *et-klik brugerreaktioner* såsom *likes* og *delinger* (*shares*). At like og at dele er centrale måder hvorpå information spredes i online sociale netværk. Disse former for engagement med indhold informerer endvidere platformenes indholdssorteringsalgoritmer, hvilket øger spredningen yderligere.

Forstærket spredning gennem likes og delinger kan være drevet af koordinerede og/eller uægte aktører såsom sociale *bots*. Også autentiske, menneskelige brugere kan sprede indhold af lav kvalitet. Påvirket af social indflydelse og kognitive biases kan autentiske brugere finde på at like eller dele indhold givet høje engagementstal, uden at tildele opmærksomhed til at vurdere indholdets nøjagtighed eller kvalitet. Både uægte og autentiske brugerreaktioner forstærker spredningen af misinformation online, og underminerer *folkemængdernes visdom* (*wisdom of crowds*): Høje engagementmål indikerer ikke nødvendigvis høj kvalitet. Mens inflationen af engagementsmål er en let tilgængelig manipulationsstrategi, der underminerer folkemængdernes visdom, har forskning endnu ikke i vid udstrækning studeret forstærkningen af lavkvalitetsindhold gennem likes og delinger. En væsentlig årsag er, at data om et-klik brugerreaktioner er ikke-trivielle at indsamle.

Hoveddelen af denne afhandling består af tre forskningsartikler. Fra forskellige vinkler adresserer disse artikler trusler mod folkemængdernes visdom. Artiklerne har som fælles mål at forbedre den (epistemiske) kvalitet af den information, der bliver forstærket på sociale medier. Artikel I og II studerer metoder til at *opdage* uægte, koordineret engagementsinflation og mistænkelige sammenhænge i reaktionsdata. Disse artikler er baserede på computersimulerede data fra en agent-baseret model (Artikel I) og nye empiriske data indsamlet live fra Twitter med et program skrevet med det formål at overvinde datamanglen for et-klik brugerreaktioner (Artikel II). Artikel III studerer adfærdsinterventioner baseret på *friktion* for at *forhindre* forstærkning af lavkvalitetsindhold analyseret med en agent-baseret model.

Et indledende kapitel går forud for artiklerne. Indledningen beskriver afhandlingens tværfaglige grundlag, og præsenterer centrale begreber og teorier, der efterfølgende fungerer som et stillads for præsentationen af hver artikel. Indledningen diskuterer nøglebegreber såsom sociale medier, informationsspredning, opmærksomhedsøkonomi, kvalitetsbegreber (f.eks. misinformation og disinformation) og forstærkning. Disse begreber betragtes teoretisk ud fra et folkeafstemningsperspektiv. Dette perspektiv bruges til at vurdere, hvordan det sociale medielandskab udfordrer folkemængdernes visdom gennem kognitive biaser, social indflydelse og tilstedeværelsen af uægte aktører, koordinering og indflydelsesoperationer.

Et metodeafsnit forklarer og begrunder modelleringsvalg og antagelser vedrørende agentbaseret modellering og empirisk dataindsamling fra Twitter. Metodeafsnittet uddyber yderligere databehandling og -analyse, og afsluttes med en beskrivelse af de indsamlede data samt en diskussion af etiske overvejelser omkring dataindsamling og censur. Efterfølgende gennemgår introduktionskapitlet relateret arbejde indenfor koordinering, bot-detektion og adfærdsmæssige interventioner. Afslutningsvis beskrives implementeringsmuligheder af de foreslåede procedurer og fremtidige forskningsperspektiver.

Zusammenfassung

Diese Dissertation untersucht Möglichkeiten, die *Verstärkung* (*amplification*) von *minderwertigen Inhalten*, wie Misinformation, in sozialen Medien einzudämmen. Agentenbasierte Modelle und Daten von der Social-Media-Plattform Twitter werden zur Untersuchung verwendet. Die Arbeit konzentriert sich explizit auf die Verstärkung durch *Ein-Klick-Nutzer-Reaktionen* (*one-click user reactions*) wie das *Liken* und *Teilen* (*likes and shares*). Liken und Teilen sind zentrale Wege, mit denen sich Informationen in einem sozialen Netzwerk verbreiten, während sie die algorithmische Sortierung von Inhalten auf Plattformen informieren und die Reichweite weiter erhöhen.

Die Verstärkung durch Liken und Teilen kann von koordinierten und/oder nicht authentischen Akteuren wie sozialen Bots vorangetrieben werden. Aber auch authentische menschliche Nutzer können minderwertige Inhalte verbreiten. Angesichts von sozialem Einfluss und kognitiver Vorurteile beschäftigen sich authentische Nutzer mögicherweise mit Posts mit hohem Engagement (oft gelikt und geteilt) und der Bewertung von Genauigkeit oder Qualität wird wenig bis gar keine Aufmerksamkeit gewidmet. Sowohl unauthentische als auch authentische Dynamiken verbreiten Misinformationen online und untergraben die *Weisheit der Vielen (wisdom of crowds)*: Hohes Engagement weist nicht zuverlässig auf hohe Qualität hin. Während die Inflation von Engagement-Metriken eine leicht verfügbare Manipulationsstrategie ist, die den Weisheit-der-Vielen-Effekt untergräbt, muss die Forschung die Verstärkung von minderwertigen Inhalten durch Liken und Teilen noch ausgiebig untersuchen. Ein Hauptgrund dafür ist, dass Daten zu Ein-Klick-Nutzer-Reaktionen nicht einfach zu sammeln sind.

Der Hauptteil dieser Arbeit besteht aus drei Forschungsartikeln. Aus verschiedenen Blickwinkeln sprechen diese Artikel Bedrohungen für die Weisheit der Vielen an. Sie teilen das Ziel, die (epistemische) Qualität von Informationen, die in sozialen Medien amplifiziert werden, zu verbessern. In den Artikeln I und II werden computer-basierte Methoden untersucht, um nicht authentische, koordinierte Inflation von Engagement-Metriken und verdächtige Korrelationen in Reaktionsdaten *zu erkennen*. Dieser Teil der Arbeit basiert auf computersimulierten Daten aus einem agentenbasierten Modell (Artikel I) und neuartigen empirischen Daten, die live über Twitter mit einem geskripteten Algorithmus gesammelt wurden. Das Skript wurde mit dem Ziel geschrieben, den Datenmangel bei Reaktionsdaten zu überwinden (Artikel II). Artikel III untersucht Verhaltensinterventionen, die auf *Friktion* (*friction*) basieren, um die Verstärkung von Inhalten von geringer Qualität *zu verhindern*. Friktion wurde mit Hilfe eines agentenbasierten Modells analysiert.

Den Artikeln ist ein einleitendes Kapitel vorangestellt. Die Einleitung beschreibt die interdisziplinäre Grundlage der Arbeit und stellt Schlüsselkonzepte und Theorien vor, die anschließend als Gerüst für die Präsentation jedes Artikels dienen. In der Einführung werden Konzepte wie soziale Medien, Informationsverbreitung, Aufmerksamkeitsökonomie, Qualitätsbegriffe (z. B. Misinformation und Desinformation) und Amplifikation erörtert. Diese Konzepte werden theoretisch aus einer Abstimmungsperspektive (voting perspective) betrachtet. Diese Perspektive wird verwendet, um zu bewerten, wie die heutige Social-Media-Landschaft die Weisheit der Vielen—mathematisch basierend auf dem *Condorcet Jury Theorem*—herausfordert: durch kognitive Vorurteile, sozialen Einfluss und das Vorhandensein nicht authentischer Akteure, Koordination und strategischer Einflussnahme.

Ein Abschnitt zur Methodik erläutert und begründet Modellierungsentscheidungen und -annahmen in Bezug auf agentenbasierte Modellierung und empirische Datenerfassung von Twitter. Der Methodenteil geht zusätzlich auf die Datenverarbeitung und -analyse ein und schließt mit einer Beschreibung der Daten und einer Diskussion ethischer Überlegungen zu Datenerhebung und Zensur. Anschließend behandelt das Einführungskapitel verwandte Arbeiten zu Koordination, Bot-Erkennung und Verhaltensinterventionen. Abschließend werden Umsetzungsmöglichkeiten der vorgeschlagenen Methoden und zukünftige Forschungswege beschrieben.

Acknowledgements

I am extremely grateful to my supervisors, Vincent F. Hendricks and Rasmus K. Rendsvig, for invaluable feedback and for trusting in my abilities, especially in times when I did not. Thank you for your extraordinary kindness and academic guidance throughout. I would furthermore like to thank Irina Shklovski, Yong-Yeol Ahn, and Olivier Roy for accepting to be members of the thesis committee. I also gratefully acknowledge the financial support of the Carlsberg Foundation throughout.

To continue, I would like to thank my co-authors, Alessandro Flammini, Vincent F. Hendricks, Filippo Menczer, Rasmus K. Rendsvig, and Jacob Stærk-Østergaard. It was a privilege to collaborate with and learn from all of you! Special thanks to Rasmus, who is a co-author of *all* the articles included in this thesis. Thank you for your mentoring and friendship, expertise and dedication.

Many, many thanks to *everyone* currently and formerly affiliated with the *Center for Information and Bubble Studies* (CIBS). Your diverse backgrounds and perspectives, academic openness, and our formal and informal gatherings (read: great parties) made CIBS a special and warm place.

During my Ph.D., I had the great pleasure of teaching *Philosophy of Science* and *(Mis)Information and Democracy*. Thank you, Thor Grünbaum, Vincent F. Hendricks, and Kristian Høyer Toft, for your trust and mentorship during teaching. I learned a lot.

My Ph.D. experience was enriched by a research visit at the *Observatory on Social Media* (OSoMe, tellingly pronounced *awesome*), Indiana University, Bloomington. The collaboration with Fil and Alessandro in person and later via Zoom was full of enthusiasm and appreciation. Many thanks to Bao Tran Truong, who kindly and patiently introduced me to work on which Article III in this thesis builds. Thank you to everyone at the lab, and to Vincent, Rasmus, and Fil for arranging the stay and hosting me.

I thank my friends in Copenhagen. Thank you, Anne-Sophie, Valentyna, and Victoria, for our conversations, drinks, and shared laughs in- and outside of work. Special thanks to Gaia, Fabienne, Josi, and Mareike. I could always count on you! Thank you for being there, in the good and not-so-good times.

Throughout, I felt loved and supported by family and friends outside Copenhagen. Thank you for visits—home and abroad—and long phone calls: Maike, Sophie, Franzi, Pia, Lisa, David, Annie, Michéle, and my brother Jonas.

Despite much of my Ph.D. taking place during COVID-19 and lockdowns, I enjoyed meaningful conversations with many more people than explicitly mentioned here, whom in some way or another helped completing this thesis. Thank you.

I am thanking my parents in German: Ein sehr besonderer Dank gilt meinen Eltern, Bärbel und Klaus Jahn. Ihr habt mich während meines akademischen Werdegangs in jeder erdenklichen Weise unterstützt. Ihr wart immer für mich da, egal ob digital oder während eurer zahlreichen Besuche in Mannheim, Bayreuth, München, Bath, Shanghai und Kopenhagen. Und ich kam immer gerne nach Hause. Ohne eure Liebe und die *absolute* Abwesenheit jedmöglichen Drucks oder Erwartungen gäbe es diese Doktorarbeit nicht.

Lastly, Mikkel: Thank you for your endless excitement, intellectual sparring, and remarkable help. Thank you for listening, understanding, and believing in me. You are my person. And you were before I knew where this reference came from.

Laura Jahn
Copenhagen, February 2023

Introduction

1 Introduction and Problem Statement

It costs a few US Dollars to buy 100 likes on Twitter. Sold through readily available vendors, an experimental purchase of 100 likes catapulted a tweet to the *Top* feed of `#dkpol`[1], the main Twittersphere for discussing Danish politics. There it stayed for several hours. This true tale illustrates that posts attracting engagement in the form of *one-click user reactions* such as likes are algorithmically sorted to bring selected items to the attention of users. Beyond the Danish Twittersphere, content-sorting algorithms affect the information diets of users in more than 200 countries, with an estimated 396.5 million monthly Twitter users [103], of which 100.9 million are in Europe [142].

Our small experiment had no intent to mislead or manipulate. However, posts containing misinformation[2] may similarly be amplified through engagements such as *likes* and *shares*. Coordinated groups of accounts may engage in influence operations, manipulating ranking algorithms to amplify content and inflate popularity while evading detection [18, 132].

The role of social media in two recent global crises exemplifies the seriousness of the threat: First, the COVID-19 pandemic was accompanied by an *infodemic* of misinformation, where false and harmful information was spread and amplified through social networks, both intentionally and unintentionally [41]. Second, amidst Russia's invasion of Ukraine, Russia seeks to produce a sense of support for their aggression through political influence campaigns that spread misleading content and amplify authentic posts by users that are consistent with Russia's preferred political narratives [8]. In both cases, people worldwide may be mislead about global issues of broad importance for health, wealth, and security.

Crucially, it is *not* only coordinated and/or inauthentic actors that are involved in amplifying misinformation. Authentic human users are prone to be misled. Users customarily fall prey to, for example, social influence and cognitive biases, and engage with high-engagement posts allocating little to no attention to assess accuracy or quality [15, 100, 147]. Both inauthentic

[1] The web-archived Top feed of `#dkpol` on June 24 2022: `https://tinyurl.com/mtndfb3n`

[2] Note that we discuss the notions of quality, misinformation, and related concepts in Sec. 3.2

and authentic dynamics amplify misinformation online and undermine *the wisdom of crowds* [126]: High engagement does not reliably point to high quality.

While engagement metric inflation is a readily available manipulation strategy, research has yet to extensively study the amplification of low-quality content through likes and shares. A major reason is that *data* on one-click user reactions (such as likes and shares) is non-trivial to collect.

From different angles, the articles in this article-based thesis address *threats* to the wisdom of crowds. They share the goal of improving the (epistemic) *quality of information* spread on social media. All articles set out to curb the *amplification of misinformation* through one-click user reactions, be it through benign actors, such as authentic human users, or malevolent actors, such as social bots.

On the one hand, the articles study methods to *detect* and flag inauthentic, coordinated metric inflation and suspicious correlations in reactions data. This part of the thesis is based on computer-simulated data from an agent-based model (**Article I**) and novel data live-collected through Twitter with a scripted algorithm (**Article II**). This script was written with the purpose of overcoming the data shortage on one-click user reactions. On the other hand, the thesis studies behavioral interventions based on *friction* to *prevent* the amplification of low-quality content analyzed with an agent-based model.

The remainder of this first chapter of the dissertation is structured as follows: We[3] start by discussing the subject fields and the research project to which this thesis contributes in Sec. 2, followed by introducing *key concepts and theories* in Sec. 3. We describe and define notions while providing an overview of concepts including *social media*, *information diffusion*, and *attention economy*, followed by surveying notions of *quality* and the users' role in online *amplification*. We then consider these concepts from a *voting perspective* and discuss how today's social media landscape *challenges* the *wisdom of crowds* through *cognitive biases*, *social influence*, and the presence of *inauthentic actors, coordination*, and *influence operations*. The key concepts discussed in this section will resurface throughout this introductory chapter.

In Sec. 4, *overviews of the articles* that form part of this thesis are provided, alongside detailed *author contribution* descriptions. The methodology is introduced in Sec. 5. We elaborate on the two main methods—*agent-based modeling* and *empirical data collection* from Twitter. This is followed by subsections laying out *data processing and analysis* steps as well as *data descriptions* and *ethical considerations*. We conclude in Sec. 7.

[3] The first-person plural is used despite the PhD student being the sole author of this introduction.

2 Subject Fields

Working towards *curbing the amplification of misinformation on social media*, this thesis is of an *applied* nature. The present work contributes to solving the aforementioned problem, grown in relevance with social media enabling the sharing of information at unprecedented speeds and scales. The relatively young yet growing and constantly mutating research subject—social media—has been studied in a great number of papers and across disciplines relating to information quality.

As a consequence of the problem it tries to solve, this thesis is also *interdisciplinary*. Various subject fields analyze mechanisms of information disorders and try to find ways to reduce the spread of low-quality content online. A non-exhaustive list (in no particular order) of scientific disciplines includes (computational) social science, web science, network science, machine learning, data science, computer science, information science, journalism, misinformation research, psychology, behavioral economics, and philosophy [13, 66, 74, 110, 125, 143, 144, 159, 160, 161].

Research spans across disciplines not least because of the many forms and contexts of misinformation, complex features of online environments, different data and method requirements to understand the phenomenon [67, 110], to name a few. It is important to bear in mind that conducting interdisciplinary research also comes with costs. Costs and risks comprise efforts to stay alert about publications in various fields, and decreased productivity as interacting with reviewers and collaborators from across fields may be challenging [72, 122]. In the best case, diverse approaches complement each other and tackle different parts of a shared problem. For example, fighting misinformation might benefit from both advances in bot detection *and* from media literacy research. Approaches from a variety of research avenues, rich toolkits, and multiple stakeholders—besides researchers, most importantly the platforms themselves and policy makers—are promising to jointly contribute to and supplement solutions to the epistemic chaos [159, 161].

At the interdisciplinary *Center for Information and Bubble Studies* (CIBS), based at the Department of Communication, Faculty of Humanities, University of Copenhagen, this thesis contributes to the broader research project *Attention Economics and Informational Social Influence*. This projects sets out to investigate the structural properties, epistemological conditions, and democratic repercussions of the relationship between human interactions, attention and influence. The papers in this thesis were written in collaboration with current and former members of CIBS—Rasmus K. Rendsvig, Jacob Stærk-Østergaard, and Vincent F. Hendricks—and in collaboration with Filippo Menczer and Alessandro Flammini from the Observatory on Social Media (OSoMe) at Indiana University, Bloomington, USA.

The following sections—key concepts and theories in Sec. 3, methodology in Sec. 5, and related work in Sec. 6—introduce notions, modeling perspectives, and formal methods drawing from different fields, substantiating the interdisciplinary subject areas both in which this thesis is situated and to which it contributes.

3 Key Concepts and Theories

This section introduces key concepts and theories from relevant subject fields (cf. Sec. 2). We begin with a brief overview of *social media, information diffusion, and attention economy*, followed by surveying notions of the *quality* of online content and the users' role in its *amplification*. We then consider these concepts from a *voting perspective* and discuss how today's social media landscape *challenges* the *wisdom of crowds* through *cognitive biases*, *social influence*, and the presence of *inauthentic actors, coordination, and influence operations*. The key concepts discussed in this section will resurface throughout this introduction and serve as a scaffold in an overview of each article (introduced in Sec. 4). Isolated, simplified, or abstracted notions of these concepts will play roles in the motivation, problem descriptions, methods, model assumptions, and discussions of the articles that follow.

3.1 Social Media, Information Diffusion, and Attention Economy

With the advent of Web 2.0 of the 21st century, the internet has taken a turn to the *social* [87, 143]. Social networking sites facilitate the production of shared digital content and the creation of connections between users' online social presences and social networks offline. Early social networking sites (e.g., MySpace and Facebook) were, among others, followed by microblogging platforms (e.g., Twitter), media sharing sites (e.g., TikTok and Instagram), and messenger services (e.g., WhatsApp) [143]. In more and more accessible ways, these services enable users to engage in social interaction, a significant supplement to the web's former individualistic wheels [87]. The resulting networks allow for the flow of information, goods, or support through community structures [117].

The relationships between users connected on social media platforms may be plausibly formalized through graphs consisting of nodes connected by edges or links given a network interpretation [110]. How *information diffuses* between users depends on the structure of a given social network. For example, strongly clustered groups of users (communities) reinforce content within and obstruct diffusion between communities, resulting in polarization and echo chambers; thereby, content diffuses successfully through affirmation and reinforcement from multiple, different sources, a phenomenon known as complex contagion [14, 20, 73, 89, 110, 149]. At the same time, social networking sites enable content to spread to millions of users at unprecedented scales and speeds. It is by far not only entertaining content

that spreads. Content consumed on social media serves as a principal gateway to information and shapes public opinion [19]. Almost any content may be accessed, produced, and relayed to a large audience without editorial judgment or fact-checking [4, 159].

Such information abundance consumes the *attention* of users [32]. In this information-rich world, attention as a cognitive resource becomes scarce, and users cannot allocate attention such that they consume everything; attention is a zero-sum game [120]. For social media platforms, attention also is a highly profitable asset [53]. Platforms are interested in continuous user *engagement*, as more traffic means more data harvest, driving an *attention economy*: Content producers, such as users or media outlets, generate posts on which other users spend their attention. This expenditure of attention results in engagement with the posts and increased traffic on the platform, which in turn generates user data. Social media platforms and tech firms then sell the collected and analyzed data as targeted advertising packages to interested advertisers, essentially capitalizing users' attention. Online behavior may also translate into behavior data, which is sold to companies to predict the future behavior of different segments of users and to map the potential for influencing said segments [53] (also referred to as *surveillance capitalism* [146, 163]).

Social media platforms deploy *algorithms* to sort content and optimize engagement metrics, affecting what content is presented to users in their news feeds [143]. These algorithms routinely exhibit a popularity bias, that is, algorithms favoring the visibility of content that has already logged a substantial amount of engagement, such as shares, likes, or views [18, 121]. Together with algorithmic sorting, *biases* among users contribute to the spread of high-engagement content. Examples of such biases and heuristics include: Confirmation bias [12, 92], a disconnect between what users deem accurate and what they deem shareable [100, 102], automatic habits to share the most engaging content [15], and the illusory effect increasing the perceived reliability of content through repetition [31, 54, 69]. Social and cognitive biases and heuristics are discussed again in Sec. 3.5. Together with the aforementioned network effects, this may result in a reinforcement of the spread of highly engaging content, affecting the quality of what becomes popular content.

3.2 Quality of Information Spread Online

The content that spreads on social media platforms is often highly engaging. Ideally, high-quality content, such as truthful pieces of news (e.g., in contexts where quality is equated with true and timely), or a funny meme (e.g., when quality is entertainment value), is what engages users. However, engaging posts are not necessarily of high quality.

This thesis *does not aim to classify* social media content as *good* or *bad* or *high* or *low quality*. For an overview of automated methods for assessing the truthfulness of online data, see for example Lozano et al. [80], for a discussion of manual fact-checking solutions, see for exam-

ple Ruffo et al. [110]. This section will briefly digress into reviewing the discussion around *misinformation* and *disinformation*. The concepts, and the information disorders they capture, serve to motivate the articles that form part of this thesis. These concepts are relevant as the central problem which the thesis tries to tackle, deals with *curbing the amplification* of *content that may be classified as mis- or disinformation*.

Information disorders such as misinformation and disinformation are often used interchangeably, with shifting and overlapping definitions [5, 15, 42, 50, 127, 161]. Definitions of misinformation usually agree that misinformation has a misleading nature [13, 42, 50, 110, 123, 124, 144]. Some scholars further state that misinformation is false or inaccurate by definition (e.g., [50]), while others argue that misinformation may indeed be true (but misleading) or false (e.g., [123]). Often, misinformation is defined even more narrowly as the *unintentional* creation or spread of misleading content, not *meant* to cause harm [13, 42, 50, 110, 123, 144]. Disinformation, on the other hand, may be distinguished by intent and motive [4, 13, 71, 123, 125]: it is often defined as the creation or spread of misleading, inaccurate or deceptive content with decisive actions to mislead the public and in some respect cause harms [128, 42].

Sometimes, the term disinformation serves as an umbrella term including fake news, propaganda, and misinformation [60]. Other times, misinformation is used as the umbrella term, instrumentalized to include a plethora of information disorders. Twitter, for example, defines misinformation as misleading content that has been confirmed to be false by third-party experts or information that is shared in a deceptive or confusing manner [141]. This makes disinformation a subset of misinformation [34, 50], spread deliberatively and purposefully misleading. Similarly advertently misleading (sometimes referred to as subtypes of disinformation) are *fake news*—deliberately misleading articles mimicking the look and feel of legitimate articles—, and *propaganda*—partisan information that may be true but is used to persuade people to support one political group over another [50, 53, 127].

Individual intent, however, is difficult to divine, and harm may be caused either way, whether the spread of misleading information items happens with intent or not [15, 110, 147]. Furthermore, every piece of disinformation crafted with the intent to mislead, may ultimately become misinformation once users start sharing it, without intent to mislead, but with the belief that the information they amplify is accurate. No matter whether content started out as misinformation or disinformation, it is both actors with intent and without that may contribute to the rapid amplification. Hence, irrespective of motive, users may experience exposure to low-quality content such as misinformation and disinformation that spreads through social networks [42, 161].

In the context of this thesis, we scope our considerations to curb the amplification of *misinformation*. We take misinformation as an umbrella term for misleading low-quality content (e.g. it may have low epistemic value or may fare low on a non-truth related quality), that may be spread with intent (disinformation and its subtypes) or without. High-quality content,

in contrast, may have high epistemic value or, for example, abide by journalistic standards of excellence [62], have novelty value, or capture a subjective value such as beauty. While we naturally acknowledge that quality is often not discerned in such a binary, black and white manner (high and low), the work that follows in this thesis will at times make simplifying modeling assumptions about high- and low-quality notions (cf. Sec. 4).

3.3 Amplification

Users play a decisive role in the *amplification* of content online [110]. Given the social nature of social media, the sharing of both high-quality content and misinformation affects fellow users [147]. The exposure to an information item does not happen in a vacuum. With a social media account comes a plethora of options for users to interact with and spread content. It may be in terms of posting new content, commenting on posts, or sharing and liking posts presented in the news feed—as suggested by platforms' algorithms, and/or shared by friends and accounts one follows. Such activity, in turn, influences what other users see in their news feeds. We focus on a specific kind of behavioral user engagement that contributes to amplification online: *One-click reactions* such as *likes* and *shares*.

On social media, users can usually select a reaction to a post from a short list, with their aggregate choices typically presented as a popularity metric beneath the post. One-click reactions include perhaps most famously Facebook's original 'Like', the hearts/likes on Instagram, TikTok and Twitter, and Reddit's up- and downvotes, but also sharing and retweeting on any of these platforms.

One-click reactions steer user attention. They determine the content that spreads through a follower network, exponentially growing the number of users exposed to a popular information item. One-click reactions are also a driving mechanism behind algorithmic sorting, where the *engagement metrics* are taken as scalable indicators of quality [18, 65, 128]. As such, the easy to use reactions are quick to influence the popularity of posts [144].

From a social epistemological perspective, engagement metrics may actively influence users trying to determine what is true with the help of, or in the face of, others [47]. A high retweet count is likely to be perceived as a crowd-sourced trust signal [86, 126], possibly contributing to the content's virality [29], potentially irrespective of its quality [7]. Once trending, high engagement counts in likes and shares are shown to make users more likely to engage with popular content instead of fact-checking posts (cf. Sec.3.5).

3.4 A Voting Perspective and the Wisdom of Crowds

The role of users in the amplification of online content [110] and the influential role of one-click reactions in the popularization of content through social platform algorithms of social networks [18], has encouraged the work in this thesis to consider amplification from a *voting perspective*. With users' reactions conceptualized as votes about the quality of an information item, and the outcome summarized through an engagement metric or judgement aggregator, high-quality content should ideally receive high engagement (many votes) and thus popularity. As such, platform algorithms and the influential role of engagement metrics are predicated on the *wisdom of crowds*: Under the right circumstances, combining the votes of many may be very effective in promoting the "correct" (such as high-quality) content [47, 38].

The wisdom of crowds has a mathematical basis in the *Condorcet Jury Theorem* [21]: The theorem rests on two premises: The first is that voters vote independently of each other (the independence assumption). The second is that all voters are better than random at voting correctly, that is, each voter is more than 50% likely to reach a correct judgement (the competence assumption). Then, the probability of a correct majority judgment approaches 1 as the voting group size approaches ∞ [28, 47]. The theorem implicitly assumes that voters vote sincerely, that is their competence-based belief informs their vote, in contrast to, for example voting strategically [27].

The Condorcet Jury Theorem has maintained its relevance in the context epistemic democracy and formal epistemology: It has been generalized to account for correlation among voters (e.g. due to common evidence or common causes) given certain conditions by weakening the independence assumption [70, 106]; has been augmented to not only account for majority voting over two options, but also for plurality voting over many options [75]. The wisdom of crowds equally remains relevant with regard to social media, promising more democratic participation on the information market ruled by the attention economy, and ideally promoting high-quality content in a 'marketplace of ideas' [159]. The success of the online encyclopedia Wikipedia, consistently producing high-quality articles at a large scale and relevant content gaining popularity during the Arab Spring are cases in point [79, 159]. Given the scale of social media and with it the size of voting groups casting their vote through shares and likes, it would be desirable to be able to rely on the engagement metric signal as a quality signal.

3.5 Threats to the Wisdom of Crowds

Today's social media exhibits a structure that threatens the wisdom of crowds. In many cases, groups of users are prone to epistemic problems and epistemic chaos when it comes to combining beliefs and votes into reliable engagement votes [47, 159]. The independence and competence assumptions of the Condorcet Jury Theorem are violated by what we witness online.

To an extent, the articles in this thesis each tackle a different set of these violations in an attempt to curb the spread of low-quality content and improve the reliability of the judgement aggregators.

Cognitive Biases and Bounded Rationality When users navigate through information-abundant social media environments, evidence suggests that heuristics and biases steer behavior, instead of careful deliberate cognitive processing [87]. Rationality is bounded [61, 119, 118], especially in the midst of information overload imposed by ever-renewing and never-ending news feeds driven by social acceleration through increasing production and consumption of content, depleting users' limited attention [78, 161].

Identified cognitive shortcuts, system II thinking (heuristics), and biases range into the hundreds (for an overview, see Sec. 7.1 in [150], Sec. 4 in [110], and [101]). Social media environments have been identified as fertile grounds for deployment of such, increasing the susceptibility to engage with low-quality content [161]. Examples include: Automatic, habitual sharing of content given social media's reward structure to attract social recognition [15]; sharing posts featuring cognitive hooks making content seem appealing, easy to process, emotionally engaging, and disconnected from accuracy considerations [100, 102, 110]; or sharing content confirming prior belief [92] or due to the illusory effect of repeated exposure [31, 69]. Confirmation bias and the illusory effect especially promote repeated exposure: As a key element of complex contagion, repeated engagement and exposure reinforce algorithmic biases and network communities such as echo chambers, making engaging content even more visible through more engagement [14, 18, 73, 89].

But how do these cognitive shortcuts undermine the Condorcet Jury Theorem? Dealing with socially accelerated news feeds and limited attention, users face a trade-off between the cognitively expensive operation of assessing a post's accuracy, and the cognitively cheaper but less accurate choice of liking or sharing content following these heuristics based on readily available cues (engagement metrics, of course, are also readily available cues, which we will discuss in conjunction with social influence) [18]. If accuracy is the quality we seek to track with engagements, this behavior implies that users do not vote (engage, like, or share) competently when engaging with content, violating the competence assumption of the Condorcet Jury Theorem. Similarly, these mechanisms may expose users to influence (see the following paragraph) in network communities violating the independence assumption.

Social Biases and Social Influence With social media as an ecosystem of social information, *social biases* influence online behavior: Users engage with and amplify content online in the context of other users and witness their behavior, which in turn exerts informational *social influence* [131, 133]. This type of social influence refers to the acceptance of the behavior of others as evidence about the true state of the world [26, 135].

Engagement metrics such as the number of likes represent a form of *social proof* [87]. The presence and standpoint of others, conveniently summarized tone and quantity in judgement aggregators, may influence users' attitudes and behavior [55, 131].

Ideally, social proof is truth-conducive and promotes reaching the correct conclusion about how to engage with an information item online [30] or does not have an effect [65]. However, the aggregate opinions to which users are exposed are likely not independent, potentially undermining the wisdom of crowds effect [77]: As readily available cues, engagement metrics may signal to the users high-quality properties such as importance, relevance, and accuracy by showcasing multiple exposures and judgements. Similar to cognitive shortcuts and in line with complex contagion [89], users may assume that many other users before them must have independently judged the information item well. This may result in users less likely to scrutinize the post at hand and more likely to like or share [7, 81]. Such inaccurate social proof does not act as its ideal prescribes: Experimental evidence indicates that the display of engagement metrics promotes engagement with low-quality content [7] and increases the credibility of both high- and low-quality content[81]; the higher the engagement count, the more prone users were to share less accurate posts [7]. As such, instances of social bias and social influence violate the independence assumption of the Condorcet Jury Theorem, while potentially also lowering the competence among users.

Inauthentic Actors, Coordinated Behavior, and Influence Operations Engagement metrics may be readily boosted and manipulated to amplify target content by precisely exploiting the social and cognitive biases laid out in the previous two paragraphs, as well as algorithmic biases. This potentially undermines both the competence (and sincerity) assumption (in the case of *inauthentic actors* such as *social bots*) and the independence assumption (in the case of *coordinated influence operations*). This final part of the key concepts and theories introduces the notions of *inauthentic actors*, *coordination*, and *influence operations*.

One prominent type of inauthentic actors are social bots: Social bots are at least in part controlled by software and often by a mix of software and humans. They are designed influence and to mimic human behavior [1, 95, 110, 113]. Social bots, for instance employed to create the appearance of support for a cause (astroturfing), fall into the category of inauthentic actors, together with any accounts that take on inauthentic personas on social media. Keeping in mind that the detection of bots is a difficult and complex task (see Sec. 6), the prevalence of bots on Twitter is estimated between 9 and 20% [23, 103, 145].

Bots and/or human accounts (or mixed accounts) may be deployed to coordinate actions. Coordinated behavior comprises actions and efforts of a group of users to achieve a shared goal [91]. To boost engagement metrics, coordinated behavior is particularly effective. If accounts coordinatedly channel their reactions towards the same targets, they may jointly catapult posts to be trending. The same resources spent in an uncoordinated way may have

less impact. In the context of *influence operations* on social media platforms—with the aim of shaping public opinion [93, 125, 144]—bulk coordinated reactions effectively exploit the platforms' content sorting algorithms to highlight posts to users (also referred to as attention-hacking) [84].

Coordinated inauthentic behavior undermines the competence (and sincerity) assumption and independence assumption of the Condorcet Jury Theorem. Inauthentic reactions—e.g., liking posts not in accordance with beliefs and preferences, but according to a supplied protocol—may be considered insincere violating the competence assumption. Inauthentic or not, coordinated behavior always violates the independence assumption. Examples of authentic coordinated behavior are fandoms or grass-root initiatives [56, 144].

Social media platforms have picked up on the concepts since coordinated inauthentic and harmful behavior are an emerging problem for the platforms, linked to the spread of misinformation [19]. Coordinated behavior, first mentioned by Meta/Facebook in 2018, is also discussed in conjunction with manipulation as an example of a violation of Twitter's community standards [46, 137, 139].

Research has also allocated resources to the study of the phenomenon of coordinated inauthentic behavior. There are several challenges specific to the study of this phenomenon: there is no agreed-upon definition of what exactly coordination is; it is difficult to measure how many accounts must be involved to achieve impactful coordination and manipulation; the distinction between emerging, authentic behaviors of an online crowd may be difficult to discern from orchestrated, inauthentic influence operations [125]; automated detection methods may be challenged by non-binary and more nuanced definitions of coordination and (in-)authenticity [56]. Perhaps due to these challenges, studies so far have mostly focused on detecting coordination independently of (in-)authenticity.

We review reaction-based coordination in **Article I** and **Article II**, and discuss coordination alongside different activities on social media in Sec. 6. In Sec. 5.3, we summarize how this thesis methodologically addresses the notion of coordination in **Article I** and **Article II**.

4 Article Overview and Author Contributions

The previous section concluded with an assessment of various ways the wisdom of crowds may fail. With the key concepts serving as the scaffold, we now turn to specifying how the articles contribute to curbing the amplification of low-quality content through one-click reactions. The articles address different sets of threats to the wisdom of crowds: **Article I** and **Article II** contribute to the *detection* of coordinated inauthentic amplification of low-quality content, and the removal of these actors violating the competence and independence assumptions. **Article III**, on the other hand, studies a behavioral intervention that possibly *prevents* actors from falling prey to cognitive biases and bounded rationality, social biases and social influence, thus serving the premises of the Condorcet Jury Theorem. The presentation of the articles in this section is followed by a detailed account over the methods employed in the articles (Sec. 5).

4.1 Article I

Jahn, Laura and Rendsvig, Rasmus K. and Stærk-Østergaard, Jacob. **Detecting Coordinated Inauthentic Behavior in Likes on Social Media: Proof of Concept**. Submitted to *Social Network Analysis and Mining (2022).*

Article I is a proof of concept paper that suggests a procedure to mitigate *reaction-based* coordinated inauthentic behavior. The work contributes to the detection of *coordinated inauthentic amplification* by applying computational methods developed with computer-simulated data through an agent-based model (ABM).

Article I takes as a premise that in an honest world, reactions may be informative to users when selecting posts to pay attention to: through the *wisdom of crowds*, summed reactions may help identifying relevant and high-quality content. This is nullified by coordinated inauthentic liking.

To restore wisdom-of-crowds effects, it is therefore desirable to separate the inauthentic agents from the wise crowd, and use only the latter as a voting 'jury' on the relevance of a post. To this end, we design two *jury selection procedures* (JSPs) discarding agents classified as inauthentic. The core idea is this: given a collection of votes from a voting population of agents, a JSP searches the collection for coordinated voting and from the findings classifies agents as inauthentic or authentic. Finally, the procedure returns a subset of the population—the *jury*—whose votes are tallied to serve as proxy for the epistemic quality of a post. I.e., a JSP censors a subset of the population's votes in order to restore wisdom-of-crowds effects for the remainder.

Using machine learning techniques, the two JSPs cluster binary vote data—one using a Gaussian Mixture Model (GMM JSP), one using the k-means algorithm (KM JSP)—and label agents by logistic regression with quality as the dependent variable. We evaluate the jury selection procedures with an agent-based model, and show that the GMM JSP detects more inauthentic agents, but both JSPs select juries with vastly increased correctness of vote by majority. To the best of our knowledge, only [37] attempts to flag agents given just binary vote data, collected in bi-partite structures between agents and posts, i.e., with no added information about e.g. temporal coordination. Such data is obtainable intra-platform by social media platforms.

As any paper containing a model and classification procedures, the agent-based model and jury selection procedures designed in **Article I** make various modeling assumptions. Please note that we thoroughly discuss model assumptions and ethical considerations in Sec. 5 of **Article I**. Here, we provide a brief overview of the roles the key concepts and theories introduced earlier (Sec. 3) play in this paper.

The *information diffusion* of posts into the *social network* and *limited attention* among users are not modeled explicitly in **Article I**. Instead, all agents react (vote) on all posts, thus assuming a complete social network and a non-scarce attention resource. The output data from the ABM hence produces data less sparse than an empirical dataset would be. For simplicity, we have not included *abstaining from voting* as an option or *lack of exposure due to limited attention* in the ABM. However, all steps including the majority correctness score calculation and jury selection would be unaffected; the classification may accommodate for less complete vote participation, too. This simplification is justified since we are interested in producing first simple binary data as input for our jury selection procedures unaffected by less sparse data.

The notion of *quality* is modeled as a property of each post and simplified to a binary representation of either high or low quality, on which agents vote. *High-quality posts* are thought of as relevant, well-produced, or otherwise high quality content. *Low-quality posts* are thought of as, for example, misinformation or disinformation. It is relevant that our approach assumes an agreed-upon notion of truth about the quality of posts for which a commonly acknowledged arbiter exists. This is a fundamental premise of our method: if no such notion exists, majority correctness scores lose their meaning and the assumptions of the classifiers are unmet. Such a notion of quality is of paramount importance in relation to misinformation such as fake news, where, arguably, "objective" quality criteria exists, embodied e.g. by the Principles of Journalism.

Authentic agents vote fully in accordance with their beliefs about quality, that is, they vote sincerely, independently of others, and with a competence strictly above 0.5. The competence of authentic agents is sampled such that authentic agents make mistakes and by no means always able to identify high- or low-quality posts correctly. These mistakes may be thought to stem from e.g. *cognitive biases*. As such, authentic agents failing to vote correctly

may accordingly be thought of as agents unintentionally spreading misinformation. While *social influence* exerted by (unreliable) engagement metrics serves as a main motivation for the paper, **Article I** does not model the *emergence* of social influence.

Inauthentic agents vote *insincerely* with respect to the quality property of posts, and vote based on beliefs about a property *distinct* from quality. With different patterns and varying degrees, inauthentic agents, e.g. *social bots*, coordinate their votes through a different set of properties, violating the competence and independence assumption.

From a *voting perspective*, the results of **Article I** show significant increases in majority correctness scores, providing a direct perspective on the collective epistemic practice of a group of agents, and a more conclusive perspective than just misclassification scores.

Further lending credit to the idea of jury selection procedures—and since the time of writing—Twitter research has worked on leveraging the wisdom of crowds in a different, but related context, as revealed in their crowd-based community notes [155]: The community notes (Birdwatch) aim to provide a reliable judgment of posts. Ideally, the system is effective at discerning quality notes, produced by users and made available to other users. As polarization, a lack of consensus, and a lack of collaboration actively challenge the efficacy of this crowd-based system [155, 159], the Twitter researchers have investigated algorithms that *select a subset of Birdwatch notes* that both inform understanding and viewed as helpful. The paper algorithmically filters a subset of notes to achieve a better than a supermajority voting baseline given all notes.

As a proof of concept, **Article I** provides an argument for the release of reactions data from social media platforms through a direct use-case in the fight against online misinformation.

Author Contribution

Article I is co-authored between Laura Jahn, Rasmus K. Rendsvig and Jacob Stærk-Østergaard. R. K. Rendsvig initiated the project and suggested the study's general idea to implement theoretical contributions explored in [37] in a computer simulation environment.

L. Jahn co-designed the overall structure and all details of the study. L. Jahn was the prime mover behind the production of code for the implementation of the agent-based model, data handling, and data analysis: In detail, the agent-based model was designed and implemented by Laura Jahn and R.K. Rendsvig, who also produced ABM voting data. The classification methods were designed by L. Jahn, J. Stærk-Østergaard and R.K. Rendsvig, and implemented by L. Jahn, who also produced (mis)classification results. The jury selection procedures were designed by R.K. Rendsvig and L. Jahn, and implemented by R.K. Rendsvig who also produced majority correctness score results.

L. Jahn further managed meetings with both co-authors, combined and channeled suggestions from both co-authors, and coordinated submissions and general project management.

During the project, the work received helpful feedback at conferences and seminar talks. L. Jahn prepared and presented an extended abstract of the work at the *ACM Collective Intelligence Conference 2021* (Copenhagen), and gave talks at the *2021 Social Choice and Welfare Seminar Series* at the Center for Mathematical Social Science (University of Auckland), at the *Economics Department* of the University of Copenhagen (2021), at *Social Logic* (Dolomites, 2022), at her *work-in-progress* seminar (University of Copenhagen, 2022), and at the *Observatory on Social Media* (OSoMe) at Indiana University (2022).

The paper manuscript was written by L. Jahn and R.K. Rendsvig, with L. Jahn contributing to writing of all parts of the manuscript. The manuscript was commented on by J. Stærk-Østergaard. L. Jahn and all other authors read and approved the final manuscript. A *co-author statement* is appended to this thesis (Sec. IV).

4.2 Article II

Jahn, Laura and Rendsvig, Rasmus K.. **Towards Detecting Inauthentic Coordination in Twitter Likes Data.**

Article II builds on **Article I** and develops a scripted algorithm to collect suitable data of liking and retweeting users from Twitter. The paper identifies perfectly correlated groups among likes in a case-study around the Danish political Twitter sphere collected under the hashtag `#dkpol`.

Article II is motivated by the premises that *social media* feeds typically arrange and favor posts according to user *engagement* metrics. The most ubiquitous type of engagement, and the type we study, is *likes*. As discussed in Sec. 3.3 above, users customarily take engagement metrics such as likes as a neutral proxy for quality and authority. This incentivizes *like* manipulation to *influence* public opinion through *coordinated inauthentic behavior* (CIB) and has led to the establishment of a marketplace for vendor-purchased engagement. Such CIB targeted at likes is largely unstudied as collecting suitable data about users' liking behavior is non-trivial. Beyond the scripted algorithm to collect liking data from Twitter and the 30 day dataset from `#dkpol`, **Article II** analyzes the script's performance, and identifies large clusters of perfectly correlated users, and discuss our findings in relation to CIB.

In the following paragraphs, we situate **Article II** given the key concepts and theories introduced in Sec. 3.

With focus on *amplification* driven by *users* (Sec. 3.3) unbeknownst to us beforehand, the script we present in **Article II** collects the IDs of liking (and/or retweeting) users of tweets

that satisfy a selected textual query. As such, the script takes a *domain first* perspective on data collection, rather than a *user first* perspective as other work designed to investigate coordinated inauthentic behavior often does. The data collection does not require any prior knowledge about potentially coordinated users. Nor does subsequent data analysis necessarily require the retrieval of additional account data. If so desired, additional account information may be rehydrated via public APIs. We find that the algorithm performs well and collects comprehensive datasets. In particular for high engagement tweets—receiving many likes—-the script collects a high percentage of the liking users despite the Twitter API limit only allowing to collect the most recent 100 likers per tweet.

The focus of the collected data and subsequent applications is on identifying the *effects* of CIB inflating specific tweets. Such effects are robust to changes in the evolution of *social bots* which with varying degrees of automation increasingly emulate authentic users. Our data and applications are not dependent on individual account features nor time-synchronous actions. Instead, they only depend on the like behavior towards an observed tweet given a group of users. As we did in **Article I** with synthetic data, we here, too, only use the binary matrix of users and the tweets they liked and did not like (a bi-partite graph structure). This time using empirical data collected from Twitter, we analyze the now much sparser like (vote) data. Yet, this empirical vote data mirrors many of the same patterns in a dimensionality reduced space when it comes to possibly coordinated behavior.

For the domain `#dkpol`—arguably small in an international context—the script had a reasonably low rate of missing liking users, and misses more than 10% of liking users in only 3% of cases when run continuously for 30 days. Such a targeted dataset cannot be obtained directly through any of Twitter's data access options.[4]

Our findings highlight the societal need to address potential CIB-caused misrepresentation of political views and the spread of possibly harmful *low-quality content* and *misinformation* in the online public sphere. We acknowledge the difficulty in identifying *coordinated* behavior. We apply the strictest measure (perfect coordination) to label behavior as coordinated since such behavior will also be labeled as coordinated using any less discriminating measure. From a *voting perspective*, perfectly correlated users are users that have all liked the same set of tweets, and have not liked *all the same tweets* (what we considered downvotes in **Article I**). We conclude that several large groups of users with perfectly identical liking behaviors are unlikely, suspicious, and warrant further analysis. **Article II** does not go as far as labeling users as authentically or inauthentically coordinated.

[4] **Please note that we include a disclaimer for Article II in the subsequent section, Sec. 5.2**. The disclaimer concerns Twitter API updates that affect the premises under which this paper was developed. We introduce the disclaimer in the methodology section because the update concerns the data collection method.

Author Contribution

Article II is co-authored between Laura Jahn and Rasmus. K. Rendsvig. L. Jahn initiated the project and suggested the study's general idea to validate the results from **Article I** with empirical data collected from Twitter. L. Jahn co-designed the overall structure and all details of the study. L. Jahn was a prime mover behind all theory considerations, first draft writing, and the development of the code for data collection, and data analysis. L. Jahn and R. K. Rendsvig together designed, developed, and implemented the scripts for data collection and data analysis.

L. Jahn further initiated and managed the application for *Approval of data collection and processing of personal data in the research project* that was granted by the faculty secretariat of the University of Copenhagen. L. Jahn also researched and held meetings with various data sharing platforms and identified suitable platforms offering hosting facilities in compliance with the Twitter terms and policy.

During the project, the work received helpful feedback at conferences and seminar talks. L. Jahn presented the work as a poster at *Social Informatics 2022* (University of Glasgow), and presented the work at her *work-in-progress* seminar (University of Copenhagen, 2022) and at the *Observatory on Social Media (OSoMe)* at Indiana University (2022).

The paper manuscript was written by L. Jahn and R. K. Rendsvig, and both read and approved the final manuscript. A *co-author statement* is appended to this thesis (Sec. IV).

4.3 Article III

Jahn, Laura and Rendsvig, Rasmus K. and Flammini, Alessandro and Menczer, Filippo and Hendricks, Vincent F.. **Friction Interventions to Curb the Spread of Misinformation on Social Media.**

Article III is motivated by the rise of *social media* communication platforms enabling the spread of information at unprecedented speeds and scales, and with it the proliferation of *high-engagement*, *low-quality content*. This development in today's *attention economy*—a failure of the wisdom of crowds—may be attributed to *algorithmic bias* in interplay with *cognitive biases* and *bounded rationality* as well as *social biases* and *social influence* affecting the amplification of low-quality content, violating the *competence* and *independence assumption* of the Condorcet Jury Theorem.

A suggestion is that *friction*—behavioral design interventions that generally encumber given actions, here specifically making *reacting* to content (i.e., *liking* or *sharing*) more cumbersome—might be a way to raise the quality of what is spread online. In **Article III**, we survey

related work on friction online as well as implementations across the social media landscape, and study the effects of friction prompts with and without quality-recognition learning. Experiments from an agent-based model (ABM) suggest that friction alone decreases the number of posts without improving their *quality*. On the other hand, a small amount of friction combined with learning increases the average quality of posts significantly and improves the system's capacity to discriminate between high- and low-quality posts.

In the ABM, friction is restricted to one-click reactions, such as sharing, in which agents may face a friction prompt. Hence, the intervention targets *users* when they—possibly unintentionally—*amplify* content. The intuition is that friction triggers agents to pause, potentially impeding their sharing activity. Such a pause may curb the effects of acting on the basis of *social influence cues*, *heuristics*, *system II thinking*, or *biases*, both *cognitive* and *social*. Agents may resume re-sharing the chosen post after having spent mental resources, or passed a quiz. On the other hand, agents may not resume sharing the chosen post after re-considering or failing to comply.

Agents may learn through exposure to a friction prompt, e.g., through deliberation-triggering nudges or educational quizzes that remind agents to *pay attention to quality*. The next time an agent is about to re-share a post, an agent who has learned no longer re-shares the most engaging post, but instead chooses a post to re-share based on quality. Learning thus aids (re-)gaining *competence* and relying less on heuristics but deliberatively assessing a post's quality, (re-)gaining *independence*.

We call this type of learning *quality-recognition learning*, drawing intuitions from research both on priming effects and nudges, and on testing effects and retrieval practices shown to boost learning.

Like **Article I**, **Article III** deploys an agent-based model, and makes modeling assumptions. Here, we provide a brief overview of these assumptions and of the roles the relevant key concepts and theories introduced in Sec. 3 play in this paper.

In **Article III**, we explicitly model *information diffusion* through a minimal social media model of information sharing. In the ABM, social media posts may be created or shared by agents, and appear on the news feeds of other agents. Each agent's news feed consists of a bounded number of posts, all shared by agents they follow. The bounded news feed models *limited individual attention*, which gives rise to heavy-tailed distributions of post popularity and lifetime.

Networks in the ABM are directed graphs with vertices representing agents and edges representing follower relations. To capture the characteristic presence of hubs, we construct networks using a directed variant of the Barabási-Albert preferential attachment mechanism. Similarly, to capture the characteristic presence of clustering, we generate additional closed directed triads. Networks are integrated exogenously.

As related to the notion of *quality*, posts in the ABM may vary both in quality and in how *engaging* they are. In this paper, quality models a property such as accuracy or relevance of posts. Quality is interpreted as a *non-subjective* property of posts. Engagement models the quality of a post *as perceived by agents*. The quality and engagement of a post are sampled such that low-quality posts are more likely than high-quality posts, and low-engagement posts are more likely than high-engagement posts. Quality and engagement are sampled independently to reflect that high quality and high engagement do not necessarily coincide.

While the ABM does not encode agent types (such as *inauthentic actors* like in **Article I**), low-quality posts may stem from a variety of accounts, such as authentic human users or social bots, broadly understood.

Based on the preliminary evidence suggested in **Article III**, we propose a friction intervention with a learning component about the platform's community standards, to be tested via an online field experiment. We discuss how to go about conducting a field experiment and map an experiment on community standards micro-exams as friction prompts. The proposed intervention is argued to have minimal effects on engagement and may easily be deployed at scale, as it does not require labeling of content or detection of inauthentic actors. The paper concludes with discussing possible directions of policy desiderata.

Author Contribution

Article III is co-authored between Laura Jahn, Rasmus. K. Rendsvig, Alessandro Flammini, Filippo Menczer, and Vincent F. Hendricks. V. F. Hendricks suggested the general idea to study friction related to community standards. F. Menczer suggested to study friction through an agent-based model (ABM). L. Jahn visited the *Observatory on Social Media* at the Luddy School of Informatics, Computing, and Engineering (Indiana University) to collaborate with F. Menczer and A. Flammini in person on this article, and continued the collaboration beyond the visit.

L. Jahn was the prime mover behind all implementation and design decisions, and a prime mover behind theory considerations and the main contributor to implementing the ABM code for data production and data analysis. The ABM in this paper builds on existing work by [134], but L. Jahn brought forth substantial and novel features and code to implement friction and learning in the model. L. Jahn further planned and led meetings with the international co-authors, combined and channeled suggestions from all co-authors, and coordinated general project management.

The paper manuscript was written by L. Jahn and edited by all authors. All authors read and approved the final manuscript. A *co-author statement* is appended to this thesis (Sec. IV).

5 Methodology

This section presents the main methods—agent-based models and empirical Twitter data collection—in greater detail. **Article I** and **Article III** use agent-based models (ABMs). In **Article II**, data was empirically curated through Twitter's Application Programming Interface (API). We further review the data processing and analysis employed in the articles, and present data descriptions as well as ethical considerations. Modeling assumptions are discussed throughout.

Various methods may set out to study the detection and prevention of the amplification of low-quality content. Ideally, real-time detection of influence operations or randomized controlled trials on preventive behavioral interventions may be conducted from inside a social media platform such as Twitter or Facebook. Such environments are rarely available to academic researchers. To study and detect amplifiers of low-quality content, researchers require data to develop methods. Such data may be obtained through APIs provided by social media platforms. To study behavioral design interventions, researchers need access to social media environments, ideally from within a platform ecosystem. Access and data, however, is only made available by some platforms or in insufficient quantity and quality. Extensive field experiments in an emulated social media environment to test interventions prove difficult. Given the scope of this thesis, these considerations motivate us to use ABMs and collect data through APIs.

5.1 Agent-Based Models

Agent-based modeling uses computer simulations to study (*iterated*) *interactions* between *agents* that give rise to a target phenomenon. Many disciplines widely employ agent-based models (ABMs), among them (computational) social science, history, philosophy, political science, economics, ecology and epidemiology [64].

An ABM produces synthetic data about a typically complex yet formally represented social interactive situation. The purpose of the generated ABM data may roughly fall in one or more of the following three categories:

First, such data may help to learn about the real world and explain, understand, or predict phenomena.

Second, ABM data may provide insights about mathematical models. Some mathematical models of real world social interactive situations become so complex that it is difficult to establish theorems that describe their behavior. Here, simulation data may help to provide a data-based statistical description or visualization of the model's behavior, adjust theorems, and inspire conjectures.

Third, ABM data proves useful while learning about models that take data as input. If it is difficult or costly to collect data about a real-life social interactive situation, ABM data may substitute for real-world empirical data points. The simulated data may be used to test various methods' effectiveness, data features and quantity requirements, etc., prior to collecting empirical data.

The ABM developed in **Article I** falls into the third category. Over *synthetic vote data* generated by the ABM, we test and validate the machine learning-based jury selection procedures. Validating with synthetic data in this case circumvents three main challenges in detecting coordinated inauthentic users: the lack of reproducibility, the lack of available data, and the lack of ground truth. Meanwhile, the downside is that synthetic data has limited ecological validity since it was not empirically observed. We discuss reproducibility, data availability, and ground truth issues in Sec. 1.1 of **Article I**.

The ABM developed in **Article III** falls into the first category. The ABM—a minimal, highly abstract model of information sharing—emulates the target system of real-world social media in a simplified manner. As such, the ABM is a toy model. We implemented friction and learning in order to *understand* how these additions *possibly* impact information sharing[108]. In terms of *prediction*, we discuss policy desiderata considering how potential choices on friction and learning translate into future states given the ABM results, treated as preliminary evidence.

The ABM in **Article III** also falls into the second category, as—by choosing an ABM over a purely theoretical model—we describe the model behavior through a data-based statistical method. Then, we measure the outcome of the ABM through pre-defined metrics, such as Kendall's correlation coefficient assessing the system's capability to discriminate between high- and low-quality posts. This way, we can explore the social interactive situation of information sharing that is affected by different friction and learning combinations despite the iterative nature and possibly large networks.

Agents in an ABM are defined by homo- or heterogenous behavioral rules. Rules may follow, for example, game-theoretical best response reasoning, reinforcement learning, simple conditional "if, then" statements, or random sampling. A behavioral rule usually specifies action resources: a set of possible actions performed by the agent when a rule is triggered. Behavioral rules are functions of one or more parameters, jointly defining a *parameter space*. The parameter space then is the set of all possible parameter settings. Parameter spaces may thus become high-dimensional, and may even include uncountable dimensions.

Article I defines a heterogenous set of agent types and defines behavioral rules for each type, following "if, then" statements and (random) sampling from distributions. For instance, agents were equipped with a sampled competence level and formed beliefs about sampled properties of posts. Agents took action to either upvote or downvote a post conditional on the beliefs they formed, according to a supplied protocol, or when a condition was fulfilled.

In **Article III**, agents are defined by slightly more homogenous behavioral rules, that is, the same rules apply indiscriminately to all agents. The behavior of each acting agent is probabilistically determined by a set of parameters. Agents, for example, create a new post to share with their followers with $p = 0.5$, else they re-share a post from their newsfeed. Subsequent choices (e.g. which posts to share) are informed by earlier behavior ("if, then"), or sampled. Properties of posts affecting behavior of agents are similarly sampled from probability distributions.

Agents in **Article I** do not interact with one another. In **Article III**, however, agents interact through the edges of a network emulating a social network, representing friend and follower relations. The network topology in exogenously imposed with typical characteristics from social networks and remains fixed during simulation runs. As network structures and individual behaviors are intertwined, social systems like social media networks often show nonlinear and unpredictable behavior. Through an ABM as a modeling choice, we are thus able to understand such results better as the emergence of phenomena is explicit while control over parameters is intact [154].

In order to learn what data the behavioral rules yield across a parameter space, *parameter sweeps* are performed: The ABM is run for (manually or randomly sampled) subsets of the parameter space. These subsets are considered representative of the full space with respect to model behavior. The ABM will output data for each defined parameter setting, often including robustness checks and randomness control through seeds and Monte Carlo methods (when random sampling is involved). A run is an application of the behavioral rules to a parameter setting and produces a (perhaps complex) data point. Runs may comprise simple one shot games, e.g., a single voting round, or runs may develop temporally (ticks), e.g., as in the Schelling segregation game [115]. A run is terminated given some endpoint, such as when reaching the end of a life span measured in ticks, a fixed point, or when enough data is produced.

Parameter spaces in the ABM of both **Article I** and **Article III** are large, sweeping through different noise settings, network samplings, behavioral parameters, and robustness repetitions, while controlling randomness through seeds. While **Article I** creates data through repeating single voting rounds, the ABM in **Article III** develops temporally and converges to a point where changes in the average quality grow small using exponential moving average smoothing. The stop conditions and resulting data are explicated in the articles in greater detail. The **Article I** ABM and the **Article III** ABM both use parallel computing [97], implemented in R and Python respectively, to expedite runtimes.

ABMs are used to model agents in complex yet formally represented social interactive situations. ABMs are particularly strong at modeling more complex, noisier, and "messier" properties and conditions through implementation of probabilistic choices and imperfect agents. Both ABMs in **Article I** and **Article III** make use of imperfection and control for agents

making mistakes (e.g. failing in competence, or not applying previously learned behavior). The ABM in **Article I**, aiming to create a dataset later used to classify agents, deliberately produces highly noisy dataset. Here, agents' behavioral rules largely overlap between different agent types, making the classification task more difficult.

As an exploratory tool, ABMs may help gain intuitions pertaining to the workings of a mechanism, especially as ABMs can model small, graded correlations. Eventually, ABM results based on synthetic data gain validity through careful technical implementation and documentation, parameter sweeps, robustness checks, and matching with empirical data. With empirical data, **Article II** validates some of the results from **Article I** based on synthetic data.

For further discussion, Klein et al. [64] give an exhaustive overview over agent-based modeling across disciplines.

While we are aware of existing ABM simulation packages and programs, such as *Netlogo* [152], *Laputa* [6, 94], or *Hashkat* [111], the ABMs were implemented in *R* and *Python*, respectively. **Article I**'s ABM was implemented from scratch in R in order to retain freedom in agent design. The simplicity of the encoded behavior and generated data did not need advanced features. Since the ABM in **Article III** builds on existing work by [134], we continued to build in Python.

5.2 Empirical Data Collection

Social media and the discourse carried out on the platforms provide a lens into the public sphere [17] and is one of the breeding grounds for the amplification of misinformation. To address the research question on how to curb the amplification of misinformation online, learning through empirical data gleaned from social media seems inevitable. Motivated by the work in **Article I**, **Article II** sets out to collect empirical social media data. The data collection sought after a dataset of public posts from social media, each with a given post-ID and and array of user-IDs that publicly engaged with any of the posts through a reaction, such as a like. As we in **Article I** search for systematic behavior on the user level, the data must allow that we consistently map reactions per posts to (anonymized) user-IDs.

To the best of our knowledge, none of the existing datasets include user-level data on reactions. For example, the impressive 2023 dataset of 'complete 24 hours of Twitter Data' does not contain datapoints of liking or retweeting users [103]. Neither Meta, Twitter nor Reddit supply this data with suitable scope [11, 98]. For example, Twitter's transparency reports do not include information of liking or retweeting users [140]; data access points and analytics tools like Meta's CrowdTangle by default only allow to partially collect data from public sites via post content and interaction metrics endpoints. CrowdTangle does not allow for the col-

lection of user-IDs of the users that have publicly reacted. Only Twitter provides post-ID *and* user-ID of users that have publicly engaged (liked or retweeted) with a tweet.

Twitter has long served as a vast research resource due to their data-sharing policies via a variety of free Application Programming Interfaces (APIs), among them a specific Academic Research API [17, 103, 104]. In February 2023, Twitter announced major changes to their data-sharing and API policies, most notably a discontinuation of free API access, including the Academic Research API. Sec. 5.2 discusses the implications of these events for this thesis. In the current section, we describe how Twitter's (soon discontinued) API landscape before these events has contributed to data collection and shaped the research in this thesis.

Twitter offers *filter* and *sample APIs* (together often referred to as their *streaming API*) providing *real-time* tweets filtered by a search query or a sample of all tweets. Twitter's *search APIs*, on the other hand, can access historical tweets and user account information. Both the streaming and search APIs come with different pricing models and elevated access restricted to certain user groups or paying clients [17, 104].

Among Twitter's *commercial* streaming APIs, the *Decahose API* is the only API providing reaction data: the stream lists 100% of liking user-IDs, but only of a random 10% sample of all tweets, making a targeted analysis of a specific political discourse impossible [138].

Supplementing Twitter's *free* access search APIs, the platform introduced an API for *Academic Research* [136] in 2021, granting extensive search access to the full Twitter archive of all historic tweets and supplemented by fewer rate limit restrictions. Pfeffer at *al*. [104] argue that the endpoints in the Academic Research API to date deliver the best data. Importantly for this thesis, the Academic Research API provides endpoints for researchers to look up liking and retweeting users of tweets.

We have made use of the Academic Research API to collect the liking and retweeting users. Twitter, however, does not give direct access to comprehensive lists of such IDs, but only releases the user-IDs of the 100 most recent liking/retweeting users of any single post.[5] Additionally restrictive, at most 75 such lists may be requested per 15 minutes.

One of **Article II**'s contributions is a scripted algorithm that live-collects liking and/or retweeting users balancing Twitter's request limits. Using a running survey approach, the script retrieves IDs of the most recent liking users of tweets satisfying a specified text query (e.g. a keyword or hashtag of a chosen political debate), while timing retrievals by taking into account Twitter set rate limits API for Academic Research Access. This way, we collect a comprehensive dataset. For the specifics, we refer to **Article II** (Sec. 2) for an extensive discussion of the data collection script. Below we discuss two broader data collection decisions.

[5] Twitter has since updated this endpoint to allow for the collection of more comprehensive lists. We discuss the implications in a disclaimer in Sec. 5.2.

The data collection curates *domain-specific* reaction data, e.g. around a discourse tagged with a specific hashtag, instead of a surveying preselected groups of users. When identifying correlated reactions given just tweet and like/retweet data, one grasps, firstly, which specific tweet(s) are targeted by a potential influence operation. Secondly, one may learn which users are involved in the metric inflation. This is in contrast to existing methods (e.g., [116]), where prior knowledge over the identity of users is needed to collect the retweeting user-IDs to reconstruct reaction data.

The domain we chose to survey for liking and retweeting user behavior is the Danish political Twittersphere `#dkpol` ("DenmarK POLitics"). While the fact that the authors are situated in Denmark undoubtedly played into the decision, the hashtag also captures a Twittersphere small enough to justify the ambition to collect comprehensive datasets despite Twitter's API limits. Under `#dkpol`, citizens, organizations, politicians, and journalists from across the political spectrum air, discuss, and orientate themselves about current debates in Danish politics. It is *the* centralized, place-to-be source of information on the debates of the day.

The data collection script live-collects the desired reaction data. This choice is on the one hand motivated by Twitter's API limits and Twitter's endpoint constraint limiting researchers to collect a total of 100 liking accounts per tweet for all time. If a tweet receives a total $100+n$ likes, it is impossible to collect the first n likers despite generous historical access. On the other hand, the choice is motivated by an interest to reduce *temporal bias* [83, 104, 132]. Traces of coordination may be altered or deleted [132]. Not only reactions and users accounts get deleted and/or retracted. Tweets collected from the past are usually under-represented in non-live collected datasets [104].

Addendum to Empirical Data Collection

On February 2, 2023, Twitter announced that the platform will discontinue the free access to its APIs. Twitter confirmed on February 10, 2023 that the Academic Research API is included in this discontinuation, to take effect February 13, 2023. The APIs that were freely accessible before will be replaced with a low-level of API usage for $100/month, about which details are still unclear.

Researchers across the globe have since reacted negatively to this threat to thousands of research projects. An open letter signed by hundreds of researchers, universities, and institutes appeals to Twitter and policy makers to protect access to the data which the Twitter API provides [35, 68].

While the data we have collected is of value in its accessible and archived version, we and the research community lose access to reproduce, replicate, and supplement the data. The scripted algorithm, meant as a tool for the research community to collect their own datasets,

may not be usable in its current form after February 13. At the time of writing, it is unclear what a "low-level" API usage for $100 entails, and whether volume-scaled fees (e.g., per request) apply beyond low-level usage. If the latter becomes an option to upgrade usage, the algorithm may prove helpful as it can be parametrized to spend requests smartly and frugally during live-collection.

Disclaimer for Article II

During work on **Article II** (spanning 2021 and 2022), Twitter changed their endpoints for collecting data on liking and retweeting users to allow for *pagination*. Pagination allows collecting *all* liking and retweeting users of a tweet, yet subject to request limits. With pagination, results—lists of 100 users per request—are delivered in reverse-chronological order. The limit of maximally 75 requests per 15 minutes windows remains unchanged. Together, instead of only being allowed to collect the most recent 100 liking or retweeting users per tweet, one may now collect up to 7,500 liking or retweeting users per 15 minutes on one tweet, and continue to request even more users after a 15 minutes period of pause. The first page of results typically contains the most recent users, and the last one the oldest. Prior to this change, Twitter only allowed to retrieve the first page. This restriction originally motivated our live-collection approach surveying engagement metrics of tweets to prioritize request allocation to high-engagement tweets.

The change in endpoints was implemented in January 2022, but is unfortunately not reflected in the workings of the scripted algorithm of **Article II**. The scripted algorithm is built on the premise that the *"liking users endpoint* [and retweeted-by endpoint] *limits you to a total of 100 liking accounts per tweet for all time*", which is also what is *still* stated in official API documentation (February 2023)[6]. Without reference to a version update, the specific endpoint documentation has changed to include a pagination parameter[7]. The work on **Article II** began in 2021, which is when the documentation was consulted and the idea for a resource-balancing algorithm was born. The subtle change in documentation in early 2022 and the—to date—conflicting information in the documents has led to this oversight on the side of the authors.

In an effort to improve the collected data we make available to the research community, we have since used pagination to re-collect the liking and retweeting users. Deletions possibly

[6] Twitter API documentation February 2023: `https://web.archive.org/web/20230209111342/https://developer.twitter.com/en/docs/twitter-api/tweets/likes/introduction`

[7] Twitter API endpoint documentation, 2021: `https://web.archive.org/web/20211021132828/https://developer.twitter.com/en/docs/twitter-api/tweets/likes/api-reference/get-tweets-id-liking_users`
2022: `https://web.archive.org/web/20220122234738/https://developer.twitter.com/en/docs/twitter-api/tweets/likes/api-reference/get-tweets-id-liking_users`

introduce a temporal bias here nevertheless. We include the supplemented data in the distribution of the dataset (Sec. 5.4).

In a scenario where the new, paid low-level API access offers data access options akin to the Academic Research API (see Sec. 5.2) on which this algorithm relies, there is still use for the algorithm. Since pagination does not come for free but costs requests, using pagination on one tweet in a 15-minutes window also entails (in the worst case) not live-collecting (any) liking or retweeting users of other surveyed tweets. A non-live collection induces temporal bias in datasets. The idea behind our script—balancing Twitter rate and request limits to live-collect comprehensive datasets—thus does not lose its relevance. A future version of the script may include a parameter to control of the use of pagination to assign its usage to covering high-engagement tweets attracting many likes quickly. We expect that making use of pagination will increase performance in more complete datasets, as pagination may be used as a backup option, should a tweet have received more than 100 likes/retweets since it was last scraped.

In a scenario where the new, fee-based low-level API access (see Sec. 5.2) can be leveled up to the functionality of the Academic Research API endpoints through a volume-scaled fee subject to request usage, our script yet again retains its usefulness. The script spends requests frugally. There is a hard limit of one request per tweet, spent when like/retweet count thresholds are triggered. If pagination was used, requests may be spent as long as there are results to be returned, potentially quickly growing expensive for high-engagement tweets.

5.3 Data Processing and Analysis

This thesis uses simple methods from statistics, machine learning, and network science to analyze synthetic and empirical data. Context and exact methods are introduced separately in all articles. On a higher level, we here discuss some of our choices in modeling and how they fit into the current paradigm of detecting inauthentic coordination online. This section mainly pertains to **Article I** and **Article II**. We discussed **Article III**'s main method in Sec. 5.1.

In recent years, there has been a shift from individual-account-based identification of inauthentic or bad actors online to group-based detection methods. And example of the former is *Botometer* [114, 158]. Botometer's supervised, feature-based approach considers accounts one at a time. If coordinated accounts do not seem bot-like at an individual level, Botometer does not pick up on group anomalies based on suspicious similarity. Similarity is often used as a proxy for coordination (e.g., [93, 96]). Unsupervised, group-based methods are promising "in the arms race against the novel spambot" [23, 95]: Social bots evolve fast, emulate human behavior to an ever increasing extent, and may act in an orchestrated, coordinated way, inconspicuous at the individual level.

The shift towards group-based detection is accompanied by a lack of ground truth and labeled datasets on which models could be trained. Labeling data is unfeasible, if not impossible at scale, not least since coordinated tactics change rapidly [144]. Moreover, qualified guesses may be made based on suspicious similarities in behavior or profile features, but it often remains unknown whether two users' actions are authentically correlated or inauthentically coordinated—or how many fully or partially automated accounts exist in a total population [10, 16, 82, 83, 112]. We extend the discussion on the empirical problems of a lack of ground truth in **Article I**, Sec. 1.1 and **Article II**, Sec. 1.2. In **Article I**, we circumvent the problem by relying on computer-simulated data allowing us to study different degrees of coordination and (in-)authenticity. This provides us with a ground truth, which is of a purely synthetic nature.

In **Article I**, we rely on unsupervised methods that disclose coordination. We use one supervised element—logistic regression—to apply labels to agents clustered into groups. In the logistic regression, we use that authentic votes correlate with post quality (possibly allowing for noise in observing quality). Other subjective assumptions could come into play steering labeling while producing equally efficient jury selection procedures. In **Article II**, we similarly use unsupervised clustering methods detecting perfectly correlated data. A controlled experiment involving vendor-purchased likes serves as a notion of ground truth.

Besides limiting supervision, the input data requirements of the methods laid out in **Article I** and **Article II** are sparse, simple, and require little pre-processing. They are not dependent on individual account features beyond the reaction data, nor time-synchronicity in reactions, and may accommodate missing data. To exemplify the general data structure, consider n tweets or posts. The set of all observed or synthetically curated liking users or agents is given by $Likers = \cup_{k \leq n} Likers_k$, with $Likers_k$ the set of users that liked post k. With $m = |Likers|$, we then compress our data to a binary $n \times m$ vote matrix with entry values in $\{0, 1\}$, each row representing a tweet or post, each column a user or agent. In this matrix $\mathbf{L}$, the entry $\mathbf{L}_{i,j} = 1$ if user or agent i has liked tweet or post j, and 0 else. A matrix $\mathbf{L}$ may be visualized as follows:

$$\mathbf{L} = \begin{array}{c} \\ tweet_1 \\ tweet_2 \\ tweet_3 \\ \vdots \\ tweet_n \end{array} \begin{array}{c} \begin{array}{ccccc} user_1 & user_2 & user_3 & \dots & user_m \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 0 & 1 & 0 & \dots & 1 \\ 0 & 0 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix} \end{array}$$

Based on data of this structure, synthetic or empirical, both **Article I** and **Article II** *classify* users or agents as 'inauthentic' (**Article I**) or 'suspiciously correlated' (**Article II**) based on a given threshold.

Article I models various authenticity and coordination degrees; a classification threshold in the jury selection procedures controls the balance between *vox populi* (to minimize false positive errors, i.e., to preserve as many authentic agents in the jury as possible) and *precaution* (to minimize false negative errors, i.e., to eliminate as many inauthentic agents from the jury as possible). We precautiously fix that agents should be labeled authentic 4 of 5 times to be classified as authentic. This choice also preserves vox populi to a large extent. The resulting precautious jury selection procedures still exhibit low false positive misclassification errors. The approach remains flexible to emphasizing vox populi by lowering the classification threshold.

In contrast, in **Article II**, dealing with real users this time, we apply the strictest measure (perfect coordination) to label behavior as coordinated. The labeled behavior will also be labeled as coordinated using any less discriminating measure. The approach thus values vox populi and is careful with assigning coordination labels (also see the ethical considerations in Sec. 5.4).

The method in **Article II** is not designed to identify all coordinated inauthentic behavior in likes. There may very well be nuances and less than perfectly correlated inauthentic behavior. To answer whether a collection of tweet likes exhibit first signs of coordinated inauthentic behavior, we propose the method only as valid for positive answers: if this strongly discriminatory method finds such signs, then methods with lower thresholds of coordination should, too. If the method does not find signs of coordination, we would deem it fallacious to take this as evidence that no coordinated inauthentic behavior occurred.

The vendor-purchased coordination which we detect is achieved via weak ties when the data structure $\mathbf{L}$ is considered as a bipartite graph instead of a matrix. In contrast to existing work (e.g. [116]), the present approach does *not* need to filter the data for strongly tied communities, highly influential users and superspreaders, or very active or users that, e.g. like a minimum number of times within a short period. Without filtering, we are able to group users with such behaviors together. The strict measure applied in **Article II** also has computational advantages. The approach of binning identical like profiles into a cluster saves pairwise computations of distance or similarity metrics.

5.4 Data Descriptions and Ethical Considerations

Data Description

In an effort to describe the empirical material and make the datasets available to the research community in a transparent way, we follow the *datasheet* documentation approach as suggested by Gebru et al. [40] and situate the data with regard to the *FAIR* principles [153]. Together with Sec. 5.1 and 5.2, and explicit descriptions of the datasets used in each of the

articles, we hope to provide a sufficient account of the criteria for selection of the empirical material. We refer to the datasets which went into this thesis with abbreviations **Data I**, **Data II**, **Data III**. **Data I** refers to the *synthetic data* and *data analysis scripts* produced with the ABM in **Article I**. **Data II** refers to *data empirically collected* through the Twitter API and respective *data analysis code*, used in **Article II** . With **Data III**, we refer to data and script affiliated with the ABM and analysis in **Article III**. Please refer to Sec. 4 for the identification of creators of the datasets. **Data I**, **II**, and **III** were supported by the *Carlsberg Foundation*. **Data III** was supported in part by *Lilly Endowment, Inc.*, through its support for the *Indiana University Pervasive Technology Institute* providing a computing cluster.

Composition The instances (datapoints) in **Data I** represent binary up- and downvotes in $\{0, 1\}$ by agents on posts. An example of such an instance is discussed in Sec. 5.3. Per instance, labels per agent types and per post concerning its quality are available. Additional information on states was logged. For each state (3 noise levels), we produced 100 runs, each based on a random seed, over 1,000 voting rounds, producing a dataset with $3 \times 100{,}000$ (state, vote profile) pairs given 1,900 agents, in total producing 570,000,000 datapoints. Given its synthetic nature, no data is missing from this collection. Data production faced minimal errors in the parallel computing set up. In an effort to retain reproducibility, re-starts and respective seeds to control random sampling are documented in the code.

In **Data II**, the instances (datapoints) in the pre-processed datasets similarly represent binary votes in $\{0, 1\}$ by users on tweets containing `#dkpol`, namely whether a user has liked a tweet (1), or not (0). A key difference to **Data I**, however, is that "downvotes" 0 are not collected from Twitter, but complemented by us, and represent that a user has not liked a given tweet. The pre-processed dataset analyzed in **Article II** is of tweet–user dimension $13{,}243 \times 47{,}714$, producing more than 600,000,000 datapoints. **Data II** does not contain information on *private* users and only collects data from *public* user profiles. Completeness of the data is further discussed in Sec. 3.1 of **Article II**.

The pre-processed datasets in **Data II** are self-contained and contain both username and tweet-ID information, i.e. usernames and tweet-IDs do not have to be rehydrated. In case researchers undergo further analysis and look-ups of users or tweets, information may be rehydrated from Twitter. There are, however, no guarantees that users nor tweets remain on Twitter over time. The pre-processed dataset in its archival version remains self-contained (see also *Distribution* below).

We also make the *raw form* of the collected data in **Data II** (as opposed to the preprocessed form) available. The raw form of the data, in addition to usernames and tweet-IDs, contains datapoints on time of tweet creation, like count, and retweet count. Given that neither the preprocessed data nor the raw data contain tweet text, we believe that **Data II** does not contain any sensitive, offensive, insulting, threatening, or otherwise anxiety causing content.

The same holds for **Data I** and **Data III**.

As **Data II** contains user-level information (usernames) that might identify individuals, approval of data collection and processing of personal data in the research project was assessed and granted by the faculty secretariat at the University of Copenhagen (Sec. v). The approval emphasizes that the processing of personal data in the project is in accordance with the rules of the *European General Data Protection Regulation*, *Regulation 2016/679* on the protection of natural persons in regard to the processing of personal data. It was made public on the authors' university websites that the study would be undertaken. We reserve space to discuss the ethical implications of using user-level data in Sec. 5.4.

The instances in **Data III** represent 3 descriptive metrics (quality, diversity, discriminative power) per parameter combination (sweeping through values of friction, learning, and network input), each run 50 times controlled through a random seed, measured at the end of each ABM run. Jointly, this accounts for more than 40,000 resulting descriptive metrics. Additional information on fixed parameter settings was logged. The ABM may also produce extensive verbose output, making network information at each time step (used for robustness checks and experiments) explicit. Given its synthetic nature, no data is missing from this collection.

Collection Process Both **Data I** and **Data III** were collected using agent-based models implemented in *R* and *Python*. Data II was collected using the Twitter API for Academic Research. Data collection is described in detail in Sec. 2 of **Article II**. **Data II** was collected between Spring 2022 and Fall 2022.

Pre-processing/Cleaning/Labeling Data in **Data I** was cleaned of labels (stored in separate vectors) and pre-processed into bootstrapped, smaller datasubsets, on which we performed the jury selection procedures. **Data II**'s raw data was pre-processed into tweet-user matrices as described in Sec. 5.4 and in Sec. 5.3, not capturing temporal information. No pre-processing, cleaning or labeling was done to **Data III**.

Uses All three datasets have been used in the papers that form part of this thesis. These papers are referenced in the respective repositories (see *Distribution* below for links). The datasets may be used by the research community for anything related to modeling voting behavior or in understanding coordination on social media.

Distribution/Maintenance **Data I** and **Data III** are reproducible and code to do so is publicly available on the GitHub repositories *Coordinated-Inauthentic-Behavior-Likes-ABM-*

Analysis (**Data I**)[8] and *Friction-Social-Media-Model* (**Data III**)[9]. **Data II**'s code for data collection and analysis is publicly available on the GitHub repository *Get-Twitter-Likers-Data* (**Data II**)[10]. All GitHub Repositories are licensed under the terms of the GNU General Public License v3.0 (gpl-3.0). Datasets collected in **Data II** are available to the research community on the archival repository Harvard Dataverse and findable with a Document Object Identifier (DOI) [59][11]. To comply with the Twitter terms, access to the data on Harvard Dataverse is granted when researchers actively agree to the Twitter Terms of Service, Privacy Policy and Developer Policy. Date(s) of data collection, the exact search query, the utilized API and used endpoints are transparently contained in the code (and data) on Harvard Dataverse.

Additional datasets from **Data II** which have not been analyzed in this thesis are made available to the research community in the Harvard Dataverse collection and hopefully prove useful to someone in the future. Among others, we collected data filtering for the German hashtag `#bundestag` and from the Danish Twittersphere with respect to the Danish National election in Fall 2022.

Code and data are *findable*: The Harvard Dataverse collection comes with a DOI. GitHub repositories are findable through the distributed links. Code and data are *accessible* and may be downloaded. Code and data are *interoperable*: Code for reproducibility and raw data are available in standard formats such as *csv* or *pkl*, and standard coding packages. Workflows are written as *R* scripts, *jupyter noteboks*, or as *shell* or *Snakemake* scripts for *python*/*pytorch* scripts. With proper attribution, code and data are *reusable* upon agreement to the respective license (**Data I**, **Data II**, **Data II**) and, in case of **Data II**, upon actively accepting the Twitter terms. The latter is the reason why the dataset is published in a restrictive way on Harvard Dataverse. Only if interested third parties, too, actively agree to the Twitter terms, we can share the data in compliance with the Twitter terms. This process is implemented through Harvard Dataverse's guestbook feature. Code for the data collection in **Data II** contains extensive documentation and shell scripts that guide a potential user through the setup.

Data I, **Data II**, **Data III** are made available from 2023. The authors will maintain the datasets and may be contacted through their private email addresses. If others want to extend/augment/build on/contribute to the datasets and/or code, they may contact the authors or raise GitHub issues.

[8] See `https://github.com/LJ-9/Coordinated-Inauthentic-Behavior-Likes-ABM-Analysis`.
[9] See `https://github.com/LJ-9/Friction-Social-Media-Model`.
[10] See `https://github.com/humanplayer2/get-twitter-likers-data` [57], and [59].
[11] See `https://doi.org/10.7910/DVN/WRUNZD`.

Further Ethical Considerations

We acknowledge that ethical considerations and privacy concerns are connected to collecting and using social media data[12] for research. The *collection* and subsequent analysis of user-level information may identify individual persons, and the development of methods be used to *censor* users' reactions. Together, these warrant a discussion of the ethical implications.

Ethics and data collection Legally, users of social media platforms grant broad permissions pertaining to their data to be used by third parties, among them, researchers. Requiring certain epistemic capabilities, users often do not thoroughly read or understand the terms of service or may not realize that their data is public [52, 90, 162]. Due to scale, it proves infeasible to ask explicit consent of social media users to use their data for research purposes.

We assess our data collection for research purposes with respect to users' perception of *acceptable* uses of data, in addition to having received faculty approval for the collection and processing of personal data (see above).

Hemphill et al. [52] found that users' *sensitivity* (how risky a person perceives sharing their data) towards data such as screen names varies. Screen name data is perceived less sensitive than individuals' driver license numbers or medical records, but more sensitive than, for example, demographic details. Yet, the researchers found evidence that sensitivity is not a good predictor of whether users think it is acceptable to collect social media data.

Instead, research indicates that acceptability is a function of other variables, such as the identity of the *data analyst*, *data sharer characteristics*, *data use purpose*, *and data types* echoing earlier research [52, 33, 45].

In terms of the data analysts' identity, respondents were more accepting of researchers—looking to produce social benefit—using their data compared to social media companies or journalists. This is further moderated by the level of trust users have in institutions such as universities collecting their data (data sharer characteristics), even if it is for research. The acceptance rating increases, when users are more familiar with the data analysts' doing and the *data use purpose*. In an attempt to inform the users who are mainly Twitter users tweeting in Danish about Danish politics, we stated on our respective university websites that "*...we are currently collecting public data from Twitter, more specifically from #dkpol with the aim of validating bot detection methods. If you are a Twitter user and/or have any questions, please contact us at rasmus@hum.ku.dk or jahn@hum.ku.dk*". Given generally high trust levels among the Danish society [130], we believe that the type of reaction data we collected is perceived acceptable. The social media data type we collected does not resemble photos, videos, data about sexual habits, preferences and behaviors, nor posts about their friends or family members, which users are most concerned about [52].

[12] This discussion is not relevant for the synthetic data analyzed in this thesis.

Ethics and Censorship The ethical implications of our results in **Article I** and **Article II** relate to possibilities of censoring users' reactions from juries, or labeling users as acting in an inauthentic coordinated way. The studies in this thesis are on synthetic data or observational and retrospective. Hence, no users were censored as a result. Yet, even if not implemented, any flagging of behavior, such as the suppression of reactions on social media, raises the moral dilemma between upholding freedom of speech and preventing harm caused by misinformation.

Generally, we find that the suppression of coordinated inauthentic behavior as used by attention hackers is defendable, justified by the aim to combat misinformation online. In that vein, a recent study [66] with 2,564 US respondents found that the majority prefers "quashing harmful misinformation over protecting free speech", across diverse political leanings. However, in applying automated techniques based on classification, there is always a risk that misclassification occurs.

We here summarize an excerpt from an extensive discussion of censorship in Sec. 5.1 of **Article I**: we designed the jury selection procedures focusing on the two stated desiderata *vox populi* and *precaution*. Vox populi implies a desire to not unrightfully censor individuals, and is opposed to precaution against allowing inauthentic behavior: the most precautious model censors all, while the model that preserves most voices censors none. Given our ABM and its parameters, employing *ends-justify-the-means* reasoning, and taking the correct evaluation of posts' quality to be the primary end, we find it worth compromising vox populi over deprioritizing precaution. Results illustrated that deprioritizing precaution quickly threatens the wisdom-of-crowds effect as few inauthentic agents in the jury drastically lower the majority correctness score. Yet, there is a compromise to strike with vox populi by allowing small fractions of authentic agents to be labelled as inauthentic is absorbed by the wisdom of crowds exhibited by even a small jury of only authentic agents.

The methods for initial exploration of empirical data in **Article II** may risk unjustly labeling users due to behavioral correlation with strongly coordinated groups of users. We thus chose a vox-populi valuing classification threshold while conservatively assessing the empirical data. Nevertheless, we strongly recommend that the methods here are taken as a first step towards fact-checking content and users and not as a final verdict about specific individual users. We are transparent about the identified perfectly correlated liking behavior in user clusters unrelated to our purchases. We conclude that one cannot infer from *correlation* to *coordination* nor causally state that the clusters are involved in coordinated inauthentic behavior without further analysis.

As an alternative to seeking the most ethically justifiable balance between vox populi and precaution, the methods employed in **Article I** and **Article II** may be extended beyond binary classification. In the hope to lower harm through misclassification, a non-binary classification [56, 93] may serve the nuanced coordination phenomenon task more adequately, possibly sacrificing simplicity and automation.

Due to the risk of unrightful censorship, we would always suggest that users are made aware of censorship decisions that concern them and are given the option to appeal. This, of course, also allows accounts used in influence operations to appeal, but appeal adds non-trivial *friction* to e.g. large bot collectives.

The complexity of the moral dilemma between content moderation and free speech may provide an argument in favor of behavioral intervention measures as a promising tool to *prevent* (instead of *detect* and *censor*) the amplification of low-quality content. A possible implementation of the friction intervention in **Article III** is suggested to have minimal effects on engagement (as it is to occur rarely). It may easily be deployed at scale not requiring labeling of content or detection of bad actors, circumventing this specific ethical dilemma. Naturally, other ethical considerations are relevant, such as accessibility, e.g., the provision of accessible text in the users' preferred language and mode (visual or audio). Friction interventions, such as quizzes with a text prompt, however, are accessible internet content, e.g. in contrast to picture-based CAPTCHAs.

6 Related Work

Research on topics related to this thesis has been growing and moving fast. In this section, we situate the thesis' contribution in a broader research context within *coordination*, *bot detection*, and *behavioral interventions*. Like this thesis, these three research fields relate to curbing amplification online by studying users (human or not)—but in a *broader* sense than by one-click reactions. In each of the articles in this thesis, we cover work *closely* related to and relevant for the papers.

Coordination

The detection of coordination is a relatively young research field. As far as we know, no systematic review has been published on this topic as of yet. Broadly, most recent advances share the feature that they build on leveraging unexpected similarities between social media users as a proxy for coordination (e.g., [24, 93, 96, 116]). Yet studies differ in the exact forms of coordination they study, such as coordinated pushing of selected hashtags, URLs, images, or coordinated commenting. The studied form of coordination directly translates into diverse input data requirements and affects methodological choices. Below, we review these contributions. We review coordination studies directly targeted at *reactions* in **Article I** and **Article II**.

Several approaches rely on more than one coordination type, and in general, the literature is not straightforwardly parceled into meaningful disjoint compartments. One distinction which allows for meaningful partitioning of approaches is whether or not the approaches

use *explicit temporal data*. Any social media dataset implicitly uses some aspect of time; at least by reflecting moments in constantly changing online activity at the time of extraction. A social media dataset may, however, also feature explicit timestamp data to capture the temporal development of continuous postings and changes. If the timestamps of data points are ignored in analyses, we say that the method only makes *implicit* use of temporal data (Sec. 6) (as this thesis does). Alternatively, methods may make explicit use of temporal data, such as time-ordered activity logs, synchronicity in activity streams, or activity peaks in defined time intervals (Sec. 6).

The existing work specifically focused on coordinated inauthentic behavior to inflate reaction metrics, such as likes and retweets, is discussed in **Article I** and **Article II**. This body of work includes both implicit and explicit uses of temporal data.

Implicit use of temporal data Early work by Ahmed and Abulaish [2] identifies coordinated clusters of users given a series of posts containing, e.g., hashtag sequences, URLs or images, incorporating temporality implicitly. The authors investigate Twitter [Facebook] users that exhibit similar behavior with regard to tweeting hashtags, URLs, or mentions [friendship requests, posts, page likes, URLs shared]. They consider the activities of 160 [165] manually identified spam profiles (based on high activity of spam link sharing) and 145 [155] normal user profiles. After training and analyzing a supervised classifier, the authors further explore coordination between users and identify spam *campaigns* on Facebook and Twitter. Applying graph-based methods to find coordinated clusters of users, they draw edges between spam users that exhibit similar behavior on the basis of a similarity matrix between users. Similarities are computed from previously extracted features such as friends/followers, page/hashtags, and URLs.

Al-Khateeb and Agrawal [3] study group-level coordination on Twitter based on a dataset of 1,361 tweets and 588 users logging follows, mentions, replies, and tweets. Their approach analyzes tweets to label suspicious users based on e.g., unusually high tweet frequency, large number of identical tweets, or similar use of hashtags to direct attention to a link. Graph analysis and clustering of the suspicious accounts' friends and followers relations disclose subnetworks with group-level coordination (such as dense connections or multiple bots following one central human account) among accounts that share certain behaviors and amplify the distributions of links.

Nizzoli et al. [93] investigate a subset of Twitter users and build a user similarity network, connecting nodes based on varying degrees of similarity (i.e., applying a soft threshold) between user-based characteristics. Examples of user-based characteristics are hashtag sets, retweet sets, or sets of following and retweeted accounts. Out of more than 1 million distinct Twitter users and more than 11 million tweets containing hashtags in relation to the UK 2019 general election, the authors select a subset of most active 1% users in terms of retweeting (super-

spreaders), amounting to 10, 782 users. Pairwise similarity is then determined via similarity metrics such as Cosine or Jaccard. The similarities between users result in an undirected user-user graph; filtering and clustering returns coordinated communities passed on to content analyses or NLP. Adapting Nizzoli et al. [93]'s approach and dataset, Hristakieva et al. [56] extend the work and combine the analysis of coordination and propaganda scores.

Similarly, in a range of case studies, Pacheco et al. [96] define coordinated Twitter action types, e.g. use of a hashtag or images, and create a bipartite graph between users and the action (e.g. a hashtag). These action types are used to project the graph onto a user-user graph preserving the user nodes and creating edges given similarly standard similarity measures. After applying a hard similarity threshold and ignoring resulting singletons (nodes no longer connected as their mutual similarity measure was below the threshold), connected components create clusters of coordinated users. For different cases, filtered graphs then contain between 50 and 7, 879 users. The researchers also consider timestamp data explicitly and retweeting as a coordinated action type in two of their case studies, which we discuss in Sec. 6 and in **Article II**, respectively.

Kirn and Hinders [63] attempt to identify influence operation campaigns in a dataset of 60 million tweets collected during the early Covid-19 pandemic (reduced to 7 million tweets, to including only the top hashtags).They pre-process tweet *texts* into a standardized form by vectorizing them into individual tweet vectors. These vectors are collected in a term-tweet matrix and factorized disclosing latent structures of underlying narratives (tweet storms) hidden in the overall matrix. A user-user network is then built by creating edges between users that co-occur in the same tweet storms at least a number of times. Such users pushing stories through social media form an isolated network of accounts that are densely connected (applying ridge count thresholding) and are potentially part of an IO campaign, Kirn and Hinders argue.

Explicit use of temporal data Several approaches to coordination detection make explicit use of temporal timestamp data.

Spambot detection leverages similarities in synchronicity of users' timeline activity streams given Twitter data [16, 22, 24, 25]. As a point of departure, users are selected upon engagement with a specified hashtag or keyword. Chavoshi et al. [16] assumes that authentic human users do not exhibit highly synchronously timed activities over extended periods of time. The study looks for time-correlated groups given sequences of actions like posting, tweeting, retweeting, liking, and deleting on Twitter, each associated with a corresponding timestamp. The work searches for spikes in action sequences around certain timestamps indicating an increased magnitude of coordination.

The approach of Cresci et al. [22, 24] retrieves users' Twitter timeline as a sequence of time-ordered data representing users' actions such as retweeting and posting in terms of types, or the use of hashtags in terms of content, encoded in character strings, called "digital DNA" sequences. To quantify similarity among users and to label spambots, the authors plot the users' longest common substrings (LCS) in digital DNA sequences as a curve. They then split users into two groups, given a peak in the first derivate plot (when the original LCS curve notably drops) flagging users to the left of the splitting points as suspiciously coordinated [24]. In later works, this method finds application in identifying coordinated groups of bots tweeting about low-value stocks by exploiting the popularity of high-value ones, scraping tweets that contain a specific so-called *cashtag* (similar in filtering nature to a hashtag, but referring to a specific stock, prefixed with a $ instead of a #).

Also considering cashtags, Pacheco et al. [96] study synchronous activity on Twitter that aims at synthetically increasing prices of cheaply acquired stocks. Given a cashtag, the authors follow the rationale that accounts tweeting at least 8 times in the same 30-minutes interval may be coordinated. These thresholds were chosen to balance false positive errors and computation time. A bipartite graph is built between users and 30-minute intervals given the users' tweet timestamps falling in such a time interval, and weighted and projected onto a user-user network using the Cosine similarity. Keeping the nodes with the highest similarity score (top 0.5%) yields coordinated connected components of users.

Magelisnki et al. [82] and Weber et al. [148] consider multiple timing-based coordination dimensions that generalize to many online social networks and are readily available from public social media APIs: coordination through the use of URLs, hashtags and mentions in posts.

Magelinski et al. [82] exclude those action types that require pre-processing such as text or images. Based on 9.9 million tweets from the American Twittersphere discussing opening up after COVID-19 lockdowns, the authors build user-user coordination graphs and draw connections between two users when they engage in synchronous action. Synchronous action in this setting refers to the use of same action type in a five minute time window (in later works, Ng et al. [91] scrutinize how to choose an optimal time window without capturing coincidental synchronicity). An edge is more heavily weighted, the more often users exhibit such behavior, exposing dense cluster of users pointing to groups of suspicious users upon analysis. They analyze the graphs by layering and clustering them into a multi-view graph, where each layer considers an action type or combination of action types (e.g. edges are drawn given synchronous use of a hashtag *and* URL). Their work rests on the assumption that synchronization is unlikely to be seen by genuine political organizers. However, the authors are admitting to false positive edges between users posting opposing views when only looking at one action types such as hashtag.

Using only metadata and temporal data, Weber et al. [148] extract highly coordinated communities from latent coordination networks. In these networks, accounts are connected

based on shared *actions*, such as posting the same URL, mention or hashtag, or retweeting the same tweet, and not necessarily on friendship/follower relationships. For each temporal window, the authors construct and aggregate coordination networks, identify highly coordinated communities around influential nodes, and analyze coordination strategies. The methods are applied to two Twitter datasets comprising, respectively, >100,000 and >1.5 million tweets from > 20,000 and > 1,300 accounts. The first dataset was curated via keywords during a 2018 Australian election. The authors treated official political accounts as ground truth for coordination, assuming coordinated political influence techniques at play. The second dataset was published by Twitter and contains accounts Twitter believes were connected to influence operations.

Vargas et al. [144] focus on separating known influence operations (shared by Twitter data archives) from authentic communities. Based on activities around posting co-hashtags, co-mentions, co-retweets, or co-URLs, the authors leverage patterns in daily and weekly coordination using sliding window time series prediction in a supervised model trained on influence campaigns. The authors acknowledge that coordination behavior differs for each campaign, making it harder to learn generalizable patterns and detection on unseen influence operations.

Focusing on URLs only, Giglietto et al. [44, 43] study synchronous link sharing behavior of public and 'influential' Facebook pages surrounding recent Italian elections. Given datasets of several hundred thousands URL shares and timestamp entries obtained through Facebook's API *CrowdTangle*, the authors label pages as coordinated. Coordinated pages are highly publicly engaged in link sharing that is both frequent (that is, above the 90th percentile in activity frequency of all link sharers or active more than a defined number of times) and near-simultaneous (that is, within a time window < 30 seconds). The resulting coordinated networks between suspicious pages and links are shown to contain several problematic domains and overlap with pages blacklisted with Italian fact-checkers.

Bot Detection

There is extensive research on the detection of social media bots. Several papers review the research published in the field (e.g., [51, 95, 110]). These articles help us to briefly summarize the state-of-the-art in this section.

Orabi et al.'s [95] systematic review preambles with comparing bot detection research to a cat and mouse game. Researchers and social media platforms develop methods to catch bots, while bots evolve to evade detection [156]. These dynamics are helpful in understanding why bot detection is a fast-moving and growing research field. It further underlines why a broad definition of *social bots* (Sec. 3.5) is useful when talking about bots, in addition to refined taxonomies which may age rapidly. For proposed taxonomies, see Sec. 3 in [95] and Sec. 3 in [49].

Most bot detection methods rely on supervised machine learning. These classifiers are trained on labeled datasets and predict labels on unseen data based on known features. Feature spaces can grow large and leverage behavior-based and/or content-based characteristics. Fewer detection methods use unsupervised machine learning, graph-based approaches, crowdsourcing, or anomaly-based detection methods or combinations thereof. Yet, the number of unsupervised and graph-based methods is increasing not least because of the shift towards group-based detection and the interest in coordination (Sec. 6).

Measuring the effectiveness of bot detection methods is a difficult task. Most methods and datasets are not (made) publicly available, and there are technical costs in comparing methods. Reported performance of supervised methods is sensitive to training and testing datasets. Performance on unseen datasets may vary greatly and does not generalize well to new datasets [51]. This may partly be due to evolving bots. Hays et al. [51], however, found evidence that diverse and intransparent data collection and labeling procedures contribute to this result. The publicly available supervised-learning tool *Botometer* [114, 158, 157] transparently adopts more datasets for training and testing and makes the labeled datasets available for the research community in a publicly accessible repository.[13] Most methods have been developed with Twitter data due to their data sharing policies. This poses questions about how well these methods fare with data from other platforms [95].

Bot detection models may never become the single sufficient method to fight misinformation given the cat and mouse dynamics between bot detection and bot evolution. Detection methods may be supplemented with other means to prevent the amplification of low quality online, such as behavioral interventions.

Behavioral Interventions

Supplementing the *detection of bots* and *coordination*, there is a growing research body on *behavioral interventions* to target competences and behaviors of users to curb the spread of misinformation. In **Article III**, we gather preliminary evidence on the effectiveness of the behavioral design intervention *friction*. While we do not carry out an online field experiment, we map the key ingredients of an experiment to test a friction intervention where learning is leveraged through quizzes about a platform's community standards. Accordingly, the research field on behavioral interventions is relevant in this section.

We closely review studies on friction interventions in **Article III**. In this section, we briefly summarize the broader state-of-the-art of the related work. A recent expert review [67] and a systematic review [161] provide an overview of online behavioral design measures.

[13] Indiana University's Bot Repository, a resourceful, centralized repository of annotated datasets of Twitter social bots: `https://botometer.osome.iu.edu/bot-repository/datasets.html`

Behavioral interventions are an attractive tool to fight misinformation as they circumvent the moral dilemma and trade-off between content moderation and free speech [66, 67]. Evidence that these interventions may slow down the spread of misinformation accumulates. Behavioral design measures promise to steer online user behavior away from acting on the basis of *social influence cues*, *heuristics*, *system II thinking*, or *biases*, both *cognitive* and *social* [67]. Different disciplines, such as cognitive sciences, psychology, and misinformation research have looked at interventions such as nudging (paternalistically steering people's decisions). Kozyreva et al. [67] have put together a toolbox[14] categorizing 42 published, peer-reviewed studies into 10 intervention types, falling in three partially overlapping categories: *Nudges* including accuracy prompts, friction, and social norms; *Boosting* including debunking, inoculation, lateral reading, rebuttals of scientific denialism, self-reflection tools, and media-literacy tips; and *Refutation strategies* including warnings, labels, rebuttals of scientific denialism, lateral reading, and inoculation. Ziemer et al.'s [161] review of 254 studies propose a slightly more fine-grained grouping: For them, *boosting* merely contains literacy interventions, while e.g. rebuttals fall under their category of *fact-checking*. The authors categorize self-reflection tools as *identity management*. Ziemer et al. similarly group *nudges*, but split refutation strategies into two groups: *fact-checking* and *inoculation*.

The two review studies agree that there is an overrepresentation of research conducted in Western countries and limited knowledge about long-term effects to derive from the empirical evidence [67, 161]. Researchers rarely test the longevity of interventions. This is possibly due to local experiment setups or tools like Amazon's *Mechanical Turk* being subject to time and resource constraints. In a majority of studies, little attribution is given to theoretical foundations such as basic theory on learning or reasoning to explain intervention effects. Pennycook et al. [101] provide an extensive overview over theory.

Studies also remain difficult to compare as they fall into heterogenous research paradigms when it comes to test stimuli and outcome variables [67]. Studies on refutation strategies, for example, fall into a *misinformation-correction paradigm* where participants are presented with a correction derived from the *content* of posts. The measure of success is whether participants believe and remember corrections conveyed through warning labels or rebuttal notes. Later research in the *headline-discernment paradigm*—arguably most relevant in the context of one-click reactions—conducts experiments on e.g. accuracy prompts, friction, or inoculation games. Studies in this paradigm measure how well participants can discern between true and false headlines and how that is reflected in their sharing behavior. In the last paradigm, the *skill-adoption paradigm*, research focuses on how well participants learn skills to evaluate information veracity. To study lateral learning, for example, experiments often resort to stimuli such as entire websites instead of merely posts [67].

[14] Also available as an interactive table supplementing [67] `https://interventionstoolbox.mpib-berlin.mpg.de/index.html`

These methodological considerations are relevant in order to measure effectiveness and reproducibility. For example, hypothetical judgements about social media sharing may not correspond to actual sharing; asking users about accuracy to assess their discernment quality nudges users to consider accuracy and undermines inferences [99, 100, 101]. Methodology becomes especially relevant when research does not corroborate earlier findings on, for example, accuracy nudges [39, 109] or inoculation games [88]. Behavioral research in online environments and surrounding misinformation, however, is also a dynamic and volatile object of study. Stories and posts used in experiments may age rapidly making replication more difficult. In light of the recent COVID-19 pandemic, more tangible outcome measures such as vaccine uptake also qualify as success measures—instead of outcomes such as distal measures of knowledge or attitudes [151].

7 Concluding Remarks

Taken together, the articles in this thesis contribute to curbing the *amplification of misinformation* through *one-click user reactions* such as *likes* and *shares*. Using different methods and considering the problem from different angles, the papers address *threats* to the *wisdom of crowds* on social media, and hence to the (epistemic) *quality of information* spread on social media.

The articles study methods to *detect* and flag inauthentic, coordinated metric inflation and suspicious correlations in reactions data. This part of the thesis is based on computer-simulated data from an agent-based model and data carefully live-collected through Twitter with a scripted algorithm. The thesis further studies behavioral interventions (*friction*) to *prevent* the amplification of low-quality content, analyzed with an agent-based model.

The main methods employed by this thesis are agent-based models and empirical Twitter data collection. The data collected from Twitter through a scripted algorithm is novel: to the best of our knowledge, none of the existing datasets include user-level data on reactions. This thesis uses simple methods from statistics, machine learning, and network science to analyze synthetic and empirical data. It carefully discusses modeling decisions such as classification alongside data descriptions and ethical considerations. At last, we have situated the work in broader research fields on coordination, bot detection, and behavioral interventions.

We conclude by discussing the real-world relevance of our findings and future research.

Given the thesis' *applied* nature, we discuss possible implementations of the suggested methods and interventions. The jury selection and detection procedures proposed in **Article I** and **Article II** rely on sparse, simple input data, available to and controlled by the social media platforms. Twitter has begun to look into leveraging the wisdom of crowds in a related

way, lending relevance and credit to the idea of jury selection procedures laid out in **Article I**. Twitter's research concerns their crowd-based community notes [155] and investigates algorithms that select a subset of high-quality notes to achieve a better than a supermajority voting baseline given all notes. **Article II**'s first steps towards identifying coordinated likes only use unsupervised clustering based on very sparse input data. Such methods may prove employable inside platforms as a first conservative filter without needing to label content.

The friction intervention proposed in **Article III** requires no input data, would have minimal effects on engagement, and may easily be deployed at scale, not requiring classification and labeling of content as high- or low-quality of content or detection of inauthentic actors.

We hope this thesis inspires future research. Research addressing the amplification of misinformation online may continue to look into discerning inauthentic and authentic coordinated activity [56, 144]. **Article II** has left us with hypotheses worth considering given the identified clusters. We are currently working on a follow-up paper that analyzes the suspiciously correlated users concerning automated and bot-like activity. In the last period of this Ph.D. and before the Twitter API closed down, we collected data to supplement **Data II**, including a new monthlong dataset collected during the Danish National Election 2022. We retrieved *Botometer* (v4 and lite) scores for all liking and retweeting users, re-collected the liking and retweeting users once more using pagination, and looked up tweets, like and retweet counts one more time. The supplemented data and scripts for analysis and collection have been made available alongside **Data II** on Harvard Dataverse [58].

The supplementary data may prove useful in assessing the authenticity of user groups. In a similar vein to Hristakieva et al. [56], one avenue for future research is to consult correlations between Botometer scores of users and their membership in coordinated groups of users. Following Torres-Lugo et al. [132], another direction is to analyze whether we observe significant drops in like and retweet counts, or whether the data exhibits deleted/suspended users indicating short-lived influence campaigns.

Future research may also look into relationships between correlated user clusters and the tweets they amplify. Tweets may be classified manually or at the source level. These considerations may further be moderated by domain characteristics of `#dkpol` and similar Danish hashtags. Other domains or hashtags may not resemble such a contained public environment (mainly through language, size, and relevance of Danish politics). An interesting direction of further research may be to experiment with samples that vary in observation period length, and to systematically sample keywords and hashtags.

Another direction of future work—initiated in collaboration with members of the Economics department at the University of Copenhagen—similarly further explores the correlated user groups found in **Data II**. In an attempt to understand the correlation structures of users,

we are looking at the robustness of using only 2 eigenvectors to disclose > 2 clusters [9, 76]. These clusters are also readily visible as block-diagonal matrix structures when re-ordering correlation plots based on an angular ordering of the first two eigenvectors [36].

Naturally following from **Article III**, future research may consider testing friction in field experiments. Future results of behavioral experiments testing the proposed friction strategy on community standards provide further arguments for implementation—relevant for both policy makers and social media platforms.

The decisions of platforms and policy makers will affect future misinformation research significantly. Since Twitter was bought by Elon Musk, the company's misinformation and data sharing policies have created worries. Among other events, Twitter has been rolling back policies aimed at tackling misinformation related to COVID-19 [107] and reinstated many anti-vaccine influencers. These accounts are estimated to account for 35% of all re-shares of COVID-19-related misinformation. Some of the previously detected and banned accounts now appear verified with a blue checkmark, obtained by subscribing to Twitter Blue [105].

The discontinuation of the Academic Research API is undoubtedly among the most notable recent events, preventing future research starkly as data access is cut [68]. The reactions to the change in Twitter's API policy has highlighted that much research has entered into a dependent relationship with Twitter. Meta recently announced that they will soon make their *Influence Operations Research Archive* available to researchers. It is still being determined what exact data this archive provides, yet it might compensate and aid research hit by the loss of Twitter data.

However, even in a scenario where researchers eventually retain access to Twitter data and access to other platforms, the events have brought a larger issue to the fields' attention: Despite the pivotal role and associated threats of social media in today's society, it is social media platforms which are in control of the research conducted on these topics. Policymakers seek to address this issue. The 2022 European *Digital Service Act* (DSA)—in force from 2024 onwards—requires platforms to share data with researchers studying systemic risks [129]. It remains unclear for the time being how readily accessible such data shall have to be [68]. And it remains to be seen how policymakers will treat the platforms' responsibility for content shared online, moderation efforts in light of free speech, and their reservations to being arbiters of truth [48, 85, 125].

References

[1] Abokhodair, Norah and Yoo, Daisy and McDonald, David W. Dissecting a social Botnet: Growth, content and influence in Twitter. *CSCW 2015 - Proceedings of the 2015 ACM International Conference on Computer-Supported Cooperative Work and Social Computing*, pages 839–851, 2015.

[2] Ahmed, Faraz and Abulaish, Muhammad. A generic statistical approach for spam detection in Online Social Networks. *Computer Communications*, 36(10-11):1120–1129, 6 2013.

[3] Al-Khateeb, Samer and Agarwal, Nitin. Understanding Strategic Information Manoevres in Network Media to Advance Cyber Operations: A Case Study Analysing Pro-Russian Seperatists' Cyber Information Operations in Crimean Water Crisis. *Journal on Baltic Security*, 2(1), 2016.

[4] Allcott, Hunt and Gentzkow, Matthew. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36, 2017.

[5] Altay, Sacha and Berriche, Manon and Acerbi, Alberto. Misinformation on misinformation: Conceptual and methodological challenges. *Social Media + Society*, 9(1), 2023.

[6] Angere, Staffan. Knowledge in a social network. *Synthese*, pages 167–203, 2010.

[7] Avram, Mihai and Micallef, Nicholas and Patil, Sameer and Menczer, Filippo. Exposure to social engagement metrics increases vulnerability to misinformation. *The Harvard Kennedy School Misinformation Review*, 1(5), July 2020.

[8] Bacio Terracino, Julio and Matasick, Craig. Disinformation and Russia's war of aggression against Ukraine: Threats and governance responses. `https://www.who.int/directo,r-general/speeches/detail/munich-security-conference`, 3 November 2022. Accessed: 2023-02-18.

[9] Behrisch, Michael and Bach, Benjamin and Henry Riche, Nathalie and Schreck, Tobias and Fekete, Jean-Daniel. Matrix reordering methods for table and network visualization. In *Computer Graphics Forum*, volume 35, pages 693–716. Wiley Online Library, 2016.

[10] Beutel, Alex and Xu, Wanhong and Guruswami, Venkatesan and Palow, Christopher and Faloutsos, Christos. Copycatch: Stopping group attacks by spotting lockstep behavior in social networks. In *Proceedings of the 22nd international conference on World Wide Web*, pages 119–130, 2013.

[11] Bliss, Nadya and Bradley, Elizabeth and Garland, Joshua and Menczer, Filippo and Ruston, Scott and Starbird, Kate and Wiggins, Chris. An Agenda for Disinformation Research. Quadrennial paper, CRA Computing Community Consortium (CCC), 2020.

[12] Brugnoli, Emanuele and Cinelli, Matteo and Quattrociocchi, Walter and Scala, Antonio. Recursive patterns in online echo chambers. *Scientific Reports*, 9(1):20118, 2019.

[13] Calo, Ryan and Coward, Chris and Spiro, Emma S and Starbird, Kate and West, Jevin D. How do you solve a problem like misinformation? *Science Advances*, 7(50), 2021.

[14] Centola, Damon and Macy, Michael. Complex contagions and the weakness of long ties. *American journal of Sociology*, 113(3):702–734, 2007.

[15] Ceylan, Gizem and Anderson, Ian A. and Wood, Wendy. Sharing of misinformation is habitual, not just lazy or biased. *Proceedings of the National Academy of Sciences*, 120(4), 2023.

[16] Chavoshi, Nikan and Hamooni, Hossein and Mueen, Abdullah. DeBot: Twitter bot detection via warped correlation. *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 817–822, 2017.

[17] Chen, Kaiping and Duan, Zening and Yang, Sijia. Twitter as research data: Tools, costs, skill sets, and lessons learned. *Politics and the Life Sciences*, 41(1):114–130, 2022.

[18] Ciampaglia, Giovanni Luca and Nematzadeh, Azadeh and Menczer, Filippo and Flammini, Alessandro. How algorithmic popularity bias hinders or promotes quality. *Scientific Reports*, 8(1):1–7, 2018.

[19] Cinelli, Matteo and Cresci, Stefano and Quattrociocchi, Walter and Tesconi, Maurizio and Zola, Paola. Coordinated Inauthentic Behavior and Information Spreading on Twitter. *Decision Support Systems*, 2022.

[20] Cinelli, Matteo and De Francisci Morales, Gianmarco and Galeazzi, Alessandro and Quattrociocchi, Walter and Starnini, Michele. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9), 2021.

[21] Condorcet , M.J.A.N.C. Marquis de. *Essai sur l'Application de l'Analyse à la Probabilité des Décisions Rendues à la Pluralité des Voix*. Paris, 1785.

[22] Cresci, Stefano and Di Pietro, Roberto and Petrocchi, Marinella and Spognardi, Angelo and Tesconi, Maurizio. DNA-Inspired Online Behavioral Modeling and Its Application to Spambot Detection. *IEEE Intelligent Systems*, 31:58–64, 9 2016.

[23] Cresci, Stefano and Di Pietro, Roberto and Petrocchi, Marinella and Spognardi, Angelo and Tesconi, Maurizio. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the 26th international conference on world wide web companion*, pages 963–972, 2017.

[24] Cresci, Stefano and Di Pietro, Roberto and Petrocchi, Marinella and Spognardi, Angelo and Tesconi, Maurizio. Social Fingerprinting: Detection of Spambot Groups Through DNA-Inspired Behavioral Modeling. *IEEE Transactions on Dependable and Secure Computing*, 15:561–576, 7 2018.

[25] Cresci, Stefano and Lillo, Fabrizio and Regoli, Daniele and Tardelli, Serena and Tesconi, Maurizio. Cashtag piggybacking: Uncovering spam and bot activity in stock microblogs on twitter. *ACM Transactions on the Web*, 13, 4 2019.

[26] Deutsch, Morton and Gerard, Harold B. A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology*, 51(3):629, 1955.

[27] Dietrich, Franz and Spiekermann, Kai. Epistemic Democracy with Defensible Premises. *Economics and Philosophy*, 29(1):87–120, 2013.

[28] Dietrich, Franz and Spiekermann, Kai. Jury Theorems. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2022 edition, 2022.

[29] Dutta, Hridoy Sankar and Chetan, Aditya and Joshi, Brihi and Chakraborty, Tanmoy. Retweet us, we will retweet you: Spotting collusive retweeters involved in blackmarket services. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 242–249. IEEE, 2018.

[30] Epstein, Ziv and Lin, Hause and Pennycook, Gordon and Rand, David. How many others have shared this? Experimentally investigating the effects of social cues on engagement, misinformation, and unpredictability on social media. 2022.

[31] Fazio, Lisa K and Brashier, Nadia M and Payne, B Keith and Marsh, Elizabeth J. Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, 144(5):993, 2015.

[32] Festré, Agnès and Garrouste, Pierre. The 'Economics of Attention': A History of Economic Thought Perspective. *Œconomia. History, Methodology, Philosophy*, (5-1):3–36, 2015.

[33] Fiesler, Casey and Proferes, Nicholas. "Participant" Perceptions of Twitter Research Ethics. *Social Media+ Society*, 4(1), 2018.

[34] Floridi, Luciano. Is semantic information meaningful data? *Philosophy and Phenomenological Research*, 70(2):351–370, 2005.

[35] Coalition for Independent Technology Research. Letter: Imposing fees to access the twitter api threatens public-interest research. `https://independenttechresearch.org/letter-twitter-api-access-threatens-public-interest-research/`, February 6 2023. Accessed: 2023-02-11.

[36] Friendly, Michael. Corrgrams: Exploratory displays for correlation matrices. *The American Statistician*, 56(4):316–324, 2002.

[37] Galeazzi, Paolo and Rendsvig, Rasmus K. and Slavkovik, Marija. Improving Judgment Reliability in Social Networks via Jury Theorems. In Patrick Blackburn, Emiliano Lorini, and Meiyun Guo, editors, *Logic, Rationality, and Interaction (LORI 2019)*, volume 11813 of *Lecture Notes in Computer Science*, pages 230–243. Springer, 2019.

[38] Galton, Francis. Vox Populi. *Nature*, 75:450–451, 1907.

[39] Gavin, Lyndsay and McChesney, Jenna and Tong, Anson and Sherlock, Joseph and Foster, Lori and Tomsa, Sergiu. Fighting the spread of covid-19 misinformation in kyrgyzstan, india, and the united states: How replicable are accuracy nudge interventions? *Technology, Mind, and Behavior*, 2022.

[40] Gebru, Timnit and Morgenstern, Jamie and Vecchione, Briana and Vaughan, Jennifer Wortman and Wallach, Hanna and Iii, Hal Daumé and Crawford, Kate. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[41] Ghebreyesus, Tedros Adhanom. Speech at the Munich Security Conference. `https://www.who.int/directo,r-general/speeches/detail/munich-security-conference`, 15 February 2020. Accessed: 2023-02-18.

[42] Gibbons, Andrew and Carson, Andrea. What is misinformation and disinformation? understanding multi-stakeholders' perspectives in the asia pacific. *Australian Journal of Political Science*, 57(3):231–247, 2022.

[43] Giglietto, Fabio and Righetti, Nicola and Rossi, Luca and Marino, Giada. Coordinated Link Sharing Behavior as a Signal to Surface Sources of Problematic Information on Facebook. *ACM International Conference Proceeding Series*, pages 85–91, 2020.

[44] Giglietto, Fabio and Righetti, Nicola and Rossi, Luca and Marino, Giada. It takes a village to manipulate the media: coordinated link sharing behavior during 2018 and 2019 Italian elections. *Information Communication and Society*, 23(6):867–891, 2020.

[45] Gilbert, Sarah and Vitak, Jessica and Shilton, Katie. Measuring americans' comfort with research uses of their social media data. *Social Media + Society*, 7(3), 2021.

[46] Gleicher, Nathaniel. Coordinated Inauthentic Behavior Explained. `https://about.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/`, December 6, 2018. Accessed: 2023-02-18.

[47] Goldman, Alvin and O'Connor, Cailin. Social Epistemology. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2021 edition, 2021.

[48] Goldman, Eric. Content moderation remedies. *Mich. Tech. L. Rev.*, 28:1, 2021.

[49] Gorwa, Robert and Guilbeault, Douglas. Unpacking the social media bot: A typology to guide research and policy. *Policy & Internet*, 12(2):225–248, 2020.

[50] Guess, Andrew M and Lyons, Benjamin A. Misinformation, disinformation, and online propaganda. In Nathaniel Persily and Joshua A Tucker, editors, *Social Media and Democracy: The State of the Field, Prospects for Reform*. Cambridge University Press Cambridge, 2020.

[51] Hays, Chris and Schutzman, Zachary and Raghavan, Manish and Walk, Erin and Zimmer, Philipp. Simplistic collection and labeling practices limit the utility of benchmark datasets for twitter bot detection. 2023.

[52] Hemphill, Libby and Schöpke-Gonzalez, Angela and Panda, Anmol. Comparative sensitivity of social media data and their acceptable use in research. *Scientific Data*, 9(1):1–14, 2022.

[53] Hendricks, Vincent F. and Mehlsen, Camilla. *The Ministry of Truth: BigTech's Influence on Facts, Feelings and Fiction*. Springer Nature, 2022.

[54] Hills, Thomas T. The dark side of information proliferation. *Perspectives on Psychological Science*, 14(3):323–330, 2019.

[55] Hilverda, Femke and Kuttschreuter, Margôt and Giebels, Ellen. The effect of online social proof regarding organic food: comments and likes on facebook. *Frontiers in Communication*, 3:30, 2018.

[56] Hristakieva, Kristina and Cresci, Stefano and Da San Martino, Giovanni and Conti, Mauro and Nakov, Preslav. The spread of propaganda by coordinated communities on social media. In *14th ACM Web Science Conference 2022*, WebSci '22, pages 191–201, New York, NY, USA, 2022. Association for Computing Machinery.

[57] Jahn, Laura and Rendsvig, Rasmus K. Get-twitter-likers-data. `https://github.com/humanplayer2/get-twitter-likers-data/`, 2022.

[58] Jahn, Laura and Rendsvig, Rasmus K. Supplementary Data for "Twitter User Reactions Data (Liking and Retweeting Users)". `https://dataverse.harvard.edu/dataverse/twitter-likers`, 2023. https://doi.org/10.7910/DVN/RWPZUN.

[59] Jahn, Laura and Rendsvig, Rasmus K. Twitter User Reactions Data (Liking and Retweeting Users). `https://dataverse.harvard.edu/dataverse/twitter-likers`, 2023. https://doi.org/10.7910/DVN/WRUNZD.

[60] Jungherr, Andreas and Schroeder, Ralph. Disinformation and the structural transformations of the public arena: Addressing the actual challenges to democracy. *Social Media+ Society*, 7(1), 2021.

[61] Kahneman, Daniel and Slovic, Stewart Paul and Slovic, Paul and Tversky, Amos. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge uUiversity Press, 1982.

[62] Kille, Leighton Walter. Committee of Concerned Journalists: The rinciples of journalism. *The Journalistic Resource*, 2009.

[63] Kirn, Spencer Lee and Hinders, Mark K. Ridge count thresholding to uncover coordinated networks during onset of the Covid-19 pandemic. *Social Network Analysis and Mining*, 12, 12 2022.

[64] Klein, Dominik and Marx, Johannes and Fischbach, Kai. Agent-Based Modeling in Social Science, History, and Philosophy: An Introduction. *Historical Social Research*, 43(1):7–27, 2018.

[65] Koch, Timo K. and Frischlich, Lena and Lermer, Eva. Effects of fact-checking warning labels and social endorsement cues on climate change fake news credibility and engagement on social media. *Journal of Applied Social Psychology*, pages 1–13, 2023.

[66] Kozyreva, Anastasia and Herzog, Stefan M. and Lewandowsky, Stephan and Hertwig, Ralph and Lorenz-Spreen, Philipp and Leiser, Mark and Reifler, Jason. Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences*, 120(7), 2023.

[67] Kozyreva, Anastasia and Lorenz-Spreen, Philipp and Herzog, Stefan and Ecker, Ullrich and Lewandowsky, Stephan and Hertwig, Ralph. Toolbox of Interventions Against Online Misinformation and Manipulation. 2022.

[68] Kupferschmidt, Kai. Twitter's plan to cut off free data access evokes 'fair amount of panic' among scientists. *ScienceInsider*, 2023.

[69] Lacassagne, Doris and Béna, Jérémy and Corneille, Olivier. Is Earth a perfect square? Repetition increases the perceived truth of highly implausible statements.

[70] Ladha, Krishna K. Information pooling through majority-rule voting: Condorcet's jury theorem with correlated votes. *Journal of Economic Behavior & Organization*, 26(3):353–372, 1995.

[71] Lazer, David M. J. and Baum, Matthew A. and Benkler, Yochai and Berinsky, Adam J. and Greenhill, Kelly M. and Menczer, Filippo and Metzger, Miriam J. and Nyhan, Brendan and Pennycook, Gordon and Rothschild, David and Schudson, Michael and Sloman, Steven A. and Sunstein, Cass R. and Thorson, Emily A. and Watts, Duncan J. and Zittrain, Jonathan L. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[72] Leahey, Erin. The perks and perils of interdisciplinary research. *European Review*, 26(S2):S55–S67, 2018.

[73] Sune Lehmann and Yong-Yeol Ahn. Spreading in social systems: Reflections. pages 351–358. Springer, 2018.

[74] Leonelli, Sabina. Scientific Research and Big Data. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2020 edition, 2020.

[75] List, Christian and Goodin, Robert E. Epistemic democracy: Generalizing the condorcet jury theorem. *Journal of Political Philosophy*, 9(3), 2001.

[76] Li Liu, Douglas M Hawkins, Sujoy Ghosh, and S Stanley Young. Robust singular value decomposition analysis of microarray data. *Proceedings of the National Academy of Sciences*, 100(23):13167–13172, 2003.

[77] Lorenz, Jan and Rauhut, Heiko and Schweitzer, Frank and Helbing, Dirk. How social influence can undermine the wisdom of crowd effect. *Proceedings of the national academy of sciences*, 108(22):9020–9025, 2011.

[78] Lorenz-Spreen, Philipp and Mønsted, Bjarke Mørch and Hövel, Philipp and Lehmann, Sune. Accelerating dynamics of collective attention. *Nature communications*, 10(1):1–9, 2019.

[79] Lorenz-Spreen, Philipp and Oswald, Lisa and Lewandowsky, Stephan and Hertwig, Ralph. A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature human behaviour*, pages 1–28, 2022.

[80] Lozano, Marianela García and Brynielsson, Joel and Franke, Ulrik and Rosell, Magnus and Tjörnhammar, Edward and Varga, Stefan and Vlassov, Vladimir. Veracity assessment of online data. *Decision Support Systems*, 129:113132, 2020.

[81] Luo, Mufan and Hancock, Jeffrey T and Markowitz, David M. Credibility perceptions and detection accuracy of fake news headlines on social media: Effects of truth-bias and endorsement cues. *Communication Research*, 49(2):171–195, 2022.

[82] Magelinski, Thomas and Ng, Lynnette and Carley, Kathleen. Synchronized Action Framework for Detection of Coordination on Social Media. *Journal of Online Trust and Safety*, 1, 2 2022.

[83] Martini, Franziska and Samula, Paul and Keller, Tobias R. and Klinger, Ulrike. Bot, or not? Comparing three methods for detecting social bots in five political discourses. *Big Data and Society*, 8, 2021.

[84] Matthews, Jeanna and Goerzen, Matthew. Black hat trolling, white hat trolling, and hacking the attention landscape. *The Web Conference 2019 – Companion of the World Wide Web Conference, WWW 2019*, 2:523–528, 2019.

[85] McCabe, David. Supreme Court Poised to Reconsider Key Tenets of Online Speech, January 23 2022. Accessed: 2023-01-23.

[86] Metaxas, P. T. and Mustafaraj, E. and Wong, K. and Zeng, L. and O'Keefe, M. and Finn, S. What Do Retweets Indicate? Results from User Survey and Meta-Review of Research. *Proceedings of the 9th International Conference on Web and Social Media, ICWSM 2015*, pages 658–661, 2015.

[87] Metzger, Miriam J and Flanagin, Andrew J and Medders, Ryan B. Social and heuristic approaches to credibility evaluation online. *Journal of Communication*, 60(3):413–439, 2010.

[88] Modirrousta-Galian, Ariana and Higham, Philip A. Gamified inoculation interventions do not improve discrimination between true and fake news: Reanalyzing existing research with receiver operating characteristic analysis, Aug 2022.

[89] Mønsted, Bjarke and Sapieżyński, Piotr and Ferrara, Emilio and Lehmann, Sune. Evidence of complex contagion of information in social media: An experiment using twitter bots. *PloS one*, 12(9):e0184148, 2017.

[90] Nadon, Guillaume and Feilberg, Marcus and Johansen, Mathias and Shklovski, Irina. In the user we trust: Unrealistic expectations of facebook's privacy mechanisms. In *Proceedings of the 9th International Conference on Social Media and Society*, pages 138–149, 2018.

[91] Lynnette Hui Xian Ng and Kathleen M Carley. Online coordination: methods and comparative case studies of coordinated groups across four events in the united states. In *14th ACM Web Science Conference 2022*, pages 12–21, 2022.

[92] Nickerson, Raymond S. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.

[93] Nizzoli, Leonardo and Tardelli, Serena and Avvenuti, Marco and Cresci, Stefano and Tesconi, Maurizio. Coordinated Behavior on Social Media in 2019 UK General Election. In *Proc. International AAAI Conference on Web and Social Media (ICWSM)*, volume 15, pages 443–454, 2021.

[94] Olsson, Erik J. A Bayesian Simulation Model of Group Deliberation and Polarization. In *Bayesian Argumentation*, pages 113–133. Springer, 2013.

[95] Orabi, Mariam and Mouheb, Djedjiga and Al Aghbari, Zaher and Kamel, Ibrahim. Detection of Bots in Social Media: A Systematic Review. *Information Processing and Management*, 57, 2020.

[96] Pacheco, Diogo and Hui, Pik-Mai and Torres-Lugo, Christopher and Truong, Bao Tran and Flammini, Alessandro and Menczer, Filippo. Uncovering Coordinated Networks on Social Media: Methods and Case Studies. In *Proc. International AAAI Conference on Web and Social Media (ICWSM)*, volume 15, pages 455–466, 2021.

[97] Parry, Hazel R. Agent-based modeling, large-scale simulations. *Complex Social and Behavioral Systems: Game Theory and Agent-Based Models*, pages 913–926, 2020.

[98] Pasquetto, Irene V. and Swire-Thompson, Briony and others. Tackling misinformation: What researchers could do with social media data. *HKS Misinformation Review*, 1(8), 2020.

[99] Pennycook, Gordon and Binnendyk, Jabin and Newton, Christie and Rand, David G. A practical guide to doing behavioral research on fake news and misinformation. *Collabra: Psychology*, 7(1):25293, 2021.

[100] Pennycook, Gordon and Epstein, Ziv and Mosleh, Mohsen and Arechar, Antonio A. and Eckles, Dean and Rand, David G. Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855):590–595, 2021.

[101] Pennycook, Gordon and Rand, David G. The psychology of fake news. *Trends in Cognitive Sciences*, 25(5):388–402, 2021.

[102] Pennycook, Gordon and Rand, David G. Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications*, 13(1):1–12, 2022.

[103] Pfeffer, Juergen and Matter, Daniel and Jaidka, Kokil and Varol, Onur and Mashhadi, Afra and Lasser, Jana and Assenmacher, Dennis and Wu, Siqi and Yang, Diyi and Brantner, Cornelia and others. Just another day on twitter: A complete 24 hours of twitter data. 2023.

[104] Pfeffer, Juergen and Mooseder, Angelina and Hammer, Luca and Stritzel, Oliver and Garcia, David. This Sample seems to be good enough! Assessing Coverage and Temporal Reliability of Twitter's Academic API. *Proceedings of the International Conference on Web and Social Media, ICWSM 2023*, forthcoming, 2023.

[105] Pierri, Francesco and DeVerna, Matthew R and Yang, Kai-Cheng and Axelrod, David and Bryden, John and Menczer, Filippo. One year of COVID-19 vaccine misinformation on Twitter. *Journal of Medical Internet Research*, Forthcoming 2023.

[106] Pivato, Marcus. Epistemic democracy with correlated voters. *Journal of Mathematical Economics*, 72:51–69, 2017.

[107] Reuters. Twitter rolls back covid misinformation policy, November 29 2022. Accessed: 2023-01-02.

[108] Reutlinger, Alexander and Hangleiter, Dominik and Hartmann, Stephan. Understanding (with) toy models. *The British Journal for the Philosophy of Science*, 2018.

[109] Roozenbeek, Jon and Freeman, Alexandra LJ and van der Linden, Sander. How accurate are accuracy-nudge interventions? a preregistered direct replication of pennycook et al.(2020). *Psychological science*, 32(7):1169–1178, 2021.

[110] Ruffo, Giancarlo and Semeraro, Alfonso and Giachanou, Anastasia and Rosso, Paolo. Studying fake news spreading, polarisation dynamics, and manipulation by bots: A tale of networks and language. *Computer Science Review*, 47:100531, 2023.

[111] Ryczko, Kevin and Domurad, Adam and Buhagiar, Nicholas and Tamblyn, Isaac. Hashkat: large-scale simulations of online social networks. *Social Network Analysis and Mining*, 7(1):1–13, 2017.

[112] Samper-Escalante, Luis Daniel and Loyola-González, Octavio and Monroy, Raúl and Medina-Pérez, Miguel Angel. Bot Datasets on Twitter: Analysis and Challenges.

[113] Sayyadiharikandeh, Mohsen and Varol, Onur and Yang, Kai-Cheng and Flammini, Alessandro and Menczer, Filippo. Detection of novel social bots by ensembles of specialized classifiers. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*. ACM, 2020.

[114] Sayyadiharikandeh, Mohsen and Varol, Onur and Yang, Kai-Cheng and Flammini, Alessandro and Menczer, Filippo. Detection of novel social bots by ensembles of specialized classifiers. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM)*. ACM, 2020.

[115] Schelling, Thomas C. Models of Segregation. *The American Economic Review*, 59(2):488–493, 1969.

[116] Schoch, David and Keller, Franziska B. and Stier, Sebastian and Yang, Jung Hwan. Coordination patterns reveal online political astroturfing across the world. *Scientific Reports*, 12, 2022.

[117] Shklovski, Irina. Our digital social life. In *Routledge Handbook of Digital Media and Communication*, pages 126–142. Routledge, 2020.

[118] Herbert A Simon. A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, pages 99–118, 1955.

[119] Simon, Herbert A. Models of Man. 1957.

[120] Simon, Herbert A and others. Designing organizations for an information-rich world. In Martin Greenberger, editor, *Computers, communications, and the public interest*, pages 37–72. The John Hopkins Press, Baltimore and London, 1971.

[121] Sîrbu, Alina and Pedreschi, Dino and Giannotti, Fosca and Kertész, János. Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model. *PloS one*, 14(3):e0213246, 2019.

[122] Smaldino, Paul E and O'Connor, Cailin. Interdisciplinarity can aid the spread of better methods between scientific communities. *Collective Intelligence*, 1(2):26339137221131816, 2022.

[123] Søe, Sille Obelitz. A unified account of information, misinformation, and disinformation. *Synthese*, 198(6):5929–5949, 2021.

[124] Song, Yunya and Wang, Sai and Xu, Qian. Fighting misinformation on social media: effects of evidence type and presentation mode. *Health Education Research*, 37(3):185–198, 2022.

[125] Starbird, Kate and Arif, Ahmer and Wilson, Tom. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction*, 3:1–26, 2019.

[126] Surowiecki, James. *The wisdom of crowds*. Anchor, 2005.

[127] Tandoc Jr, Edson C and Lim, Zheng Wei and Ling, Richard. Defining "Fake News": A Typology of Scholarly Definitions. *Digital journalism*, 6(2):137–153, 2018.

[128] The European Commission. Tackling online disinformation: a European Approach. `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018DC0236`. Accessed: 2022-12-22.

[129] The European Union. The Digital Services Act: ensuring a safe and accountable online environment. `https://tinyurl.com/u94fbmfz`. Accessed: 2022-12-22.

[130] Tinggaard, Gert. Researcher: Denmark's world-record level of trust is helping us in the fight against Corona. *Science Nordic*, 2020.

[131] Tormala, Zakary L and DeSensi, Victoria L and Clarkson, Joshua J and Rucker, Derek D. Beyond attitude consensus: The social context of persuasion and resistance. *Journal of Experimental Social Psychology*, 45(1):149–154, 2009.

[132] Torres-Lugo, Christopher and Pote, Manita and Nwala, Alexander and Menczer, Filippo. Manipulating Twitter Through Deletions. In *Proceedings of the 16th International AAAI Conference on Web and Social Media (ICWSM)*, 2022.

[133] Traberg, Cecilie S and Roozenbeek, Jon and van der Linden, Sander. Psychological inoculation against misinformation: Current evidence and future directions. *The ANNALS of the American Academy of Political and Social Science*, 700(1):136–151, 2022.

[134] Truong, Bao Tran and Lou, Xiaodan and Flammini, Alessandro and Menczer, Filippo. Vulnerabilities of the Online Public Square to Manipulation. 2023.

[135] Turner, John C and Oakes, Penelope J. The significance of the social identity concept for social psychology with reference to individualism, interactionism and social influence. *British Journal of Social Psychology*, 25(3):237–252, 1986.

[136] Twitter. Academic Research access. `https://developer.twitter.com/en/products/twitter-api/academic-research`. Accessed: 2023-02-09.

[137] Twitter. Coordinated harmful activity. `https://help.twitter.com/en/rules-and-policies/coordinated-harmful-activity`. Accessed: 2022-14-22.

[138] Twitter. Decahose API. `https://developer.twitter.com/en/docs/twitter-api/enterprise/decahose-api/overview/streaming-likes`. Accessed: 2022-09-10.

[139] Twitter. Platform manipulation and spam policy. `https://help.twitter.com/en/rules-and-policies/platform-manipulation`. Accessed: 2022-14-22.

[140] Twitter. Twitter Moderation Research Consortium. `https://transparency.twitter.com/en/reports/information-operations.html`. Accessed: 2022-07-05.

[141] Twitter, Inc. How we address misinformtion on Twitter. `https://help.twitter.com/en/resources/addressing-misleading-info`, 2023. Accessed: 2022-14-22.

[142] Twitter Transparency. AMARS in the EU. `https://transparency.twitter.com/en/reports/amars-in-the-eu.html`, 2023. Accessed: 2023-02-18.

[143] Vallor, Shannon. Social Networking and Ethics. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2022 edition, 2022.

[144] Vargas, Luis and Emami, Patrick and Traynor, Patrick. On the detection of disinformation campaign activity with network analysis. In *Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop*, CCSW'20, pages 133–146, 2020.

[145] Varol, Onur and Ferrara, Emilio and Davis, Clayton and Menczer, Filippo and Flammini, Alessandro. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 280–289, 2017.

[146] Vestergaard, Mads. *Digital Threats to Democracy: Studies in Digital Politics and Digitalization Policy-Making*. University of Copenhagen, 2021.

[147] Vosoughi, Soroush and Roy, Deb and Aral, Sinan. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[148] Weber, Derek and Neumann, Frank. Amplifying influence through coordinated behaviour in social networks. *Social Network Analysis and Mining*, 11(1):1–42, 2021.

[149] Weng, Lilian and Menczer, Filippo and Ahn, Yong-Yeol. Virality prediction and community structure in social networks. *Scientific reports*, 3(1):1–6, 2013.

[150] Wheeler, Gregory. Bounded Rationality. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2020 edition, 2020.

[151] Whitehead, Hannah S. and French, Clare E. and Caldwell, Deborah M. and Letley, Louise and Mounier-Jack, Sandra. A systematic review of communication interventions for countering vaccine misinformation. *Vaccine*, 2023.

[152] Wilensky, Uri. NetLogo, 1999. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.

[153] Wilkinson, Mark D and Dumontier, Michel and Aalbersberg, IJsbrand Jan and Appleton, Gabrielle and Axton, Myles and Baak, Arie and Blomberg, Niklas and Boiten, Jan-Willem and da Silva Santos, Luiz Bonino and Bourne, Philip E and others. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

[154] Will, Meike and Groeneveld, Jürgen and Frank, Karin and Müller, Birgit. Combining social network analysis and agent-based modelling to explore dynamics of human interaction: A review. *Socio-Environmental Systems Modelling*, 2:16325–16325, 2020.

[155] Wojcik, Stefan and Hilgard, Sophie and Judd, Nick and Mocanu, Delia and Ragain, Stephen and Hunzaker, MB and Coleman, Keith and Baxter, Jay. Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation. 2022.

[156] Yan, Harry Yaojun and Yang, Kai-Cheng. The landscape of social bot research: a critical appraisal. 2022.

[157] Yang, Kai-Cheng and Ferrara, Emilio and Menczer, Filippo. Botometer 101: Social bot practicum for computational social scientists. 2022.

[158] Yang, Kai-Cheng and Varol, Onur and Hui, Pik-Mai and Menczer, Filippo. Scalable and generalizable social bot detection through data selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1096–1103, 2020.

[159] Yasseri, Taha and Menczer, Filippo. Can crowdsourcing rescue the social marketplace of ideas? *Communications of the Association for Computing Machinery*, forthcoming 2023.

[160] Zhou, Xinyi and Zafarani, Reza. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40, 2020.

[161] Ziemer, Carolin-Theresa and Rothmund, Tobias. Psychological underpinnings of disinformation countermeasures: A systematic scoping review. 2022.

[162] Zimmer, Michael. "But the data is already public": on the ethics of research in Facebook. In *The Ethics of Information Technologies*, pages 229–241. Routledge, 2020.

[163] Zuboff, Shoshana. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs, New York, 2019.

Papers

Paper 1

Detecting Coordinated Inauthentic Behavior in Likes on Social Media: Proof of Concept

Laura Jahn[1], Rasmus K. Rendsvig[1], and Jacob Stærk-Østergaard[1,2]

[1]Center for Information and Bubble Studies, Department of Communication, University of Copenhagen
[2]Animal Welfare and Disease Control, Department of Veterinary and Animal Science, University of Copenhagen

Abstract

Coordinated inauthentic behavior is used as a tool on social media to shape public opinion by elevating or suppressing topics using systematic engagements—e.g. through 'likes' or similar reactions. In an honest world, reactions may be informative to users when selecting on what to spend their attention: through the wisdom of crowds, summed reactions may help identifying relevant and high-quality content. This is nullified by coordinated inauthentic liking. To restore wisdom-of-crowds effects, it is therefore desirable to separate the inauthentic agents from the wise crowd, and use only the latter as a voting 'jury' on the relevance of a post. To this end, we design two *jury selection procedures* (JSPs) that discard agents classified as inauthentic. Using machine learning techniques, both cluster on binary vote data—one using a Gaussian Mixture Model (GMM JSP), one the k-means algorithm (KM JSP)—and label agents by logistic regression. We evaluate the jury selection procedures with an agent-based model, and show that the GMM JSP detects more inauthentic agents, but both JSPs select juries with vastly increased correctness of vote by majority. This proof of concept provides an argument for the release of reactions data from social media platforms through a direct use-case in the fight against online misinformation.

Keywords: Coordinated inauthentic behavior, bot detection, social media, wisdom of crowds, simulation, agent-based modeling

1 Introduction

In April 2022, we bought 100 Twitter likes for 3.85 USD through a readily accessible website. These 100 likes sufficed to catapult the liked tweet to the top of the *Top* feed of `#dkpol`, the main Twittersphere for discussing Danish politics. There, it stayed for several hours.[1] This illustrates that Twitter's content sorting algorithm may be easily hacked to bring selected items to users' attention using only likes.

Our tweet was clearly marked as off-topic for `#dkpol`, but could have been misinformation. Our "inauthentic likes" could thus have been used with the intent to mislead or manipulate—and this would not be uncommon: when deploying *influence operations* (IOs) on social media platforms to shape public opinion (Nizzoli et al. 2021), a central strategy is to exploit the platforms' content sorting algorithms to highlight posts to users, a process known as *attention hacking* (Goerzen and Matthews 2019). Attention hacking through likes requires coordination of likes to maximize effect. As the liking behavior does not reflect authentic personal beliefs, it is an example of so-called *coordinated inauthentic behavior* (CIB) (Pacheco et al. 2021; Schoch et al. 2022; Nizzoli et al. 2021).[2] Coordinated inauthentic behavior may be exhibited by humans and bots alike.

Liking is an engagement type common across social media platforms, but as different platforms use different labels, we refer to *reactions*, understood as one-click engagements where users may select one option from a short pre-defined list as their 'reaction' to a post, with users' choices typically summed and presented as a quantified metric beneath the item. Reactions include perhaps most famously Facebook's original 'Like' and their now five other reaction emojis, the hearts/likes on Instagram, TikTok and Twitter, and Reddit's up- and downvotes Weber and Neumann (2021). Importantly, all these reactions inform the platforms' algorithmic content sorting, and thus steer users' attention.

In an honest world, reactions may be informative in steering attention: through the wisdom-of-crowds, summed reactions may help identify relevant, well-produced, or otherwise high quality content as attention-worthy, so it may be presented to users at the top of their news feed (Bhadani et al. 2022). Alas, that reactions serve as attention-steering exactly makes them—along with other quantified attention metrics (Giglietto et al. 2020b)—a target candidate for influence operations that spread misinformation based on coordinated inauthentic behavior (CIB-based IOs). Accounts (often bots) used to hack users' attention simulate authentic interest in a topic through reacting to social media posts (Goerzen and Matthews 2019). While not actively posting content, they seek to elevate or suppress specific topics in the public perception, flood platforms with misinformation, and boost narratives counter to an authentic public interest (Takacs and McCulloh 2019). The identification of such

[1]When the hashtag was viewed in private browser tab without being logged or when logged in with a new Twitter profile. The tweet was clearly marked as a test, and published by an account with almost no network or activity.

[2]As many platforms' sorting algorithms assign higher rank to posts that many users have engaged with—e.g., through liking, upvoting, sharing, retweeting or commenting—attention hacking influence operations orchestrate coordinated engagements through coordinated inauthentic behavior to maximize their effect (Nizzoli et al. 2021).

computational propaganda is difficult as modern bots mask their identity, mimicking human behavior to an increased extent (Beatson et al. 2021; Bradshaw and Howard 2017).

When CIB-based IOs target reactions, the wisdom-of-crowds effect is lost. Scholars have called for ways to promote the Internet's potential to strengthen rather than diminish democratic virtues (Lazer et al. 2018), e.g., by redesigning online environments to enable informed choice of attention expenditure by providing transparent crowd-sourced voting systems (Lorenz-Spreen et al. 2020). Here, current implementations of reactions are in the ballpark, yet strongly flawed as they may be hacked by CIB-based IOs. Adopting exactly a voting perspective, this paper develops a computational approach to detect and remove CIB influence on reactions, with the aim to restore reactions' wisdom-of-crowds effects.

Detecting and removing coordinated inauthentic behavior targeted to reactions is a neglected area of research (perhaps partially because relevant data is difficult for researchers to obtain despite often being public, a topic we return to below and in the concluding remarks). In general, computational approaches to combat CIB have not been studied extensively (Nizzoli et al. 2021). Recent research has explored user information-based coordination such as account handle sharing, content-based coordination (e.g., synchronized co-posting of images, hashtags, text, and links), attention metric-based coordination such as co-retweeting, or timing-based coordination (Kirn and Hinders 2022; Pacheco et al. 2021; Nizzoli et al. 2021; Giglietto et al. 2020b,a; Grimme, Assenmacher, and Adam 2018; Weber and Neumann 2021). Despite reactions being a commonly adopted and an easily manipulatable mechanism, research on CIB more narrowly targeted at reactions is quite scarce. Borderlining relevancy are studies on purchased likes not of posts, but of *pages [followers]* on Facebook [Twitter] (Ikram et al. 2017; De Cristofaro et al. 2014; Beutel et al. 2013) [(Aggarwal and Kumaraguru 2015)]. This stream of work tries to understand the modus operandi of page like farms [follower farms] (De Cristofaro et al. 2014) [(Aggarwal and Kumaraguru 2015)] and develops supervised classification models based on demographic, temporal, and social characteristics (Ikram et al. 2017) [(Aggarwal and Kumaraguru 2015)]. Here, notably, Ikram et al. (2017) find that their bot classifier has difficulty detecting page like farms that mimick regular like-spreading over longer timespans, and conclude that Beutel et al. (2013)'s unsupervised approach to detect page like farms—even developed with data from inside Facebook—yielded large false positive errors.[3] Directly about reactions to posts is Torres-Lugo et al. (forthcoming 2022)'s study of metric inflation through strategic deletions on Twitter. They analyze coordination in repetitive *(un)liking* on *deleted* tweets in influence operations that seek to bypass daily anti-flooding tweeting limits. From a curation point of view, looking at unlikes is a very smart move, as this data is in fact available to purchase from Twitter. Alas, the approach is inapplicable to tweets that remain online, such as those central to CIB-based IOs that push narrative through political astroturfing (Schoch et al. 2022).

[3]Also in the closely related field of bot detection has the detection of bots that are mainly designed to engage through reactions gone unstudied, again perhaps due to data restrictions. For a systematic review of the bot detection literature, see (Orabi et al. 2020).

1.1 A Voting and Simulation Approach to Coordinated Inauthentic Behavior

To study CIB targeted at reactions, we methodologically take a voting perspective on reactions and a computer simulation approach to validate the proposed methods.

With the voting perspective, we conceptualize reactions as votes about the epistemic quality of an information item. We restrict attention to a two-reaction case, with one reaction interpreted as a vote *for* the item being of high quality, the other a vote against. We adopt this voting perspective as it allows us to clearly explicate a structure of reactions as binary voting, to specify different patterns and varying degrees of coordination (Nizzoli et al. 2021), and to define and quantify the aptitude of a group of users with respect to tracking quality.

Further, it allows us to draw on intuitions from the *Condorcet Jury Theorem*[4] (Condorcet 1785): while many weakly competent authentic judgments may lead to a highly accurate collective judgment through simple majority vote, such positive wisdom-of-crowds effects may be counteracted by the non-independence exhibited by coordinated inauthentic behavior.

The latter motivates the paper's fundamental approach to counter CIB influence, namely to design *jury selection procedures* (JSPs). The core idea is this: given a collection of votes from a voting population of agents, a JSP searches the collection for coordinated voting and from the findings classifies agents as inauthentic or authentic, before finally returning a subset of the population—the *jury*—whose votes are tallied to determine the epistemic quality of a post. I.e., a JSP censors a subset of the population's votes in order to restore wisdom-of-crowds effects for the remainder.

Methodologically, the paper is also a computer simulation paper. We develop an agent-based model (ABM) in which agents vote on the quality of fictitious posts. The ABM includes agents that vote authentically—in accordance with their private beliefs about the quality of the post and the assumptions of the Condorcet Jury Theorem—and some that do not, either by voting only inauthentically or coordinatedly inauthentically. Over synthetic vote data generated by the ABM, we test and validate the machine learning-based JSPs that we develop.

Validating with synthetic data circumvents three main challenges in detecting coordinated inauthentic users (lacking reproducibility, lacking data availability, and lacking ground truth), while suffering the downside that synthetic data has limited ecological validity. First, empirical social media studies of bots remain problematic to replicate and

[4]When all jurors vote *independently* and are *better than random* at voting correctly, the probability of a correct majority judgment approaches 1 as the jury size approaches ∞.

reproduce due to a time-sensitivity of the relevant data (Martini et al. 2021; Samper-Escalante et al. 2021; Bebensee, Nazarov, and Zhang 2021). Attempts to collect the same data twice are likely to fail, as traces of coordination may be altered or deleted after an influence operation was concluded. While e.g. Twitter grants generous academic research access to historic tweets through their API, accounts involved in CIB may evade detection as they are no longer retrievable in their original appearance (Torres-Lugo et al. forthcoming 2022). The shortcomings in data reproducibility make CIB/bot detection frameworks difficult to compare, as these typically require live data access (Martini et al. 2021). Data and analyses of the methods proposed here are time-insensitive and reproducible (cf. Data Availability Statement and Supplementary Material[5]).

Second, data availability limits research. Large scale studies may simply be impossible due to data access restrictions (Martini et al. 2021; Bliss et al. 2020; Pasquetto, Swire-Thompson et al. 2020). Specifically data concerning users' reactions is very difficult for researchers to obtain: none of the currently existing datasets include it,[6] and neither Meta, Twitter nor Reddit supply this data in necessary scope (Bliss et al. 2020; Pasquetto, Swire-Thompson et al. 2020). We outline data collection strategies in connection with empirical validation of our methods in the concluding remarks. Data from an ABM can be (re)synthesized in any quantity.

Third, there is an issue with lacking ground truth as researchers do not have access to the empirical truth about accounts involved in coordinated inauthentic behavior. Qualified guesses can be made based on suspicious similarities in behavior or profile features, but *de facto*, it remains unknown whether two users' actions are authentically correlated or inauthentically coordinated, or how many fully or partially automated accounts exist in a total population (Magelinski, Ng, and Carley 2022; Martini et al. 2021; Samper-Escalante et al. 2021; Chavoshi, Hamooni, and Mueen 2017; Beutel et al. 2013). Specifically for reaction-based CIB, it seems infeasible to create a labeled dataset that even *approximates* the ground truth: labeling accounts individually e.g. via crowd-sourcing or the well-established bot classifier *Botometer* will likely fail as i) single accounts will often seem inconspicuous unless looked at in concert at a collective level (Magelinski, Ng, and Carley 2022; Grimme, Assenmacher, and Adam 2018; Yang et al. 2019, 2020a),[7] and ii) collective level labeling is impossible due to current data restrictions as reactions data is available only in severely limited quantities, if at all.[8]

By validating over an ABM where we specify which agents are involved in CIB, we gain transparency and a ground truth. We get precise baselines, exact measurements of the effect of our methods, and certainty about the degrees of misclassification. We elaborate on this below. Hereby, the ABM validation allows us to provide methodologically robust proof of concept for the JSP approach.

1.2 Existing Work and Contributions

Little work exists on identifying and eliminating inauthentic votes and JSPs. Galeazzi, Rendsvig, and Slavkovik (2019) suggest to remove inauthentic influence by identifying an independent jury via the χ^2 test of independence. Their model takes sharing-induced diffusion in social networks as evolving crowdvoting. Their main results pertain to JSP time-complexity, with their least requiring suggestion still exponential in the jury size (a direct consequence of using χ^2). In addition, we find that the number of data points required for χ^2 application (see Sec. 3) makes their JSPs practically inapplicable and computationally unservicable. A performance comparison with their bot detecting scheme is therefore impossible beyond contrasting data requirements.

The central goal of this paper is to develop jury selection procedures that raise the correctness of vote by majority of juries, complementing (Galeazzi, Rendsvig, and Slavkovik 2019). The methodological voting perspective allows us to define a metric of success for the methods we develop: majority correctness scores (MCSs). Majority correctness scores give a direct perspective on the collective epistemic practice of a group of agents, providing a more conclusive perspective than misclassification scores. Beyond raising majority correctness scores, we desire *accurate* JSPs that minimize misclassification of i. authentic agents as inauthentic and ii. inauthentic agents as authentic (i.e., minimize i. false positive and ii. false negative errors). The first values *vox populi* and penalizes censorship (Shao et al. 2018), while the second is a *precautionary principle* against inauthentic influence. Further, we desire *feasible* JSPs that use only data that is obtainable by social media platforms and that requires little to no preprocessing, have few to no supervised elements (Orabi et al. 2020; Grimme, Assenmacher, and Adam 2018), and have reasonable complexity.

This paper develops two JSPs, evaluated with respect to vote data generated by the agent-based model. The ABM is presented in Sec. 2 where varying baseline agent populations' majority correctness scores (MCSs) are inspected, on which inauthentic activity has a substantial negative impact.

The core JSP machinery is presented in Sec. 3. Each jury selection procedure invokes a classifier method that decomposes the ABM data into singular values (SVD), applies a clustering strategy (either a Gaussian Mixture Model (GMM) or the k-means algorithm (KM)), and labels agents using a non-standard application of logistic regression on the qualitative property of the post voted on. In related work, vote data—such as US congress roll call data—has been suc-

[5]Code to reproduce and analyse the data can be found at the public GitHub repository *Coordinated-Inauthentic-Behavior-Likes-ABM-Analysis* at https://github.com/LJ-9/Coordinated-Inauthentic-Behavior-Likes-ABM-Analysis

[6]See e.g. Indiana University's Bot Repository, a resourceful, centralized repository of annotated datasets of Twitter social bots: https://botometer.osome.iu.edu/bot-repository/datasets.html.

[7]*Botometer*'s feature-based approach considers accounts one at a time and does therefore not pick up on group anomalies based on suspicious similarity (Yang et al. 2019, 2020a).

[8]Twitter is the only platform that offers *any* access, with limitations of 75 requests per 15 minutes, each granting only the most recent 100 liking users of a single tweet. See https://developer.twitter.com/en/docs/twitter-api/tweets/likes/api-reference/get-tweets-id-liking_users.

cessfully grouped employing dimensionality reduction, e.g., (Yang et al. 2020b; Porter et al. 2005; Sirovich 2003; Poole 2000). Our approach is novel in applying such methods in the realm of digital propaganda using simple binary input data. Dimensionality reduction and clustering methods have so far been applied to less sparse data structures, such as HTTP-level traffic patterns (Suchacka 2019; Suchacka and Iwański 2020), textual data of tweets (Kirn and Hinders 2022), or rich datasets with behavior-based features (number of friends/followers, mentions and hashtags, etc.) like in detection of spam bots on social media sites (e.g., (Ahmed and Abulaish 2013)). A systematic review on detection of bots on social media (Orabi et al. 2020) further discusses unsupervised methods, e.g. (Chavoshi, Hamooni, and Mueen 2017; Chen and Subramanian 2018), yet to our knowledge only Galeazzi, Rendsvig, and Slavkovik (2019) attempt to flag agents given just binary vote data (i.e., with no added information about e.g. temporal coordination as in (Beutel et al. 2013; Grimme, Assenmacher, and Adam 2018; Magelinski, Ng, and Carley 2022; Pacheco et al. 2021; Schoch et al. 2022)) obtainable intra-platform by social media sites.

In Sec. 4, we define and evaluate the GMM and KM jury selection procedures. We show that both are highly successful, as they select juries that have vastly increased majority correctness scores compared to baseline juries. Moreover, the GMM JSP outperforms the KM JSP with respect to its accurate and particularly precautious results. Sec. 5 summarizes the main findings and discusses ethical considerations, model assumptions, and data collection.

Technically, we contribute a novel, reactions-based approach to detect CIB, implemented in two variants evaluated to have positive effects over synthetic ABM data, thus showing proof of concept. Societally, the proof of concept provides a direct argument to be raised to social media platform to open access to reactions data: the data is necessary to evaluate, tweak and deploy promising methods (i.e., JSPs) to combat coordinated inauthentic behavior and thus to inhibit the spread of misinformation.

2 Agent-Based Model (ABM)

We evaluate the two jury selection procedures over data generated by the following agent-based model. A model *run* consists of a fixed set of agents partitioned into agent types (see below), and a sequence of independent *voting rounds*. Each round concerns a given post (which we do not explicitly represent) and whether the post is of high or low quality, on which agents vote $\{1, -1\}$ (1 for high, -1 for low). We think of these votes as users' reactions, and call 1 an *upvote* and -1 a *downvote*.

Agents are either *authentic* or *inauthentic*. We formally define the agents types in Sec. 2.2 below. We think of authentic agents as regular social media users that use their up- and downvotes to inform about post quality (e.g., analogously to Metaxas et al. (2015) who showed that by retweeting, users on Twitter signal trust in the message). Authentic agents vote independently according only to their competence-based beliefs about post quality: they satisfy the assumptions of the Condorcet Jury Theorem. Inauthentic agents do not: with different patterns and varying degrees, they coordinate their votes through properties distinct from quality. On social media, inauthentic behavior can both be witnessed among human controlled and automated accounts. Given the scale of influence operations, it is relevant to think about inauthentic behavior in terms of so-called *social bots*: "*Computer programs designed to use social networks by simulating how humans communicate and interact with each other*" (Abokhodair, Yoo, and McDonald 2015). The design of our inauthentic agents draws inspiration from the social bot classes *astroturfing bots* (that create "*the appearance of widespread support for a candidate or opinion*" (Ratkiewicz et al. 2011)) and *influence bots* ("*Realistic automated identities that illicitly shape discussion*" (Subrahmanian et al. 2016)) (Orabi et al. 2020).

2.1 Post Properties, Competences and Beliefs

Let A be a finite set of agents and $I = \{1, 2, 3\}$ an index set for properties. A voting round concerns a given post, and commences with the (Monte Carlo like) sampling of a state

$$s = (p_i, C_a(p_1), B_a(p_i))_{i \in I \cup A, a \in A} \in \mathbb{R}^{3+|A|+|A|+3|A|+|A|^2}$$

where each p_i represents a property of the post, $C_a(p_1)$ is agent a's competence in evaluating whether the post has property p_1 and $B_a(p_i)$ is a's belief about whether the post has property p_i. Properties $(p_i)_{i \in I} = (p_1, p_2, p_3) \in \{-1, 1\}^3$ are sampled independently from a binomial distribution with probabilities $P(p_i = 1) = (1 - P(p_i = -1))$, given as noise levels in Sec. 2.4. Each p_a is sampled as p_3, and is a private property used by some agent types.[9] We say that the post has property p_i if $p_i = 1$, else that it does not. Property p_1 represents whether the post has high or low quality, and is the only property relevant to authentic agents. Inauthentic agents act also on additional properties, as described below.

Each agent $a \in A$ is assigned a competence $C_a(p_1)$ to determine whether the post has high quality, p_1.[10] To evaluate p_1, it is assumed that all agents are better than fair coin tosses but not perfect: $C_a(p_1) \in [0.65, 0.95]$. We chose $[0.65, 0.95]$ for $C_a(p_1)$ to expedite convergence towards a 100% MCS for authentic agents while ensuring imperfect competence. Any closed, convex subinterval of the open $(0.5, 1)$ would yield similar results w.r.t. MCS, more or less quickly. Competences are uniformly resampled each round, to capture that agents' expertise may vary from post to post. Inauthentic agents are assumed perfectly competent in evaluating properties p_2 and p_3, which they use to coordinate their actions: $C_a(p_2) = C_a(p_3) = 1$. Properties and competences probabilistically determine agents' beliefs: for all $a \in A, i \in I$, the beliefs $B_a(p_i) \in \{-1, 1\}$ are sampled with

$$C_a(p_i) = P(B_a(p_i) = p_i). \tag{1}$$

[9] We include p_a and $B_a(p_b)$ for all agents $a, b \in A$ in the state for description simplicity. In the simulation implementation, we only sampled p_a and $B_a(p_a)$ for agents a that make use of p_a.

[10] Even if p_1 is irrelevant to the agent's voting behavior. This is to simplify the implementation of the model simulation.

If $B_a(p_i) = 1$, then a believes that the post has property p_i, else a believes it does not.[1] If $B_a(p_i) = p_i$, then a's belief about p_i is correct. Hence, (1) states that the probability of agent a's beliefs about p_i being correct equates a's competence with respect to p_i. For two rounds and their states s and s', all sampling is independent, and in each state s, each $C_a(p_1)$ is independent from $C_b(p_1)$, $a \neq b$. No correlations between properties are assumed due to the interpretations of p_2 and p_3, stated below.

2.2 Agent Types

We define 10 agent types. Each agent type is a behavior-defining function that maps an agent's beliefs to votes. The set of agent types is $\{\mathtt{A}, \mathtt{B}_i, \mathtt{D}_i, \mathtt{L}_i\}_{i\in\{\uparrow,\downarrow,\updownarrow\}}$, each defined and described below. A *population* is a map $\mathcal{P} : A \longrightarrow \{\mathtt{A}, \mathtt{B}_i, \mathtt{D}_i, \mathtt{L}_i\}_{i\in\{\uparrow,\downarrow,\updownarrow\}}$ that assigns each agent an agent type.

Intuitively, $\{\mathtt{A}, \mathtt{B}_i, \mathtt{D}_i, \mathtt{L}_i\}_{i\in\{\uparrow,\downarrow,\updownarrow\}}$ contains the following agent types: $\mathtt{A}$ is the *authentic* agent type, and the inauthentic agents come in three types that incorporate different patterns of coordination—*boosters* $\mathtt{B}_i$, *distorters* $\mathtt{D}_i$, and *lone wolfs* $\mathtt{L}_i$. Each inauthentic type votes based on beliefs about a property *distinct* from quality. Boosters and distorters vote respectively given properties p_2 and p_3 to coordinate their inauthentic behavior in-group. Lone wolfs do not coordinate. Each group contains three sub-types: one main to our story which *upvotes on cue* ($i = \uparrow$), and two auxiliary that *downvote on cue* ($i = \downarrow$) or *both up-and downvote on cue* ($i = \updownarrow$). We include the auxiliary sub-types to create a more noisy—and thus harder to maneuver—setting for the JSPs. We hope the notation is mnemonically helpful rather than distracting.

Throughout, the largest population is $\mathcal{P}_{\mathtt{Full}}$, defined for an agent set A, $|A| = 1900$, with 1000 agents assigned to $\mathtt{A}$ and 100 agents to each $\mathtt{X} \in \{\mathtt{B}_i, \mathtt{D}_i, \mathtt{L}_i\}_{i\in\{\uparrow,\downarrow,\updownarrow\}}$. This size and ratio allows for flexibly choosing subpopulations with sizes large enough to produce robust votes. We mainly study subpopulations (restrictions) of $\mathcal{P}_{\mathtt{Full}}$. We specify these subpopulations by stating the size of the pre-image of the agent types (which is sufficient as precise agent identity will not matter), where we write $|\mathtt{X}|$ for $|\mathcal{P}_{\mathtt{Full}}{}^{-1}(\mathtt{X})|$ for agent type $\mathtt{X}$. The four main subpopulations are subsets of either 1000 agents ($\mathcal{P}_{\mathtt{All}}$ containing all agents types, with 100 agents of each type) or 200 agents ($\mathcal{P}_{\mathtt{B}_\uparrow}$, $\mathcal{P}_{\mathtt{D}_\uparrow}$ and $\mathcal{P}_{\mathtt{L}_\uparrow}$ each with 100 authentic agents and 100 agents of either type $\mathtt{B}_\uparrow$, $\mathtt{D}_\uparrow$ or $\mathtt{L}_\uparrow$). Thus, let $\mathcal{P}_{\mathtt{All}}$ be the restriction of $\mathcal{P}_{\mathtt{Full}}$ with $|\mathtt{X}| = 100$ for each $\mathtt{X} \in \{\mathtt{A}, \mathtt{B}_i, \mathtt{D}_i, \mathtt{L}_i\}_{i\in\{\uparrow,\downarrow,\updownarrow\}}$, let $\mathcal{P}_{\mathtt{B}_\uparrow}$ be the restriction of $\mathcal{P}_{\mathtt{Full}}$ with $|\mathtt{A}| = |\mathtt{B}_\uparrow| = 100$ and $|\mathtt{X}| = 0$ for $\mathtt{X} \in \{\mathtt{B}_i, \mathtt{D}_i, \mathtt{L}_i\}_{i\in\{\uparrow,\downarrow,\updownarrow\}} \setminus \{\mathtt{B}_\uparrow\}$, and let $\mathcal{P}_{\mathtt{D}_\uparrow}$ and $\mathcal{P}_{\mathtt{L}_\uparrow}$ be given as $\mathcal{P}_{\mathtt{B}_\uparrow}$ replacing $\mathtt{B}_\uparrow$ with respectively $\mathtt{D}_\uparrow$ and $\mathtt{L}_\uparrow$. We may further specify subpopulations of $\mathcal{P}_{\mathtt{All}}$, $\mathcal{P}_{\mathtt{B}_\uparrow}$, $\mathcal{P}_{\mathtt{D}_\uparrow}$ and $\mathcal{P}_{\mathtt{L}_\uparrow}$ like we specify subpopulations of $\mathcal{P}_{\mathtt{Full}}$. These restrictions mainly serve to describe what happens when we reduce the number of authentic agents. We write e.g. "$\mathcal{P}_{\mathtt{B}_\uparrow}$ for $|\mathtt{A}| = 25$" to mean the subpopulations of $\mathcal{P}_{\mathtt{B}_\uparrow}$ with 125 agents in total, 25 of them authentic.

[1] Hence, agents never suspend judgment, even on properties irrelevant to their voting behavior. Superfluous beliefs have no effects, and are only to simplify implementation.

Authentic Agents. Authentic agents—agents a of type $\mathtt{A}$—correspond to those assumed in the Condorcet Jury Theorem: they vote fully in accordance with their beliefs about quality (p_1), independently of others, and with a competence strictly above 0.5. The vote of an authentic agent a in state s is $\mathtt{A}(a, s) \in \{-1, 1\}$, given by the following table:

	$B_a(p_1) = 1$	$B_a(p_1) = -1$
$\mathtt{A}(a, s)$	1	-1

In this and the below tables, row index ($\mathtt{A}(a, s)$) denotes the agent type and the cell content denotes the action taken in the circumstances specified in the column index (e.g. $B_a(p_1) = 1$).

Boosters. Boosters vote in a coordinated partisan fashion, aiming to swing the majority vote in a direction given by p_2, irrespective of quality (p_1). Hence boosters exhibit CIB. In social media terms, we think of p_2 as disconnected from quality (p_1), but as representing that the post, e.g., originates from a specific source, expresses a given viewpoint, or—taking booster agents as bots—as tagged for special action by a handler.

The main *Upvote Booster* $\mathtt{B}_\uparrow$ has as goal to boost and amplify p_2 posts: they upvote ("Yes, the post has p_1") if they believe the post has property p_2, and else vote authentically (to hide their inauthentic activities). For auxiliaries, the *Downvote Booster* $\mathtt{B}_\downarrow$ 'inverts' $\mathtt{B}_\uparrow$: $\mathtt{B}_\downarrow$ demotes non-p_2 posts by downvoting if they believe the post does not have p_2, and else vote authentically, while the *Both Booster* $\mathtt{B}_\updownarrow$ combine the inauthentic behaviors of $\mathtt{B}_\uparrow$ and $\mathtt{B}_\downarrow$ by always voting according to p_2, and never authentically. The vote of an agent a of type $\mathtt{B}_{i\in\{\uparrow,\downarrow,\updownarrow\}}$ in state s is $\mathtt{B}_i(a, s)$ given by

	$B_a(p_2) = 1$	$B_a(p_2) = -1$
$\mathtt{B}_\uparrow(a, s)$	1	$\mathtt{A}(a, s)$
$\mathtt{B}_\downarrow(a, s)$	$\mathtt{A}(a, s)$	-1
$\mathtt{B}_\updownarrow(a, s)$	1	-1

The table also refers to the authentic agent type $\mathtt{A}$ to make it visually explicit in which cases the Up- and Downvote Boosters behave authentically.

Distorters. Distorters seek to create noise among the votes by, on cue, voting against their beliefs about quality. They vote in a coordinated, but non-partisan fashion: triggered by p_3, they vote contrary to their private beliefs about quality (p_1). As with p_2, we think of p_3 as encoding a property of the post distinct from quality, such as, e.g., tag, source or viewpoint. The $\mathtt{D}$ agents seek to water down the majority view and damper public impressions of consensus, thus exhibiting one form of *concern trolling* (Goerzen and Matthews 2019).

The main *Upvote Distorter* $\mathtt{D}_\uparrow$ votes authentically (to hide) except when they believe the post has p_3 but not p_1: then they distort by voting contrary to their belief about p_1 (e.g., they upvote low quality posts of a given viewpoint to dampen consensus impressions). For auxiliaries, the *Downvote Distorter* $\mathtt{D}_\downarrow$ 'inverts' $\mathtt{D}_\uparrow$: they vote authentically except when believing the post has both p_1 and p_3; then they distort by voting contrary to their beliefs about quality. The *Both Distorter* $\mathtt{D}_\updownarrow$ join the inauthentic behaviors of $\mathtt{D}_\uparrow$ and $\mathtt{D}_\downarrow$: if

they believe the post has p_3, then they vote contrary to their p_1 beliefs (e.g., to always sow distrust about content from a given source, or of a given viewpoint). The vote of an agent a of type $\mathtt{D}_{i\in\{\uparrow,\downarrow,\updownarrow\}}$ in state s is $\mathtt{D}_i(a,s)$ given by

	$B_a(p_3)=1$, and $B_a(p_1)=1$	$B_a(p_3)=1$, and $B_a(p_1)=-1$	$B_a(p_3)=-1$
$\mathtt{D}_\uparrow(a,s)$	$B_a(p_1)$	$-1\cdot B_a(p_1)$	$\mathtt{A}(a,s)$
$\mathtt{D}_\downarrow(a,s)$	$-1\cdot B_a(p_1)$	$B_a(p_1)$	$\mathtt{A}(a,s)$
$\mathtt{D}_\updownarrow(a,s)$	$-1\cdot B_a(p_1)$	$-1\cdot B_a(p_1)$	$\mathtt{A}(a,s)$

Lone Wolfs. Lone wolfs also create noise among the votes by voting against their beliefs about quality. They do so exactly as the distorters, but without coordination through p_3. We interpret these agents as individual users that—cued by a personal property—upvote contra their beliefs about quality ($\mathtt{L}_\uparrow$, main, *Upvote Lone Wolf*), e.g., out of sympathy, downvote contra their beliefs about quality ($\mathtt{L}_\downarrow$, aux., *Downvote Lone Wolf*), e.g., out of anger or spite, or both ($\mathtt{L}_\updownarrow$, aux., *Both Lone Wolf*).

Instead of voting given shared property p_3, a lone wolf, i.e., an agent a of type $\mathtt{L}_{i\in\{\uparrow,\downarrow,\updownarrow\}}$, votes on a personal property $p_a \in \{-1,1\}$, believing $B_a(p_a) \in \{-1,1\}$ with $P(B_a(p_a)=p_a)=1$. For all $a,b\in A$, properties p_a, p_b are sampled as p_3, but if $a\neq b$, p_a and p_b are sampled independently. The voting rules for each $\mathtt{L}_i$, $i\in\{\uparrow,\downarrow,\updownarrow\}$, is obtained by replacing p_3 with p_a in the table for $\mathtt{D}_i$.

2.3 Majority Vote and Correctness

We are interested in how agent populations' votes fair with respect to *majority correctness*, both before (baseline experiments) and after we have applied our two jury selection procedures. A *jury* is a set of agents $J\subseteq A$. Let $(v_a)_{a\in J}$, $v_a\in\{1,-1\}$ be a *voting profile* of J with respect to the post. The *majority vote* of $(v_a)_{a\in J}$ is whichever of 1 and -1 that gets more votes, tie-breaking to 1, giving the post the benefit of doubt. I.e., the majority vote of $(v_a)_{a\in J}$ is -1 if $\sum_{a\in J} v_a < 0$, else 1. The majority vote is *correct* if it equals the post's quality, $p_1\in\{1,-1\}$. Finally, the *majority correctness score* (MCS) of a jury over a set of voting rounds is the percentage of correct majority votes of the jury in those rounds. The MCS of a jury is a measure of its competence with respect to tracking quality, and is the jury performance indicator of interest in this paper.

2.4 Parameters and Generated Dataset

Using R to implement the ABM,[2] we chose three *noise level* parameter combinations for the sampling of properties:

	$P(p_1=1)$	$P(p_2=1)$	$P(p_3=1)$
LOW	0.75	0.75	0.9
MID	0.75	0.5	0.5
HIGH	0.75	0.1	0.1

[2]We implemented the ABM from the ground up to retain freedom in agent design and as the simplicity of the encoded behavior and generated data do not invoke advanced features of existing ABM simulation packages and programs, such as Netlogo (Wilensky 1999), Laputa (Angere 2010; Olsson 2013), or Hashkat (Ryczko et al. 2017).

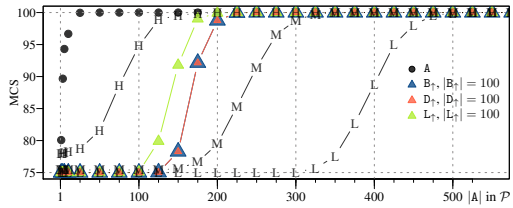


Figure 1: Baseline MCS of $\mathcal{P}$ with $|\mathtt{A}|$ = 1,3,5,10,25,50,..., for $p_2=p_3=1$. Colored: populations with single inauthentic type in low noise. L/M/H: populations with multiple inauthentic types $\mathcal{P}=\{\mathtt{B}_i,\mathtt{D}_i,\mathtt{L}_i\}_{i\in\{\uparrow,\downarrow,\updownarrow\}}$, with $|\mathtt{B}_i|=|\mathtt{D}_i|=|\mathtt{L}_i|=100$ in low (L), mid (M), high (H) noise. Find MCS of $\mathcal{P}_{\mathtt{All}}$, $\mathcal{P}_{\mathtt{B}_\uparrow}$, $\mathcal{P}_{\mathtt{D}_\uparrow}$, $\mathcal{P}_{\mathtt{L}_\uparrow}$ at $|\mathtt{A}|=100$.

These noise levels were chosen to produce different voting patterns, and to introduce varying degrees of coordination and correlations among votes, in turn producing three levels of difficulty for vote-based agent classification. Quality (p_1) is fixed across levels, leaving authentic agents unaffected. Inauthentic agents perform less (coordinated) inauthentic activities in higher levels, as p_2 and p_3 decrease. They thus mimic authentic agents more (more noise), raising the difficulty of classification. The sampling of p_2 is asymmetric to avoid mirrored results in low and mid noise for $\mathtt{B}_\downarrow$ and $\mathtt{B}_\uparrow$ given that booster agents solely rely on p_2. We chose a symmetric setup for $P(p_3=1)$ as distorters and lone wolfs' votes are not solely determined by p_3, but influenced by the sampling of p_1, too, hence making completely mirrored votes less likely. For each noise level, we performed 100 runs, each based on a random seed and with voting rounds $r=1000$, producing a dataset with $3\times 100,000$ (state, vote profile) pairs. Each was done for $\mathcal{P}_{\mathtt{Full}}$, thus counting 1900 agents: 1000 authentic and 100 of each inauthentic type. Sec. 2.5 displays diverse population ratios that explore the effect of authentic agents in minority and majority on MCS. Throughout, results are based on and evaluated against a datasubset with $r=500$. As all runs and rounds are independent, choosing fewer or more voting rounds is without problem. Other values of r are mentioned explicitly when robustness checks are discussed.

2.5 Baseline Majority Correctness Scores

To showcase varying populations' behaviors, we illustrate two sets of baseline MCS results in Figures 1 and 2.

Figure 1 shows 7 populations' MCSs as a function of the number $|\mathtt{A}|$ of authentic agents in the population. As expected from the Condorcet Jury Theorem, the MCS of authentic agents alone converges to 100%, with 25 agents sufficing. This is representative for all noise levels, as noise does not affect authentic agents. Figure 1 is filtered to rounds with $p_2=p_3=1$, so $\mathtt{B}_\uparrow$ and $\mathtt{D}_\uparrow$ are 'actively inauthentic' (and both always upvote). Given this filter, the figure is representative for all noise levels for $\mathtt{B}_\uparrow$ and $\mathtt{D}_\uparrow$. The effect of $\mathtt{L}_\uparrow$ is level specific (but unaffected by the filter).

We make three observations concerning the main, upvot-

ing inauthentic agents $\mathtt{B}_\uparrow, \mathtt{D}_\uparrow$ and $\mathtt{L}_\uparrow$ of Figure 1. **First**, the left-most part of Figure 1 shows populations with $|\mathtt{A}| = 1$, a very hospitable environment for inauthentic activity. Here, each of $\mathtt{B}_\uparrow, \mathtt{D}_\uparrow$ and $\mathtt{L}_\uparrow$ exhibit a MCS of 75%. This is an artifact of how their behavior interacts with the sampling frequency for p_1. For $\mathtt{B}_\uparrow$ and $\mathtt{D}_\uparrow$, the MCS of 75% follows as Figure 1 is filtered for $p_2 = p_3 = 1$, and only contains rounds where both always upvote. As $p_1 = 75\%$, they are thus correct 75% of the time. Though not all $\mathtt{L}_\uparrow$ always upvote in these rounds, they do so individually with a 96.5% chance (assuming average $C_a(p_1) = 0.8$). As a group, they thus sway the majority vote to 1 with high probability, again correct with 75%. **Second**, $\mathtt{B}_\uparrow, \mathtt{D}_\uparrow$ and $\mathtt{L}_\uparrow$ each exhibit their maximal lowering effect on the MCS while $|\mathtt{A}| = 100$. This is a motivating factor in focusing on populations with $|\mathtt{A}| = 100$, the MCSs of which we return to in Table 2. **Third**, for $|\mathtt{A}| > 100$, $\mathtt{B}_\uparrow$ and $\mathtt{D}_\uparrow$ negatively influence the MCS identically, as both upvote in the shown rounds, with their effect declining from a MCS of 75% at $|\mathtt{A}| = 125$ to an MCS of 100% by $|\mathtt{A}| = 225$. For $|\mathtt{A}| > 100$, to form an incorrect majority, inauthentic agents must be 'aided' by authentic agents that happen to vote incorrectly. The probability that enough such exist to overcome the correctly voting authentic agents drops as $|\mathtt{A}|$ grows. With $|\mathtt{A}| \geq 225$, $\mathtt{B}_\uparrow$ and $\mathtt{D}_\uparrow$ are seen to have lost all effect. $\mathtt{L}_\uparrow$ have a less robust effect, as they vote in an uncoordinated fashion, and are thus more quickly outnumbered by authentic agents' votes.

Finally, the effect of the 900 inauthentic agents jointly drops with higher noise levels, i.e., with decreased activity. In the high activity case (low noise), the 900 inauthentic agents seem 'overwhelmed' already by between 325 and 475 authentic agents. This is correct on the aggregate level, but 900 inauthentic agents do not equate 900 inauthentic actions: given the filter, some types act authentically always ($\mathtt{B}_\downarrow$) or sometimes ($\mathtt{D}_\uparrow, \mathtt{D}_\downarrow, \mathtt{L}_\uparrow, \mathtt{L}_\downarrow, \mathtt{L}_\updownarrow$). Additionally, some types partially cancel each other (e.g., $\mathtt{D}_\uparrow$ and $\mathtt{L}_\downarrow$) or even themselves (e.g., $\mathtt{D}_\updownarrow$) out.

Figure 2 shows MCS summary plots of all inauthentic agent types in isolation and jointly, as a function of $|\mathtt{A}|$, not filtered for properties. As noise increases, the figure evinces how inauthentic agents' impact on MCS decreases. Note how in low noise, agent types $\mathtt{D}_\updownarrow, \mathtt{D}_\uparrow, \mathtt{L}_\updownarrow$, and $\mathtt{L}_\uparrow$ are more effective than $\mathtt{B}_i$ for each $i \in \{\uparrow, \downarrow, \updownarrow\}$ in lowering the MCS as the former agent types directly counteract correct majority voting concerning quality (p_1). The picture flips in the high noise level given how p_1, p_2, and p_3 are sampled (Sec. 2.4).

3 Classification

Our jury selection procedures (GMM and KM JSPs) classify the set of agents into two agent groups: authentic and inauthentic. Each jury selection procedure invokes a classifier method that decomposes the ABM data into singular values (SVD), applies a clustering strategy—either a Gaussian Mixture Model (GMM) or the k-means algorithm (KM)—and labels agents using logistic regression on the quality property p_1 of the post voted on.

We assume p_1 known, as we know of the general setting: the agents vote on quality. We do not assume knowledge

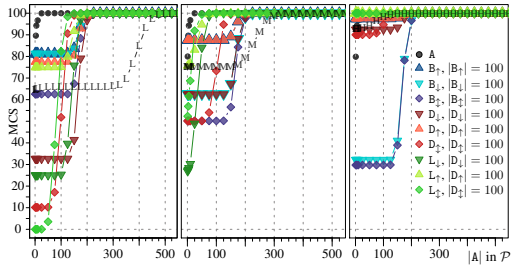


Figure 2: Baseline MCS of $\mathcal{P}$ with $|\mathtt{A}| = 1,3,5,10,25,50,...$ for low (left), mid (mid), high (right) noise and all agent types. L/M/H: $\mathcal{P} = \{\mathtt{B}_i, \mathtt{D}_i, \mathtt{L}_i\}_{i \in \{\uparrow, \downarrow, \updownarrow\}}$, $|\mathtt{B}_i| = |\mathtt{D}_i| = |\mathtt{L}_i| = 100$ in low (L), mid (M), high (H) noise. Find MCS of $\mathcal{P}_{\mathtt{All}}$, $\mathcal{P}_{\mathtt{B}_\uparrow}$, $\mathcal{P}_{\mathtt{D}_\uparrow}$, $\mathcal{P}_{\mathtt{L}_\uparrow}$ at $|\mathtt{A}| = 100$.

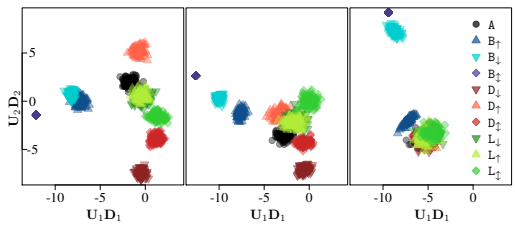


Figure 3: Exemplary, representative scatterplot of $\mathbf{U}_q\mathbf{D}_q$ for $n = 1000$ for $\mathcal{P}_{\mathtt{All}}$ and $r = 500$, in low (left), mid (mid), high (right) noise.

of p_2 and p_3, or even of their existence. The input dataset consists of binary votes of n agents over a given number of voting rounds r, where $r > n$ is not a requirement regarding the machinery. Yet the more observations r, the better we cluster. Data requirements are thus feasible, in contrast to the χ^2 test suggested by Galeazzi, Rendsvig, and Slavkovik (2019) that requires p_1 known plus at least 1 observation for each of the 2^n possible voting round outcomes.

For each ABM run, the classification analysis is performed on five resampled (with replacement) datasets with $r = 500$ and n either 1000 for $\mathcal{P}_{\mathtt{All}}$ or 200 for $\mathcal{P}_{\mathtt{B}_\uparrow}$, $\mathcal{P}_{\mathtt{D}_\uparrow}$, and $\mathcal{P}_{\mathtt{L}_\uparrow}$. For each of the bootstrapped datasets, we calculate the Singular Value Decomposition (SVD) $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ of the $n \times n$ sample correlation matrix $\mathbf{X}$ of the vote data. For clustering, we consider the first $q = 2$ dimensions' eigenvectors, i.e., the first two columns of the $n \times p$ orthogonal matrix $\mathbf{U}$ where $n = p$, weighted with the corresponding eigenvalue collected in the diagonal $p \times p$ matrix $\mathbf{D}$. Hence, we cluster on the q partial components $\mathbf{U}_q\mathbf{D}_q$ (Hastie, Tibshirani, and Friedman 2009). Figure 3 shows the scatterplots of $\mathbf{U}_q\mathbf{D}_q$, illustrating more blurred clustering environments as the noise level increases from low to high.

We contrast the probabilistic Gaussian Mixture Model (GMM) and the deterministic k-means (KM) algorithm for clustering the components. The soft clustering GMM is more

memory-intensive, while the hard clustering KM algorithm is faster. We choose GMM and k-means as they are among the simplest, most well-known, and most efficiently implementable unsupervised clustering methods (Hastie, Tibshirani, and Friedman 2009). Both cluster the weighted eigenvectors into k groups, $k = 2, ..., 20$ (in testing, 20 proved sufficient as upper bound). In the GMM, k is chosen by maximizing the log-likelihood according to the Bayesian Information Criterion (BIC). The BIC penalizes the number of parameters more heavily than Akaikes Information Criterion, aiming for a model fit with fewer parameters to avoid overfitting (Scrucca et al. 2016). In the KM algorithm, k is estimated with the gap statistic, which compares the change in the within-cluster dispersion with that under a reference null distribution (Tibshirani, Walther, and Hastie 2001).

Having clustered the data into k groups, the mean vote per voting round of those agents clustered together—i.e., the row sums of k subsets of the vote data, viz. k $r \times 1$ vectors—are used in a logistic regression model with the two-level factor p_1 as the response variable. Put differently, the k coefficients refer to the clusters' mean vote per voting round given the quality of posts. To select those clusters comprising inauthentic agents, we add the lasso penalty term to the optimization, $\sum_{j=1}^{k} ||\beta||_j$ with k predictor variables (clusters), as implemented in the R package `glmnet` (Friedman, Hastie, and Tibshirani 2010). Coefficients consequently shrunk to 0 when regressing on p_1 receive the label 'inauthentic'. Coefficients *not* shrunk to 0 receive the label 'authentic'. Lasso regularization was chosen over the ridge regularization as the former shrinks coefficients to 0 and thereby imposes sparseness. In contrast, the ridge penalty never fully removes variables. Coefficients shrunk to 0 accordingly do not play an important role when regressing on p_1 and therefore receive the label 'inauthentic'. These labels are then forwarded to the agents found in each cluster.

Note that logistic regression is applied in a non-standard way. In this paper, the goal of the logistic regression is *not* to predict each vote per voting round into the categorical dependent variable p_1, in contrast to classical approaches where the dependent variable describes the classes in which one is interested. We seek to classify each agent as inauthentic or authentic which we do via hard classification through shrinking components to 0 and an additional labeling step. This makes traditional classifier metrics like a receiver operating characteristic curve (ROC curve) and a corresponding area under the curve score (AUC score) inapplicable. Instead, Sec. 3.1 discusses false positive and false negative classification errors to transparently and separately assess the two desiderata vox populi and precaution.

Once the classification analysis is completed on all 5 bootstrapped datasets, each agent has been classified as either 'authentic' or 'inauthentic' 5 times. Only if an agent received the 'authentic' label at least 4 out of 5 times, the overall 'authentic' label will be granted. Else, the agent is overall classified as 'inauthentic'. The $4/5$ classification threshold was fixed pragmatically to balance runtime efficiency and precaution against inauthentic influence. Simple majority would exhibit less precaution and more vox populi, while a $19/20$ classification threshold (95%) would heavily increase runtime. We discuss this modeling choice further in the final remarks (Sec. 5.2).

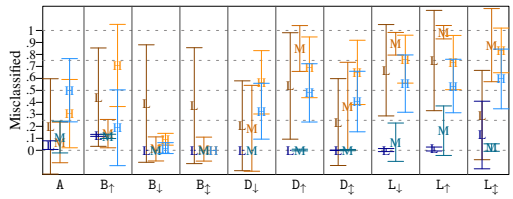


Figure 4: GMM (blue) and KM (orange) mean misclassification and standard deviation in $\mathcal{P}_{\texttt{All}}$ in low (L), mid (M), and high (H) noise.

3.1 Classification Results

In order to evaluate the GMM and KM classifier methods, we first inspect the misclassification results for $\mathcal{P}_{\texttt{All}}$ for $r = 500$, second comment on selected robustness observations, and third examine classifier accuracy for smaller subpopulations $\mathcal{P}_{\texttt{B}\uparrow}$, $\mathcal{P}_{\texttt{D}\uparrow}$, and $\mathcal{P}_{\texttt{L}\uparrow}$ for $r = 500$.

First, in population $\mathcal{P}_{\texttt{All}}$, GMM classifies well in the low (mid) noise case, accurately misclassifying only 4% (11%) of authentic agents as inauthentic, and 3% (4%) of inauthentic agents as authentic (Table 1), exhibiting both vox populi and precaution. As expected, classifier accuracy reduces in the high noise case given that inauthentic agents hide and mimic authentic behavior, i.e., often vote authentically and are accordingly difficult to detect. However, here the inauthentic agents' impact on majority correctness scores is limited (Figure 2) despite a 35% false negative error. Indeed, as Figure 2 suggests, it is $\texttt{B}_{\downarrow}$ and $\texttt{B}_{\updownarrow}$ agents that negatively affect the MCS to the largest extent in the high noise case, and both GMM and KM identify these agent groups accurately as inauthentic (Figure 4). Moreover, classification results in Figure 4 and Table 1 show how GMM outperforms KM in all noise levels with regard to identifying inauthentic agents, exhibiting less false negative misclassification. Thus, GMM overall clusters more precautiously than KM.

Second, robustness checks given $\mathcal{P}_{\texttt{All}}$ show differences between GMM and KM. Results based on fewer observations ($r = 250$ instead of $r = 500$) affect false negative errors less for GMM, but notably for KM. E.g., GMM still does not misclassify any booster or distorter agents, while KM's false negative errors ($^{\text{Mean}}_{(\text{SD})}$) rise in low noise as follows

	$\texttt{B}_{\uparrow}$		$\texttt{B}_{\downarrow}$		$\texttt{B}_{\updownarrow}$		$\texttt{D}_{\uparrow}$		$\texttt{D}_{\downarrow}$		$\texttt{D}_{\updownarrow}$	
r	500	250	500	250	500	250	500	250	500	250	500	250
KM	.44 (.41)	.56 (.43)	.39 (.5)	.52 (.5)	.37 (.48)	.51 (.5)	.54 (.44)	.68 (.4)	.39 (.5)	.52 (.5)	.24 (.36)	.34 (.41)

with similar trends observable in mid and high noise. Based on more observations ($r = 750$ ($r = 1000$) instead of $r = 500$), both GMM and KM classify all inauthentic agent types but $\texttt{L}_{\updownarrow}$ more precautiously in difficult high noise environments. Misclassification in high noise for the main inauthentic agent types improve thusly:

		LOW				MID				HIGH			
		All	$B_\uparrow$	$D_\uparrow$	$L_\uparrow$	All	$B_\uparrow$	$D_\uparrow$	$L_\uparrow$	All	$B_\uparrow$	$D_\uparrow$	$L_\uparrow$
GMM	Authentic	.04 (.04)	—	—	.03 (.05)	.11 (.13)	—	.00 (.008)	.04 (.03)	.5 (.26)	.01 (.02)	.01 (.02)	.09 (.06)
	Inauthentic	.03 (.03)	.13 (.00)	—	—	.04 (.04)	.13 (.00)	—	.03 (.12)	.35 (.17)	.13 (.002)	—	.81 (.11)
KM	Authentic	.2 (.4)	—	—	.002 (.006)	.07 (.17)	.01 (.03)	.00 (.01)	.016 (.016)	.31 (.29)	.03 (.004)	.05 (.07)	.01 (.03)
	Inauthentic	.43 (.28)	.13 (.00)	.05 (.06)	.00 (.001)	.48 (.12)	.13 (.003)	.05 (.06)	.06 (.11)	.56 (.16)	.13 (.01)	.44 (.41)	.97 (.07)

Table 1: GMM and KM mean misclassification and standard deviation ($^{\text{Mean}}_{(\text{SD})}$; '$-$' is short for $^{.00}_{(.00)}$) of authentic (G) and inauthentic (N) agents in populations $\mathcal{P}_{\texttt{All}}$, $\mathcal{P}_{\texttt{B}_\uparrow}$, $\mathcal{P}_{\texttt{D}_\uparrow}$ and $\mathcal{P}_{\texttt{L}_\uparrow}$, for each noise level. E.g., for low noise in $\mathcal{P}_{\texttt{D}_\uparrow}$, GMM perfectly classifies, while KM misclassifies no authentic agents, but 5% inauthentic agents (SD = 6%).

	$B_\uparrow$			$D_\uparrow$			$L_\uparrow$		
r	500	750	1000	500	750	1000	500	750	1000
GMM	.19 (.32)	.09 (.2)	.11 (.24)	.47 (.24)	.29 (.19)	.25 (.16)	.54 (.22)	.36 (.2)	.33 (.17)
KM	.7 (.34)	.5 (.39)	.49 (.38)	.69 (.25)	.46 (.26)	.43 (.28)	.73 (.23)	.53 (.24)	.51 (.25)

However, in the same case, false positive errors increase: authentic agent misclassification worsens from $^{.5}_{(.26)}$ to $^{.67}_{(.24)}$ ($^{.73}_{(.19)}$) for GMM, and from $^{.3}_{(.29)}$ to $^{.48}_{(.3)}$ ($^{.54}_{(.32)}$) for KM. In contrast to GMM, KM demonstrates robustness shortcomings as its classification accuracy notably improves with increased $|\texttt{A}| = 1000$. This difference is pronounced in false positive errors in low (mid) noise levels for $r = 500$: KM improves from mean misclassification $^{.2}_{(.4)}$ to $^{.00}_{(.00)}$ ($^{.07}_{(.17)}$ to $^{.01}_{(.01)}$), while GMM misclassification changes from $^{.04}_{(.04)}$ to $^{.1}_{(.06)}$ ($^{.11}_{(.13)}$ to $^{.1}_{(.07)}$). Neither notably improves false positive errors in the high noise case, but the misclassification exhibits lower SD (KM: $^{.3}_{(.29)}$ to $^{.31}_{(.15)}$, GMM: $^{.5}_{(.26)}$ to $^{.47}_{(.16)}$). Full robustness results can be produced with the Supplementary Material[3].

Third, in small sub-populations $\mathcal{P}_{\texttt{B}_\uparrow}$ and $\mathcal{P}_{\texttt{D}_\uparrow}$, we accurately classify inauthentic agents as such without significant costs of false positive errors: GMM weakly outperforms KM throughout (it does at least as good everywhere, and sometimes strictly better), and strictly outperforms KM in the high noise case (it does strictly better everywhere), cf. Table 1. In $\mathcal{P}_{\texttt{L}_\uparrow}$, both perform well in low and mid noise cases, however, fail to accurately distinguish between A and $\texttt{L}_\uparrow$ in the high noise case, causing large false negative errors. Yet again, $\texttt{L}_\uparrow$ agents do not have a robust effect in watering down MCSs (Figure 1), given their uncoordinated and camouflaging behavior.

4 Jury Selection and Majority Correctness

The GMM and KM classifications of agents as authentic or inauthentic directly provide *jury selection procedures* (JSPs): select the largest jury that includes only agents classified as authentic. This defines the GMM and KM JSPs.

Evaluation Conditions To evaluate the GMM and KM JSPs, we compare the majority correctness scores of the juries they select from $\mathcal{P}_{\texttt{All}}, \mathcal{P}_{\texttt{B}_\uparrow}, \mathcal{P}_{\texttt{D}_\uparrow}$ and $\mathcal{P}_{\texttt{L}_\uparrow}$ with $r = 500$. The low number of authentic agents and rounds result in more diffuse clustering environments and situations in which inauthentic agents have strong negative effects on MCSs (cf. Sec. 2.5).

Expected Juries For each JSP, population, and noise level, we produce 3 expected juries—the average, best and worst cases—based on mean misclassification scores and standard deviations. Let $\mathcal{P} : A \longrightarrow \{\texttt{A}, \texttt{B}_i, \texttt{D}_i, \texttt{L}_i\}_{i\in\{\uparrow,\downarrow,\updownarrow\}}$ be a population. Assume that for each agent type $\texttt{X} \in \{\texttt{A}, \texttt{B}_i, \texttt{D}_i, \texttt{L}_i\}_{i\in\{\uparrow,\downarrow,\updownarrow\}}$, we have a misclassification score $\delta_\texttt{X}$. As a JSP removes agents classified as inauthentic, the selected jury will contain $(1 - \delta_\texttt{A})|\texttt{A}|$ authentic agents and $\delta_\texttt{Y}|\texttt{Y}|$ inauthentic agents, for each inauthentic agent type $\texttt{Y} \in \{\texttt{B}_i, \texttt{D}_i, \texttt{L}_i\}_{i\in\{\uparrow,\downarrow,\updownarrow\}}$. As specific agent identity does not matter for behavior and thus for MCSs, it is not important exactly *which* agents of each type are removed, only the percentage is important. We implemented selecting agents as juries as follows:

For each agent type $\texttt{X} \in \{\texttt{A}, \texttt{B}_i, \texttt{D}_i, \texttt{L}_i\}_{i\in\{\uparrow,\downarrow,\updownarrow\}}$, let $\texttt{X}^{-1}$ denote the agents of that type, i.e., $\mathcal{P}^{-1}(\texttt{X}) \subseteq A$. Enumerate each such $\texttt{X}^{-1}$ so that $\texttt{X}^{-1} = \{x_1, ..., x_{|\texttt{X}^{-1}|}\}$. Let $\delta_\texttt{X}$ represent a misclassification score, and let $\lfloor\cdot\rfloor$ be the floor function which rounds down to closest integer. Given this, we define the agents of type X to *keep* in the jury to be

$$\texttt{X}(\delta_\texttt{X}) = \{x_k \in \texttt{X}^{-1} : k \leq \min(\lfloor|\texttt{X}^{-1}| \cdot \delta_\texttt{X}\rfloor, |\texttt{X}^{-1}|)\}.$$

I.e., $\texttt{X}(\delta_\texttt{X})$ contains the first $(1 - \delta_\texttt{X})$ percent (rounded down to closest integer) of $\texttt{X}^{-1}$. The min operation is needed as $\delta_\texttt{X}$ may exceed 1 if standard deviation is added.

To obtain the average, best and worst case expected juries, we either directly use $\delta_\texttt{X}$ equal to the mean misclassification score $M_\texttt{X}$ of X, or add or deduct two standard deviations. I.e., we use $\delta_\texttt{X} = M_\texttt{X}$ for the average, $\delta_\texttt{X} = \max(0, M_\texttt{X} - 2\sigma_\texttt{X})$ for the best, and $\delta_\texttt{X} = \min(M_\texttt{X} + 2\sigma_\texttt{X}, 1)$ for the worst case jury. The average [best / worst] expected jury of $\mathcal{P}$ is then the set of agents

$$\texttt{G}(1 - \delta_\texttt{G}) \cup \bigcup_{\texttt{Y}\in\{\texttt{A}_i,\texttt{D}_i,\texttt{L}_i\}_{i\in\{\uparrow,\downarrow,\updownarrow\}}} \texttt{Y}(\delta_\texttt{Y})$$

with $\delta_\texttt{X} = M_\texttt{X}$ [$\delta_\texttt{X} = \max(0, M_\texttt{X} - 2\sigma_\texttt{X})$ / $\delta_\texttt{X} = \min(M_\texttt{X} + 2\sigma_\texttt{X}, 1)$] for each $\texttt{X} \in \{\texttt{A}, \texttt{B}_i, \texttt{D}_i, \texttt{L}_i\}_{i\in\{\uparrow,\downarrow,\updownarrow\}}$.

I.e., the average expected jury contains the mean percentage of authentic agents classified as authentic plus the mean percentage of each inauthentic type misclassified as authentic. The best and worst cases are similar, just factoring in standard deviation.

[3] See the public GitHub repository *Coordinated-Inauthentic-Behavior-Likes-ABM-Analysis* at https://github.com/LJ-9/Coordinated-Inauthentic-Behavior-Likes-ABM-Analysis

	FILTER: NONE				FILTER: $p_2{=}p_3{=}1$				FILTER: NONE						FILTER: $p_2{=}p_3{=}1$					
	$\mathtt{B}_\uparrow$	GMM		KM	$\mathtt{B}_\uparrow$	GMM		KM	$\mathtt{D}_\uparrow$	GMM			KM		$\mathtt{D}_\uparrow$	GMM			KM	
	Base	B	A	W	Base	B	A	W	Base	B	A	W	B	A	W	Base	B	A	W	B	A	W
LOW	81.11 (1.99)	–	–	–	74.83 (2.75)	–	–	–	77.43 (2.18)	–	–	–	—	—	—	74.83 (2.75)	–	–	–	—	—	—
MID	87.50 (1.39)	–	–	–	75.41 (3.8)	–	–	–	87.62 (1.6)	–	–	–	—	—	—	75.41 (3.8)	–	–	–	—	—	—
HIGH	97.58 (.73)	–	–	–	77.17 (21.54)	–	–	–	97.59 (.73)	–	–	–	—	99.89 (.14)	97.59 (.73)	77.17 (21.54)	–	–	–	—	99.20 (3.52)	77.17 (21.54)
	$\mathtt{L}_\uparrow$	GMM		KM	$\mathtt{L}_\uparrow$	GMM		KM	All	GMM			KM		All	GMM			KM	
	Base	B	A	W	Base	B	A	W	Base	B	A	W	B	A	W	Base	B	A	W	B	A	W
LOW	74.94 (2.25)	–	–	–	74.93 (2.77)	–	–	–	66.21 (2.18)	–	–	92.41 (1.27)	—	82.19 (1.98)	64.21 (2.27)	74.83 (2.75)	–	–	89.84 (1.81)	—	74.83 (2.75)	74.83 (2.75)
MID	99.98 (.07)	–	–	–	99.99 (.07)	–	–	–	75.12 (2.08)	–	–	–	100.00 (.02)	97.07 (.77)	92.99 (1.1)	75.41 (3.8)	–	–	–	99.99 (.07)	89.54 (2.8)	78.43 (3.72)
HIGH	—	–	–	–	—	–	–	–	98.37 (.45)	–	–	99.90 (.14)	—	—	99.95 (.1)	94.18 (11.34)	–	–	91.02 (12.79)	—	—	94.92 (10.74)

Table 2: Mean MCS and standard deviation ($\frac{\text{Mean}}{(\text{SD})}$; '$-$' is short for $\frac{100.00}{(0.00)}$) for populations $\mathcal{P}_{\mathtt{B}_\uparrow}, \mathcal{P}_{\mathtt{D}_\uparrow}, \mathcal{P}_{\mathtt{L}_\uparrow}$ and $\mathcal{P}_{\mathtt{All}}$ in their baseline form (Base) and in the best (B), average (A) and worst (W) cases for each of the GMM and KM JSPs, for $r = 500$ either unfiltered or filtered so $\mathtt{B}_\uparrow$ and $\mathtt{D}_\uparrow$ are active ($p_2 = p_3 = 1$). Each of the 8 sub-tables (with tinted upper left corner for population subscript) allows i) comparisons of a population's MCS with those of the JSPs' best, average and worst case juries, and ii) comparisons of the MCSs of the two KM and GMM JSPs.

Jury Results Table 2 summarizes GMM and KM JSP's mean majority correctness scores and SD for populations $\mathcal{P}_{\mathtt{B}_\uparrow}, \mathcal{P}_{\mathtt{D}_\uparrow}, \mathcal{P}_{\mathtt{L}_\uparrow}$ and $\mathcal{P}_{\mathtt{All}}$ in key conditions. In $\mathcal{P}_{\mathtt{B}_\uparrow}$, KM and GMM JSPs result in MCSs that clearly show how removing agents classified as inauthentic from the baseline jury suffices to yield perfect MCSs, despite 13% misclassification among inauthentic agents by both KM and GMM (Table 1). Similar observations hold for $\mathcal{P}_{\mathtt{D}_\uparrow}$, where the GMM JSP achieves maximum MCS. The setback in the KM high noise case is explained by difficulties in distinguishing authentic from inauthentic agents: The non-precautious misclassification of inauthentic agents as authentic forecloses the jury to achieve a higher MCS when the inauthentic agents are activated. For GMM and KM JSPs in $\mathcal{P}_{\mathtt{L}_\uparrow}$, we might expect lower MCSs given rather substantial misclassification numbers in the high noise case (Table 1). Yet, we observe perfect MCSs, explained by the non-coordinated way that $\mathtt{L}_\uparrow$ act inauthentically. Hence, difficulties classifying this subgroup for both classifiers are mitigated by its limited effect on lowering MCSs. Note how $\mathcal{P}_{\mathtt{L}_\uparrow}$ is unaffected by the filter in Table 1: $\mathtt{L}_\uparrow$ agents act on their individual beliefs about quality, i.e., they act uncoordinated on their personal property, which cannot be filtered for per voting round without changing the population size.

Assessing JSPs given $\mathcal{P}_{\mathtt{All}}$, we show the MCSs achieved through the GMM method strictly dominate those from KM in low and mid noise cases, both in terms of mean value and SD. Moreover, the GMM JSP strictly outperforms baseline juries in all noise cases when looking at average and best juries. Merely in high noise, worst case, we observe that neither the GMM nor the KM JSP outperforms the baseline jury.

5 Concluding Remarks

Influence or information operations such as coordinated inauthentic behavior (CIB), e.g. performed by attention hacking bots, shape public opinion by elevating or suppressing topics through coordinatedly up- or downvoting social media posts, mimicking authentic behavior to avoid detection, nullifying online voting judgments' reliability. To restore wisdom-of-crowds effects, this paper designed two accurate and feasible *jury selection procedures* (JSPs) that discard agents classified as inauthentic from the voting jury.

Comparing the GMM and KM JSPs, the main difference is accuracy: The GMM JSP detects more inauthentic agents, exhibiting smaller false negative errors and hence more precaution. Both JSPs select juries with vastly increased *majority correctness scores* (MCSs), with preponderantly better scores for the GMM JSP. Overall, the application of either almost fully restores wisdom-of-crowds effects, despite the presence of inauthentic agents. In the low and mid noise cases, inauthentic agents strongly affect the baseline MCSs negatively, but both JSPs successfully eliminate this effect. Only in populations with a high degree of hiding (i.e., high noise, where inauthentic agents act mainly authentically), the JSPs do not significantly increase MCSs. However, in these cases the inauthentic agents also exhibit negligible negative effects on MCSs.

The latter highlights a trade-off for inauthentic attention hacking behavior: attention hackers must balance their accounts' activity to, on the one hand, hide their true identity by acting authentically, and, on the other, act in a coordinated manner to sway the majority vote. We believe this may be exploited in designing attention hack resistant social media vote systems. Employing JSPs means inauthentic actors must hide more often, raising the cost of influence for the attention hackers that handle them. Further, JSPs could be combined with a user reputation system that only publicly displays a user's vote if the user has logged enough (ignored) votes. Beyond raising bot startup costs, this may provide early data for JSPs.

We round off with a discussion of ethical considerations, model assumptions, and data collection.

5.1 Ethical Considerations

Any suppression of information in public fora raises ethical concerns about censorship. The suppression of reactions to

social media posts is no different. Generally, we find that the suppression of coordinated inauthentic behavior as used by attention hackers is defendable, justified by the aim to combat misinformation online. We omit further discussion of this point. However, in applying automated techniques based on classification, there is always a risk that misclassification occurs. If the classification is used for censorship—as is the case here—misclassification may then lead to unrightful censorship.

The JSPs risk unjustified censorship on two points: the unrightful censorship of individuals due to behavioral correlation with inauthentic agents, and the unrightful censorship of groups due to an authentic disagreement with the notion of quality assumed by JSP deployers.

Concerning individuals, then we designed the JSPs with a focus on the two stated desiderata *vox populi* (to minimize false positive errors, i.e., to preserve as many authentic agents as possible) and *precaution* (to minimize false negative errors, i.e., to eliminate as many inauthentic agents as possible). Vox populi implies a desire to not unrightfully censor individuals, but is opposed by precaution: the most precautious model censors all, while the model that preserves most voices censors none. Given our ABM and its parameters, employing *ends-justify-the-means* reasoning, and taking the correct evaluation of posts' quality to be the primary end, we find it worth compromising vox populi over deprioritizing precaution: as illustrated in Figures 1 and 2, deprioritizing precaution quickly threaten the wisdom-of-crowds effect as few inauthentic agents in the jury drastically lower the majority correctness score, while compromising with vox populi by allowing small fractions of authentic agents to be labelled as inauthentic is—with respect to MCS—absorbed by the wisdom of crowds exhibited by even a small jury of only authentic agents.

In the classification, the balance between vox populi and precaution is controlled by the classification threshold. As classification threshold, we precautiously chose that agents should be labelled authentic 4 of 5 times to be classified as authentic. This choice did not cause tremendous collateral damage to vox populi. While we deem especially the GMM JSP a precautious method, it still exhibits low ($< .11$) false positive misclassification errors throughout, except for $\mathcal{P}_{\texttt{All}}$ in high noise. The KM JSP, similarly shows low false positive errors ($< .1$) except for $\mathcal{P}_{\texttt{All}}$ in low and high noise (cf. Table 1). The approach remains flexible to emphasizing vox populi further by lowering the $4/5$ classification threshold.

Concerning group censorship, it is relevant that our approach assumes an agreed-upon notion of *truth about the quality of posts* for which a commonly acknowledged arbiter exists. This is a fundamental premise of our method: if no such notion exists, majority correctness scores loose their meaning and the assumptions of the classifiers are unmet. Such a notion of quality is of paramount importance in relation to fake news, where, arguably, "objective" quality exists, embodied e.g. by the Principles of Journalism. However, the criteria for what constitutes quality may lead to marginalization of groups. E.g., sympathizers of Alex Jones and InfoWars might be marginalized by censorship if quality is equated with adhering to the Principles of Journalism, or sympathizers of the black feminist Combahee River Collective may be marginalized if quality is equated with adhering to ideals of the National Association for the Advancement of Colored People of the 1970s. Therefore, the notion of quality used in applications should be carefully defined, and preferably made open to the public e.g. by inclusion in community standards or terms and conditions of social media platforms.

Due to the risk of unrightful censorship, we would always suggest that users are made aware of censorship decisions that concern them and are given the option to appeal. This, of course, also allows accounts used in IOs to appeal, but appeal adds a non-trivial maintenance cost to e.g. large bot collectives.

5.2 Assumptions of the ABM and Classification

While our contribution hopefully serves as a proof of concept for jury selection procedures as a tool to counter reaction-oriented CIB-based IOs, the simulated environment is not in a one-to-one correspondence with the plethora of environments found on social media platforms. We discuss how modeling choices relate to social media platforms and how assumptions may be relaxed, first concerning the ABM, then the classification.

On social media platforms, it is likely that human users at times vote inauthentically to a low degree that should not be penalized by censorship. Such inauthentic voting violates the ABM's assumptions about authentic agents who vote given only their competence-based beliefs. Our classification results indicate, however, that the authenticity assumption may be relaxed. *Lone wolfs* in the high noise case behave *almost* authentically, and may be interpreted as generally, but not fully, authentic, uncoordinated users. These agents are further—by the GMM JSP—often *misclassified* as authentic in high noise (cf. L sections of Figure 4 and Table 1), but correctly classified for low and mid noise, which indicates that the GMM method may be tuned to tolerate a degree of uncoordinated, inauthentic behavior.

Further, on social media, vote participation is not complete: most users do not react to most posts. For simplicity, we have not included abstaining as an option in the ABM, but all steps including MCS calculation and jury selection would be unaffected. As we return to below, also the classification can accommodate for a less complete vote participation.

Concerning classification, disciplines not directly related to social media applications and misinformation research show how dimensionality reduction and SVD procedures can be applied to empirical data to disclose coordinated voting groups and patterns: US Congress roll call votes have been clustered based on scores similar to the weighted eigenvectors used in this paper (Yang et al. 2020b; Porter et al. 2005; Sirovich 2003; Poole 2000). SVD Scatterplots of votes as suggested by Porter et al. (2005), for instance, provide proxies for party stance. While Yang et al. (2020b) explore roll call vote data only 1-dimensionally, we expand the application and cluster on 2 partial components; both their and our applications can be generalized to more dimensions to increase precision in less exposing vote environments. Moving towards social media applications, this can

become relevant for votes with not only binary but several options from which to choose, such as vote data reflecting Facebook's 6 reactions.

We rely on unsupervised methods that disclose coordination that go unnoticed by supervised methods that take only features of individual accounts into consideration (Khaund et al. 2022; Orabi et al. 2020; Grimme, Assenmacher, and Adam 2018; Cresci et al. 2017). We add a single supervised learning element—logistic regression—to apply labels to agent clusters found by the unsupervised steps. In the logistic regression, we have used that authentic votes correlate with post quality (possibly allowing for noise in observing quality). Other subjective assumptions could be used to steer labeling while producing equally efficient jury selection procedures.

Besides limiting supervision, the input data needs of the GMM and KM jury selection procedures are vastly more feasible than Galeazzi, Rendsvig, and Slavkovik (2019)'s: we rely on 500 observations, where the χ^2 test would require at least 2^{1000} for our population $\mathcal{P}_{\texttt{All}}$. In empirical application, obtaining 500 votes of one user group may still be a challenge. A mitigating factor is that the proposed JSPs can accommodate for missing data, and, for validation, only the inauthentic agents need to be fixed over several voting rounds, while the authentic agents may vary, as these vote independently. Thus, we can lift the assumption that all agents are always presented with, and vote, on every post.

5.3 Empirical Validation and The Release of Reactions Data

Empirical data—in contrast to simulated data—to further validate jury selection procedures remains difficult to obtain (Bliss et al. 2020; Pasquetto, Swire-Thompson et al. 2020; Torres-Lugo et al. forthcoming 2022). Among the platforms that provide APIs for academic purposes, only Twitter releases user-IDs of (public) profiles that have clicked the like-button. However, while Twitter provides generous academic access to historical data for researchers, the platform does not allow to automatically scrape *comprehensive* lists of users that have liked, but only releases the user-IDs of the 100 *most recent* liking users of any single post. Additionally, lists of liking users may be requested at most 75 times per 15 minutes. For small-scale Twitter environments where posts receive few likes, these restrictions may be balanced by using a suitably timed algorithm. However, for large political hashtags like `#MakeAmericaGreatAgain` or `#Brexit` where CIB-based IOs may be feared to be in play, the current data restrictions make it practically impossible to obtain a complete picture of liking behavior.

The proof of concept for JSPs provided in this paper provides a direct use case for reactions data in the fight against online misinformation. The data is necessary to evaluate, tweak and deploy the suggested methods. The paper thus provides a direct argument for a more comprehensive release of and access to reactions data to researchers, e.g. under full anonymization and non-disclosure agreements or via open API access to publicly available data.

Statements and Declarations

The authors have no relevant financial, non-financial nor competing interests to disclose.

Code for data generation and analysis (c.f. Data Availability Statement) is submitted alongside this manuscript.

The authors blind the Author Contribution Statement and Funding Information during peer review and refer to the separate title page for this information.

References

Abokhodair, N.; Yoo, D.; and McDonald, D. W. 2015. Dissecting a social Botnet: Growth, content and influence in Twitter. *CSCW 2015 - Proceedings of the 2015 ACM International Conference on Computer-Supported Cooperative Work and Social Computing*, 839–851.

Aggarwal, A.; and Kumaraguru, P. 2015. What they do in shadows: Twitter underground follower market. In *2015 13th Annual Conference on Privacy, Security and Trust (PST)*, 93–100. IEEE.

Ahmed, F.; and Abulaish, M. 2013. A generic statistical approach for spam detection in Online Social Networks. *Computer Communications*, 36(10-11): 1120–1129.

Angere, S. 2010. Knowledge in a social network. *Synthese*, 167–203.

Beatson, O.; Gibson, R.; Cunill, M. C.; and Elliot, M. 2021. Automation on Twitter: Measuring the Effectiveness of Approaches to Bot Detection. *Social Science Computer Review*, 1–20.

Bebensee, B.; Nazarov, N.; and Zhang, B.-T. 2021. Leveraging node neighborhoods and egograph topology for better bot detection in social graphs. *Social Network Analysis and Mining*, 11(1): 1–14.

Beutel, A.; Xu, W.; Guruswami, V.; Palow, C.; and Faloutsos, C. 2013. Copycatch: Stopping group attacks by spotting lockstep behavior in social networks. In *Proceedings of the 22nd international conference on World Wide Web*, 119–130.

Bhadani, S.; Yamaya, S.; Flammini, A.; Menczer, F.; Ciampaglia, G. L.; and Nyhan, B. 2022. Political audience diversity and news reliability in algorithmic ranking. *Nature Human Behaviour*.

Bliss, N.; Bradley, E.; Garland, J.; Menczer, F.; Ruston, S.; Starbird, K.; and Wiggins, C. 2020. An Agenda for Disinformation Research. Quadrennial paper, CRA Computing Community Consortium (CCC).

Bradshaw, S.; and Howard, P. N. 2017. Troops, Trolls and Troublemakers: A Global Inventory of Organized Social Media Manipulation. *Computational Propaganda Research Project*, 2017(12): 1–37.

Chavoshi, N.; Hamooni, H.; and Mueen, A. 2017. DeBot: Twitter Bot Detection via Warped Correlation. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 817–822.

Chen, Z.; and Subramanian, D. 2018. An Unsupervised Approach to Detect Spam Campaigns that Use Botnets on Twitter.

Condorcet, M. M. d. 1785. *Essai sur l'Application de l'Analyse à la Probabilité des Décisions Rendues à la Pluralité des Voix*. Paris.

Cresci, S.; Di Pietro, R.; Petrocchi, M.; Spognardi, A.; and Tesconi, M. 2017. The Paradigm-Shift of Social Spambots: Evidence, Theories, and Tools for the Arms Race. In *Proceedings of the 26th international conference on world wide web companion*, 963–972.

De Cristofaro, E.; Friedman, A.; Jourjon, G.; Kaafar, M. A.; and Shafiq, M. Z. 2014. Paying for likes? Understanding Facebook Like Fraud Using Honeypots. In *Proceedings of the 2014 Conference on Internet Measurement Conference*, 129–136.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1): 1–22.

Galeazzi, P.; Rendsvig, R. K.; and Slavkovik, M. 2019. Improving Judgment Reliability in Social Networks via Jury Theorems. In Blackburn, P.; Lorini, E.; and Guo, M., eds., *Logic, Rationality, and Interaction (LORI 2019)*, volume 11813 of *Lecture Notes in Computer Science*, 230–243. Springer.

Giglietto, F.; Righetti, N.; Rossi, L.; and Marino, G. 2020a. Coordinated Link Sharing Behavior as a Signal to Surface Sources of Problematic Information on Facebook. *ACM International Conference Proceeding Series*, 85–91.

Giglietto, F.; Righetti, N.; Rossi, L.; and Marino, G. 2020b. It takes a village to manipulate the media: coordinated link sharing behavior during 2018 and 2019 Italian elections. *Information Communication and Society*, 23(6): 867–891.

Goerzen, M.; and Matthews, J. 2019. Black hat trolling, white hat trolling, and hacking the attention landscape. *The Web Conference 2019 – Companion of the World Wide Web Conference, WWW 2019*, 2: 523–528.

Grimme, C.; Assenmacher, D.; and Adam, L. 2018. Changing Perspectives: Is It Sufficient to Detect Social Bots? In Meiselwitz, G., ed., *Social Computing and Social Media. User Experience and Behavior*, 445–461. Cham: Springer International Publishing.

Hastie, T.; Tibshirani, R.; and Friedman, J. 2009. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2nd edition.

Ikram, M.; Onwuzurike, L.; Farooqi, S.; Cristofaro, E. D.; Friedman, A.; Jourjon, G.; Kaafar, M. A.; and Shafiq, M. Z. 2017. Measuring, Characterizing, and Detecting Facebook Like Farms. *ACM Transactions on Privacy and Security (TOPS)*, 20(4): 1–28.

Khaund, T.; Kirdemir, B.; Agarwal, N.; Liu, H.; and Morstatter, F. 2022. Social Bots and Their Coordination During Online Campaigns: A Survey. *IEEE Transactions on Computational Social Systems*, 9(2): 530–545.

Kirn, S. L.; and Hinders, M. K. 2022. Ridge count thresholding to uncover coordinated networks during onset of the Covid-19 pandemic. *Social Network Analysis and Mining*, 12.

Lazer, D. M. J.; Baum, M. A.; Benkler, Y.; Berinsky, A. J.; Greenhill, K. M.; Menczer, F.; Metzger, M. J.; Nyhan, B.; Pennycook, G.; Rothschild, D.; Schudson, M.; Sloman, S. A.; Sunstein, C. R.; Thorson, E. A.; Watts, D. J.; and Zittrain, J. L. 2018. The science of fake news. *Science*, 359(6380): 1094–1096.

Lorenz-Spreen, P.; Lewandowsky, S.; Sunstein, C. R.; and Hertwig, R. 2020. How behavioural sciences can promote truth, autonomy and democratic discourse online. *Nature Human Behaviour*, 4(11): 1102–1109.

Magelinski, T.; Ng, L.; and Carley, K. 2022. Synchronized Action Framework for Detection of Coordination on Social Media. *Journal of Online Trust and Safety*, 1.

Martini, F.; Samula, P.; Keller, T. R.; and Klinger, U. 2021. Bot, or not? Comparing three methods for detecting social bots in five political discourses. *Big Data and Society*, 8.

Metaxas, P. T.; Mustafaraj, E.; Wong, K.; Zeng, L.; O'Keefe, M.; and Finn, S. 2015. What Do Retweets Indicate? Results from User Survey and Meta-Review of Research. *Proceedings of the 9th International Conference on Web and Social Media, ICWSM 2015*, 658–661.

Nizzoli, L.; Tardelli, S.; Avvenuti, M.; Cresci, S.; and Tesconi, M. 2021. Coordinated Behavior on Social Media in 2019 UK General Election. In *Proc. International AAAI Conference on Web and Social Media (ICWSM)*, volume 15, 443–454.

Olsson, E. J. 2013. A Bayesian Simulation Model of Group Deliberation and Polarization. In *Bayesian Argumentation*, 113–133. Springer.

Orabi, M.; Mouheb, D.; Al Aghbari, Z.; and Kamel, I. 2020. Detection of Bots in Social Media: A Systematic Review. *Information Processing and Management*, 57.

Pacheco, D.; Hui, P.-M.; Torres-Lugo, C.; Truong, B. T.; Flammini, A.; and Menczer, F. 2021. Uncovering Coordinated Networks on Social Media: Methods and Case Studies. In *Proc. International AAAI Conference on Web and Social Media (ICWSM)*, volume 15, 455–466.

Pasquetto, I. V.; Swire-Thompson, B.; et al. 2020. Tackling misinformation: What researchers could do with social media data. *HKS Misinformation Review*, 1(8).

Poole, K. T. 2000. Nonparametric Unfolding of Binary Choice Data. *Political Analysis*, 8(3): 211–237.

Porter, M. A.; Mucha, P. J.; Newman, M. E.; and Warmbrand, C. M. 2005. A network analysis of committees in the U.S. House of Representatives. *Proceedings of the National Academy of Sciences of the United States of America*, 102(20): 7057–7062.

Ratkiewicz, J.; Meiss, M.; Conover, M.; Gonçalves, B.; Flammini, A.; and Menczer, F. 2011. Detecting and Tracking Political Abuse in Social Media. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 297.

Ryczko, K.; Domurad, A.; Buhagiar, N.; and Tamblyn, I. 2017. Hashkat: large-scale simulations of online social networks. *Social Network Analysis and Mining*, 7(1): 1–13.

Samper-Escalante, L. D.; Loyola-González, O.; Monroy, R.; and Medina-Pérez, M. A. 2021. Bot Datasets on Twitter: Analysis and Challenges. *Applied Sciences*, 11(9): 4105.

Schoch, D.; Keller, F. B.; Stier, S.; and Yang, J. H. 2022. Coordination patterns reveal online political astroturfing across the world. *Scientific Reports*, 12.

Scrucca, L.; Fop, M.; Murphy, T. B.; and Raftery, A. E. 2016. mclust 5: Clustering, Classification and Density Estimation using Gaussian Finite Mixture Models. *The R Journal*, 8(1): 289–317.

Shao, C.; Ciampaglia, G. L.; Varol, O.; Yang, K. C.; Flammini, A.; and Menczer, F. 2018. The spread of low-credibility content by social bots. *Nature Communications*, 9(1).

Sirovich, L. 2003. A pattern analysis of the second Rehnquist U.S. Supreme Court. *Proceedings of the National Academy of Sciences of the United States of America*, 100(13): 7432–7437.

Subrahmanian, V. S.; Azaria, A.; Durst, S.; Kagan, V.; Galstyan, A.; Lerman, K.; Zhu, L.; Ferrara, E.; Flammini, A.; and Menczer, F. 2016. The DARPA Twitter Bot Challenge. *Computer*, 49(6): 38–46.

Suchacka, G. 2019. Improving Clustering Of Web Bot And Human Sessions By Applying Principal Component Analysis. *Communications of the ECMS*, 33(1).

Suchacka, G.; and Iwański, J. 2020. Identifying legitimate Web users and bots with different traffic profiles – an Information Bottleneck approach. *Knowledge-Based Systems*, 197: 1–18.

Takacs, R.; and McCulloh, I. 2019. Dormant bots in social media: Twitter and the 2018 U.S. senate election. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2019*, 796–800.

Tibshirani, R.; Walther, G.; and Hastie, T. 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society Series B*, 63(Part 2): 411–423.

Torres-Lugo, C.; Pote, M.; Nwala, A.; and Menczer, F. forthcoming 2022. Manipulating Twitter Through Deletions. In *Proceedings of the 16th International AAAI Conference on Web and Social Media (ICWSM)*.

Weber, D.; and Neumann, F. 2021. Amplifying influence through coordinated behaviour in social networks. *Social Network Analysis and Mining*, 11(1): 1–42.

Wilensky, U. 1999. NetLogo. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.

Yang, K. C.; Varol, O.; Davis, C. A.; Ferrara, E.; Flammini, A.; and Menczer, F. 2019. Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, 1: 48–61.

Yang, K.-C.; Varol, O.; Hui, P.-M.; and Menczer, F. 2020a. Scalable and Generalizable Social Bot Detection through Data Selection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 1096–1103.

Yang, V. C.; Abrams, D. M.; Kernell, G.; and Motter, A. E. 2020b. Why Are U.S. Parties So Polarized? A "Satisficing" Dynamical Model. *SIAM Review*, 62(3): 646–657.

Paper II

Towards Detecting Inauthentic Coordination in Twitter Likes Data

Laura Jahn and Rasmus K. Rendsvig

Center for Information and Bubble Studies, Department of Communication, University of Copenhagen

ABSTRACT

Social media feeds typically favor posts according to user engagement. The most ubiquitous type of engagement (and the type we study) is *likes*. Users customarily take engagement metrics such as likes as a neutral proxy for quality and authority. This incentivizes like manipulation to influence public opinion through *coordinated inauthentic behavior* (CIB). CIB targeted at likes is largely unstudied as collecting suitable data about users' liking behavior is non-trivial. This paper contributes a scripted algorithm to collect suitable liking data from Twitter and a collected 30 day dataset of liking data from the Danish political Twittersphere `#dkpol`, over which we analyze the script's performance. Using only the binary matrix of users and the tweets they liked, we identify large clusters of perfectly correlated users, and discuss our findings in relation to CIB.

KEYWORDS

Novel digital data, political opinion dynamics, social media, coordinated inauthentic behavior, bot detection

1 INTRODUCTION

Algorithmically curated social media feeds favor posts according to user engagement. The most ubiquitous type of engagement (and the type we study) is *likes* [49]. A post—a tweet, a shared news article, a video, a meme, etc.—may be highlighted e.g. by being placed highly on users' news feeds. Users customarily take engagement metrics such as likes as a neutral proxy for quality and authority [33, 49]. This incentivizes *influence operations* to misrepresent, mislead or manipulate opinion dynamics online [41]. Such media manipulation tactics have been labeled *coordinated inauthentic behavior* (CIB) [21, 22, 25, 38, 43, 44]. Influence operations and CIB may thus shape public opinion and political discourse through *attention hacking*, the act of exploiting platforms' content sorting algorithms to highlight certain information items to users. This highlights the societal need to address CIB-caused misrepresentation of political views and the spread of harmful low-quality content and misinformation in the online public sphere [33].

To effectively push narratives on social media, influence operations resort to *coordinated* groups of accounts rather than individual accounts [34, 35]. This has, for example, led to the establishment of a marketplace for vendor-purchased engagement [30, 49] and metric inflation through coordinated social bots. The behavior dictated by an influence operation is labeled *inauthentic* as it may not reflect the personal beliefs of the instructed user accounts, as these accounts may be run by algorithmic amplifiers such as automated bots or humans according to a supplied protocol [18].

CIB targeted at one-click reactions such as likes is largely unstudied as collecting data about users' liking behavior around a specific political discourse is non-trivial due to the lack of access to platform data for researchers or severe API rate restrictions that prevent collecting comprehensive datasets. The first main contribution of this paper is a script to collect comprehensive data on liking users from Twitter. The second main contribution is a dataset collected with the script. The dataset contains a month-long survey of liking user behavior from the Danish political Twittersphere, collected through the hashtag `#dkpol` ("DenmarK POLitics"). Under this hashtag, citizens, organizations, politicians and journalists from across the political spectrum air, discuss and orientate themselves about current debates in Danish politics. It is *the* centralized, place-to-be source of information on the debates of the day. The hashtag thus seems a likely candidate for inauthentic coordination, if one seeks to increase the Danish public sphere's attention on some topic. We use the dataset first to evaluate the effectiveness of the script, and second as basis for a case study of liking users behavior with the aim to determine if the simple liking data has sufficient structure to serve as an entry point for the detection of CIB. We argue that it does.

Using a running survey approach, the script retrieves IDs of the most recent liking users of tweets satisfy a specified text query (e.g. a keyword or hashtag of a chosen political debate), timing retrievals by taking into account Twitter set rate limits of the public v2 API for Academic Research Access. The script can retrieve far more comprehensive sets of liking user IDs than are available through the default public and commercial tools of the Twitter APIs and Decahose stream. To the best of our knowledge, the resulting data is the first to contain comprehensive collections of user-IDs of liking users. The dataset thus advances the specialized field of studying one-click reaction-based CIB.

The script's point of departure for data collection is the survey of an online *discourse* around a *domain* (e.g. a hashtag) instead of a survey of a preselected group of users. Hence, data collection does not require any prior knowledge about potentially coordinated users nor does subsequent data analysis necessarily require the retrieval of additional account data. When identifying coordination of likes given such concise data, one immediately grasps firstly which specific tweet(s) a potential influence operation is targeted at, and secondly which users are involved in the metric inflation (this is in contrast to existing methods for collecting retweeting user IDs, cf. Sec. 1.1 below). If desired, additional account information may then be rehydrated via public APIs. The focus of the collected data and following applications is thus rather on identifying the *effects* of CIB inflating specific tweets. These effects may be more robust to changes in the evolution of algorithmic amplifiers, social bots and cyborgs, that with varying degrees of automation increasingly emulate authentic users. Our data and applications are not dependent on individual account features nor time-synchronous actions but only on the like behavior towards an observed tweet.

We analyse the dataset in a case study of `#dkpol`, mainly to illustrate that the liking behavior data has sufficient structure to serve as a point of entry for detection of CIB. Preprocessing the data points into a simple binary and sparse tweet/like matrix suffices to detect like-coordinated accounts without relying on textual, temporal, nor training data (see Sec. 3.2), a topic that has previously gone unstudied. We undertake two simple analyses: First, we group users by the toughest clustering criteria of complete equality of their like profiles. Under this very strict criteria, we identify several large perfectly correlated groups, including likes we purchased from online vendors. Notably, we detect the vendor-purchased CIB and more perfectly correlated groups of users despite the users not being particularly active (one like suffices), so without any requirement that they have liked aggressively. Second, we show that these groups can be visualized using the first two dimensions in a dimensionality-reduced space using the first two eigenvectors of a Singular Value Decomposition of the tweet/like matrix.

Given a lack of ground truth, we cannot be sure the perfectly coordinated clusters we detect (other than the vendor-purchased groups) are artifacts of CIB. We do believe that the natural correlation is unlikely enough that the groupings raise red flags, warranting further inspection, out of scope of this case study. Our methods may thus serve as pre-studies for bot detection and the application of fact checkers [35].

We make our resources available to the research community, including the raw datapoints complemented with timestamp data (tweet text must be rehydrated per Twitter data sharing policies) and pre-processed user-like data matrices, the scripts used for data collection, for data preprocessing, for evaluation of the completeness of a collected dataset, and for clustering and visualization. Data and scripts are available on Harvard Dataverse [32] and the data collection script is additionally available at the public GitHub repository *Get-Twitter-Likers-Data* [31].

1.1 Related Work

Social media users have a plethora of available action types [36], many of which may be used in coordinated fashions. E.g., users may coordinate using a specific hashtag, posting a specific URL, tweet, image or mention, or coordinate replies, shares or reactions to existing content. As coordination is not visible when inspecting accounts in isolation, research on CIB has turned to study the collective behavior of groups, with similarities between users serving as a proxy for coordination. Studies have analysed similarities between users posting similar *content* [6, 7, 35, 44], users having similar *friends and followers* [41], and having similarly *timed activities* (e.g., [15, 16, 23–25, 36, 50]). Few studies have looked directly at coordination in one-click reactions such as liking.

Liking is a one-click engagement where users may select one option from a short pre-defined list as their 'reaction' to a post, with users' choices typically summed and presented as a quantified metric beneath the item. Reactions include perhaps most famously Facebook's original 'Like', the hearts/likes on Instagram, TikTok and Twitter, and Reddit's up- and downvotes. Sharing and retweeting may also be taken as a one-click reaction on any of these platforms.

Importantly, these reactions inform the platforms' algorithmic content sorting, thus steering users' attention. With attention metrics such as likes being widely used as a proxy for quality and authority, manipulating like counts becomes incentivized for the sake of increased exposure, influence, and financial gain [49]. High engagement counts may be perceived as a trust signal about the content [40] and as a positive crowd reaction aiding content to broadcast and to trend [20]. Once trending, high engagement counts in likes and shares make users more likely to engage with low-credibility content instead of fact-checking questionable posts [9]. Scholars have stressed that to fight disinformation campaigns, it is less effective to look at the pushed content (e.g., hashtags, URLs, memes, etc.) and more effective to look at the coordinated content pushing *behaviors* [35].

Related work on coordinated retweeting. To push stories online, retweeting and inflation of the retweet metric attracts manipulation. Several recent papers look at retweeting as a coordination dimension.

Dutta et al. [20] investigate non-synchronized, collusive retweeters ($n < 1,500$) involved with *blackmarket* services.

Such collusive retweeters re-share the tweets of other blackmarket customers to earn credits. The authors use a human annotated dataset and supervised machine learning methods leveraging features such as, e.g., user activity or social network characteristics to distinguish between *customers* and *genuine retweeters*, later extended to detect *paying customers* [19]. Building on these works, Arora et al. [8] analyze user representations to improve the performance of detecting blackmarket customers while Chetan et al. [13] develop an unsupervised approach to detect collusive blackmarket retweeters leveraging, for example, the merit of tweets and timing of retweets analyzed through a bipartite tweet-user graph.

Schoch et al. [46] study time-synchronous co-retweeting (and co-tweeting) as a trace of coordination to detect astroturfing campaigns given a dataset released by Twitter consisting of tweets by accounts that Twitter classified as being involved in hidden information campaigns. The authors filtered the data and only looked at campaigns with more than $50,000$ tweets and users that tweeted at least 10 times in the observation period. They do so by analyzing timing and centralization of coordination. The approach rests on the assumption that it seems implausible that repeated co-retweeting and co-tweeting happens without centralized coordination (e.g., one actor controlling multiple accounts) in a small time window of 1 minute up to 8 hours. Increasing the temporal window beyond that yields higher false positive rates in flagging astroturfing accounts. The study builds a co-(re)tweeting graph by drawing an edge between two users that (re)tweet the same post within a minute, but only if this can be observed more than 10 times. While the authors rightfully claim that co-retweeters and co-tweeters can be rehydrated from a Twitter dataset, it remains a necessity that one has selected a list of users prior to dataset construction. Some knowledge over the presence of astroturfers is hence necessary a priori: Their approach presupposes to have a list of (suspicious) users instead of embarking on detection given an observable effect.

Similarly concerned with co-retweeting, Graham et al. [26] searches for evidence of bots in > 25 million retweets of > 2.5 million tweets, collected over the course of 10 days, containing COVID-related hashtags. The authors create a user-user '*bot-like*' *co-retweet network* of $> 5,000$ Twitter accounts that *frequently* co-retweeet the same tweets within a time window as small as 1 second, followed by manual inspection of the connected components.

Pacheco et al. [44] take a high number of overlapping retweets (co-retweeting) as a coordination trace and construct a bipartite network between retweeting accounts and retweeted messages, filtering for accounts that logged at least 10 retweets. The authors represent users with TD-IDF weighted vectors containing the retweeted IDs. The weighting discounts the contributions of popular tweets. The projected co-retweet graph is then established via the cosine similarity between the account vectors. Using a hard threshold, they only keep the most suspicious 0.5% edges leaving them with a coordinated set of users. The analysis is conducted on an anonymized dataset from DARPA SocialSim containing identified Russian disinformation campaigns, collected from Twitter using English and Arabic keywords. Messages that were identified as coordinated are no longer publicly available.

Interested in how well network communities hide from coordination detection, Weber et al. [50] study retweets using a latent coordination network. When members of a group retweet each others' posts, detection of the involved accounts becomes easy, as the accounts are connected via an edge. The larger the detected coordinated community, the greater the likelihood that members would retweet other members. Notably, the authors find that large groupings of accounts in the Twitter curated dataset, believed to be involved in influence operations, hide well with low internal retweet ratios, and that also official political accounts seem to refrain from being involved in self-retweeting.

Adopting the network approach [41, 44, 50], Tardelli et al. [48] model *evolving* coordinated retweet communities. This work explores that users may belong to different coordinated groups at different points in time. Using the Jaccard similarity measure, the authors compare influx and outflux into and out of communities at each time step. The resulting temporal networks and dynamic community detection identifies many coordinated communities and highlights the relevance of temporal nuances of coordination.

Instead of leveraging graph-based techniques, Mazza et al. [39] only require the timestamps of retweets and the retweeted tweets for each account, and not, e.g. full user timelines. Their work investigates temporal and synchronous retweeting patterns. The collected data spans short of 10 million Italian retweets from > 1.4 million distinct users collected over the course of two weeks. The collected data is filtered for human-like retweet activity between 2 and 50 times per day and excludes fully automated, benign retweet bots with high retweeting activity, resulting in a dataset with $63,762$ distinct users. Manual annotation of a subset of the data ($1,000$ users) serves as a ground truth. Given a user and their retweet history, the authors first visualize different temporal retweet patterns by plotting the timestamp of the original tweet against the timestamp of the retweet in a scatterplot. With a granularity of seconds, the authors compress timestamp data into per user time series vectors containing time information if the user retweeted a given tweet at a given time, and 0 otherwise. The resulting series remains sparse as users usually only retweet once every few minutes.

To reduce sparsity, the data is then compressed employing a sequence compression scheme. Using automatic unsupervised feature extraction, the work exploits that synchronous and coordinated users will be grouped densely together in the feature space, in contrast to heterogeneous human behavior. The authors apply dimensionality reduction techniques and deep neural networks and eventually hierarchical and density-based clustering. Users that are clustered and not treated as noise (i.e., not clustered) are labeled as bots. Users clustered together are then thought of as bots acting in a coordinated and synchronous fashion.

Related work on coordinated liking. Despite likes being a commonly adopted and an easily manipulatable mechanism, research on CIB more narrowly targeted at likes is quite scarce:

Border-lining relevancy are studies on purchased likes not of posts, but of *pages* and *followers* on Facebook and Twitter [5, 10, 17, 30]. Studying page like or follower farms [5, 17], these works develop supervised classifiers using demographic, explicitly temporal, and social characteristics [5, 30]. Notably, Ikram et al. [30] find their bot classifier has difficulty detecting like farms that mimick regular like-spreading over longer timespans, i.e. deliver likes slowly, without high temporal synchronization, and with lower like counts per account.

Beutel et al. [10] study coordinated and time-synchronized attempts to inflate likes on Facebook pages. Their unsupervised method, developed with data from inside Facebook, detects ill-gotten likes from groups of users that coordinate to like the same page around the same time, leveraging temporal data explicitly. The authors follow a graph-based approach, draw a bipartite graph between users and pages noting down the time at which each edge was created. They then apply co-clustering looking for users liking the same pages at around the same time. Since [10]'s approach depends on timing and is designed to detect synchronous likes in a "single burst of time", [30] find that [10]'s approach, too, suffers large false positive errors in detecting liking accounts that mimick regular users and deliver likes more slowly.

While the Facebook like button is the same whether it regards a page or a post, page likes inflation differs in the mechanism from post like inflation. Liking a page on Facebook entails "following" the account, subscribing to new account posts. Thus, this kind of coordinated metric inflation may not catapult a single *post* to the top of an algorithmically curated newsfeed but creates the illusion of a popular *account*.

Directly about reactions to posts is Torres-Lugo et al.'s [49] study of metric inflation through strategic deletions on Twitter. They analyze coordination in repetitive *(un)liking* on *deleted* tweets in influence operations that seek to bypass daily anti-flooding tweeting limits. From a collection point of view, looking at unlikes is a smart move, as this data is in fact available to purchase from Twitter. Alas, the approach is inapplicable to tweets that remain online, such as those central to CIB-based influence operations that push narratives through political astroturfing [46].

Also in the related field of bot detection has the detection of bots designed to engage through reactions gone unstudied, perhaps due to data restrictions. For a systematic review of the bot detection literature, see [43].

1.2 Empirical Problems

Group-based detection methods are promising "in the arms race against the novel social spambot" [14]. Yet empirical research meets challenges in this domain. The following three problems highlight the need for a feasible data collection script and findable datasets for researchers to develop and test methods to address CIB targeted at reactions online.

Time-sensitivity. First, empirical social media studies of coordinated online accounts remain problematic to replicate and reproduce due to time-sensitivity of the relevant data [37]. Attempts to collect the same data twice are likely to fail, as traces of coordination may be altered or deleted after an influence operation was concluded. While e.g. Twitter grants generous academic research access to historic tweets through their API, accounts involved in CIB may evade detection e.g. by changing handle, so they are no longer retrievable in their original appearance [49]. The shortcomings in data reproducibility make CIB/bot detection frameworks difficult to compare, as these typically require live data access [37].

Data availability. Second, data availability limits research [11, 25, 37, 45]. Large scale studies may simply be impossible due to data access restrictions [11, 37, 45]. Specifically data concerning users' reactions is very difficult for researchers to obtain: none of the currently existing datasets include it,[1] Twitter's transparency reports do not include information of liking or retweeting users [4], and neither Meta, Twitter nor Reddit supply this data in necessary scope [11, 45].

Among the platforms with APIs for academic purposes, only Twitter releases user-IDs of (public) profiles that have liked or retweeted a given tweet. Twitter does not give direct access to *comprehensive* lists of such IDs, but only releases the user-IDs of the 100 *most recent* liking/retweeting users of any single post. Further restrictive, at most 75 such lists may be requested per 15 minutes. For some Twitter environments, these restrictions may be balanced by using a suitably timed algorithm, cf. below. For huge political hashtags like `#MakeAmericaGreatAgain` or `#Brexit` where CIB-based influence operations may most be feared to be in play, current

[1] See e.g. Indiana University's Bot Repository, a resourceful, centralized repository of annotated datasets of Twitter social bots [1].

data restrictions make it practically impossible to obtain a complete picture of liking and retweeting behavior. Twitter's commercial Decahose API stream lists 100% of liking user-IDs, but only of a *random* 10% sample of all tweets, making a targeted analysis of a specific political discourse impossible [2].

Ground truth. Third, there is an issue with lacking ground truth as researchers have no access to the empirical truth about accounts engaged in coordinated inauthentic behavior. Qualified guesses can be made based on suspicious similarities in behavior or profile features, but *de facto*, it remains unknown whether two users' actions are authentically correlated or inauthentically coordinated, or how many (partially) automated accounts exist in a total population [10, 12, 36, 37].

Specifically for reaction-based CIB, it seems infeasible to create a labeled dataset that even *approximates* the ground truth: labeling accounts individually e.g. via crowd-sourcing or the well-established bot classifier *Botometer* will likely fail as single accounts will often seem inconspicuous [39]. *Botometer*'s feature-based approach considers accounts one at a time and does therefore not pick up on group anomalies based on suspicious similarity [52, 53]. Especially when it comes to coordinated liking behavior, Botometer's feature "favourites_count" (the number of likes a user has delivered) predicts less bot-like behavior, the higher the count is [53], thus undermining the attempt to identify coordinated liking. For purposes of studying liking behavior in concert at a collective level [27, 36, 52, 53], data availability restrictions make collective labeling impossible.

Instead of relying on (an approximation of) a ground truth, groups of users may be labeled as suspicious, e.g. in terms of graph structure [10, 36], contextually validated via manual inspection and individual confession by the original poster [27, 39], through NLP of the content promoted [12, 41], or compared to behavior of experimental vendor-purchased metric inflation [5, 30], as we do in the case study in Sec. 3.

2 DATA COLLECTION

To collect a comprehensive dataset needed to identify coordinated inauthentic liking behavior, we scripted an algorithm that makes effective use of the data limits set by Twitter. Here, we aim to give an intuition of the implementation and workings of the data collection algorithm. We then present its pseudocode.

2.1 Data Collection Script: Intuition

In short, the script surveys Twitter for tweets falling under a *textual query* during a live *observation period* (e.g. 30 days). During the observation period, with a fixed time interval p (e.g. every 5 min.), the script executes a *pull*. Each pull loop contains four steps:

(1) It logs tweets posted since the last pull that satisfy the query, and their current number of likes (*like count*).
(2) It updates the logged like count of previously logged tweets. Only tweets that are recent enough are tracked in this way (e.g., posted within the last 48 hours).
(3) For each logged tweet, it compares the tweet's new like count to its like count *at the last pull where its liking users were requested* (0 if the liking users have never been requested). Call the numerical difference between these two like counts the tweet's *delta*.
(4) It requests the 100 most recent liking users of the top n tweets with the highest delta above a set threshold (e.g., has minimum 25 new likes).

At the end of the observation period and once every logged tweet is no longer tracked, the liking users of all logged tweets is requested a final time (in timed batches). The script also allows pulling retweeting users in the pull loop. The logic is the same. Pulling liking and retweeting users draws on separate pools of request resources.

To raise the chance of a complete data set—one that has not missed any liking users—it is preferable to set the tweet track time as long as possible, the pull interval p as short as possible, and the number of top n tweets checked to its maximum. Alas, this will often lead to request request shortage.

Twitter's request limits entail that the parameters of the script have to be balanced carefully. For example, a query with 10.000 new tweets a day, each tweet tracked for 24 hours at 5 minute intervals uses 8.640.000 tweet-requests over a 30 day period. Twitter allows 10.000.000. The same parameters but a query with 12.000 tweets/24h uses 10.3680.000 tweet-requests. Hence, the pull interval and the track time must balanced with respect to the query volume. Additionally, the pull interval (p) and the number of requests used per pull (n) must also be balanced with respect to the liking frequency *and* the activity under the query. Given the 75 likers-requests available per 15 minutes, there are two extremes (if one plays it safe; see further below): a short pull interval of $p = \frac{15\,\text{min.}}{75} =$ 12 seconds, each pull getting the likers the top $n = 1$ tweet and a long pull interval of $p = 15$ min., each pull getting the likers of the top $n = 75$ tweets. The former lowers the risk of missing out on likers during rapid hours, but burns through many more tweet-requests per hour, counting against the 10.000.000 limit. Long pull intervals, on the other hand, raise the risk of missing put on liking users.

The script allows extending the Twitter request resources by the inclusion of multiple bearer tokens. If working in a team where multiple members have Academic Research access to Twitter, all their bearer tokens may be included. The script then cycles through them, using one per pull loop.

Finally, the pull loop is written in Python 3, and is run through a shell script that resumes it from the point of failure

in case of Twitter connection errors, e.g. caused by an overuse of requests or a network disruptions. This means the script allows *not* playing safe with request resources, most notably with the pull interval p and the number of likers-requests used per pull, n. Playing it unsafe allows for some flexibility. One may e.g. set $p = 3$ min. and $n = 30$ if one trusts that the actual distribution of tweets and likes is unlikely to break the request limit but wants to readily sacrifice more than the safe amount of requests in case of an activity surge.

2.2 Script: Details and Pseudocode

The algorithm is parameterized by three time periods. First, $observationtime$ is the length of data collection (e.g. 24 hours, or 1 month), without restriction: with properly set parameters, one can span 1 month, after which request limits reset, making it extendable. The $observationtime$ starts at a point in time ($startpoint$). Second, $pullinterval$ defines a sleep period between the conclusion of one pull and the initiation of the next. The shorter it is, the finer the temporal resolution and the lower the risk of missing any liking users, but also the higher the request usage. Third, $tracktime$ specifies how long a tweet is monitored for new likes and retweets after it is posted (e.g., each tweet is tracked for 1 hour, or 48 hours). To collect full data for all tweets posted in $observationtime$, the total scraping time amounts to $observationtime + tracktime$.

The algorithm is split into two steps, Alg. 1 and Alg. 2, with Alg. 1 undertaking most of the work, and collects data from Twitter using the Academic Research access API (ARA). ARA provides significant data scraping resources to researchers that are, however, subject to rate limits and request caps specified by Twitter in an advance to manage server requests. Among others, but most notably, requesting liking users from ARA always returns the most recent 100 liking users of a given tweet in question. Furthermore, this request can only be made $req.rate.lim = 75$ times per 15 minutes. As tweets routinely get more that 100 likes in total, a dataset that contains an as complete as possible set of identifiable liking users must live-log liking users runningly.

This is accomplished in Alg. 1, which runs from $startpoint$ to $endpoint := startpoint + observationtime + tracktime$. At $endpoint$, Alg. 2 runs. It completes a final harvest of liking users by requesting the 100 most recent liking users from all logged tweets. This is especially relevant for those tweets with low like counts de-prioritized in Alg. 1.

Between $startpoint$ and $endpoint$, Alg. 1 performs a *pull* every $pullinterval$ seconds. A pull at time t outputs a dataframe L_t of tweet-IDs and their liking users. Further, it continuously outputs dataframes T_t that contain tweets, like count, retweet count, and meta-data including time of origin, text, posting user, language etc. Alg. 1 and Alg. 2 require the input parameters in Table 1.

$keyword$	Keyword(s) or hashtag(s). e.g. `#dkpol`.
$token$, $\|token\|$	ARA Twitter Authentication Bearer Token, number of tokens. More than 1 is possible. More raise request limits.
$startpoint$	Date and time to start data collection. Must be in the past. E.g now, minus 10 seconds.
$observationtime$	Observation period. E.g., 1 hour, or 60 days.
$tracktime$	How long to track each tweet for new likes. E.g., 48 hours,. Longer periods use up rate limit more quickly.
$pullinterval$	Sleep interval between pull completion and next pull. E.g. 300 seconds. Shorter interval use up rate limit more quickly.
$min.delta$	How many new likes must a tweet have gotten before we request its liking users? To play safe, satisfy $min.delta + min.delta \leq req.rate.lim$.
$top.n$	Determines from how many tweets to request likers per pull. To play safe, satisfy $top - n \leq rlim \cdot \frac{pullinterval}{15 \cdot 60 sec} \cdot \|token\|$.
$min.likes$	Minimum like (retweet) count of tweets to be considered for final harvest. E.g. 1 or 10.
$req.rate.lim$	Twitter rate limit: 75 requests per 15 min. for liking and retweeting users each.

Table 1: Input parameters for Algorithms 1 and 2.

3 CASE STUDY: DATA COLLECTION AND ANALYSIS OF THE DANISH TWITTERSPHERE

To study both the performance of the contributed script and the usefulness of the resulting dataset to address CIB, we analyze a case study of the Danish political Twittersphere.

3.1 Dataset: Parameters, Completeness and Descriptive Statistics

The dataset used in this paper was collected using the described script, without manual intervention during its runtime. The text query was "#dkpol -is:retweet", meaning that the script sought tweets falling under `#dkpol`, excluding retweets. Two bearer tokens were used, doubling the request resources available. The observation period started the afternoon of May 25th, 2022 and was 30 days long. Tweets

Algorithm 1 Main loop of algorithm to retrieve liking users from Twitter

Input: $keyword, token, startpoint, observationtime, pullinterval, tracktime, min.delta, top.n$
Output: T_t, L_t for $t \in pullpoints := \{t \leq endpoint : t = startpoint + k \cdot pullinterval \text{ for a } k \in \mathbb{N}\}$
if exists file *log* **then**
  load *log* // to resume from error
else
  $log \leftarrow \emptyset$ // start empty dataframe with columns $tweet, like.count, like.count.last$ to track tweets' like count now and last their likers were pulled
end if
while true **do**
  if $sys.time = t$ for some $t \in pullpoints$ // if now is a time to pull **then**
    $T_t, L_t \leftarrow \emptyset$ // start empty dataframes for tweets and their metadata, and for liking users
    $start = \begin{cases} startpoint \textbf{ if } t - tracktime < startpoint \\ t - tracktime \textbf{ else} \end{cases}$
    $end = \begin{cases} t \textbf{ if } t < startpoint + observationtime \\ startpoint + observationtime \textbf{ else} \end{cases}$
    $T_t \leftarrow$ get_tweets($keyword, start, end, token$) // pull tweets (incl. $like.count$) under $keyword$ posted between $start$ and end, auth. with $token$
    save T_t // save to file with timestamp
    $log \leftarrow$ update_log_1(log, T_t) // For $tweet$ in T_t: if $tweet$ is not in log, append it with $like.count$ from T_t and $like.count.last = 0$; else update $tweet$'s $like.count$ in log to its $like.count$ in T_t
    $candidates \leftarrow$ find_candidates($log, min.delta$) // return list of all $tweet$ in log for which $delta := like.count - like.count.last \geq min.delta$
    sort $candidates$ by $delta$ in descending order // introduce retrieval priority.
    $top \leftarrow candidates[0 : top\_n - 1]$ // restrict to top_n tweets with highest $delta$.
    for $tweet$ in top **do**
      $L_t \leftarrow$ get_likers($tweet, token$) // pull 100 most recent likers
      $log \leftarrow$ update_log_2(log, T_t) // update $tweet$'s $like.count.last$ in log to its $like.count$ in T_t
    end for
    save L_t // save to file with timestamp
    save log // save to file
  else
    break
  end if
end while

Algorithm 2 Final harvest to retrieve liking users from Twitter

Input: $token, min.likers, req.rate.lim, \boldsymbol{T} = \{T_t :$ output dataframe of Alg. 1$\}$
Output: L_{final}
$L_{final} \leftarrow \emptyset$ // Start empty dataframe columns $tweet, last.likers$
$all \leftarrow$ all_tweets($\boldsymbol{T}$) // Load and concatenate all T_t. For duplicates, keep tweets with highest $like.count$. Subset columns of all to $tweet$ and $like.count$, rows to those with $like.count \geq min.likers$
$counter \leftarrow 0$
for $tweet$ in all **do**
  $L_{final} \leftarrow$ get_likers($tweet, token$) // Pull 100 most recent likers, append to L_{final}
  $counter = counter + 1$
  if $counter \geq req.rate.limit \cdot |token|$ **then**
    $counter = 0$
    sleep for 15 minutes // Reset request limits
  end if
end for
save L_{final} // Save to external file

were tracked for 48 hours, as prior tests had shown that liking activity on almost all tweets under `#dkpol` stops before 48 hours after posting. The interval between pulls was 5 minutes, and each pull requests the liking users of the top $top.n = 36$ tweets with $min.delta = 3$.

Following the observation period, we requested the last 100 most recent liking users of all tweets that had at least 10 likes. We used this limit to strongly diminish the amount of tweets in the final check, with the justification that that so little total liking activity would most likely not be hurtful coordinated inauthentic behavior. In total, the script collected 47,714 liking user IDs for 13,243 tweets. While this case study focuses on liking behavior, the published dataset contains retweeting user IDs as well.

To assess completeness of the dataset, we compare the number of collected likers (for those tweets subject to final harvest collection) to the maximum like count a tweet has logged during the tracktime for each tweet, i.e. 48 hours, see Fig 1.

First, we see both positive *and negative* deviation numbers. Positive deviation is expected: the script cannot collect more than 100 liking users per tweet per pull interval. For testing, we used vendor-purchased likes on clearly-marked test-tweets. We highlight the targeted tweets in Fig 1. As we purchased some batches of more than 100 likes and these were placed almost simultaneously (some vendors place likes more slowly), we miss out on collecting them. To detect CIB,

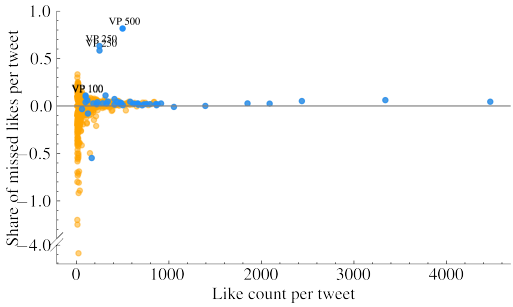


Figure 1: Missed likes per tweet, as share of its maximal like count, arranged by like count in ascending order. Dots represent tweets. Labels "VP n" are on tweets for which we vendor-purchased n likes. Blue marks the tweets of the 50 largest bins of perfectly correlated likes (cf. Sec 3.2).

this is not necessarily a problem: temporal detection methods leveraging time-synchronous user behavior to detect coordination can easily identify such behavior. Negative deviation indicates that the script has collected more liking users than the *like count* suggests. This happens when likes are retracted, the liking profiles are deleted,[2] or a tweet attracted likes post tracktime, which we collected in the final harvest.

Second, we find that for high engagement tweets, the script performs well and collects most of the liking users. In contrast, for very low engagement tweets, the script is more prone to miss out on more than 10% of users. This is due to the algorithm prioritizing tweets that get traction by allocating requests to collect the growing sets of likers.

Third, and to complement the plots in Fig. 1, for 39.98% of 6702 tweets, the script collects exactly as many liking users as the like count suggests. For 93.7% of the 6702 tweets, the script collects numbers of likers that fall within 10 of the like count. If considering negative deviation only, in 96.6% of 6702 tweets, the script deviates negatively 10% or less. If considering positive deviation only, in 97.06% of 6702 tweets, the script deviates positively 10% or less. I.e., in 97% of cases, the script seemingly collects 90%+ of liking users.

3.2 Analysis: Perfect Correlation

In this case study, we make use of very simple user data: a binary matrix containing a row for each tweet and column for each user, each cell marked 1 if the user liked the tweet, else 0. Again, the dataset contains temporal data as well, but we ignore it here, as we are mainly interested in seeking patterns in like behavior alone.

Assume we have observed n tweets. Let $Likers_k$, $k \leq n$, be the set of users observed to have liked tweet k, so $Likers = \cup_{k \leq n} Likers_k$ is the set of all observed liking users. With $m = |Likers|$, we then compress our data to a binary $n \times m$ matrix with entry values in $\{0, 1\}$, each row representing a tweet, each column a user. With this matrix called $\mathbf{L}$, the entry $\mathbf{L}_{i,j} = 1$ if user i has liked tweet j, and 0 else. Henceforth, we hence identify user i with the row $\mathbf{L}_{*,i}$ that contains their like profile. In this case study, $\mathbf{L}$ is of dimension $13,243 \times 47,714$.

We seek to group users as exhibiting coordinated liking behavior if their like profiles are sufficiently similar, according to some measure. Existing work routinely projects bipartite data structures (which $\mathbf{L}$ is) onto a user-user similarity graph using a distance or similarity metric (e.g., [41, 44]) or develops algorithms to detect dense subgraphs to identify anomalous groups of nodes (e.g., [29, 47]). Here, we apply the strictest measure: we group two users if, and only if, they exhibit *exactly the same like behavior*. This is equivalent to grouping users that have cosine similarity 1, Jaccard similarity 1, or Hamming distance 0.

We apply this strictest measure as behavior labeled as coordinated will also be labeled as coordinated using any less discriminating measure. The approach thus is precautious with regard to labeling coordinated users. The method is not designed to identifying all coordinated inauthentic behavior in likes. There may very well be nuances and less than perfectly correlated inauthentic behavior. To answer whether a collection of tweet likes exhibits first signs of CIB, we propose the method only as valid for positive answers: if this strongly discriminatory methods finds such signs, then methods with lower bars for coordination should, too. If the method does not find such signs, we would deem it fallacious to take this as evidence that no CIB occurred.

To group users with identical like profiles, we worst case have to pair-wise compare all users, i.e. undertake $\frac{47,713^2 - 47,714}{2}$ comparisons. To avoid as many of these comparisons as possible, we sort users into bins: we initiate a list with one bin containing the first user. For every later user, we compare them with one user from each bin in the list of bins, checking larger bins first, and stopping to place them in the first bin that provides a perfect match. If no such bin exists, we add a new bin for the user in the end of the list. We find only 25,806 bins.

49.9% of users are sorted into bins of size 1. Filtering for bins of at least size 50 (as smaller bins are negligible in impact for CIB), we find 50 bins with 13,018 out of 47,714 users. Put differently, 27.28% of users are in a group with at least 49 others that share the exact same like behavior across all 13,243 tweets. These 27.28% like most often only 1 tweet,

[2]These are both actions genuine users may take, but are also often observed with vendor-purchased metric inflation [49].

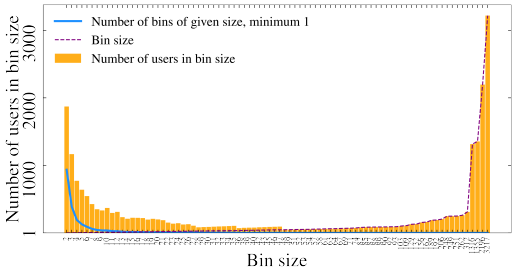


Figure 2: Bins with at least two users, the number of users in bins of each size, and the number of bins of each size. E.g., the left-most bar shows there are $\sim$ 2000 users (yellow bar) distributed over $\sim$ 1000 bins (solid blue line) of size 2 (dotted purple line), while the right-most shows there are $\sim$ 3217 users distributed over 1 bin of size 3217. The number of bins of size n drops to 1 at $n = 48$.

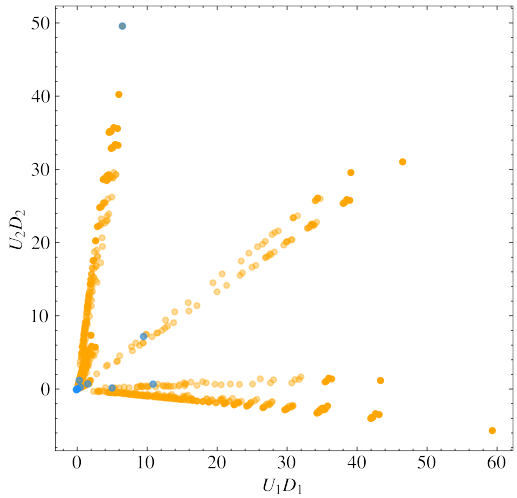


Figure 3: Scatterplot of $\mathbf{U}_q\mathbf{D}_q$. Top 50 perfectly correlated bins of users overlap perfectly with one another in clusters colored in blue. Bins of vendor-purchased likers are among the bottom left groups of clusters.

sometimes 2. In the largest bin, 3,217 users are perfectly correlated liking the same tweet.

Collected in their own bins, we find the users behind the likes we purchased from online vendors. We refer to Fig. 2 for an overview of the magnitude of bins.

We find several bins of users with perfectly identical liking behaviors unrelated to our purchases. We cannot conclude from *correlation* to *coordination* to state these bins contain users engaged in coordinated inauthentic behavior. We do find the larger bins suspicious and in warrant of further analysis, cf. the discussion in Sec. 4.

We find the larger bins suspicious as we find it unlikely that the correlation has arisen without coordination. E.g., rate the probability of each bin as being non-coordinated using the following charitable assumptions (charitable to favor the odds of large bins): Assume that the probability that any two users share the exact same like profile without being coordinated is $c = .95$. For simplicity and charity, ignore that this probability attaches to every unordered pair of users in a bin, and let the probability that a bin B of size $|B|$ occurred without coordination be $P(B) = c^{|B|-1}$, i.e., the probability that $|B|-1$ users pairwise and independently correlated with the same user i from B. This probability drops drastically with the growth of B:

$\lvert B\rvert =$	2	10	50	60	75	100	200
$P(B) =$	.95	.63	.08	.05	.02	.006	$3.69 \cdot 10^{-5}$

These (fictitious) probabilities do not mean that it is unlikely that e.g. 60 users liked the same tweet—but that it is unlikely that they all liked or did not like *all the same tweets*. Even under charitable conditions, bins larger than 60 quickly seem highly unlikely. We further discuss the implications of our results in Sec. 4.

Singular Value Decomposition. To visualize and locate the identified bins among all users, we turn to plotting the data in a dimensionality-reduced space: With dimensionality reduction, user behavior often exhibits a clustered structure, for example, separating bots and humans in labeled bot datasets [42, 53], disclosing synchronous clusters of retweeters [23] (later used in baseline experiments by [8, 13, 20]), revealing generally correlated groups such as polarized groups of users [51] among users writing Twitter Birdwatch notes, or coordinated clusters of agents as in [33] given computer-simulated data.

We calculate the singular value decomposition (SVD) $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ of the $m \times m$ sample correlation matrix $\mathbf{X}$ of the data in matrix $\mathbf{L}$. We consider the first $q = 2$ dimensions' eigenvectors, i.e., the first two columns of the $n \times p$ orthogonal matrix $\mathbf{U}$ where $n = p$, weighted with the corresponding eigenvalue collected in the diagonal $p \times p$ matrix $\mathbf{D}$ [28]. We plot the scatterplot of $\mathbf{U}_q\mathbf{D}_q$ in Figure 3. In the plot, each dot represents a liking user. While we color-coded the users placed in the largest 50 bins, they may also be discerned through their darker shade that stems from many dots perfectly overlapping one another. The SVD and the scatterplot

thus picks up on correlation and the vendor-purchased metric inflation. As an alternative route, note that clustering on these first two eigenvectors (e.g. using a Gaussian Mixture Model as done in [33]) picks up on the inauthentically coordinated users we know of, too.

4 CONCLUDING REMARKS: DISCUSSION & ETHICAL CONSIDERATIONS

Data collection discussion. The script we have presented here is designed to collect the IDs of liking (and/or retweeting) users of tweets that satisfy a selected textual query. As such, the script takes a *domain first* perspective on data collection, rather than a *user first* perspective as most other work designed to investigate coordinated inauthentic behavior.

The dataset presented in this paper is collected around the domain of the Danish political Twittersphere, found under `#dkpol`. For this domain, using the parameters described and two bearer tokens, the script had a reasonably low rate of missing liking users, and misses more than 10% of liking users in only 3% of cases when run continuously for 30 days. Such a targeted dataset cannot be obtained directly through any of Twitter's data access options.

In an international context, `#dkpol` is a small domain. With the same parameters and number of bearer tokens, the script would indubitably fare less well on much larger domains. For larger domains with more intense liking activity, it would be interesting to study the script's performance with more bearer tokens and far more aggressive pull parameters, such as much lower *pullinterval*. As data retrieval from Twitter is not instantaneous (especially when it comes to updating the like count of a large batch of tweets), we suspect that a satisfactory data collection will involve multiple machines running the script in parallel, each tracking a subset of tweets assigned to them (e.g. using tweet ID *modulo* k for k machines).

Another, and favorable, option for obtaining the data on one-click reactions would be if Twitter or other social media platforms made this data available to the research community. We hope that the case study in this paper—where even a crude and strict analysis raises red flags for CIB—may be used as an argument that one-click reaction data is relevant in the study of coordinated inauthentic behavior and thus in the arms race against online misinformation to ultimately put pressure on the social media industry to release data.

Analysis discussion. In our case study, the controlled CIB through vendor-purchased likes is grouped into distinct bins that we can match to our tweets. The coordination here is achieved through weak ties in our bipartite graph structure $\mathbf{L}$. We complement, for example, Weber et al.'s [50] approach focused on coordination through strong ties. As [50] acknowledges as an open issue and we show, coordination may take place along weak ties. With our like-based approach, we provide first steps towards a measure to detect such. In contrast to existing work (e.g. [46]), the present like-based approach does *not* need to filter the data for strongly tied communities, highly influential users and superspreaders, or very active or users that, e.g. like a minimum number of times within a short period. Without filtering, we are able to group users with such behaviors together.

Our analysis made use of vendor-purchased likes. Purchasing engagement metric inflations violates Twitter's platform manipulation and spam policy [3], which defines "platform manipulation as using Twitter to engage in bulk, aggressive, or deceptive activity that misleads others and/or disrupts their experience.". We created two Twitter accounts that in the name of the research center with which the authors are affiliated ('CIBS1' (@CIBS110) and 'CIBS2' (@CIBS22)) posted 6 tweets with text '*Research test tweet n/6. Apologies for spamming #dkpol.*' for $n = 1, ..., 6$. We inflated the like count for these 6 tweets. We acknowledge that the coordinated inflation of these tweets might have disrupted the experience of Twitter users. To the best of our assessment, the amplification of these tweets does not comprise *harmful* coordinated activity nor was it deceptive or commercially-motivated, but declared a research motivation. Ethically, we thus believe that the benefits of studying coordinated inauthentic behavior outweigh the minimal disruptions we have caused to Twitter users by violating Twitter's manipulation and spam policy.

Unrelated to our purchases, we further find and visualize several large groups of users with perfectly correlated, identical liking behaviors—similarly achieved through weak ties. We have no ground truth about whether the suspected accounts beyond our test are naturally correlated and not inauthentically coordinated, yet we believe that natural correlation is unlikely enough that such groupings are red flags for CIB, and warrant further inspection, out of scope of this case study. Our methods may thus serve as pre-studies for bot detection and the application of fact checkers [35]. Further, the dataset and explorative case study may serve as a point of departure for future research to explore the correlation structures among liking users and the development of novel detection methods.

Censorship. Any flagging of behavior in public fora raises ethical concerns about censorship. The classification of reactions such as likes and retweets to tweets is no different. Generally, we find that the flagging of coordinated behavior used by inauthentic attention hackers is defendable, justified by the aim to combat misinformation online. We omit further

discussion of this point. However, in applying automated techniques, there is always a risk of misclassification. If a technique is used for censorship, this may lead to unrightful labeling. The methods for initial exploration proposed here may then risk unjustified labeling users due to behavioral correlation with strongly coordinated groups of users. We strongly recommend that the methods here are taken as a first step towards fact-checking content and users and not as a final verdict about specific individual users.

Data collection approval. Approval of data collection and processing of personal data in the research project was granted by the faculty secretariat of the university of Copenhagen. The approval emphasizes that the processing of personal data in the project is in accordance with the rules of the European General Data Protection Regulation, Regulation 2016/679 on the protection of natural persons with regard to the processing of personal data. That the study would be undertaken was made public on the authors' university websites.

Datasets and code availability. Dataset and code are made available for the research community [32], hosted on the archival repository Harvard Dataverse that provides a Document Object Identifier (DOI) for better findability. To comply with the Twitter terms, access to the data on Harvard Dataverse is granted when researchers actively agree to the Twitter Terms of Service, Privacy Policy and Developer Policy. The data collection code is also available on the public GitHub repository *Get-Twtter-Likers-Data* [31].

REFERENCES

[1] Bot Repository. https://botometer.osome.iu.edu/bot-repository/datasets.html. Accessed: 2022-01-15.

[2] Decahose API. https://developer.twitter.com/en/docs/twitter-api/enterprise/decahose-api/overview/streaming-likes. Accessed: 2022-09-10.

[3] Platform manipulationa and spam policy. https://help.twitter.com/en/rules-and-policies/platform-manipulation. Accessed: 2022-03-01.

[4] Twitter Moderation Research Consortium. https://transparency.twitter.com/en/reports/information-operations.html. Accessed: 2022-07-05.

[5] A. Aggarwal and P. Kumaraguru. What they do in shadows: Twitter underground follower market. In *2015 13th Annual Conference on Privacy, Security and Trust (PST)*, pages 93–100. IEEE, 2015.

[6] F. Ahmed and M. Abulaish. A generic statistical approach for spam detection in Online Social Networks. *Computer Communications*, 36(10-11):1120–1129, 2013.

[7] S. Al-Khateeb and N. Agarwal. Understanding strategic information manoeuvres in network media to advance cyber operations: A case study analysing pro-russian separatists'cyber information operations in crimean water crisis. *Journal on Baltic Security*, 2(1), 2016.

[8] U. Arora, H. S. Dutta, B. Joshi, A. Chetan, and T. Chakraborty. Analyzing and detecting collusive users involved in blackmarket retweeting activities. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–24, 2020.

[9] M. Avram, N. Micallef, S. Patil, and F. Menczer. Exposure to social engagement metrics increases vulnerability to misinformation. *The Harvard Kennedy School Misinformation Review*, 1(5), July 2020.

[10] A. Beutel, W. Xu, V. Guruswami, C. Palow, and C. Faloutsos. Copycatch: Stopping group attacks by spotting lockstep behavior in social networks. In *Proceedings of the 22nd international conference on World Wide Web*, pages 119–130, 2013.

[11] N. Bliss, E. Bradley, J. Garland, F. Menczer, S. Ruston, K. Starbird, and C. Wiggins. An Agenda for Disinformation Research. Quadrennial paper, CRA Computing Community Consortium (CCC), 2020.

[12] N. Chavoshi, H. Hamooni, and A. Mueen. DeBot: Twitter bot detection via warped correlation. *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 817–822, 2017.

[13] A. Chetan, B. Joshi, H. S. Dutta, and T. Chakraborty. Corerank: Ranking to detect users involved in blackmarket-based collusive retweeting activities. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 330–338, 2019.

[14] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the 26th international conference on world wide web companion*, pages 963–972, 2017.

[15] S. Cresci, R. D. Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi. Dna-inspired online behavioral modeling and its application to spambot detection. *IEEE Intelligent Systems*, 31:58–64, 9 2016.

[16] S. Cresci, R. D. Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi. Social fingerprinting: Detection of spambot groups through dna-inspired behavioral modeling. *IEEE Transactions on Dependable and Secure Computing*, 15:561–576, 7 2018.

[17] E. De Cristofaro, A. Friedman, G. Jourjon, M. A. Kaafar, and M. Z. Shafiq. Paying for likes? Understanding Facebook Like Fraud Using Honeypots. In *Proceedings of the 2014 Conference on Internet Measurement Conference*, pages 129–136, 2014.

[18] Z. Duan, J. Li, J. Lukito, K.-C. Yang, F. Chen, D. V. Shah, and S. Yang. Algorithmic agents in the hybrid media system: Social bots, selective amplification, and partisan news about covid-19. *Human Communication Research*, 2022.

[19] H. S. Dutta and T. Chakraborty. Blackmarket-driven collusion among retweeters-analysis, detection, and characterization. *IEEE Transactions on Information Forensics and Security*, 15:1935–1944, 2020.

[20] H. S. Dutta, A. Chetan, B. Joshi, and T. Chakraborty. Retweet us, we will retweet you: Spotting collusive retweeters involved in blackmarket services. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 242–249. IEEE, 2018.

[21] E. Ferrara. Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday*, 22(8), 2017.

[22] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini. The Rise of Social Bots. *Commun. ACM*, 59(7):96–104, 2016.

[23] M. Giatsoglou, D. Chatzakou, N. Shah, A. Beutel, C. Faloutsos, and A. Vakali. Nd-sync: Detecting synchronized fraud activities. In *Advances in Knowledge Discovery and Data Mining: 19th Pacific-Asia Conference, PAKDD 2015, Ho Chi Minh City, Vietnam, May 19-22, 2015, Proceedings, Part II 19*, pages 201–214. Springer, 2015.

[24] F. Giglietto, N. Righetti, L. Rossi, and G. Marino. Coordinated Link Sharing Behavior as a Signal to Surface Sources of Problematic Information on Facebook. *ACM International Conference Proceeding Series*, pages 85–91, 2020.

[25] F. Giglietto, N. Righetti, L. Rossi, and G. Marino. It takes a village to manipulate the media: coordinated link sharing behavior during 2018 and 2019 Italian elections. *Information Communication and Society*, 23(6):867–891, 2020.

[26] T. Graham, A. Bruns, G. Zhu, and R. Campbell. Like a virus, 2020.

[27] C. Grimme, D. Assenmacher, and L. Adam. Changing Perspectives: Is It Sufficient to Detect Social Bots? In G. Meiselwitz, editor, *Social Computing and Social Media. User Experience and Behavior*, pages 445–461, Cham, 2018. Springer International Publishing.

[28] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer, 2nd edition, 2009.

[29] B. Hooi, K. Shin, H. Lamba, and C. Faloutsos. Telltail: Fast scoring and detection of dense subgraphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4150–4157, 2020.

[30] M. Ikram, L. Onwuzurike, S. Farooqi, E. D. Cristofaro, A. Friedman, G. Jourjon, M. A. Kaafar, and M. Z. Shafiq. Measuring, characterizing, and detecting Facebook like farms. *ACM Transactions on Privacy and Security (TOPS)*, 20(4):1–28, 2017.

[31] L. Jahn and R. K. Rendsvig. Get-twitter-likers-data. https://github.com/humanplayer2/get-twitter-likers-data/, 2022.

[32] L. Jahn and R. K. Rendsvig. Twitter User Reactions Data (Liking and Retweeting Users). https://dataverse.harvard.edu/dataverse/twitter-likers, 2023. https://doi.org/10.7910/DVN/WRUNZD.

[33] L. Jahn, R. K. Rendsvig, and J. Stærk-Østergaard. Detecting Coordinated Inauthentic Behavior in Likes on Social Media: Proof of Concept. *Under Review*, 2022.

[34] T. Khaund, B. Kirdemir, N. Agarwal, H. Liu, and F. Morstatter. Social Bots and Their Coordination During Online Campaigns: A Survey. *IEEE Transactions on Computational Social Systems*, 9(2):530–545, 2022.

[35] S. L. Kirn and M. K. Hinders. Ridge count thresholding to uncover coordinated networks during onset of the covid-19 pandemic. *Social Network Analysis and Mining*, 12, 12 2022.

[36] T. Magelinski, L. Ng, and K. Carley. Synchronized Action Framework for Detection of Coordination on Social Media. *Journal of Online Trust and Safety*, 1, 2 2022.

[37] F. Martini, P. Samula, T. R. Keller, and U. Klinger. Bot, or not? Comparing three methods for detecting social bots in five political discourses. *Big Data and Society*, 8, 2021.

[38] J. Matthews and M. Goerzen. Black hat trolling, white hat trolling, and hacking the attention landscape. *The Web Conference 2019 – Companion of the World Wide Web Conference, WWW 2019*, 2:523–528, 2019.

[39] M. Mazza, S. Cresci, M. Avvenuti, W. Quattrociocchi, and M. Tesconi. Rtbust: Exploiting Temporal Patterns for Botnet Detection on Twitter. In *Proceedings of the 10th ACM conference on web science*, pages 183–192, 2019.

[40] P. T. Metaxas, E. Mustafaraj, K. Wong, L. Zeng, M. O'Keefe, and S. Finn. What Do Retweets Indicate? Results from User Survey and Meta-Review of Research. *Proceedings of the 9th International Conference on Web and Social Media, ICWSM 2015*, pages 658–661, 2015.

[41] L. Nizzoli, S. Tardelli, M. Avvenuti, S. Cresci, and M. Tesconi. Coordinated Behavior on Social Media in 2019 UK General Election. In *Proc. International AAAI Conference on Web and Social Media (ICWSM)*, volume 15, pages 443–454, 2021.

[42] A. C. Nwala, A. Flammini, and F. Menczer. A general language for modeling social media account behavior. Preprint, 2022.

[43] M. Orabi, D. Mouheb, Z. Al Aghbari, and I. Kamel. Detection of Bots in Social Media: A Systematic Review. *Information Processing and Management*, 57, 2020.

[44] D. Pacheco, P.-M. Hui, C. Torres-Lugo, B. T. Truong, A. Flammini, and F. Menczer. Uncovering Coordinated Networks on Social Media: Methods and Case Studies. In *Proc. International AAAI Conference on Web and Social Media (ICWSM)*, volume 15, pages 455–466, 2021.

[45] I. V. Pasquetto, B. Swire-Thompson, et al. Tackling misinformation: What researchers could do with social media data. *HKS Misinformation Review*, 1(8), 2020.

[46] D. Schoch, F. B. Keller, S. Stier, and J. H. Yang. Coordination patterns reveal online political astroturfing across the world. *Scientific Reports*, 12, 12 2022.

[47] K. Shin, B. Hooi, J. Kim, and C. Faloutsos. Densealert: Incremental dense-subtensor detection in tensor streams. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1057–1066, 2017.

[48] Tardelli, Serena and Nizzoli, Leonardo and Tesconi, Maurizio and Conti, Mauro and Nakov, Preslav and Martino, Giovanni Da San and Cresci, Stefano. Temporal dynamics of coordinated online behavior: Stability, archetypes, and influence. 2023.

[49] C. Torres-Lugo, M. Pote, A. Nwala, and F. Menczer. Manipulating Twitter Through Deletions. In *Proceedings of the 16th International AAAI Conference on Web and Social Media (ICWSM)*, 2022.

[50] D. Weber and F. Neumann. Amplifying influence through coordinated behaviour in social networks. *Social Network Analysis and Mining*, 11(1):1–42, 2021.

[51] S. Wojcik, S. Hilgard, N. Judd, D. Mocanu, S. Ragain, M. Hunzaker, K. Coleman, and J. Baxter. Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation. 2022.

[52] K. C. Yang, O. Varol, C. A. Davis, E. Ferrara, A. Flammini, and F. Menczer. Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, 1:48–61, 1 2019.

[53] K.-C. Yang, O. Varol, P.-M. Hui, and F. Menczer. Scalable and generalizable social bot detection through data selection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1096–1103, 2020.

Paper III

Friction Interventions to Curb the Spread of Misinformation on Social Media

Laura Jahn[1], Rasmus K. Rendsvig[1], Alessandro Flammini[2], Filippo Menczer[2], and Vincent F. Hendricks[1]

[1]Center for Information and Bubble Studies, Department of Communication, University of Copenhagen, Denmark
[2]Observatory on Social Media, Indiana University, Bloomington, USA

Abstract

Social media communication platforms have enabled the spread of information at unprecedented speeds and scales, and with it the proliferation of high-engagement, low-quality content. A suggestion is that *friction*—behavioral design measures that make the sharing of content more cumbersome—might be a way to raise the quality of what is spread online. Here, we study the effects of friction prompts with and without quality-recognition learning components. Experiments from an agent-based model suggest that friction alone decreases the number of posts without improving their quality. On the other hand, a small amount of friction combined with learning increases the average quality of posts significantly. Based on this preliminary evidence, we propose a friction intervention with a learning component about the platform's community standards, to be tested via a field experiment. The proposed intervention would have minimal effects on engagement and may easily be deployed at scale as it does not require labeling of content or detection of bad actors.

1 Introduction

With the advent of Web 2.0, the spread of misinformation online has been recognized as a global, societal threat to democracy, eroding trust in mainstream news sources, authorities, experts, and other socio-political institutions [1, 2, 3, 4]. Social media communication platforms have enabled the sharing of information at unprecedented speeds and scales, and with it the proliferation of not only misinformation, but also other low-quality and harmful content such as hate speech, cyberbullying, and malware [2, 5]. The focal question of this paper is whether adding a bit of *friction* to the sharing process might mitigate the spread of low-quality and malicious content. Large social media platforms amplify the so-called attention economy, where abundant information competes for scarce attention [1, 6, 7]. Through this competition, one would hope for accurate information to emerge from the interactions among many users by combining *independent* opinions according to the *wisdom of crowds* [8]. Alas, scholars have demonstrated that engaging yet false content gets shared more and travels faster [9].

Socio-cognitive biases and algorithmic sorting both contribute to the spread of high-engagement content over high-quality content. Strategies such as accepting new information

if it comes from multiple sources [10] or through posts that have been shared many times fail because the aggregate opinions to which we are exposed online are not necessarily independent [11]. This is boosted by confirmation bias [12], a disconnect between what users deem accurate and what they deem shareable [13, 14], and automatic habits to share the most engaging content [15]. The illusory effect [16] further increases vulnerability to misinformation in social media by increasing the perceived truth value of low-quality content through repetition [17, 18].

Algorithmic biases in the content sorting algorithms of social media platforms also prioritize high engagement [19], increasing the exposure of low-quality content in user news feeds [20]. One-click reactions, such as "Like" or "Share," are a driving mechanism behind algorithmic sorting since they are easy to use and quick to influence the popularity of posts. Users can select a reaction to a post from a short list, with their aggregate choices typically presented as a popularity metric beneath the post. These reactions steer user attention. For example, a high retweet count is likely to be perceived as a crowd-sourced trust signal [8, 21], possibly contributing to the content's virality [22] irrespective of its quality [23].

One corollary of these dynamics is an incentive for influence operations based on inauthentic behaviors, such as coordinated liking [24, 25, 26, 27, 28, 29, 30, 31, 32].

Scholars have called for ways to promote the Internet's potential to strengthen rather than diminish democratic virtues and public debate [33] and to leverage the economics of information for protection rather than for misguidance [34]. A relatively recent idea is to improve the quality of what is shared online by introducing *friction* on social media. The hope is to curb the spread of harmful content and misinformation by making it more difficult to share or like content online [1, 35, 34].

In the context of online interactions, *behavioral friction*, in general, denotes "any unnecessary retardation of a process that delays the user accomplishing a desired action" [36]. The more friction, the lower the chances that the user will complete an action. While reducing friction is generally deemed desirable in user interface design, some protective friction, like CAPTCHAs, may be useful [36, 37, 38].

Friction added to otherwise one-click sharing and liking will make the spread of both harmful and benign content more cumbersome and time-consuming. Friction is thought to prompt a more deliberate approach to sharing or liking content. Examples include exposing users to a contextual label [37], impeding the completion of an action with a prompt asking the user to reflect [13], exacting micro-payments, or requiring users to spend mental resources through micro-exams such as quizzes and puzzles [3, 5]. Such friction strategies promise to deliver socio-political benefits by supporting cognitive autonomy while increasing the cognitive burden of sharing low-quality content [38].

In this paper, we explore ways in which friction may positively affect quality in a social media environment. We study the effects of friction prompts with and without quality-recognition learning components through an agent-based model (ABM). Our experiments suggest that friction alone decreases the number of posts without improving their quality. On the other hand, a small amount of friction combined with learning increases the average quality of posts significantly. Inspired by this preliminary evidence, we propose a friction intervention where learning is leveraged through quizzes about a platform's community standards. We map the key ingredients of a field experiment to test the idea. Lawmakers could create policies to facilitate larger-scale studies and incentivize platforms to test scalable friction strategies.

2 Related Work

Borrowing terminology from Tomalin [36], the type of friction of interest here is *non-elective* for users—users cannot control exposure. This characteristic is also present in *sludges* and *dark patterns.* Sludges generally refer to excessive friction, bad almost by definition and with a clear negative valence (e.g., bureaucratic form-filling). Dark patterns coerce, steer, or deceive people into making unintended and potentially harmful choices [39]. The types of friction we consider in this paper aim to obtain social benefits. They are easily distinguishable from sludges and dark patterns as they are *overt*, neither *deceptive* nor *accidental*, but *intended*, *protective*, and *non-commercial* [36]. Friction strategies that share these characteristics may be *impeding* or *distracting.* For example, friction can impede action by letting users complete an action only after a micro-exam is passed. Certain nudges provide non-impeding friction; deliberation-promoting nudges, for example, are distracting but leave all options available to the user [39].

Recent research suggests that non-elective, overt, intended, protective friction is a promising tool to boost the accuracy and quality of information shared online. Adding as little friction as having users pause to think before sharing may prevent misinformation proliferation on social media: in a set of online experiments, participants who were asked to explain why a headline was true or false were less likely to share false information compared to control participants [40]. In an effort to nudge users to consciously reflect on tweet content, Bhuiyan et al. [41] developed a browser extension that introduced a distracting emphasis on high-quality content and greyed out posts from low-quality sources. This raised the accuracy of tweet credibility assessments.

Pennycook et al. [14, 13] see potential in reminding users of accuracy. They prompted experiment participants to rate the accuracy of a news headline before scrolling through an artificial social media news feed. Participants subsequently shared higher-quality content than a control group. The authors suggest to translate their findings into attention-based interventions subtly reminding users of accuracy to slow down the sharing of low-quality content online. Similarly, checking for accuracy assisted the fight against the illusory effect. This was demonstrated in a set of experiments that prompted participants to behave like fact checkers by asking them for initial truth ratings at first exposure [42].

Priming critical thinking made users less prone to trusting, liking, and sharing fake news about climate change on Facebook [43]. Reminding users of critical thinking was accomplished through considerations about news evaluation guidelines, using questions to help identify fake news as articulated in the Facebook Help Center (e.g., "Does the information in the post seem believable?").

Time pressure has further been shown to negatively influence the ability to distinguish true and false headlines [44]. Friction—as a means to reduce time pressure—may actively improve the discrimination of accurate and false information.

Lastly, friction has been deemed beneficial when applied to mass-sharing. Model-based work by Jackson et al. [45] shows that caps on depth (how many times messages can be forwarded) or breadth (the number of others to whom messages can be forwarded) improve the ratio of true to false messages, assuming messages mutate at every instance of re-sharing (deliberately or inadvertently).

Most interventions by social media platforms to date do not impose restrictions on sharing [46]. They tend to use redirection (suggesting content from authoritative sources such as the WHO during the COVID-19 pandemic) and content labeling (exposing users to additional context), thus preserving user choice and autonomy. Examples of friction interventions by various platforms include:

- Twitter has implemented protective, non-elective friction that distracts or impedes. The platform has introduced caps on automated tweeting [34] and uses a distracting label to pause users about to share state-affiliated media URLs [47]. With limited success, they tested replacing retweets with "quote tweets," requiring users to comment before they could share a post [48]. Finally, Twitter conducted a promising randomized controlled trial to curb offensive behavior, where users were asked to review replies in which harmful language was detected [49].

- Facebook has established policies to provide context labels from fact checkers. The platform reduces the distribution of, and engagement with, misinformation from repeated offenders by reducing the reach and visibility of their posts. While this intervention leads to a decrease in engagement with the offender in the short term, it can be compensated by an increase in the offender's posts and followers. Furthermore, the limitation in reach can be reversed by deleting flagged posts [50].

- WhatsApp has taken first steps to counter the virality of misinformation by limiting the forwarding of messages to at most five contacts simultaneously [51].

- Instagram has introduced a distractive but non-impeding anti-bullying label that prompts users to pause by asking them "Are you sure you want to post this?" to curb abuse on the platform [52].

3 Proof of Concept: Agent-Based Model on Friction

To study the impact of friction, we add mechanisms for friction and learning to *SimSoM*, a minimal open-source agent-based model of information sharing in social media [35, 53]. The code for the augmented model and its analysis is available on the public GitHub repository *Friction-Social-Media-Model*,[1] making all results reproducible.

In the augmented ABM, *posts* are interpreted as pieces of information, such as images, links, hashtags, or phrases. These may be created or shared by *agents*, and appear on the *news feeds* of agents. Each agent's news feed consists of a bounded number of posts, all shared by agents they *follow*. The bounded news feed models limited individual attention, which gives rise to heavy-tailed distributions of post popularity and lifetime consistent with empirical data [54].

The ABM runs in discrete time. At each time step, some agents are activated. Each chooses to either introduce a new post into the network, or share an existing post from its news feed. The new or re-shared post then appears on the news feeds of the agent's followers. A time step is interpreted as a social media session where only a sample of all users are online simultaneously.

Posts may vary in *quality* and in how *engaging* they are. Quality models some property such as accuracy or relevance of posts. Engagement models the quality of a post as *perceived*

[1] See `https://github.com/LJ-9/Friction-Social-Media-Model`.

by agents. The quality and engagement of a post are sampled such that low-quality posts are more likely than high-quality posts, and low-engagement posts are more likely than high-engagement posts. Quality and engagement are sampled independently to reflect that high quality and high engagement do not necessarily coincide [9].

While the ABM does not encode agent types, low-quality posts may be thought to stem from a variety of accounts, such as authentic human users, social bots, cyborgs, or algorithmic amplifiers [55], broadly understood.

3.1 Information Diffusion

At each time step, N agents are randomly selected to act in sequence. With probability $p = 0.5$, a sampled agent i posts a new message, otherwise i re-shares a post from their news feed. Call the first the *post*-scenario and the latter the *share*-scenario (see Fig. 1). This modeling choice reflects the approximate average ratio of original tweets (vs. retweets) per agent, as measured in a large-scale sample of English-language tweets [56]. The new or re-shared post is added to the news feeds of i's followers. Each agent's feed contains the α most recent messages posted or re-shared by those they follow, i.e. their friends; if a feed exceeds α posts, the oldest is discarded. Although social media platforms do not usually sort posts in strict reverse chronological order, this is a reasonable simplifying assumption because all platforms give high priority to recent posts. The parameter α models the number of posts viewed in a news feed during a session, and represents the finite attention of the agents. Limited attention has been explored in previous work [54] and measured empirically on a social media mobile app as the number of times that a user scrolls at least 500 pixels through their feed and then stops for at least one second during an active session (idle time less than 30 minutes) [57]. Following this measure, we adopt $\alpha = 15$.

Posts differ in quality and in how engaging they are. Both the quality q and the engagement e of a post are defined in the unit interval. We independently draw a post's quality and engagement from the normalized probability density function $P(x) = \frac{(1-x)}{\int_0^1 (1-x)dx} = 2(1-x)$ with $x = \{q, e\}$.[2] This simple linear distribution reflects the intuition that high-quality and high-engagement information are more rare.

Initially, an agent cannot discern the quality of posts in their feed, but only the perceived quality, that is the engagement e. We assume that the probability that an agent re-shares a post is proportional to the post's engagement. More explicitly, let M_i be the feed of i ($|M_i| = \alpha$). The probability of post $m \in M_i$ being selected is $P(m) = e(m)/\sum_{j \in M_i} e(j)$ where $e(m)$ is the engagement of post m. This models cognitive bias: high-engagement posts will appear at a higher rate in agents' news feeds, further improving the odds of getting spread more. While the engagement e of each post is fixed and does not change when a post gets shared more often, a news feed may contain duplicates of the same post: this happens, e.g., if an agent follows two others that share the same post. This further increases the chances that the duplicated message is re-shared, implicitly modeling an algorithmic bias that amplifies popular messages.

[2] The sampling is implemented using inverse transform sampling, given the cumulative distribution function $C(x) = \int_0^1 P(x)dx = \int_0^1 2(1-x)dx$.

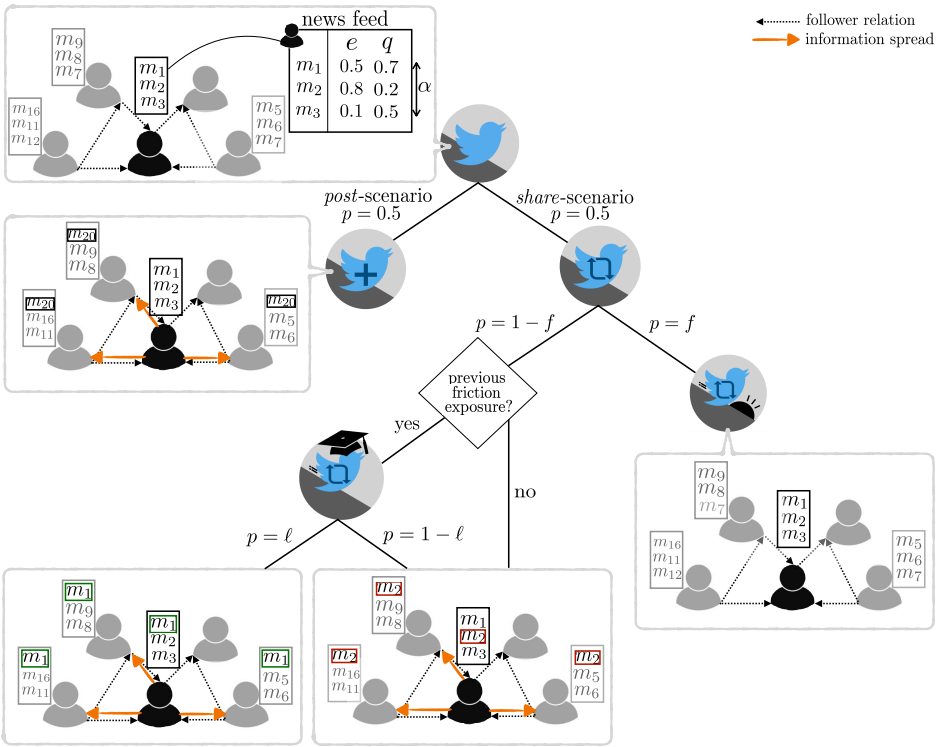


Figure 1: **Information diffusion process.** Each node has a news feed of size α, containing messages recently posted or re-shared by friends. The follower relation is illustrated by dashed arrows pointing from an agent to their friends. Information travels from agents to their followers, along the orange arrows (in the opposite direction of the follow relations). At each time step, a subset of agents act (here, the central black node). With probability 0.5, an acting agent posts a new message (here, m_{20}), else they share a message from their news feed. The new message appears at the top of the followers' news feeds and the existing messages are shifted down. If α messages were already present in a feed, the oldest one is discarded. If the agent shares, with probability f, they are exposed to friction and prevented from sharing. Else, with probability $1 - f$, they scan their feed and share a post in their feed selected with probability proportional to the post's engagement (here, m_2)—unless the agent has been exposed to friction earlier and learned (with probability ℓ): in that case, the agent instead selects the post to share with probability proportional to the post's quality (here, m_1).

3.1.1 Friction and Learning

In our model, friction is restricted to the *share*-scenario, in which agents may face a friction prompt. The intuition is that friction triggers agents to pause, potentially impeding their sharing activity. Agents may resume re-sharing the chosen post after having spent mental resources, or passed a quiz. On the other hand, agents may not resume sharing the chosen post after re-considering or failing to comply. The probability that agents in the *share*-scenario are *exposed* to friction and either *reconsider* or *fail* to comply is captured by the parameter $f \in [0, 1]$ (see Fig. 1). Therefore the probability that an agent is prevented from re-sharing due to friction is $0.5 \cdot f$. The simulation records that an agent has been exposed to the friction prompt.

Agents may learn through exposure to a friction prompt, e.g., through deliberation-triggering nudges or educational quizzes that remind agents to pay attention to quality. The next time an agent is about to re-share a post, an agent who has learned no longer re-shares the most engaging post, but instead chooses a post to re-share based on quality.

We call this type of learning *quality-recognition learning*, drawing intuition from research both on priming effects and nudges [58, 13, 43], and on testing effects and retrieval practices shown to boost learning [59, 60, 61]. Learning through priming and nudging takes place without conscious guidance. In contrast, testing and retrieval of previously absorbed knowledge takes place consciously. Yet, both serve as learning events. Which mechanism takes precedence depends on the design of the friction prompt. Learning in the ABM and the model implementation (described formally below) is kept at a general level and allows for both intuitions, as the outcome—learning to recognize quality—does not change. The implemented friction strategy may thus fall into the *headline-discernment paradigm* (discerning true and false headlines and indicating willingness to share them) and *skill-adoption paradigm*(learning the skills and strategies required to evaluate information quality) of research on behavioral interventions [62].

Agents who have previously been exposed to friction prompts learn to discriminate between engagement and quality. Without prior exposure to friction, no learning happens. Formally, with probability ℓ, an agent i previously exposed to a friction prompt selects a message $m \in M_i$ with probability $P(m) = q(m)/\sum_{j \in M_i} q(j)$ where $q(m)$ is the quality of the post m. If i's feed is only populated with posts having $q = 0$, i refrains from sharing any post and does not act at all. Through the parameter ℓ, the model does not assume that agents are always able to apply what they have learned from one-time exposure to friction; they may still make mistakes or forget.

3.1.2 Networks

Networks in the ABM are directed graphs with vertices representing agents and edges representing follower relations. If there is an edge from agent i to agent j, we say that i *follows* j. Agent i then pays attention to content shared by j. Each network consists of $N = 1{,}000$ agents, and structurally mimics online social networks. To capture the characteristic presence of hubs, we construct networks using a directed variant of the Barabási-Albert preferential attachment mechanism [63]. A network is initialized with $m = 3$ fully connected nodes and is grown by attaching new nodes each with m outgoing edges that are preferentially attached to existing nodes with high in-degree. Put differently, agents with many followers (high in-degree) attract more followers. This results in scale-free properties and few, highly influential agents, e.g., influencers, mirroring the well-documented Matthews effect in action as it relates

to follow relationships [64, 65]. We also wish to capture the characteristic presence of clustering (directed triadic closure [66]). To this end, we add edges by randomly sampling the friend of a friend of a target node and have the target node follow the sampled agent. This has the effect of generating directed triads. We do this until the undirected clustering coefficient reaches 0.29, as measured in a large sample from the empirical Twitter follower network [67]. Networks are generated before a simulation run starts and do not change during a run.

3.1.3 Descriptive Metrics and Simulations Runs

To quantify the effect of friction and learning, our simulation tracks the changes in two metrics that capture desirable properties of an online social network: the average quality of posts in the network and the system's capacity to discriminate information on the basis of its quality.

Ideally, one wants the system to discriminate against low-quality posts by reaching a strong correlation between quality and popularity: the higher the quality of a post, the more widely it should be shared among agents. We capture this discriminative power by measuring Kendall's rank correlation between popularity and quality of posts. At the end of each run, the *popularity* of posts is measured by the number of times posts have been shared or re-shared. Kendall's correlation coefficient τ is computed by creating two rankings of posts—one ranking based on the quality each post is assigned, the other based on popularity—and counting the number of post pairs for which the two rankings are concordant or discordant, properly accounting for ties [68].

Formally, τ is calculated as follows, with n_c the number of concordant pairs, n_d the number of discordant pairs, n_t^q the number of ties only in quality, and n_t^p the number of ties only in popularity:

$$\tau = \frac{n_c - n_d}{\sqrt{(n_c + n_d + n_t^q)(n_c + n_d + n_t^p)}}.$$

High τ indicates that we find posts that share the same position in each respective ranking, i.e., high-quality posts are more popular and low-quality posts are less popular at the end of a run, granting the system discriminative power; in the extreme case $\tau = 1$, the two rankings are completely concordant. Small τ signifies a lack of quality discrimination by the network.

The measurements of average quality and discriminative power are averaged across simulation runs. The overall quality of an information ecosystem is given by average quality $\hat{q}_T$, measured at the end of a run. Each simulation run halts at time T such that the exponential moving average quality of posts in the network's feeds stabilizes. At each time step t, we measure the average quality across all posts visible through the feeds of all the agents as $q = \frac{1}{aN} \sum_{i=1}^{N} \sum_{m \in M_i} q_{im}$, where q_{im} is the quality of the mth post of agent i's feed. We compute the exponential moving average $\hat{q}_t = \rho \cdot \hat{q}_{t-1} + (1 - \rho) q_t$, with smoothing factor $\rho = 0.99$. We define convergence by concluding the run at time T such that $\mid \hat{q}_T - \hat{q}_{T-1} \mid < \varepsilon$ with $\varepsilon = 10^{-5}$. Robustness tests showed no systematic changes in the trend of average quality with smaller ρ nor smaller ε, but only longer runtimes. We record each post's quality and popularity (how often a post is shared or re-shared) at time T.

We average $\hat{q}_T$ and τ across a set of five sampled networks, for each parameter combination of induced friction f and learning capability ℓ. Our parameter combinations are all values for f and ℓ in increments of 0.01 until 0.2 and in increments of 0.1 after. Each of these combinations is run 10 times per network. We thus analyze 50 runs for each of 813 combinations of f and ℓ.

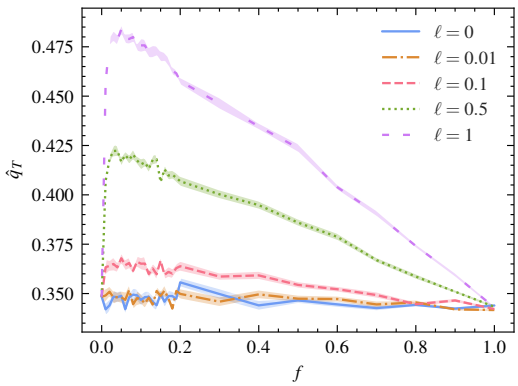


Figure 2: Average post quality $\hat{q}_T$ as a function of friction probability f in networks with $N = 1000$, for different probabilities of learning ℓ. Shaded areas indicate standard errors.

3.2 Results

To assess how friction and associated learning affect information diffusion of high- and low-quality posts in the networks, we plot average quality $\hat{q}_T$ in Fig. 2 and discriminative power τ in Fig. 3.

Fig. 2 summarizes how friction affects average quality: in the absence of learning, we do not observe an increase in quality when solely adding friction to the network. This is plausible, as quality and engagement values of posts are sampled independently, and friction is triggered independently of the quality or engagement of the post. The findings suggest a significant increase in average quality when $f = 0.1$, (i.e., the probability for an agent being prevented from re-sharing is $0.5 \cdot 0.1 = 0.05$) and friction-exposed agents only rarely learn (e.g., $\ell = 0.1$). Higher learning probabilities boost quality significantly. For $\ell = 0.5$, $\hat{q}_T \approx 0.41$ and for $\ell = 1$, $\hat{q}_T \approx 0.47$—an increase by more than one third. Lowering friction to $f = 0.01$, so agents are prevented from re-sharing content with probability $0.5 \cdot 0.01 = 0.005$, still significantly increases quality as long as agents are capable of minimal learning.

Fig. 2 also reveals that increasing the probability of friction exposure to more than $f = 0.1$ does not result in higher quality, no matter how well agents learn. These findings suggest that in the presence of learning, only a little friction provides the best conditions for agents to apply learned behavior, even when accounting for forgetting or making mistakes ($\ell < 1$). The reason for quality to drop with high f is that agents cannot benefit from newly won awareness of posts' quality when they are deprived from sharing. Therefore, combining rare friction prompts with quality-recognition learning is far more desirable than extensively restricting sharing.

There is a trade-off between average quality and diversity in posts. The maximum quality is obtained when the best post is featured on all agent news feeds, which kills diversity. As an artifact of friction preventing re-sharing, diversity increases with friction in the model. In the presence of both high quality and diversity, τ lets us assess how well the network can discriminate between high and low-quality (cf. Fig. 3). We again find that friction combined

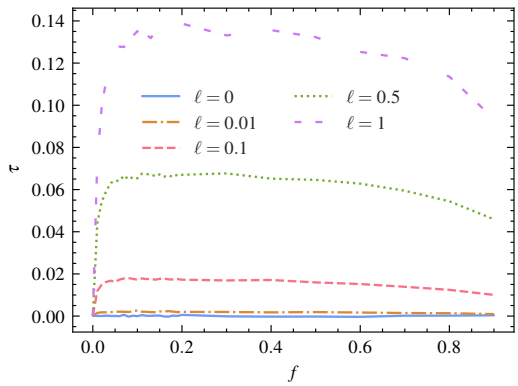


Figure 3: Average discriminative power τ as a function of friction probability f in networks with $N = 1000$, for different probabilities of learning ℓ. Shaded areas indicate standard errors. The Kendall rank coefficient is not defined for $f = 1$, as in this case nothing gets re-shared and the popularity ranking is all ties.

with learning significantly increases the system's discriminative power.

If agents fail to learn, friction yields no improvements in discriminative power ($\tau \approx 0$). With induced friction levels as little as $f = 0.1$ and associated learning $\ell = 0.1$, we find higher correlation between post popularity and quality. Discriminative power is highest for higher ℓ and f between 0.1 and 0.2: τ increases up to 0.139 at $f = 0.2$. Increasing friction to more than $f = 0.2$, does not further improve the network's discriminative power, as agents are more likely to be stopped from employing their learned awareness of quality.

4 Experiment Proposal: Friction by Community Standards Prompts

Preliminary results from the agent-based model suggest that a small amount of friction combined with quality-recognition learning increases the average quality of posts significantly. Both the average quality and the network's power to discriminate between high- and low-quality posts increase with just a bit of friction even with far-from-perfect learning. On the other hand, excessive friction, even with perfect learning, has a negative effect on average quality as agents are prevented from applying their learnings.

These preliminary findings suggest an idea for a concrete friction prompt: quizzing users about community standards to promote their quality recognition before they can react to content. To the best of our knowledge, such a friction strategy has not been subject to extensive field experiments. In an experimental environment, users would be faced with randomly assigned micro-exams (e.g., multiple choice) when they are about to share or like on a platform. The exams could involve questions about the platform's governing community standards, thus helping users learn about such norms [5].

This kind of field experiment may contribute to research on testing effects [59, 60, 61] and complement work on nudges and priming [13, 43]. We also refer to Pennycook et al. [69] for a comprehensive guide on behavioral experiments related to misinformation and fake news. In experimental work on community standards, the display of norms in online discussions on Reddit increased compliance and prevented unruly and harassing conversations [70]. The idea of impeding quizzes has most famously been established with CAPTCHA tests. Quizzes have been successfully used in the comment section of a public Norwegian broadcaster. These quizzes that test a reader's understanding of an article if they wanted to comment produced a respectful and productive conversation [71]. Lutzke et al. [43] also showed that the combination of reminding users of critical thinking and questions taken from community standards may help participants identify fake news. In contrast to previous work [49, 41, 50], the friction strategy we suggest does not rely on detection of bad actors nor on classification and labeling of content as high- or low-quality.

As both an educational and preventive measure, arbitrarily recurring tests may act as bulwarks against information vandalism, hate speech, and misinformation, explicitly mentioned in the community standards of the particular platform. We hypothesize that these effects may prove more effective than just labeling low-quality sources, a measure that is unable to counter the illusory effect [72]. As a gateway, the proposed friction strategy might have both short-term and long-term social benefits: In the short term, it slows down the spread of both high-quality and malevolent online content while educating users. The quiz may serve the same function as an accuracy prompt asking the user to pause, think, and learn to recognize quality. In the long run, the intervention might have preventive effects as users will have engaged with, and reflected upon, what they have agreed to when signing up on the social platform. In this way, users may be less prone to first-time engagement with harmful content [5].

The central research question for the proposed study is: *Does user engagement with low-quality posts drop when prompted with friction strategies incorporating quality-recognition learning based on community standards?* Engagement with low-quality posts can be measured by reactions such as liking and sharing, and compared to user engagement with high-quality content. Among the many possible violations of common community standards, a focal point could be on the proliferation of misinformation. This would complement recent work [49] where friction strategies are shown to slightly but significantly decrease offensive replies on Twitter.

As in the experiments by Pennycook et al. [73], our research question zeros in on one-click reactions instead of written replies [49]. Sticking to one-click reactions restricts the focus of the study to engagement actions that have a clear interpretation about support for the original post. One does not need to invoke NLP techniques or human annotation to determine the tone of comments/replies. As part of the experimental set-up, one may investigate how the community standard friction strategy affects sharing behavior and average quality of shared content, compared to a neutral friction strategy (e.g., ticking a box) and to a control group not exposed to friction. Field experiments comparing community standards versus neutral friction may explore whether it is possible to isolate the effect of learning on the rate of misinformation spread. In the following, we briefly map some design avenues for the proposed field experiments.

4.1 Intervention Design

The experimental setup will emulate a social media feed in which participants engage with posts featuring text and potentially links to stories, presented through photos and headlines. Participants will be asked to interact with (like and/or share) the posts as they would on their preferred platform. The friction intervention will take the form of a prompt during this process. We propose to mold a prompt around Twitter's existing community standards [74]. In contrast to Facebook, LinkedIn, and Instagram, Twitter specifically formulates and instructs users about what to look out for in terms of misinformation. Exemplary formulations in their community standards may be used as inspiration for the design of the prompt: "Is the content shared in a deceptive manner or with false context?" and "Is the content likely to result in widespread confusion on public issues, impact public safety, or cause serious harm?" Multiple-choice quizzes may accordingly query the participants about, say, definitions and examples of misinformation, risks and consequences of misinformation, or risks and consequences of violating the standards.

Exposure to friction in the intervention group may be either upon engagement with a misinformation-labeled tweet or randomly upon engagement with either a true or misinformation-labeled post. The friction strategy tested by Katsaros et al. [49] against offensive replies on Twitter was prompted as users were about to send an offensive reply. Users were not randomly reminded of the platform's community standards, but only upon violating community standards. Our proposed design revolves instead around random reminders of community standards. This has the advantage of removing the need to classify posts. In addition, the design is intended to promote fluency in the community standards rather than just awareness of particular sanctioning clauses. Both intervention and control groups should be set up to test for causality [75], rather than exposing all study participants to the designed intervention as in Avram et al. [23].

4.2 Experimental Platform

Testing friction strategies in a real environment poses difficulties [76]. Independent researchers do not have access to a platform such as Twitter to design and test real-time interventions in randomized controlled trials. This approach is therefore neither feasible for nor reproducible by researchers not employed by Twitter. One may instead approach a proof of concept in an experiment designed and carried out in a simulated social media environment. Available options include Amazon Mechanical Turk [77, 40], Volunteer Science [78], games such as Fakey [79], or open-source software such as the Mock Social Media Tool [80]. Each environment yields different levels of realism, and experiments may be informed by empirical data about user activity and social network structure guiding online information sharing [51]. However, complete ecological validity (like in Katsaros et al. [49]) is most likely not feasible for this project. One reason will be restrictions to the experimental feed, as discussed next.

One has to carefully consider *what feed* to show to participants. The higher the desired ecological validity, the more sizeable the task of labeling tweets to determine intervention effects. One extreme is working solely with users' own feeds. While this yields the highest ecological validity, it comes with two downsides. First, one cannot be sure that users are exposed to misinformation at all, making the size of a sufficient data sample unknown. Second, all posts users have seen, or at least engaged with, must be labeled as misinformation or not. Labeling has either limited accuracy when derived from lists of low-credibility sources [81, 3], or requires a lot of manual labor.

Alternatively, one may work with synthetically curated feeds. A set of posts could be (partially) curated by researchers. These posts could be shown to all participants, thus ensuring that all users are shown the same misinformation posts while restricting the set of posts that must be labeled. The curated posts may be shown either within a fully synthetic feed or embedded in each participant's own Twitter feed. To obtain a set of curated posts, one may consider the following options:

- Create fictional true and false headlines/tweets and embed them in the look and feel of a Twitter feed.
- Use an already curated sample of true and false headlines or posts and present them with the look and feel of a Twitter feed. For example, as Sultan et al. [44], we could use the pretested collection of headlines of accurate (taken from mainstream sources) and inaccurate (according to fact-checking websites) content curated by Pennycook et al. [69]. Alternatively, we could reuse the headlines by Fazio [40] or Pennycook et al. [73].[3]
- Reuse known misinformation tweets, such as (i) false claims collected from Politifact [83] or (ii) the COVID-19 misinformation dataset [82]. To sample true tweets, one may supplement with tweets from reliable sources such as @nytimes.

Ecological validity may suffer if the synthetic feed is not personalized or not relevant to the participants. If resources allow, one could tailor studies to a selection of likely participant interests and scrape recent tweets accordingly. Exposing participants to injected misinformation and low-credibility content may raise valid worries about ethical implications. Pennycook et al. [69] remind researchers to be aware of the effect of exposing participants to offensive content and to provide mental health support during the study. The experimental design will have to weigh the advantages and limitations of both real and synthetic feeds.

5 Conclusion and Policy Desiderata

The past two years have seen a significant increase in public interventions to counter influence operations by platforms such as Facebook and Twitter [46]. The most common interventions are redirection and content labeling. These only contextualize or correct potentially harmful content rather than more harsh approaches that remove it or (shadow-)ban users whospread it.

Unfortunately, social media platforms rarely disclose whether their interventions have been tested for effectiveness. The lack of transparency around the testing and implementation process is a missed opportunity to build public trust and can make it difficult for researchers to study the effectiveness of different countermeasures.

Especially the effectiveness of content labeling is challenged by research findings [42]: While the common advice for dealing with fake news is to consider the source, people often struggle to remember sources [72]. Tagging only some false news stories as "false" may boost the perceived accuracy of untagged stories due to an implied truth effect [84].

The prevalence of redirection and content labeling as dominant interventions has significant policy implications. These interventions generally distract rather than impede, and preserve user choice and autonomy by allowing sensitive content to continue to circulate, offering only

[3]Headlines by Pennycook et al. [73] are from the Harvard Global Health Institute and contain COVID-19-related misinformation. While these come with a clear label [82], the content will potentially have lost topicality and/or have been deleted, threatening the study's validity.

counterspeech as an alternative. While less intrusive than other methods, it also places a greater burden on users to themselves manage threats from influence operations. This leaves users that ignore labels or do not follow redirections vulnerable.

Friction that educates users through randomly triggered quizzes is non-elective, overt, protective and non-commercial, but does impede users more than redirection or labeling does. We believe a successful implementation would impede little, as our results show that in the presence of learning, only a little friction provides the best conditions for agents to apply learned behavior. Besides prompting users to reflect and learn, we see several advantages in such a friction strategy in the fight against misinformation.

A first advantage is that friction prompts do not require prior classifications. The development of detection and classification models to spot low-quality sources, social bots, coordinated inauthentic behavior, and low-quality posts is resource-expensive and difficult: Research suffers under data shortages [85, 86, 87, 76], both platform internal and external research suffers from the lack of ground truth [86, 88], supervised methods are limited in their ability to detect coordinated groups [89, 90], and the emulation of human behavior by social bots makes binary classifiers difficult to apply without worries about censorship [3]. Friction prompts not only bypass these challenges, but may present an unbiased, easily scalable strategy [62] that actively adds costs for inauthentic actors and social bots.

Second, community standards may gain more attention and reverence. As the friction prompts require that authentic users familiarize themselves with the community standards of the platform of choice, the resulting learning should ease the platform's burden of enforcing the standards. Likewise, having users that are largely well-informed about the community standards may encourage platforms to a transparent practice and a consistent enforcement of e.g. their principles of content moderation. Hopefully, learning about and enforcing community standards will stimulate public debate and political conversation pertaining to the online public space, democratic ambition, freedom of expression, privacy, and user rights—all themes that tech giants themselves routinely demand paying increased attention to.

We conclude with a call to further explore, test, and experiment with friction that educates users on community standards and triggers quality-recognition learning on social media platforms such as Twitter. As also required by the recent Digital Services Act (DSA) of the EU [91], we strongly encourage any platforms that undertake such or other intervention strategies to share their usually proprietary, publicly unavailable data and tools with researchers so they may aid in assessing the interventions' efficiency.

References

[1] Hendricks, Vincent F. and Mehlsen, Camilla, *The Ministry of Truth: BigTech's Influence on Facts, Feelings and Fiction.* Springer Nature, 2022.

[2] Bliss, Nadya and Bradley, Elizabeth and Garland, Joshua and Menczer, Filippo and Ruston, Scott and Starbird, Kate and Wiggins, Chris, "An agenda for disinformation research," CRA Computing Community Consortium (CCC), Quadrennial Paper, 2020.

[3] Shao, Chengcheng and Ciampaglia, Giovanni Luca and Varol, Onur and Yang, Kai Cheng and Flammini, Alessandro and Menczer, Filippo, "The spread of low-credibility content by social bots," *Nature Communications*, vol. 9, no. 1, 2018.

[4] World Economic Forum, "The global risk report," *World Economic Forum (weforum.org)*, 2018. [Online]. Available: http://www3.weforum.org/docs/WEF_GRR18_Report.pdf

[5] Hendricks, Vincent F., "Turning the Tables: Using BigTech community standards as friction strategies," *OECD The Forum Network*, vol. December 20, 2021. [Online]. Available: https://www.oecd-forum.org/posts/turning-the-tables-using-bigtech-community-standards-as-friction-strategies

[6] Simon, Herbert A and others, "Designing organizations for an information-rich world," *Computers, Communications, and the Public Interest*, vol. 72, p. 37, 1971.

[7] Ciampaglia, Giovanni Luca and Flammini, Alessandro and Menczer, Filippo, "The production of information in the attention economy," *Scientific Reports*, vol. 5, no. 1, pp. 1–6, 2015.

[8] Surowiecki, James, *The wisdom of crowds*. Anchor, 2005.

[9] Vosoughi, Soroush and Roy, Deb and Aral, Sinan, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[10] Centola, Damon and Macy, Michael, "Complex contagions and the weakness of long ties," *American journal of Sociology*, vol. 113, no. 3, pp. 702–734, 2007.

[11] Lorenz, Jan and Rauhut, Heiko and Schweitzer, Frank and Helbing, Dirk, "How social influence can undermine the wisdom of crowd effect," *Proceedings of the National Academy of Sciences*, vol. 108, no. 22, pp. 9020–9025, 2011.

[12] Nickerson, Raymond S, "Confirmation bias: A ubiquitous phenomenon in many guises," *Review of General Psychology*, vol. 2, no. 2, pp. 175–220, 1998.

[13] Pennycook, Gordon and Rand, David G, "Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation," *Nature Communications*, vol. 13, no. 1, pp. 1–12, 2022.

[14] Pennycook, Gordon and Epstein, Ziv and Mosleh, Mohsen and Arechar, Antonio A. and Eckles, Dean and Rand, David G., "Shifting attention to accuracy can reduce misinformation online," *Nature*, vol. 592, no. 7855, pp. 590–595, 2021.

[15] Ceylan, Gizem and Anderson, Ian A. and Wood, Wendy, "Sharing of misinformation is habitual, not just lazy or biased," *Proceedings of the National Academy of Sciences*, vol. 120, no. 4, 2023.

[16] Lacassagne, Doris and Béna, Jérémy and Corneille, Olivier, "Is earth a perfect square? repetition increases the perceived truth of highly implausible statements," *Cognition*, vol. 223, p. 105052, 2022.

[17] Hills, Thomas T, "The dark side of information proliferation," *Perspectives on Psychological Science*, vol. 14, no. 3, pp. 323–330, 2019.

[18] Fazio, Lisa K and Brashier, Nadia M and Payne, B Keith and Marsh, Elizabeth J, "Knowledge does not protect against illusory truth." *Journal of Experimental Psychology: General*, vol. 144, no. 5, p. 993, 2015.

[19] Nikolov, Dimitar and Lalmas, Mounia and Flammini, Alessandro and Menczer, Filippo, "Quantifying biases in online information exposure," *Journal of the Association for Information Science and Technology*, vol. 70, no. 3, pp. 218–229, 2019.

[20] Ciampaglia, Giovanni Luca and Nematzadeh, Azadeh and Menczer, Filippo and Flammini, Alessandro, "How algorithmic popularity bias hinders or promotes quality," *Scientific Reports*, vol. 8, no. 1, pp. 1–7, 2018.

[21] Metaxas, P. T. and Mustafaraj, E. and Wong, K. and Zeng, L. and O'Keefe, M. and Finn, S., "What Do Retweets Indicate? Results from User Survey and Meta-Review of Research," *Proceedings of the 9th International Conference on Web and Social Media, ICWSM 2015*, pp. 658–661, 2015.

[22] Dutta, Hridoy Sankar and Chetan, Aditya and Joshi, Brihi and Chakraborty, Tanmoy, "Retweet us, we will retweet you: Spotting collusive retweeters involved in blackmarket services," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2018, pp. 242–249.

[23] Avram, Mihai and Micallef, Nicholas and Patil, Sameer and Menczer, Filippo, "Exposure to social engagement metrics increases vulnerability to misinformation," *The Harvard Kennedy School Misinformation Review*, vol. 1, no. 5, 2020.

[24] Torres-Lugo, Christopher and Pote, Manita and Nwala, Alexander and Menczer, Filippo, "Manipulating Twitter Through Deletions," in *Proceedings of the 16th International AAAI Conference on Web and Social Media (ICWSM)*, 2022.

[25] Nizzoli, Leonardo and Tardelli, Serena and Avvenuti, Marco and Cresci, Stefano and Tesconi, Maurizio, "Coordinated Behavior on Social Media in 2019 UK General Election," in *Proc. International AAAI Conference on Web and Social Media (ICWSM)*, vol. 15, no. 1, 2021, pp. 443–454.

[26] Orabi, Mariam and Mouheb, Djedjiga and Al Aghbari, Zaher and Kamel, Ibrahim, "Detection of Bots in Social Media: A Systematic Review," *Information Processing and Management*, vol. 57, 2020.

[27] Goerzen, Matthew and Matthews, Jeanna, "Black hat trolling, white hat trolling, and hacking the attention landscape," *The Web Conference 2019 – Companion of the World Wide Web Conference, WWW 2019*, vol. 2, pp. 523–528, 2019.

[28] Ferrara, Emilio, "Disinformation and social bot operations in the run up to the 2017 French presidential election," *First Monday*, vol. 22, no. 8, 2017.

[29] Ferrara, Emilio and Varol, Onur and Davis, Clayton and Menczer, Filippo and Flammini, Alessandro, "The Rise of Social Bots," *Commun. ACM*, vol. 59, no. 7, pp. 96–104, 2016.

[30] Takacs, Richard and McCulloh, Ian, "Dormant bots in social media: Twitter and the 2018 U.S. senate election," in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2019*, 2019, pp. 796–800.

[31] Pacheco, Diogo and Hui, Pik-Mai and Torres-Lugo, Christopher and Bao Tran, Truong and Flammini, Alessandro and Menczer, Filippo, "Uncovering Coordinated Networks on Social Media: Methods and Case Studies," in *Proc. International AAAI Conference on Web and Social Media (ICWSM)*, vol. 15, no. 1, 2021, pp. 455–466.

[32] Duan, Zening and Li, Jianing and Lukito, Josephine and Yang, Kai-Cheng and Chen, Fan and Shah, Dhavan V and Yang, Sijia, "Algorithmic agents in the hybrid media system: Social bots, selective amplification, and partisan news about covid-19," *Human Communication Research*, 2022.

[33] Lazer, David M. J. and Baum, Matthew A. and Benkler, Yochai and Berinsky, Adam J. and Greenhill, Kelly M. and Menczer, Filippo and Metzger, Miriam J. and Nyhan, Brendan and Pennycook, Gordon and Rothschild, David and Schudson, Michael and Sloman, Steven A. and Sunstein, Cass R. and Thorson, Emily A. and Watts, Duncan J. and Zittrain, Jonathan L., "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

[34] Menczer, Filippo and Hills, Thomas, "The attention economy," *Scientific American*, vol. 323, no. 6, pp. 54–61, Dec 2020.

[35] Truong, Bao Tran and Lou, Xiaodan and Flammini, Alessandro and Menczer, Filippo, "Vulnerabilities of the online public square to manipulation," 2023.

[36] Tomalin, Marcus, "Rethinking online friction in the information society," *Journal of Information Technology*, vol. January 2022, 2022.

[37] Ressa, Maria and Schaake, Marietje and Halgand-Mishra, Delphine and de Villars, Iris and Domino, Jenny and Shefet, Dan, "Policy framework: Working group on infodemics," *Forum on Information and Democracy*, 2018.

[38] Goodman, Ellen P, "Digital information fidelity and friction," *Knight First Amendment Institute at Columbia University*, 2020.

[39] Sunstein, Cass R., "Sludge Audits," *Behavioural Public Policy*, pp. 1–20, 2020.

[40] Fazio, Lisa, "Pausing to consider why a headline is true or false can help reduce the sharing of false news," *Harvard Kennedy School Misinformation Review*, vol. 1, no. 2, pp. 1–8, 2020.

[41] Bhuiyan, Md Momen and Zhang, Kexin and Vick, Kelsey and Horning, Michael A. and Mitra, Tanushree, "Feedreflect: A tool for nudging users to assess news credibility on twitter," in *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, ser. CSCW '18, 2018, pp. 205–208.

[42] Brashier, Nadia M and Eliseev, Emmaline Drew and Marsh, Elizabeth J, "An initial accuracy focus prevents illusory truth," *Cognition*, vol. 194, p. 104054, 2020.

[43] Lutzke, Lauren and Drummond, Caitlin and Slovic, Paul and Árvai, Joseph, "Priming critical thinking: Simple interventions limit the influence of fake news about climate change on Facebook," *Global Environmental Change*, vol. 58, p. 101964, 2019.

[44] Sultan, Mubashir and Tump, Alan N and Geers, Michael and Lorenz-Spreen, Philipp and Herzog, Stefan M and Kurvers, Ralf HJM, "Time pressure reduces misinformation discrimination ability but does not alter response bias," *Scientific Reports*, vol. 12, no. 1, pp. 1–12, 2022.

[45] Jackson, Matthew O and Malladi, Suraj and McAdams, David, "Learning through the grapevine and the impact of the breadth and depth of social networks," *Proceedings of the National Academy of Sciences*, vol. 119, no. 34, p. e2205549119, 2022.

[46] Yadav, Kamya, "Platform Interventions: How Social Media Counters Influence Operations," https://carnegieendowment.org/2021/01/25/platform-interventions-how-social-media-counters-influence-operations-pub-83698, accessed: 2022-12-06.

[47] Twitter, "About government and state-affiliated media account labels on Twitter," https://help.twitter.com/en/rules-and-policies/state-affiliated, accessed: 2023-01-19.

[48] ABP News Bureau, "Good Old Retweet Button Is Back On Twitter! Netizens Welcome The Decision With Funniest Memes," https://help.twitter.com/en/rules-and-policies/state-affiliated, 2020, accessed: 2022-11-10.

[49] Katsaros, Matthew and Yang, Kathy and Fratamico, Lauren, "Reconsidering tweets: Intervening during tweet creation decreases offensive content," pp. 477–487, 2022.

[50] Théro, Héloïse and Vincent, Emmanuel M, "Investigating Facebook's interventions against accounts that repeatedly share misinformation," *Information Processing & Management*, vol. 59, no. 2, p. 102804, 2022.

[51] de Freitas Melo, Philipe and Vieira, Carolina Coimbra and Garimella, Kiran and de Melo, Pedro OS Vaz and Benevenuto, Fabrício, "Can whatsapp counter misinformation by limiting message forwarding?" in *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8*. Springer, 2020, pp. 372–384.

[52] Lee, Dave, "Instagram now asks bullies: 'Are you sure?'," *BBC news*, 2019. [Online]. Available: https://www.bbc.com/news/technology-48916828

[53] Truong, Bao Tran and Lou, Xiaodan and Flammini, Alessandro and Menczer, Filippo, "SimSoM: A Simulator of Social Media." [Online]. Available: https://github.com/osome-iu/SimSoM

[54] Weng, Lilian and Flammini, Alessandro and Vespignani, Alessandro and Menczer, Fillipo, "Competition among memes in a world with limited attention," *Scientific Reports*, vol. 2, no. 1, pp. 1–9, 2012.

[55] Yan, Harry Yaojun and Yang, Kai-Cheng, "The landscape of social bot research: a critical appraisal," *OSF Preprints*, 2022.

[56] Alshaabi, Thayer and Dewhurst, David Rushing and Minot, Joshua R and Arnold, Michael V and Adams, Jane L and Danforth, Christopher M and Dodds, Peter Sheridan, "The growing amplification of social media: Measuring temporal and social contagion dynamics for over 150 languages on twitter for 2009–2020," *EPJ Data Science*, vol. 10, no. 1, p. 15, 2021.

[57] Qiu, Xiaoyan and FM Oliveira, Diego and Sahami Shirazi, Alireza and Flammini, Alessandro and Menczer, Filippo, "Limited individual attention and online virality of low-quality information," *Nature Human Behaviour*, vol. 1, no. 7, pp. 1–7, 2017.

[58] Weingarten, Evan and Chen, Qijia and McAdams, Maxwell and Yi, Jessica and Hepler, Justin and Albarracín, Dolores, "From primed concepts to action: A meta-analysis of the behavioral effects of incidentally presented words." *Psychological Bulletin*, vol. 142, no. 5, p. 472, 2016.

[59] Paul, Annie Murphy, "Researchers find that frequent tests can boost learning," *Scientific American*, vol. 313, no. 2, pp. 1–7, 2015.

[60] Rowland, Christopher A, "The effect of testing versus restudy on retention: a meta-analytic review of the testing effect." *Psychological Bulletin*, vol. 140, no. 6, p. 1432, 2014.

[61] Endres, Tino and Renkl, Alexander, "Mechanisms behind the testing effect: an empirical investigation of retrieval practice in meaningful learning," *Frontiers in Psychology*, vol. 6, p. 1054, 2015.

[62] Kozyreva, Anastasia and Lorenz-Spreen, Philipp and Herzog, Stefan and Ecker, Ullrich and Lewandowsky, Stephan and Hertwig, Ralph, "Toolbox of interventions against online misinformation and manipulation," 2022.

[63] Barabási, Albert-László and Albert, Réka, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[64] Merton, Robert K, "The Matthew Effect," *Medical Journal of Australia*, vol. 1, no. 13, pp. 552–553, 1968.

[65] Perc, Matjaž, "The Matthew effect in empirical data," *Journal of the Royal Society Interface*, vol. 11, no. 98, 2014.

[66] Weng, Lilian and Ratkiewicz, Jacob and Perra, Nicola and Gonçalves, Bruno and Castillo, Carlos and Bonchi, Francesco and Schifanella, Rossano and Menczer, Filippo and Flammini, Alessandro, "The role of information diffusion in the evolution of social networks," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2013, pp. 356–364.

[67] Nikolov, D and Flammini, A and Menczer, F, "Right and left, partisanship predicts (asymmetric) vulnerability to misinformation," *The Harvard Kennedy School Misinformation Review*, 2021.

[68] Kendall, Maurice G, "The treatment of ties in ranking problems," *Biometrika*, vol. 33, no. 3, pp. 239–251, 1945.

[69] Pennycook, Gordon and Binnendyk, Jabin and Newton, Christie and Rand, David G, "A practical guide to doing behavioral research on fake news and misinformation," *Collabra: Psychology*, vol. 7, no. 1, p. 25293, 2021.

[70] Matias, J Nathan, "Preventing harassment and increasing group participation through social norms in 2,190 online science discussions," *Proceedings of the National Academy of Sciences*, vol. 116, no. 20, pp. 9785–9789, 2019.

[71] Lichtermann, Jospeh, "This site is "taking the edge off rant mode" by making readers pass a quiz before commenting," https://tinyurl.com/ye4u4f9n, accessed: 2022-12-05.

[72] Henkel, Linda A and Mattson, Mark E, "Reading is believing: The truth effect and source credibility," *Consciousness and Cognition*, vol. 20, no. 4, pp. 1705–1721, 2011.

[73] Pennycook, Gordon and McPhetres, Jonathon and Zhang, Yunhao and Lu, Jackson G and Rand, David G, "Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy nudge intervention," Mar 2020.

[74] Twitter, "The Twitter Rules," https://help.twitter.com/en/rules-and-policies/manipulated-media, 2023, accessed: 2023-02-20.

[75] Epstein, Ziv and Lin, Hause and Pennycook, Gordon and Rand, David, "How many others have shared this? Experimentally investigating the effects of social cues on engagement, misinformation, and unpredictability on social media," 2022.

[76] Pasquetto, Irene V. and Swire-Thompson, Briony and others, "Tackling misinformation: What researchers could do with social media data," *HKS Misinformation Review*, vol. 1, no. 8, 2020.

[77] Pennycook, Gordon, "Fake news fast and slow," *Journal of Experimental Psychology: General*, pp. 1–18, 2019.

[78] Radford, Jason and Pilny, Andy and Reichelmann, Ashley and Keegan, Brian and Welles, Brooke Foucault and Hoye, Jefferson and Ognyanova, Katherine and Meleis, Waleed and Lazer, David, "Volunteer science: An online laboratory for experiments in social psychology," *Social Psychology Quarterly*, vol. 79, no. 4, pp. 376–396, 2016.

[79] Micallef, Nicholas and Avram, Mihai and Menczer, Filippo and Patil, Sameer, "Fakey: A game intervention to improve news literacy on social media," *Proc. ACM Human-Computer Interaction*, vol. 5, no. CSCW1, p. 6, 2021.

[80] Jagayat, Arvin and Gurkaran Boparai, Carson Pun and Becky L. Choma, "Mock social media website tool." [Online]. Available: https://docs.studysocial.media

[81] Lin, Hause and Lasser, Jana and Lewandowsky, Stephan and Cole, Rocky and Gully, Andrew and Rand, David and Pennycook, Gordon, "High level of agreement across different news domain quality ratings," 2022.

[82] Shahi, Gautam Kishore and Dirkson, Anne and Majchrzak, Tim A., "An exploratory study of covid-19 misinformation on twitter," *Online Social Networks and Media*, vol. 22, p. 100104, 2021.

[83] Wang, Yichen and Han, Richard and Lehman, Tamara and Lv, Qin and Mishra, Shivakant, "Analyzing behavioral changes of twitter users after exposure to misinformation," *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Nov 2021.

[84] Pennycook, Gordon and Bear, Adam and Collins, Evan T and Rand, David G, "The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings," *Management Science*, vol. 66, no. 11, pp. 4944–4957, 2020.

[85] Coalition for Independent Technology Research, "Letter: Imposing fees to access the twitter api threatens public-interest research," https://independenttechresearch.org/letter-twitter-api-access-threatens-public-interest-research/, February 6 2023, accessed: 2023-02-11.

[86] Martini, Franziska and Samula, Paul and Keller, Tobias R. and Klinger, Ulrike, "Bot, or not? Comparing three methods for detecting social bots in five political discourses," *Big Data and Society*, vol. 8, 2021.

[87] Bliss, Nadya and Bradley, Elizabeth and Garland, Joshua and Menczer, Filippo and Ruston, Scott and Starbird, Kate and Wiggins, Chris, "An Agenda for Disinformation Research," CRA Computing Community Consortium (CCC), Quadrennial Paper, 2020.

[88] Magelinski, Thomas and Ng, Lynnette and Carley, Kathleen, "Synchronized Action Framework for Detection of Coordination on Social Media," *Journal of Online Trust and Safety*, vol. 1, 2 2022.

[89] Yang, Kai Cheng and Varol, Onur and Davis, Clayton A. and Ferrara, Emilio and Flammini, Alessandro and Menczer, Filippo, "Arming the public with artificial intelligence to counter social bots," *Human Behavior and Emerging Technologies*, vol. 1, pp. 48–61, 1 2019.

[90] Yang, Kai-Cheng and Varol, Onur and Hui, Pik-Mai and Menczer, Filippo, "Scalable and generalizable social bot detection through data selection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 1096–1103.

[91] The European Union, "The digital services act: ensuring a safe and accountable online environment," https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en, accessed: 2022-12-22.

Appendices

Appendix 1

Co-Author Statements

Co-author statement

PhD student	Laura Jahn
Date of birth	04.07.1993
Faculty (Department)	Center for Information and Bubble Studies, Department of Communication, Faculty of Humanities

"Attribution of authorship should in general be based on criteria a-d adopted from the Vancouver guidelines , and all individuals who meet these criteria should be recognized as authors:

A. Substantial contributions to the conception or design of the work, or the acquisition, analysis, or interpretation of data for the work, and
B. drafting the work or revising it critically for important intellectual content, and
C. final approval of the version to be published, and
D. agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved."

Article/paper/chapter/manuscript
This co-authorship declaration applies to the following:

*Title	*Detecting Coordinated Inauthentic Behavior in Likes on Social Media: Proof of Concept*
*Author(s)	*Laura Jahn, Rasmus K. Rendsvig, Jacob Stærk-Østergaard*
Journal	*Submitted to "Social Network Analysis and Mining"*
Volume (no)	
Start page	
End page	

Contributions to the paper/manuscript made by the PhD student

What was the role of the PhD student in designing the study?

Co-author statement

PhD student: Laura Jahn

Date of birth: 04.07.1993

R. K. Rendsvig suggested the study's general idea. L. Jahn co-designed the overall structure and all details of the study.

How did the PhD student participate in data collection and/or development of theory?

L. Jahn was the prime mover behind all production of code for the implementation of the agent-based model, data handling, and data analysis.

Which part of the manuscript did the PhD student write or contribute to?

All parts. The paper manuscript was written by L. Jahn and R.K. Rendsvig, and was commented on by J. Stærk-Østergaard.

Did the PhD student read and comment on the final manuscript?

Yes. All authors read and approved the final manuscript.

Signatures

If an article/ paper/chapter/manuscript is written in collaboration with three or less researchers (including the PhD student), all researchers must sign the statement. However, if an article has more than three authors the statement may be signed by a representative sample, cf. article 12, section 4 and 5 of the Ministerial Order No. 1039, 27 August 2013. A representative sample consists of minimum three authors, which is comprised of the first author, the corresponding author, the senior author, and 1-2 authors (preferably international/non-supervisor authors).

By their signature, the authors agree that the article/paper/chapter/manuscript will be included as a part of the PhD thesis made by the PhD student mentioned above.

Date 31.01.23 Name Laura Jahn Signature 

Co-author statement

PhD student Laura Jahn

Date of birth 04.07.1993

Date 31.01.23 Name Rasmus. K. Rendsvig Signature 

Date 31.01.23 Name Jacob Stærk-Østergaard Signature 

Co-author statement

PhD student	Laura Jahn
Date of birth	04.07.1993
Faculty (Department)	Center for Information and Bubble Studies, Department of Communication, Faculty of Humanities

"Attribution of authorship should in general be based on criteria a-d adopted from the Vancouver guidelines , and all individuals who meet these criteria should be recognized as authors:

A. Substantial contributions to the conception or design of the work, or the acquisition, analysis, or interpretation of data for the work, and
B. drafting the work or revising it critically for important intellectual content, and
C. final approval of the version to be published, and
D. agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved."

Article/paper/chapter/manuscript
This co-authorship declaration applies to the following:

*Title	*Towards Detecting Inauthentic Coordination on Twitter Likes Data*
*Author(s)	*Laura Jahn, Rasmus K. Rendsvig*
Journal	
Volume (no)	
Start page	
End page	

Contributions to the paper/manuscript made by the PhD student

What was the role of the PhD student in designing the study?

Co-author statement

PhD student Laura Jahn

Date of birth 04.07.1993

L. Jahn suggested the study's general idea. L. Jahn co-designed the overall structure and all details of the study.

How did the PhD student participate in data collection and/or development of theory?

L. Jahn was a prime mover behind all theory considerations and the development of the code for data collection, and data analysis.

Which part of the manuscript did the PhD student write or contribute to?

All parts. The paper manuscript was written by L. Jahn and R. K. Rendsvig.

Did the PhD student read and comment on the final manuscript?

Yes. All authors read and approved the final manuscript.

Signatures

If an article/ paper/chapter/manuscript is written in collaboration with three or less researchers (including the PhD student), all researchers must sign the statement. However, if an article has more than three authors the statement may be signed by a representative sample, cf. article 12, section 4 and 5 of the Ministerial Order No. 1039, 27 August 2013. A representative sample consists of minimum three authors, which is comprised of the first author, the corresponding author, the senior author, and 1-2 authors (preferably international/non-supervisor authors).

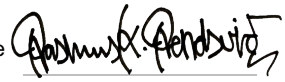
By their signature, the authors agree that the article/paper/chapter/manuscript will be included as a part of the PhD thesis made by the PhD student mentioned above.

Date 31.01.23 Name Laura Jahn Signature 

Co-author statement

PhD student Laura Jahn

Date of birth 04.07.1993

Date 31.01.23 Name *Rasmus. K. Rendsvig* Signature 

Co-author statement

PhD student	Laura Jahn
Date of birth	04.07.1993
Faculty (Department)	Center for Information and Bubble Studies, Department of Communication, Faculty of Humanities

"Attribution of authorship should in general be based on criteria a-d adopted from the Vancouver guidelines , and all individuals who meet these criteria should be recognized as authors:

A. Substantial contributions to the conception or design of the work, or the acquisition, analysis, or interpretation of data for the work, and
B. drafting the work or revising it critically for important intellectual content, and
C. final approval of the version to be published, and
D. agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved."

Article/paper/chapter/manuscript
This co-authorship declaration applies to the following:

*Title	*Friction Interventions to Curb the Spread of Misinformation on Social Media*
*Author(s)	*Laura Jahn, Rasmus K. Rendsvig, Alessandro Flammini, Filippo Menczer and Vincent F. Hendricks*
Journal	
Volume (no)	
Start page	
End page	

Contributions to the paper/manuscript made by the PhD student

What was the role of the PhD student in designing the study?

Co-author statement

PhD student Laura Jahn

Date of birth 04.07.1993

V. F. Hendricks suggested the general idea to study friction. F. Menczer suggested to study friction through an agent-based model (ABM). L. Jahn was the prime mover behind all implementation and design decisions.

How did the PhD student participate in data collection and/or development of theory?

L. Jahn was a prime mover behind theory considerations and the main contributor to implementing the ABM code for data production and data analysis.

Which part of the manuscript did the PhD student write or contribute to?

All parts. The paper manuscript was written by L. Jahn and edited by all authors.

Did the PhD student read and comment on the final manuscript?

Yes. All authors read and approved the final manuscript.

Signatures

If an article/ paper/chapter/manuscript is written in collaboration with three or less researchers (including the PhD student), all researchers must sign the statement. However, if an article has more than three authors the statement may be signed by a representative sample, cf. article 12, section 4 and 5 of the Ministerial Order No. 1039, 27 August 2013. A representative sample consists of minimum three authors, which is comprised of the first author, the corresponding author, the senior author, and 1-2 authors (preferably international/non-supervisor authors).

By their signature, the authors agree that the article/paper/chapter/manuscript will be included as a part of the PhD thesis made by the PhD student mentioned above.

Date 31.01.23 Name Laura Jahn Signature 

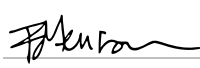
Co-author statement

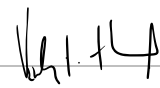
PhD student Laura Jahn

Date of birth 04.07.1993

Date 15.02.23 Name *Rasmus. K. Rendsvig* Signature 

Date 15.02.23 Name *Alessandro Flammini* Signature 

Date 15.02.23 Name *Filippo Menczer* Signature 

Date 15.02.23 Name *Vincent F. Hendricks* Signature 

Appendix II

Data Collection Approval

UNIVERSITY OF COPENHAGEN
FACULTY OF HUMANITIES

Approval of collection and processing of personal data in the research project:

"Exploration of Coordinated Behavior on Twitter"

File number: 514-0112/22-400

PhD Student, Laura Jahn, Department of Communication, has applied for approval of the processing of personal data in connection with the above project on the 3 March 2022

The application shows that the primary purpose of the project is research.

The project will process data from 1.000.000 people.

The Faculty Secretariat hereby approves the processing of personal data in the project. The approval emphasises the fact that the processing of personal data in the project is accordance with the rules of the General Data Protection Regulation, Regulation 2016/679 on the protection of natural persons with regard to the processing of personal data. Furthermore, importance is assigned to the researchers assessment that participating in the project does not result in a high risk to the data subjects participating.

The approval is valid from 03/03/2022 until 31/05/2023.

If the approval is not extended, the personal data must be erased, anonymised, destroyed or archived before that date. See below for further details on the conclusion of the project on the Researcher Portal at KUNet. The project manager must be aware that, after the expiry of the approval, all

09 AUGUST 2022

FACULTY OF HUMANITIES

KAREN BLIXENS PLADS 8
KØBENHAVN S

DIR +45 35 33 51 33

emil.bozard@hum.ku.dk

processing (including storage) of personal data is a violation of the General Data Protection Regulation.

The project is entered in University of Copenhagen's record of processing activities in research projects and biobanks, which is found in University of Copenhagen's electronic case and document handling system, Workzone. The Danish Data Protection Agency may request a copy of the record at any given time for verification purposes, including control of security incidents.

The approval is subject to the condition that the processing and storage of personal data in the project are in accordance with the rules and procedures described on University of Copenhagen's Researcher Portal. The project manager is responsible for compliance with the rules of the General Data Protection Regulation, including rules on the conclusion of the necessary data processing agreements and on follow-up on these agreements, where relevant. The project manager is also responsible for compliance with the current rules in connection with any disclosure of research data to others, including parties in research collaborations.

It should be stressed that the approval is only an authorisation to process personal data for statistical or scientific purposes in connection with the implementation of the project. The approval thus does not entail an obligation for public authorities, enterprises etc. to supply any data for use in the project.

Any terms and conditions laid down in accordance with other legislation must be observed.

Kind regards,

Emil Zolthan Bozard
Faculty Administration