Tesis Doctoral

Information Dynamics in online political Social Networks: *Recommender Systems and Spread of Disinformation and Polarization*

Pau Muñoz Pairet

Programa de Doctorado en Ingeniería Informática y de Telecomunicación

Dirección:

Fernando Díez Rubio

Alejandro Bellogín Kouki

Madrid, 2025

Resumen

Las redes sociales en línea, como X/Twitter, Instagram, LinkedIn, Blogger y TikTok, son utilizadas diariamente por cientos de millones de personas para conectarse globalmente y compartir información. Su ubicuidad, accesibilidad y capacidades de comunicación en tiempo real las han convertido en fuentes primarias de información, superando a los medios tradicionales. Este cambio ha posibilitado la "autocomunicación", donde los usuarios actúan como medios independientes, eludiendo los filtros editoriales tradicionales para compartir contenido no regulado de manera instantánea. Al mismo tiempo, los algoritmos de recomendación (o sistemas de recomendación, en general) son una parte esencial de estas redes sociales, permitiendo a los usuarios navegar y consumir información de manera resumida pero personalizada.

Esta combinación de autocomunicación, inmediatez, falta de regulación y mediación por sistemas de recomendación ha facilitado la aparición de fenómenos perjudiciales como la propagación de desinformación, la creación de cámaras de eco polarizadas y la amplificación de la influencia política que impulsa la radicalización. Si bien estos problemas han sido ampliamente estudiados desde una perspectiva social, su análisis desde la informática sigue siendo limitado. Esta tesis aborda esta brecha analizando la interacción entre los sistemas de recomendación y estos fenómenos sociales.

Nuestra investigación presenta hallazgos fundamentales relativos a la dinámica de la desinformación y la polarización en las redes sociales en línea. Identificamos un nuevo tipo de red de desinformación caracterizada por alta cohesión, coordinación y homogeneidad lingüística, facilitada por algoritmos de recomendación que aceleran su formación y amplifican su alcance. Además, este trabajo demuestra cómo los sistemas de recomendación basados en aprendizaje profundo exacerban la polarización al agrupar a los usuarios en comunidades ideológicamente alineadas, intensificando las divisiones políticas.

Para mitigar estos problemas, proponemos estrategias de recomendación conscientes del daño. En el caso de la desinformación, una intervención algorítmica basada en promover la diversidad estructural reduce la cohesión de las redes de desinformación sin comprometer la estabilidad de los flujos de información legítimos. En cuanto a la polarización, un algoritmo de recomendación orientado a la diversidad, que utiliza *clustering* jerárquico y la distancia de Jensen-Shannon, conecta comunidades fragmentadas, logrando reducir la polarización hasta en un 50% en escenarios electorales simulados. Estas contribuciones ofrecen estrategias prácticas para mejorar la integridad de los ecosistemas de información en línea y fomentar interacciones digitales más saludables.

Palabras clave: Recomendación, redes sociales, ciencia de redes, polarización, desinformación, influencia, teoría de la información, redes políticas

Abstract

Online social networks, such as X/Twitter, Instagram, LinkedIn, Blogger, and TikTok, are used daily by hundreds of millions of people to connect globally and share information. Their ubiquity, accessibility, and real-time communication capabilities have transformed them into primary sources of information, surpassing traditional media. This shift has enabled "self-communication", where users act as independent media outlets, bypassing traditional editorial filters to share unregulated content instantly. At the same time, recommendation algorithms (or recommender systems, in general) are a key part in these social networks for users, to navigate and consume the information in a summarized but personalized way.

This combination of self-communication, immediacy, lack of regulation, and the mediation of recommendation systems, has facilitated the emergence of harmful phenomena such as the spread of disinformation, the creation of polarized echo chambers, and the amplification of political influence that drives radicalization. While these issues have been widely studied from a social perspective, their analysis from a computer science standpoint remains limited. This thesis addresses this gap by analyzing the interplay between recommendation systems and these social phenomena.

Our research presents key findings on the dynamics of disinformation and polarization. It identifies a novel type of disinformation network characterized by high cohesion, coordination, and linguistic homogeneity, facilitated by recommendation algorithms that accelerate their formation and amplify their reach. Additionally, this work demonstrates how recommendation systems based on deep learning exacerbate polarization by clustering users into ideologically aligned groups, intensifying political divides.

To mitigate these issues, we propose harm-aware recommendation strategies. For disinformation, an algorithmic intervention based on promoting structural diversity reduces the cohesion of disinformation networks without compromising the stability of legitimate information flows. For polarization, a diversity-oriented recommendation algorithm, leveraging hierarchical clustering and Jensen-Shannon distance, bridges fragmented communities, achieving a reduction of polarization by up to 50% in simulated electoral scenarios. These contributions provide actionable strategies for improving the integrity of online information ecosystems and fostering healthier digital interactions.

Keywords: social networks, political networks, polarization, disinformation, mitigation, recommender systems, information theory

Dedication

To my mother,
the brightest mind I have ever known,
whose path to higher education was denied by the times she lived in.
Yet, through her wisdom, she became my greatest teacher.
Without her, I would not have come this far.
This work is not mine alone—it belongs to both of us.

Acknowledgements

I would like to deeply thank my mother for instilling in me a love for knowledge, a passion for discovery, and a critical sense to analyze the world. My journey on this project would never have been successful without her influence.

I am equally grateful to my father for teaching me the value of a strong work ethic as solid as steel—a person capable of enduring the hardships of the journey without complaint in pursuit of a greater goal. I would not have started without my mother's influence, nor would I have finished without my father's guidance.

I also want to thank my advisors, Alejandro and Fernando, for guiding me and teaching me how to conduct research. Scientific investigation demands rigor, precision, and a critical mindset—qualities not always present in other fields. Being a researcher requires patience, resilience, and tenacity, alongside curiosity. These human qualities were developed and fostered in me thanks to the lessons I received from my advisors, and I am deeply grateful for this, as what I have learned here will live in me throughout my life.

To my dear Lucía, who accompanied me through much of this process, understood me, and helped me think outside the box—it has been an honor to share part of this journey with you.

I would also like to extend my gratitude to the entire teaching staff of the "Escola Politècnica Superior" at the Universitat de Girona: Teo, Lluís, Eusebi, Quim, Joan, Mei, Mateu, Marta, Antoni, and the rest of the professors from the Bachelor's Degree in Computer Engineering. Polytechnic schools build engineers, and the skills I acquired during my time in Girona enabled me to get far in life, keep learning, and reach this point.

My thanks also go to my friends and my colleagues at the Information Retrieval Group for their encouragement, inspiration, and support throughout this process. And my friends and partners who helped and provided insights and overall good conversation including Raul, Carlos, Javier, Diego, Sofía, Juan, Borja, Jaime, Rodrigo and many others.

Finally, I wish to express my gratitude to all those who came before me, to all the authors cited in this work, and to those who paved the way for the start of this thesis. Standing on their shoulders has allowed me to travel this path.

Contents

List of Figures

List of Tables

Part I

Introduction and Context

“Every tool carries with it the spirit by which it has been created.”
— Werner Heisenberg, , Physics and Beyond

Chapter 1

Introduction

1.1 Motivation

The intersection of information technology and political phenomena within online social networks constitutes a critical field of study in the digital age. Living involves constant interaction and transmission of information, a dynamic that has significantly evolved with technological advancements such as the invention of the printing press and, more recently, the expansion of the Internet and social media. These developments have democratized access to information, enabling more participatory and multifaceted communication. However, they have also facilitated the emergence of complex and dynamic information ecosystems as networks of interconnected information sources such as media outlets, individual social media accounts, and other entities that interact and evolve over time. These ecosystems, naturally emerged from online social networks (ONS), are characterized by their organic and adaptive nature, with new accounts appearing, others disappearing, and continuous interactions shaping the flow, prioritization, and consumption of content. Recommendation systems play a central role in these spaces by filtering and prioritizing the content users consume, thereby influencing opinion formation and political dynamics. Recommender systems represent the engines that move information in these spaces.

In this context, recommendation systems not only optimize user experience by personalizing content, but also significantly shape ideological landscapes and political interactions online [ASIM18]. The ability of these systems to create echo chambers, where individuals are primarily exposed to information that reinforces their preexisting beliefs, has intensified political polarization and facilitated the spread of disinformation [Par11]. Additionally, the partial removal of traditional editorial controls has increased the vulnerability of social platforms to information manipulation, exacerbating social and political divisions.

The present thesis conducts an in-depth analysis of the role of recommendation systems in the formation of the most critical political phenomena within online social networks: disinformation spread and polarization. Employing a multidisciplinary approach that integrates theories from communication and political science from the perspective of computer science, our research seeks to develop a comprehensive framework to understand the complex interactions between information management, opinion formation, and political polarization in the massive online environments represented by online social networks. By addressing the existing gap in the current literature, our study intends to understand these phenomena and to offer strategies to mitigate the negative effects of political digital information ecosystems, thus promoting a more diverse and cohesive public discourse in contemporary society.

1.2 Thesis Structure

The thesis adopts an integrative approach to analyze these phenomena, focusing on information communication networks that include journalists, independent communicators, mainstream and alternative media outlets, and political actors, especially in electoral dynamics. This comprehensive analysis encompasses a variety of perspectives and sources within the information ecosystem, thereby providing a holistic understanding of the interplay between different actors and their influence on the information landscape.

Spain serves as the primary case study for this research, primarily due to the country's pronounced political polarization and the complex nature of its media landscape [RSM22]. The choice of Spain as a focal point offers a unique opportunity to explore these phenomena within a specific socio-political context, providing insights into the nuances of information dynamics in a highly polarized environment. This focus allows for a detailed exploration of the specific characteristics and impacts of digital information ecosystems in Spain, while also offering the potential to extrapolate findings to broader, global contexts.

This thesis is structured into five parts including the introduction and conclusions, each dedicated to exploring the dynamics of digital information ecosystems, with a specific focus on polarization and disinformation as the primary negative social phenomena stemming from the rise of social media. These phenomena are deeply intertwined with the role of recommender systems, which both influence their proliferation and offer potential avenues for mitigation. Throughout the thesis, information theory serves as the overarching analytical framework, providing a robust foundation for understanding and addressing the complexities of these issues, as information theory provides the tools to quantify uncertainty, diversity, and divergence in communication patterns, enabling a precise and theoretically robust measurement of polarization that accounts for the dynamic and multifaceted interactions within online communities.

The first part of the thesis introduces the **fundamental concepts** related to recommender systems, social networks and overall online political behavior, to provide an explicit **context** for the work, complemented with the **research methodology** followed throughout. In the second part of this thesis, the focus shifts to a clear definition of the various processes deemed fundamental within **digital information ecosystems**. That part aims to dissect and understand the intricacies of these ecosystems by analyzing both mainstream media and those considered alternative or disseminators of disinformation. The analysis explores the particular characteristics of political processes in social networks, recognizing their significant influence in shaping public discourse and opinion.

Hence, this second part of the thesis aims to identify and define the **key actors** responsible for the dynamics within these digital ecosystems. This includes a comprehensive examination of the **roles** played by political parties, media outlets, activists, and opinion leaders. Each of these entities contributes to shaping information landscapes in distinct ways. Political parties, for instance, use digital platforms for campaigning and propaganda, while media outlets serve as the primary conduits for news dissemination and public engagement. Activists and opinion leaders, on the other hand, leverage these platforms for advocacy, mobilization, and influencing public opinion. Also, this part offers a detailed exploration of the **activist figure**, examining their transformation from physical to digital spaces, evolving into what can be termed as the "activist influencer". Furthermore, it investigates the interactions between diverse user profiles on social networks surrounding major political events, such as elections, thereby enabling a comprehensive user modeling framework. This approach provides a nuanced understanding of the roles,

influences, and interactions of different actors within these online political ecosystems.

The third part of the thesis focuses on the interactions among these actors, aiming to identify a complete **information dynamic within social media communities** characterized by the dissemination of political content. Here, scientific and algorithmic approaches characterized by the use of statistical techniques and algorithmic simulation are employed to detect and measure these phenomena, especially focusing on the formation of disinformation networks as an evolution from individual disinformation-spreading actors on social platforms and the emergence of polarization in online social networks. In the study of polarization, we propose a **novel algorithmic approach** to measure its extent and intensity. Our algorithm is grounded in information theory, which serves as the most appropriate analytical framework for capturing the complexities of polarization, as previously stated.

This third part concludes with an **analysis of the role of recommendation systems** in modulating disinformation spread and polarization, that is, fostering or preventing these dynamics in a natural manner, expanding the scope of our study to encompass a holistic information flow model within these networks. This part thus provides an in-depth understanding of how recommendation systems influence the rise and modulation of political content dissemination and interaction within online communities.

In the fourth part of the thesis, the focus shifts toward presenting effective **algorithmic intervention strategies** designed to mitigate the adverse phenomena discussed in the preceding chapters. This part builds upon the dynamics and principles previously identified, proposing specific algorithmic interventions in the form of custom recommendation strategies grounded on information theory, that could serve as **countermeasures to reduce the spread and impact of disinformation and polarization**.

Finally, the last part of the thesis summarizes the main conclusions derived from our work, together with the identified limitations in our study, that could be addressed in the future.

1.3 Research Goals and Questions

The following research goals outline the primary objectives of this thesis, which is to understand the dynamics of information in online political social networks. Each research goal, in order to be achieved, is comprised of a set of general research questions which are answered in one or several chapters in the thesis. Each chapter focuses on one specific research topic and presents a subset of specific research questions related to the topic.

The first research goal focuses on **developing a comprehensive characterization** of the main elements that constitute an information ecosystem within a political context. This goal includes analyzing the opinion formation processes in social networks, with particular emphasis on their dynamics during electoral contexts. The aim is to understand how opinions evolve over time and what characteristic shifts occur during these periods. Additionally, it seeks to identify the key factors and actors that have the greatest potential to mobilize public opinion in such contexts, laying the groundwork for deeper investigation in subsequent research goals.

- **RG1:** Develop a valid definition of a political information ecosystem for this study, involving all of its main elements and thus allowing us to study their information dynamics.
  - **RG1.RQ1:** Which are the main information disseminators in the political context on microblogging social networks and how they develop their activity?

  - **RG1.RQ2:** What motivates political active users to perform online activism and promote narratives?
  - **RG1.RQ3:** How do this main information disseminators interact between each other during political processes?

This first research goal is fulfilled in the introductory chapters, precisely Chapters 1 to 6.

In order to effectively analyze political information ecosystems, it is essential to identify and characterize their main components. The second research goal is dedicated to conducting a **detailed analysis of the risks inherent in political information ecosystems**. Specifically, it aims to investigate how disinformation networks compare to legitimate journalism networks in terms of their structure and content patterns, as well as the mechanisms driving societal polarization in political contexts. This goal involved developing computational tools to analyze these phenomena, enabling a comprehensive exploration of user roles, network polarization, and the underlying factors contributing to its emergence across the full timeline of electoral processes.

Understanding the risks associated with political information ecosystems, particularly disinformation and its societal impact, requires a detailed and methodical analysis. To achieve this, our second research goal focuses on these challenges through a comprehensive investigation guided by the following approach: we analyze electoral processes across their full timeline—pre-campaign, campaign, election day, and post-campaign—examining the behavior and influence of disinformation disseminators in comparison to legitimate journalists. These legitimate journalists, defined as professionals employed by established media outlets and actively engaged in sharing current news on social media, serve as a baseline for contrast. This comparative framework allows us to explore the dynamics of disinformation against credible information dissemination practices within these ecosystems.

- **RG2:** Conduct an in-depth analysis of the main risk phenomena in political information ecosystems. Develop a set of tools for their computational analysis.
  - **RG2.RQ1:** How do disinformation networks behave compared to legitimate journalism networks according to network structure?
  - **RG2.RQ2:** How do information content patterns influence the structure of disinformation networks?
  - **RG2.RQ3:** How can polarization in a social network in a political context be measured accurately?
  - **RG2.RQ4:** What are the main factors contributing to the emergence of polarization?
  - **RG2.RQ5:** What is the role of different categories of users in the formation of polarization?

RG2.RQ1 and RG2.RQ2 are answered in Chapter 7. RG2.RQ3 and RG2.RQ4 are answered in Chapter 8. RG2.RQ5 is answered in Chapter 10.

Recommendation systems can significantly influence the spread of information and the formation of opinions within political ecosystems. The third goal explores their **impact and potential mitigation strategies**:

- **RG3:** Analyze the role or contribution of recommendation systems in promoting these phenomena. Propose mitigation strategies through these systems.

- **RG3.RQ1:** Which approach or type of recommendation system contributes the most to the formation of disinformation in the form of networks?
- **RG3.RQ2:** Which approach or family of recommendation system contributes the most to the rise of polarization?
- **RG3.RQ3:** How can recommendation systems help to mitigate these phenomena?
- **RG3.RQ4:** How do the proposed mitigation strategies affect recommendation accuracy?

RG3.RQ1 is answered in Chapter 9. RG3.RQ2 is answered in Chapter 10. RG3.RQ3 and RG3.RQ4 are answered in Chapters 11 and 12.

1.4 Research Results

This thesis has served to profoundly understand the dynamics of information in micro-blogging social networks. The objective of the thesis has been, therefore, to comprehend the underlying dynamics of processes and phenomena such as political polarization and the distribution of disinformation on social networks, as well as the role of content distribution systems in these processes. Consequently, the main outcome of the research has been the elucidation of these phenomena in the form of academic publications in some of the leading journals in the field.

However, it should be noted that the research process has involved a substantial amount of script development and small-scale software projects to automate the experimental part of this research, address research questions through data or algorithmic analysis of recommendation systems, among other tasks. Some of these projects have been consolidated into software solutions ready to be reused to facilitate new research.

1.4.1 Publications

Table 1.1 summarizes the publications derived from this doctoral thesis, including those that have been published as well as those currently under review.

1.4.2 Software

Table 1.2 provides an overview of the software tools developed as part of this doctoral research. Each tool has been designed to facilitate distinct aspects of social network analysis, addressing key challenges in the examination and understanding of information dynamics on the platform X. These tools were employed throughout the research process, contributing essential insights and data analysis capabilities to the findings presented in this thesis.

Table 1.1: Details of publications derived from this thesis.

Contribution Title	**Journal**	**Year**	**Chapter**
El derecho a la información en las redes sociales	Agenda 2030: teoría y práctica: una mirada constructiva desde la academia	2023	Chapters 1-2
The birth of the influencer-activist, a case study from Spain	Under Review	–	Chapter 5
Networks of Influence: Dissecting User Roles in Spain's Online Political Arena (2011-2019)	Under Review	–	Chapter 6
Modeling disinformation networks on Twitter: structure, behavior, and impact	Applied Network Science 9 (1), 4	2024	Chapter 7
Quantifying polarization in online political discourse	EPJ Data Science 13 (1), 39	2024	Chapter 8
The Role of Recommendation Algorithms in the Formation of Disinformation Networks	Under Review	–	Chapter 9
The Roots of Polarization in Online Social Networks	Under Review	–	Chapter 10
Disarming disinformation networks: breaking disinformative echo chambers with structural text diversity in recommendations	Under Review	–	Chapter 11
DepolarizeIT: An Information-theoretic approach to reduce polarization in Online Social Networks	Under Review	–	Chapter 12

Table 1.2: Details of Selected Software Tools.

Software Tool	**GitHub Repository**	**Description**	**Chapters**
X-Track	`https://github.com/paumnz/X-Track`	A tool designed to provide a general descriptive analysis of phenomena on the social network X. It analyzes sets of posts, offering an initial orientation for research by enhancing data comprehension.	6, 7, 8, 9
Polaris	`https://github.com/paumnz/Polaris`	A tool for analyzing the state of polarization on the social network X. It implements all current state-of-the-art metrics, including the metric developed in Chapter 7 of the thesis, allowing researchers to apply individual or multiple metrics to sets of posts.	7
Grafis	`https://github.com/paumnz/Grafis`	A tool for simulating information diffusion on the microblogging network X through the modeling of post recommendation algorithms and user reactions, based on real data.	8, 9, 10, 11

Chapter 2

Fundamental Concepts

2.1 Introduction

This thesis is grounded in the field of computer science, yet it adopts a highly interdisciplinary approach, drawing extensively from social sciences, particularly within the domain of computational social science. By integrating methodologies and concepts from both disciplines, the research seeks to address complex phenomena that lie at the intersection of technology and society.

Given this interdisciplinary scope, the thesis examines social science constructs such as disinformation, polarization, and social network dynamics, which are critical for understanding the broader implications of computational models and algorithms in societal contexts. To ensure clarity and establish a robust theoretical foundation for the analysis, this chapter provides precise definitions and explanations of these key concepts, thereby facilitating the reader's comprehension of the axiomatic basis upon which the study builds.

2.2 Social Networks

A social network is a structured system of relationships and interactions among individuals, groups, or entities within a defined social context. These networks are characterized by complex patterns of interdependence and exchange, where social ties—whether familial, professional, or communal—facilitate the flow of information, resources, and influence [SSRM22]. Unlike digital or online networks, traditional social networks rely on direct, face-to-face interactions and physical proximity, shaping collective behavior and identity within specific cultural, geographic, or institutional settings [SSRM22].

2.3 Online Social Networks

An online social network (OSN) is a digital platform that enables users to establish and manage virtual connections with other individuals, organizations, or groups. These networks facilitate communication, content sharing, and community-building across a wide geographic and social spectrum, allowing members to exchange information, ideas, and personal updates in real-time. OSNs often incorporate various multimedia elements, messaging systems, and algorithms that curate personalized content, fostering engagement and connectivity among users [SLL21].

Micro-blogging platforms, a subset of online social networks, are specifically designed for sharing brief, real-time updates, often in the form of short text posts, images, or links. Among these, X (formerly Twitter) is the most widely used platform, known for its rapid dissemination of information and its role in shaping public discourse and real-time news updates [Tuf17, Par11].

2.4 Information Networks

Information networks in online social platforms are the structural frameworks through which information flows and circulates among users. The network's structure—defined by how users are connected and interact—plays a key role in modulating how information moves, spreads, and reaches different parts of the network. Certain users may act as hubs or connectors, influencing the speed and reach of information dissemination across the network [WD17].

These networks are not uniform; they exhibit distinct characteristics such as efficiency, connectivity, and information distance, all of which can be measured and analyzed to understand how effectively information travels within them. Metrics such as path length, clustering, and centrality offer insights into the network's ability to facilitate rapid information exchange or sustain long-lasting communication chains. As a result, studying information networks provides a nuanced understanding of how structural properties impact the visibility, reach, and influence of information within online social environments [WD17].

2.5 Information Disseminators

Information disseminators represent a critical subset of actors within social networks, characterized by their primary function of spreading information to wide and diverse audiences. These entities—-ranging from individual influencers and content creators to organizations and automated accounts—-are instrumental in shaping the flow of information within digital spaces. Unlike passive participants, information disseminators actively curate, amplify, and propagate content, often serving as intermediaries that connect original sources to broader networks [Xin12].

The role of disseminators is particularly significant in the context of online discourse, as their actions can magnify the visibility of specific narratives, shape public debates, and influence collective behaviors. By leveraging strategies such as consistent posting, network bridging, and engagement optimization, disseminators maximize their reach and impact. This ability to direct information flow makes them pivotal actors in the dynamics of social networks, especially in contexts where the rapid spread of information—or misinformation—can have profound societal implications [BMA15, VRA18].

2.6 Political Social Networks

Political social networks are a specialized subtype of social networks focused on the domain of political discourse and engagement. These thematic networks are distinct from general social networks due to their unique dynamics, which are shaped by the nature of political interaction [EVHH11]. Within political social networks, various actors—including politicians, political parties, and advocacy groups—compete to capture the attention of large audiences, often constructing dense and expansive networks around themselves. These networks serve as platforms for these actors to broadcast and disseminate targeted

narratives, aiming to influence public opinion, mobilize supporters, and assert their presence within the political landscape [EVHH11, OCLB19].

2.7 Activism and Digital Activism

Activism is the intentional effort to bring about social, political, economic, or environmental change through organized actions, whether at the community level or on a larger scale. Traditionally, activism involves direct participation in demonstrations, campaigns, or movements that advocate for specific causes, demanding physical presence, resources, and often sustained commitment from participants [Yan16, DD19b].

Digital activism also known as cyber-activism, however, represents a distinct form of activism conducted through online platforms. Unlike traditional activism, digital activism allows individuals to engage in social and political movements with minimal physical or financial commitment, often through simple actions like posting hashtags on platforms such as X or participating in widespread digital campaigns [KU18]. This form of activism enables large-scale engagement, as the low barrier to entry attracts vast numbers of participants who, collectively, wield significant influence over public discourse. Digital activists can thus mobilize mass support and attention with relative ease, contributing to the rapid spread of messages and the potential for substantial impact without requiring physical involvement [KU18].

2.8 Propaganda and Computational Propaganda

Conventional propaganda refers to the deliberate dissemination of information, ideas, or rumors to influence public opinion and behavior in favor of a particular ideology, policy, or group. It typically employs various media forms, including print, radio, television, and posters, aiming to reach and sway audiences through controlled messaging that appeals to emotions, biases, or established beliefs [Ell21].

Computational propaganda, in contrast, is a modern evolution of these techniques, leveraging digital platforms and algorithm-driven technologies to manipulate public opinion on a larger, more targeted scale. Through the use of automated accounts (bots), data analytics, and artificial intelligence, computational propaganda enables the creation and amplification of persuasive messages across OSNs, often with precision targeting based on user data. This form of propaganda not only disseminates messages rapidly but also simulates grassroots support, distorts public discourse, and influences perceptions on political, social, or economic issues, effectively reaching and impacting large audiences in real-time [WH18].

2.9 Political Bots

Political bots are automated accounts on online social networks—software applications designed to simulate human behavior and interact with users on a large scale. The computational nature of these platforms allows for scalable, automated actions that can be deployed en masse and with minimal supervision. As public attention increasingly shifts from traditional news sources to online social networks, the presence of political bots has emerged as a natural development in digital political engagement [NHKE17, WH18].

Political bots are specifically programmed to generate engagement around particular political narratives, often amplifying specific messages to shape public discourse or influence perceptions [WH18].

They began to appear around 2011 and have since played significant roles in major political events, notably the 2016 U.S. elections and the Brexit referendum, where they contributed to the proliferation of targeted narratives on a massive scale [WH18, NHKE17].

2.10 Influencers

Influencers are individuals who, through their social media presence, hold significant sway over their followers' opinions, purchasing behaviors, and engagement with various topics. Leveraging personal branding, expertise, or charisma, influencers curate content that resonates with specific demographics, creating communities centered around shared interests such as lifestyle, fashion, technology, or social causes [vDD21].

Their influence is often rooted in the trust and relatability they establish with their audience, allowing them to promote products, ideas, or campaigns with substantial impact. As a result, influencers have become valuable assets in digital marketing and political communication alike, able to amplify messages rapidly across platforms. By shaping trends and public discourse, influencers bridge the gap between traditional media and personal, direct communication with followers, thereby enhancing their reach and persuasive power in digital society [CR20].

2.11 Information Dynamics

Information dynamics in online social networks refer to the processes and patterns through which information is exchanged, spread, and evolves over time within the platform. At the core of these dynamics is the continuous exchange of information among users, facilitated by the specific actions and interactions permitted by the platform—such as posting, sharing, commenting, or liking [SLL21].

In addition to user behavior, information dynamics are heavily influenced by the platform's underlying systems, such as recommendation algorithms and search engines, which determine the visibility and reach of content. These systems actively shape how information flows by curating what users see, suggesting content based on engagement patterns, and prioritizing certain posts or profiles over others. Together, users, platform functionalities, and these information-moving systems create a complex ecosystem where information is constantly propagated, amplified, and filtered, defining the broader dynamics of information exchange within the network [DZ23].

2.12 Disinformation and Misinformation

Disinformation refers to the intentional creation and dissemination of false or misleading information with the purpose of deceiving an audience. It is often used as a strategic tool to manipulate public opinion, destabilize societies, or influence political outcomes by spreading narratives that serve specific agendas. Disinformation is typically crafted to exploit existing biases, foster confusion, or sow distrust, making it a powerful means of shaping perceptions on social, economic, or political issues [SHW+18, KM22].

Misinformation, on the other hand, involves the sharing of inaccurate or false information without intent to deceive. Often spread by individuals who believe the content to be true, misinformation can arise from misunderstandings, misinterpretations, or a lack of verification. While it lacks the malicious intent of disinformation, misinformation can still contribute to public misunderstanding and amplify

confusion, especially when it spreads widely on social or traditional media. Both disinformation and misinformation can profoundly impact public knowledge and decision-making, albeit through different mechanisms and motivations [SHW+18].

2.13 Affective Polarization

Affective polarization refers to the deepening emotional divide between individuals aligned with opposing political groups, characterized by strong feelings of distrust, dislike, or even animosity toward members of the other side. Unlike traditional political polarization, which centers on differences in ideology or policy positions, affective polarization is rooted in emotions and social identities [XWQ+22]. Individuals experiencing affective polarization not only identify strongly with their own group but also perceive members of opposing groups as fundamentally different, often viewing them with suspicion or moral disdain [LRT22].

This emotional rift can lead to a decreased willingness to engage with or understand opposing perspectives, as well as a heightened sense of loyalty to one's own group. Affective polarization shapes social relationships and interactions, influencing people's trust in others, openness to diverse opinions, and even social connections. This trend is a significant driver of societal fragmentation, as it fosters a hostile and divided public landscape where empathy and compromise become increasingly scarce [LRT22, HT12].

2.14 Disinformation Actors and Networks

Disinformation actors are individuals or entities who intentionally spread false or misleading information to serve a political purpose. These actors can include a wide range of participants—such as journalists, influencers, activists, automated bots, or any individual—who strategically disseminate disinformation to sway public opinion, promote specific narratives, or advance particular agendas. Political motives are central to the activities of these actors, as they often seek to influence perceptions on contentious issues or support ideologically driven goals [AAB23].

Disinformation networks refer to the coordinated or loosely connected groups of disinformation actors who amplify these narratives through organized campaigns. These networks may operate with formal coordination or arise organically from informal connections, yet they share the goal of maximizing the reach and impact of their messages. By leveraging the collective efforts of multiple users, disinformation networks can significantly enhance the visibility and persuasiveness of their narratives, creating widespread influence that would be challenging for isolated actors to achieve alone [MC20, MV17, AAB23].

2.15 Echo Chambers

Echo chambers are environments, typically within social networks or online platforms, where individuals are primarily exposed to information, opinions, and beliefs that reinforce their existing views. Within these spaces, users interact mainly with others who share similar ideologies, interests, or perspectives, leading to a narrowing of viewpoints and limited exposure to diverse or opposing opinions. This selec-

tive exposure often results from personalized algorithms or intentional engagement patterns, creating a feedback loop that amplifies confirmation bias [RARFN22, EVHH11].

In echo chambers, differing perspectives are either filtered out or actively discredited, leading to a heightened sense of validation for the group's beliefs and a resistance to critical or alternative viewpoints. Over time, echo chambers can contribute to polarization, as individuals become more deeply entrenched in their views and less receptive to dialogue or compromise with opposing perspectives, ultimately shaping public discourse and influencing societal attitudes on a broad scale [RARFN22, XWQ+22].

2.16 Recommender Systems

What primarily characterizes the communication environment on social networks, beyond the immediacy, the large volume of data, or even the brief content of the messages and information units, is that this information traffic and circulation are mediated by "autonomous" algorithms primarily based on the principles of artificial intelligence. These automatic systems for filtering and transporting information to the user for consumption are divided into two sub-families: web search engines, with Google being the most well-known, and content recommendation systems. Additionally, large language models (LLMs) like ChatGPT or Claude are gaining ground and consolidating as a third sub-family in this regard in the last years. These AI-driven systems are a cornerstone in shaping user experiences by determining what content is seen and prioritized, further influencing the information landscape and potentially exacerbating issues like misinformation and polarization.

A search engine is a software system that locates information on the World Wide Web. It operates by crawling web pages, indexing their content, and using a varied set of information retrieval algorithms to respond to user queries with a ranked list of relevant results. The history of search engines is a part of the history of Internet development. It started with the Archie engine in 1990. Archie was created by Alan Emtage, Peter Deutsch, and Bill Heelan and is considered the first search engine of the internet [SFK11]. It searched files available on public FTP servers and helped the users find the files they sought. In 1993, Steve Foster and Fred Barrie released Veronica and Jughead, search engines for Gopher systems [Rau96]. Veronica (Very Easy Rodent-Oriented Net-wide Index to Computerized Archives) and Jughead (Jonzy's Universal Gopher Hierarchy Excavation And Display) helped users to search menus and files on Gopher servers, which improved information search in the early Internet period [Rau96, AR11].

The mid-1990s marked a pivotal era in the evolution of search engines, characterized by significant technological advancements and the emergence of several groundbreaking platforms. During this period, search engines evolved from rudimentary tools to more sophisticated systems capable of indexing and retrieving vast amounts of information. Among these developments was the creation of Google in 1998 by Larry Page and Sergey Brin, which revolutionized web search with its PageRank algorithm, delivering highly relevant results by analyzing the importance of web pages based on their link structure [BP98].

In this orbit, 1994 saw Brian Pinkerton's creation of WebCrawler, the first search engine to index entire web pages and allow users to search for any word. The same year, Michael Loren Mauldin developed Lycos, one of the first commercial search engines to use advanced mathematical techniques for website categorization [AR11]. AltaVista, launched in 1995 by Digital Equipment Corporation's research team, offered unparalleled speed and the ability to index numerous web pages while pioneering natural language searches for more intuitive user interactions. In 1996, Eric Brewer and Paul Gauthier introduced Inktomi, which equipped search engines like HotBot to manage large-scale data efficiently [SFK11].

A recommender system, on the other hand, is a software tool that suggests items to users based on their preferences and behaviors. It analyzes data such as user ratings, interactions, and profiles to predict and recommend products, services, or content that the user is likely to find appealing [KLPC22].

Recommendation systems represent the next level in information retrieval and user experience personalization. Unlike search engines, recommendation systems establish a personalized and automatic relationship between the application and the user. In online search engines, information—though potentially biased—is directly queried by the user; that is, the user actively accesses the search engine and performs searches (initiating the communication). In recommendation systems, an automatic relationship is established. In other words, the recommendation system can initiate communication with a user who only needs to access the application and wait for the content to be proposed to them. This is particularly evident in applications such as online social networks, where users access the platform and expect information to be presented in a "news feed" or a "list of recommended contacts" to connect with. The user typically has little control over how the content recommendation algorithm operates in these contexts. Many users with little or no technical knowledge are unaware of the recommendation dynamics, sometimes not even realizing that the content they consume is recommended by an automated system whose logic goes beyond the simple chronological presentation of information.

Regarding its technical aspects, nowadays, recommender systems can be broadly categorized into three main types: content-based, collaborative filtering, and hybrid methods. Content-based recommender systems generate recommendations based on the attributes of items and the user's past interactions with similar items. This approach builds a profile for each user based on the items characteristics they have previously rated positively [dGLM$^+$15]. One of the advantages of this strategy lies in its ability to recommend new items that are similar to what the user has liked before, thus effectively addressing the cold-start problem for items. Collaborative filtering (CF), on the other hand, makes recommendations by finding users who have similar preferences [RD22, RRS11]. There are two subtypes of CF: user-based and item-based. User-based CF recommends items that similar users have liked, while item-based CF recommends items similar to ones the user has liked. This method does not require item attribute data, making it versatile but susceptible to the cold-start problem for new users and items and sparsity issues in the user-item interaction matrix. Hybrid methods combine the strengths of content-based and collaborative filtering approaches to mitigate their limitations. These methods can be integrated in various ways, such as a weighted combination, sequential combination, or even switching between methods depending on the context [Bur02]. Hybrid systems often show improved performance and accuracy, leveraging the benefits of both underlying methods. In modern software solutions, hybrid approaches are commonly found on recommender systems [KLPC22].

Though recommender systems have existed both in research and industry since the 1990s, significant advancements have been made in the field since the recent developments and subsequent popularization of machine learning technologies, particularly with the integration of deep learning techniques [KLPC22]. Deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) or Graph Neural Networks (GNN), have been employed to capture non-evident user-item interactions and temporal patterns in data [KLPC22]. Additionally, techniques like matrix factorization, including Singular Value Decomposition (SVD) and its variants, have been extensively used to address issues of scalability and sparsity in collaborative filtering [Hut09, Bur02, KCN$^+$18, RD22].

The field's current state is dominated by deep learning-based approaches, with methods that employ

neural network architectures to model intricate user-item relationships more effectively [RD22]. Applying attention mechanisms within these models has further enhanced their capability to focus on the most pertinent parts of the data, improving recommendation quality.

In social networks and news applications, recommender systems face their particular challenges and opportunities. Social networks provide rich user interaction data, which can be leveraged to build precise user-profiles and enhance recommendation accuracy. Graph-based collaborative filtering and social influence models have shown promising results in this domain by effectively utilizing the network structure and user interactions. In news applications, the continuous influx of new content and the need for real-time recommendations pose significant challenges. Hybrid methods that combine content-based filtering with collaborative approaches are particularly effective here, as they can quickly adapt to new articles while refining recommendations based on user interaction data [RD22].

2.17 Recommendation Networks

Recommendation networks are the structured web of connections and interactions shaped by the recommendations generated by algorithms on social media platforms. These recommendation systems are designed to influence the content that users consume, presenting them with posts, media, or profiles aligned with their interests and engagement history. By determining what users see, recommendation systems play a pivotal role in shaping their exposure to ideas, topics, and other users, effectively guiding the formation of connections within the network [Ast21].

In micro-blogging platforms, for example, recommendation networks emerge not only through content suggestions but also by recommending new contacts to follow, thus fostering networks based on shared content interests [AK12]. Within the context of this thesis, a recommendation network refers to the network structure generated at a specific moment by the recommendations made by a system. This structure is based on the assumption that users will create connections by engaging with the recommended posts or accounts, shaping their social network in alignment with the algorithm's output at that point in time [AK12].

2.18 Algorithmic Intervention Strategies

Algorithmic intervention strategies refer to the deliberate reprogramming, adaptation, or comprehensive redesign of algorithms responsible for content distribution on social media platforms. These strategies acknowledge the pivotal role that algorithms play in shaping the structure and interactions within social networks, as they influence what content users see and how connections are formed [GSM+22].

Concerns about issues like disinformation, polarization, and harmful content have led to growing scrutiny of these algorithms' impact on public discourse and user behavior. Algorithmic intervention strategies aim to mitigate these risks by introducing mechanisms such as content moderation tools, verification systems to flag or validate news, and content diffusion blockers to limit the spread of illegal or harmful information [BNS18]. By implementing these adaptations, social platforms seek to curb the negative effects that can emerge within online networks, promoting a safer and more reliable digital environment while addressing pressing societal concerns about the impact of algorithm-driven content curation [FENKW22].

Chapter 3

State of the Art

3.1 Previous Research and State of The Art

3.1.1 Online social networks and political participation

Social media platforms, like X, can be seen as today's digital analog to the ancient Greek agora. These platforms function as contemporary gathering spaces where individuals globally share ideas and react to current social and political events. Just as the agora was central to civic dialogue in ancient times, X and its counterparts such as Facebook and Instagram, but also Mastodon, GAB, and Parler, offer a real-time pulse on global sentiments [Bou20, AS22]. However, the immediacy of these platforms also poses challenges, such as spreading misinformation and creating echo chambers. Despite these concerns, the rise of social media emphasizes the continuous human need for community and participation in societal conversations [Bou20].

These social media networks are steadily supplanting traditional media outlets as primary sources of information for many individuals. This transition is fueled by the immediacy, accessibility, and user-generated content that these platforms offer. Whereas traditional media often operates on scheduled timelines and involves editorial oversight, social media provides real-time updates and democratizes content creation. However, this shift also presents challenges. The lack of standardized fact-checking on many social platforms can spread all kinds of extreme or even harmful content (including misinformation) [TJLL18, Sun17, CRF+11, CLH+21, GK20]. The personalized algorithms driving content recommendations may also create echo chambers, limiting diverse perspectives [Par11]. Nonetheless, the emergency and consolidation of online social media signifies a profound change in how modern societies consume and interpret news and events.

Thus, this defining characteristic of modern social media platforms empowers users to produce and share their content, moving beyond the confines of traditional, established media [Bou20]. This democratization of information dissemination has transformed the media landscape. No longer are narratives solely shaped by politicians, journalists, and media houses; anyone with internet access can now contribute to the global conversation. This shift has led to a richer compound of voices and perspectives, fostering more vibrant discussions. It is a double-edged sword, in any case. While this *inclusivity* encourages diverse participation [WP13], it also opens the door to misinformation and an increasingly polarized discourse [KvS21]. Nevertheless, the ability for individuals to publish and circulate their content underscores a monumental shift in the dynamics of information exchange in the digital age [Bou20].

Previously named Twitter, X stands out as the principal social media platform for political discourse at the moment. This platform's succinct messaging format and global reach have made it a favored arena for politicians, activists, journalists, and citizens alike to share viewpoints, make announcements, and engage in debates [Sub21]. While other platforms might cater to different facets of social interaction, X's immediacy and broad user base uniquely suit the fast-paced world of political dialogue. The platform's influence is evident in how quickly tweets shape narratives, spark movements, or impact policy decisions [EVHH11, Sub21, Tuf17].

Consequently, given its central role in political discourse, X emerges as an optimal arena for studying polarization [GS11, Sun17]. The platform's vast user base, encompassing a myriad of ideologies and backgrounds, offers a rich compound of viewpoints [WP13]. Nevertheless, the very design of X, which often prioritizes content that evokes strong reactions, can amplify polarized stances [Sun17, VBP21, CRF+11].

Thus, while the extent to which platforms like X directly influence polarization remains contested, their potential to mitigate such divisions is undeniable. The same algorithms that may foster echo chambers can be retooled to promote diverse dialogues, exposing users to broader viewpoints [SLL21, MC21]. Furthermore, these platforms possess the tools to encourage constructive conversations, highlight common ground, and dispel myths or misinformation that can fuel divisiveness. By actively curating content that fosters understanding and leveraging features that promote dialogue over dispute, platforms can transform from being potential *accelerants* of polarization to agents of reconciliation and unity [SLL21].

Building on this premise, it becomes evident that devising strategies designed to identify and mitigate polarization is essential for two intertwined reasons. Firstly, a comprehensive understanding of polarization enables the development of active efforts to combat it, enabling researchers and policymakers to discern which interventions are most effective. By studying the outcomes of these strategies, we gain deeper insights into the mechanics and nuances of division. Secondly, and perhaps more urgently, these strategies are vital to ensure that existing rifts do not deepen. As polarization intensifies, societies risk becoming more fragmented, with increased misunderstanding and reduced collaborative potential. Therefore, by actively pursuing mitigation tactics on platforms like X, we enhance our understanding of polarization and contribute proactively to a more cohesive and harmonious digital discourse.

3.1.2 Disinformation

Disinformation refers to intentionally false or misleading information created and disseminated to harm individuals, communities, or societies, or to achieve economic or political gains [oFND18, WD17]. Unlike misinformation, which is unintentional, disinformation is inherently deliberate and often employs sophisticated strategies to maximize its impact [KCBP21, AAB23].

In the digital age, disinformation exploits the unique affordances of online platforms, leveraging their speed, scale, and personalization features to reach and influence vast audiences [ZSBK19]. Techniques include fabricated stories, conspiracy theories, pseudoscience, and misleading connections, all of which are designed to manipulate public opinion and behavior [KCBP21]. For example, during the COVID-19 pandemic, conspiracy theories about 5G technology and vaccines fueled mistrust and violence, illustrating the tangible consequences of digital disinformation campaigns [AAB23].

Disinformation also exacerbates political polarization by creating echo chambers and filter bubbles, where users are exposed only to like-minded content, reinforcing existing biases and reducing exposure

to diverse viewpoints [WD17]. This dynamic has been linked to election interference and the erosion of democratic trust globally [oFND18, ZSBK19].

Efforts to counter disinformation include educational initiatives to enhance media literacy, technological tools for automated detection, and policy interventions to hold platforms accountable. The European Commission, for instance, has introduced a self-regulatory Code of Practice for online platforms and supported the development of fact-checking tools [KCBP21, oFND18]. Despite these measures, the evolving nature of disinformation requires continuous innovation and collaboration across disciplines to mitigate its impacts effectively [AAB23].

3.1.3 Political polarization

Political polarization, also referred as affective polarization, refers to how citizens feel sympathy towards partisan in-groups and antagonism towards partisan out-groups [Wag21]. In multiparty systems, capturing the affect pattern towards multiple parties is more complex compared to two-party systems [Wag21]. This conceptualization and measure of political polarization in multiparty systems summarize the configuration of feelings towards political parties and their supporters [Wag21].

Other definitions of political polarization focus on the underlying ideological divisions between voters. Capturing the summary ideological divisions between citizens, considering how the policy content of debates shifts [BW19]. They emphasize the shifting meaning of left and right and the efforts of "issue entrepreneurs" in shaping political polarization [BW19].

Furthermore, political polarization can be understood as the expanded ideological gap between political groups and the increased interpersonal separation between individuals [BBF21]. This definition highlights the intergroup dimensions of polarization and the role of intergroup interaction in enabling and enhancing polarization [BBF21].

Overall, political polarization encompasses the affective divide between partisan in-groups and out-groups, the ideological divisions between voters, and the expanded gap and interpersonal separation between political groups and individuals. These definitions highlight the complexity of measuring and understanding political polarization in multiparty systems and the importance of considering both intra-group and intergroup contexts in the study of polarization [BBF21, BW19, Wag21].

Polarization is particularly stark on social media platforms, magnifying societal divisions in a unique digital microcosm. In these online social networks' (particular) context, political polarization can be defined in several ways. One definition focuses on the formation of echo chambers, where individuals are exposed to like-minded opinions and information, leading to increased ideological polarization [MS21, SLL21, Li22, BBF21]. This definition highlights the role of social media algorithms and the structure of online networks in shaping polarization [MS21, SLL21].

Another definition of political polarization in online social networks emphasizes ideological homophily, which refers to the tendency of individuals to connect and interact with others who share similar political beliefs [VBP21, AABM21, GAS$^+$15, ADM$^+$21]. This definition underscores the role of social network dynamics and the formation of homogeneous and segregated ideological groups in contributing to polarization [VBP21, AABM21, GAS$^+$15, ADM$^+$21].

Additionally, political polarization in online social networks can be understood as the affective divide between political groups, where supporters of different parties increasingly dislike and even loathe their opponents [ISL12, VBP21, GAS$^+$15, ISL12, ADM$^+$21]. This definition emphasizes the role of polit-

ical polarization and negative attitudes towards opposing political groups in online interactions[ISL12, VBP21, GAS+15, ISL12, ADM+21].

Furthermore, some definitions of political polarization in online social networks consider the behavioral component of polarization, which manifests in the interactions between individuals [GAS+15]. These definitions highlight the importance of analyzing the patterns of interaction and communication between individuals in understanding polarization [GAS+15].

Overall, in online social networks, political polarization can be defined as the formation of echo chambers and filter bubbles, the tendency towards ideological homophily, the affective divide between political groups, and the behavioral manifestations of polarization in online interactions. Thus, a polarization evaluation algorithm should be able to effectively capture those axioms.

Political polarization in Spain

Spain is a poignant example of the broader global trend of rising political polarization. Over recent years, the nation has grappled with deep-seated divisions on several fronts. The Catalonia independence movement is a prominent manifestation, with sharp divides between pro-independence Catalans and those wishing to remain unified with Spain. This issue has spurred heated debates in Catalonia and across the entire country, revealing divergent views on regional autonomy, identity, and the very fabric of the Spanish state [PMSGA21, B+21, RSAP19, BK22, Ast21].

Furthermore, the Spanish political landscape itself has witnessed fragmentation. Traditional parties like the Spanish Socialist Workers' Party (PSOE) and the People's Party (PP) have seen challenges from newer entities such as Podemos and Ciudadanos, each bringing distinct ideologies to the fore. The rise of Vox, a far-right party, further underscores the diversification of political thought in the country [RSAP19, Ast21].

Social issues, including immigration, the role of the Monarchy in the political system, the remembrance of the civil war and economic disparities, have also contributed to the polarized discourse. The economic crisis and subsequent austerity measures fueled discontent, leading to divergent views on economic policies and governance [RSAP19, Ast21, OA23].

Social media platforms have, in turn, become battlegrounds for these debates in Spain, echoing the sentiments of its populace and sometimes amplifying the extremities. Recognizing and understanding the nuances of Spain's polarization is pivotal for fostering dialogue and seeking potential avenues for reconciliation [RSAP19].

3.1.4 The role of recommender systems

The current mechanics of (social) content recommendation are governed by diverse algorithmic approaches, each bringing its unique perspective to how content is curated and presented to users [ASIM18]. These methodologies can be broadly categorized into user-based, content-based, deep learning, and reinforcement learning algorithms, each prioritizing different types of content and user interactions. The user-based collaborative filtering approach hinges on the principle of similarity among users' interactions and preferences. By analyzing the patterns in which users engage with content, this method recommends content that has been liked or interacted with by users with similar tastes [CLDB18]. The underlying assumption is that individuals who have shared preferences in the past will likely appreciate similar content in the future, thus fostering a personalized content discovery experience. Moving to the content-based

strategy, this focuses on the attributes of the content itself, including metadata like the author and genre, as well as the textual content. This approach tailors recommendations by aligning new content with the user's previously expressed interests in specific types of content characteristics [JSH+21]. For instance, if a user frequently engages with articles from a particular author or genre, the system is likely to recommend similar content, enhancing relevance through the content's inherent properties [JSH+21]. Deep learning systems, based on neural network architectures, delve into the complexities of user behavior and content features to unearth patterns not readily apparent to simpler algorithms. Its strength lies in the ability to process and analyze massive datasets, enabling the identification of nuanced user preferences and subtle content attributes, in the context of OSNs it is done mostly by analyzing text thereby facilitating highly personalized content recommendations [ZYST19]. Lastly, reinforcement learning adopts a dynamic learning approach through continuous trial and error. Recommendations are made, and based on the user's engagement or lack thereof, the system iteratively learns and adjusts its strategy [ACF22].

Collectively, these algorithmic families embody the multifaceted nature of content recommendation systems in social networks. Each brings its lens to bear on the recommendation process, influencing the distribution and visibility of content in ways that can significantly shape user experience and content consumption patterns on these platforms. Naturally, each of these families, due to the inherent definition of the recommendation method they employ, tends to generate a different recommendation network.

The way information is presented to the user, as demonstrated, contributes to their opinion formation process [ER15], including the manipulation of voting in politics by positioning one candidate or option above another in the ranking. This naturally affects information filtering systems, internet search engines, and especially recommendation systems, which are used almost transparently by vast numbers of users. It contributes to the emergence of various collective social phenomena.

Recommender systems in polarization

The creation and reinforcement of echo chambers by recommender systems (RSs) and social networks have become intensely scrutinized due to their potential adverse effects on user experience and societal discourse. Mansoury et al. [MAP+20] conducted a study investigating how feedback loops in RSs amplify popularity bias. Their research simulated user interactions with recommendation algorithms over multiple iterations, employing three specific algorithms: User-based Collaborative Filtering (UserCF), Item-based Collaborative Filtering (ItemCF), and Singular Value Decomposition (SVD). The study utilized a movie dataset to represent user-item interactions and found that these feedback loops significantly intensified popularity bias. This amplification led to several negative consequences, including reduced aggregate diversity, shifted user preferences, and homogenization of user groups. The process diminished the diversity of recommendations and skewed user profiles over time, making them more homogeneous. Notably, this effect was particularly pronounced for minority groups, exacerbating existing biases. Cinus et al. [CMMB22] further explored the role of people recommenders in fostering echo chambers and polarization within social media platforms. They proposed a Monte Carlo simulation framework that combined link recommendation and opinion-dynamics models to assess these effects. Their methodology involved simulating social networks with varying levels of homophily and modularity, using metrics such as the Neighbor Correlation Index (NCI) and Random Walk Controversy (RWC) to measure the presence of echo chambers and polarization. The study tested several state-of-the-art link predictors, including Personalized PageRank (PPR) and SALSA. Their findings indicated that people recommenders

can significantly increase echo chambers, particularly in networks with high initial homophily, although the impact is less pronounced in already segregated networks. They also identified differences in the extent of polarization and echo chamber formation across different recommender algorithms, with some exacerbating these effects more than others. The issue of preference amplification was addressed by Kalimeris et al. [KBKW21], who proposed a theoretical framework to study these dynamics in matrix factorization-based recommender systems. Their approach modeled the dynamics of user interactions with a video recommender system, where user preferences drifted towards recommended content over time. The simulations demonstrated that preference amplification could lead to echo chambers and reduced content diversity. However, the study also showed that mitigation strategies, such as reducing exposure to objectionable content, can enhance user engagement while preventing negative experiences associated with preference amplification. Ferrara et al. [FENKW22] examined the impact of link recommendations on network structure and minorities. Their research highlighted how specific recommendation algorithms (RAs) can disproportionately disadvantage minority groups by reinforcing majority preferences. This reinforcement can lead to reduced visibility and engagement for minority content and perspectives, further entrenching societal biases and inequalities. Their study employed offline simulations to model user interactions with RAs over multiple iterations, using the same movie dataset and RA (UserCF, ItemCF, and SVD) as Mansoury et al.

In the context of extremist content, Whittaker et al. [WLRV21] analyzed the role of RSs in amplifying such material, particularly on platforms like YouTube, Reddit, and Gab. Their methodology involved utilizing automated user accounts (bots) to interact with these platforms, generating behavioral data to observe content adjustments and recommendations. The study found that YouTube's algorithm promotes far-right extremist content after users interact with it, unlike Reddit and Gab, which showed no significant signs of amplifying such content through recommendations. This finding underscores the complexity of addressing algorithmic amplification of extremism and highlights the need for more nuanced regulatory approaches. The researchers emphasized that existing policy measures focusing solely on transparency are insufficient, advocating for a co-regulatory approach that includes industry and governmental oversight to tackle the issue effectively.

Recommender systems in disinformation

When it comes to disinformation spreading, the role of recommender systems in propagating misinformation has become a significant concern on par with their role in the echo chamber formation process. A small set of studies have examined how different recommendation algorithms contribute to the spread of false information, focusing on the mechanisms, impact, and potential mitigation strategies. Content-based filtering systems recommend items based on the similarity between items and the user's previous interactions. While these systems effectively personalize recommendations, they can also reinforce misinformation. For instance, if a user frequently engages with false or misleading content, the system will continue to recommend similar content, thus perpetuating exposure to misinformation [Par11, VBP21]. Tommasel and Menczer highlight that content-based systems can inadvertently contribute to spreading misinformation due to their focus on relevance and user interaction history [TM22]. Collaborative filtering techniques, including user-user and item-item CF, are commonly used in recommender systems. These methods rely on user similarity to make recommendations. However, collaborative filtering is prone to creating "echo chambers," where users are primarily exposed to content that aligns with their

existing beliefs, reducing exposure to diverse perspectives. Fernandez and Bellogín discuss that collaborative filtering can amplify existing biases, leading to increased dissemination of misinformation as users are less likely to encounter corrective or diverse information [FB20]. Their study analyzed the impact of matrix factorization techniques on the spread of misinformation, revealing that algorithms optimized for accuracy tend to reinforce popular content, which can include misinformation.

Hybrid recommender systems combine content-based and collaborative filtering methods to leverage both strengths. However, without careful design, hybrid systems can also perpetuate misinformation by combining the biases of both approaches. The combination can result in more sophisticated echo chambers, as these systems refine their recommendations based on a wider array of data points, further entrenching users in their information bubbles. Recent advancements in deep learning have introduced complex models like neural collaborative filtering, which can capture intricate user-item interactions. While these models enhance recommendation accuracy, they also pose risks. Deep learning models can unintentionally learn and propagate biases present in the training data. Dong et al. emphasize that these models, while powerful, require careful monitoring to prevent the spread of misinformation [DWX+22]. Studies have also explored specific algorithms and their impact on misinformation. For example, implicit matrix factorization techniques, which model user preferences based on latent factors, can inadvertently highlight popular but misleading content due to their emphasis on patterns in historical user behavior. In that aspect, Fernandez and Bellogín propose modifying existing algorithms to mitigate these effects, such as introducing diversity-promoting mechanisms and leveraging counterfactual explanations [FB20].

Evaluations of these systems have shown varied results. Some of the studies indicate that increasing the diversity of recommendations can reduce the spread of misinformation by exposing users to a broader range of perspectives. However, the effectiveness of such interventions depends on the user base and the specific implementation of the recommendation algorithm. For instance, algorithms introducing slight dissimilarities in recommended content from trusted sources can help users reconsider their beliefs, thereby countering misinformation [TGP20]. The role of recommender systems in social networks and news platforms is particularly critical. With their rich user interaction data, social networks can both exacerbate and mitigate the spread of misinformation. Techniques like graph-based collaborative filtering, which considers the network structure and user connections, have shown potential in identifying and limiting the influence of misinformation spreaders. In their study, Tommasel and Menczer used a voter model to simulate the impact of different recommendation strategies on the network, finding that diversification of interactions significantly inhibited misinformation spread [TM22, TGP20].

Recommender systems as natural enhancers of diversity

The notion that recommender systems and social networks inevitably create echo chambers has been increasingly challenged by recent research, which suggests a more nuanced understanding of these platforms' effects on information consumption and societal polarization. Dahlgren [Dah21] critically examines the concept of filter bubbles, juxtaposing it with the theory of selective exposure. While acknowledging that users of social networking sites tend to seek information aligning with their pre-existing beliefs and that personalization algorithms can create filter bubbles, narrowing the scope of available information over time, Dahlgren argues that the assumptions underlying filter bubbles often contradict findings from selective exposure research. He posits that the filter bubble thesis conflates technological effects with societal outcomes, a leap that remains unsubstantiated by empirical evidence. In his critique, Dahlgren

[Dah21] presents nine counterarguments to the filter bubble thesis. Notably, he highlights that while people seek supporting information, they sometimes prefer challenging information. Furthermore, he contends that technological personalization does not directly translate to societal polarization. Dahlgren cites studies that fail to observe the societal-level effects predicted by the filter bubble thesis in empirical research. For instance, a German study found minimal differences in search results for political queries, contradicting the notion of pervasive filter bubbles.

Jones-Jang and Chung [JJC22] provide compelling counter-evidence to the filter bubble thesis, particularly in the context of the COVID-19 pandemic. Their study reveals that social media use does not exacerbate polarization but can reduce partisan affect and vaccine hesitancy. This finding contradicts earlier claims that social media reinforces existing beliefs and deepens societal divides. The authors suggest that social media provides opportunities for incidental learning and exposure to diverse viewpoints, challenging the narrative that these platforms invariably lead to greater polarization. Interestingly, the same study contrasts the effects of social media with traditional media, which were found to reinforce partisan affect more than social media. This comparison highlights the differential impact of various media platforms on polarization, suggesting that traditional media may convey more politically tinged, one-sided views compared to the diverse perspectives available on social media. Garrett [Gar17] contributes to this discourse by emphasizing the importance of distinguishing between content exposure and engagement. He argues that while people may encounter diverse information, their engagement with content (e.g., likes, comments) often remains segmented. This distinction is essential in order to understand how false information spreads and how engagement patterns can reinforce echo chambers, even if exposure does not. Garrett contends that the notion of echo chambers is a distraction from the more pressing issue of disinformation campaigns. He argues that while algorithmic filtering may reduce exposure to cross-cutting content, it does not systematically eliminate it. Furthermore, Garrett emphasizes that the psychological tendency to seek diverse viewpoints can help preserve content diversity, even in algorithmically curated environments.

Garrett [Gar17] further differentiates between exposure and engagement echo chambers, noting that segmented engagement with content (e.g., likes, shares) can promote falsehoods more than mere exposure to diverse information. This nuanced perspective suggests that the real challenge lies not in creating impenetrable echo chambers, but in understanding and addressing the complex dynamics of user engagement with diverse content. These studies present a compelling case against the simplistic view that RSs and social networks inevitably create echo chambers. Instead, they point to a more complex reality where technological factors interact with human psychology, societal trends, and varying media landscapes. The research highlights the need for a more nuanced understanding of how digital platforms influence information consumption and societal polarization, suggesting that the effects are more complex and uniformly negative than the echo chamber narrative often implies. This body of work underscores the importance of empirical evidence in evaluating the impact of digital technologies on society and calls for a more balanced approach to addressing the challenges of the modern information ecosystem.

3.1.5 Algorithmic intervention strategies

Disinformation

Recommender systems, as they have become the main engine in shaping social media platforms' information landscape, have a strong potential to promote disinformation and toxic content and often do

amplify disinformation's spread [PS24]. The algorithmic biases inherent in recommendation systems, such as popularity bias and echo chambers, significantly contribute to this phenomenon [DRDB17]. For instance, as Edizel et al. [E$^+$20] highlight, the training data for these systems often contain inherent biases, which influence the recommendations, creating a "self-perpetuating loop" that reinforces the filter bubbles users reside in. These filter bubbles, as described by Pathak et al [PSP23], lead to intellectual isolation, where users are predominantly exposed to content that conforms to their preexisting beliefs, thus making them more susceptible to misinformation. The performance of recommendation systems is further impacted by user interactions, where frequent engagement with certain types of content prompts the recommendation system to narrow the recommendations, thereby exacerbating the exposure to disinformation [PSP23]. The empirical studies conducted by Fernández et al [FB20] illustrate the significant role recommendation systems play in the propagation of misinformation. Their research demonstrates that algorithms, particularly those based on popularity, are prone to repeatedly recommending a limited set of misinformation items, akin to the behavior of bots or viral accounts. This concentrated recommendation pattern increases the visibility of disinformation and accelerates its spread across the network. In contrast, neighbor-based recommendation methods tend to distribute misinformation more uniformly, albeit still contributing to its propagation [FB20, Par11]. The amplification of misinformation through recommender systems clearly indicates their active role in modulating information distribution.

The workshops and datasets dedicated to understanding and mitigating this impact, such as the OHARS initiative [TGZ21], underscore the urgency and importance of addressing these challenges within the recommender system community [E$^+$20]. Regarding existing research, there exist state-of-the-art approaches to help prevent the formation of echo chambers. An example state-of-the-art method to prevent the formation of echo chambers is described in [BNS18], in which authors employ a modified collaborative filtering-based approach (PrCP) where the original user-item ratings are stochastically modified, to foster more diverse recommendations while keeping relevant recommendation accuracy.

Similarly, authors in [GSM$^+$22] discuss different methods based on a reranking approach that prioritizes diversity at the cost of recommendation accuracy. In the case of MMR, one of the algorithms discussed in the previous research, such a trade-off can be controlled through λ, a parameter that defines the weighted average between the similarity of the new item with respect to the user preferences, and the similarity of the new item with respect to items already recommended to that user, creating a diverse set of recommendations for each user. Some other approaches aim at applying content-moderation to prevent the formation of echo chambers, as authors propose in [Str22]. However, the aforementioned reranking strategy to foster diversity at the cost of recommendation accuracy is one of the most employed strategies. In fact, existing literature describes multiple optimization algorithms that can be used within the recommender system to increase recommendation diversity, such as greedy optimization or integer programming optimization methods, which can be used for reranking the results of an existing recommendation algorithm to optimize diversity as a strategy for breaking disinformative echo chambers, as discussed in [AK12].

Currently, studies on the role of recommendation systems in modulating the distribution of disinformation are very limited. The most extensive study on this topic is by Phatak et al [PS24]. but it focuses solely on the distribution of disinformative content, largely ignoring the study of disinformative communities as a phenomenon in itself.

Similarly, almost all algorithmic responses to curbing disinformation on social networks from the

platforms involve the introduction of moderation and fact-checking systems, as well as increasing network diversity to prevent the creation of echo chambers. Thus, there is a significant gap in the direct treatment of disinformative echo chambers.

Polarization

Due to the growing role that social networks play in our lives and the ample evidence indicating the presence of polarization on these platforms, particularly on X, it is essential to inquire into this specific phenomenon to propose clear and applicable solutions. Naturally, these solutions can be multifaceted, with the problem being addressed from various fronts, including political or educational domains outside social networks and within the networks' design. We consider it fundamental to tackle the problem from a technical perspective, focusing on the design of social networks. As observed, the design of these platforms—specifically, the types of actions they allow and the methods and algorithms responsible for content dissemination—plays a key role. Without neglecting other possible interventions to mitigate the phenomenon, we believe that technical intervention within the social networks is essential. When addressing this phenomenon from the perspective of social networks, we can employ two distinct approaches: platform design and algorithmic strategies.

The design of the platform can have a significant effect on promoting phenomena such as polarization. Post immediacy, high volume, and brevity can influence these dynamics. However, the changes required to mitigate the phenomenon from this angle would be costly and complicated to implement, potentially leading to a dilemma. These changes could alter the design to the extent that the platform becomes a new social network distinct from the original, which users might abandon as it would be a different product. This perspective has been minimally explored in the existing literature. The state-of-the-art mainly involves the insertion of community interventions (such as those currently in place on X), suppression or hiding of violent content, content unsuitable for children, or overtly extremist content, and implementing fact-checking mechanisms. The primary proposals in the existing literature revolve around these approaches.

Most technical proposals aimed at mitigating polarization on social networks focus on intervention in content recommendation systems. The primary algorithmic interventions against polarization in recommender systems revolve around strategies that modify recommendation processes to reduce echo chambers and promote diversity. Among these approaches, the Pre-Recommendation Counter-Polarization (*PrCP*) strategy [BNS18] stands out. This method employs a collaborative filtering recommendation approach, wherein the original ratings of user-item pairs are modified through a stochastic mapping function. Specifically, the rating R of user U on item I is altered to rating R' with a probability ρ. This modification occurs prior to content recommendation, aiming to reduce network polarization significantly compared to the traditional Non-Negative Matrix Factorization approach [BNS18].

Another prominent approach is the Maximal Marginal Relevance (*MMR*) recommendation algorithm, extensively discussed in [GSM$^+$22]. Widely adopted to diversify results and mitigate the formation of echo chambers, this algorithm operates under a greedy setting. It selects recommendations by balancing two objectives: proximity to user preferences and dissimilarity from already recommended items. The trade-off between relevance and diversity is controlled by a λ parameter, which weights these factors to create a diverse yet relevant recommendation set [CG98, GSM$^+$22]. The versatility and continued exploration of this approach make it a cornerstone of current research in this domain [GSM$^+$22].

Content-based recommendation strategies also play a key role in addressing polarization. One such strategy involves content moderation, as outlined in [Str22]. This approach filters out content exceeding a defined hostility threshold τ, thereby preventing its recommendation to other users. While effective in reducing polarization, this method raises concerns about limiting user freedom and access to certain types of content [LVhLH20].

Finally, optimization-based strategies, such as those discussed in [AK12], remain a foundational component of interventions against polarization. These include methods like greedy optimization and integer programming, which re-rank the results of existing recommendation algorithms to enhance diversity. Although these strategies can be considered classical, they continue to be relevant in modern research for their ability to reduce polarization while maintaining recommendation accuracy [AK14, YB22].

3.2 Research Motivation

By reviewing the state of the art, we can conclude that, research on the role of information recommendation algorithms in modifying the political behavior of users in the online social network environment is scarce, partial, and weak. Very few algorithms have been analyzed, with approaches like deep learning or reinforcement learning practically excluded from published analyses, despite the fact that the approaches currently dominate the state of the art in both research and (especially) industry. Similarly, the analysis of micro-blogging social networks has been often restricted to specific events studied at a particular moment in time.

The published analyses' conceptual frameworks have primarily relied on the disputed concept of the echo chamber, omitting much of the relationship between these recommendation systems and other phenomena, such as the spread of disinformation or polarization. The latter, polarization, has always been treated generally, omitting its political component. Consequently, due to the lack of in-depth research, it is common to witness the ongoing debate between proponents and opponents of the hypothesis that recommendation systems generate echo chambers and polarization. Selective exposure remains a reality, and various political organizations or actors whose origins lie outside the network continue to have particular interests in mobilizing citizens and therefore radicalizing their positions, increasing polarization that can be transferred online. The emergence of echo chambers, besides being a general human tendency, is a complex phenomenon that requires a broad and enlightened perspective for its study.

Thus, there is a significant gap in the literature regarding the role of recommendation systems in the emergence of collective phenomena on social networks, such as the spread of disinformation, the formation of opinions, political influence, and the polarization of users in online debates. All these phenomena are strongly intertwined and enhanced by computational propaganda distributed by state and non-state actors.

Chapter 4

Research Methodology

4.1 Quantitative Data

4.1.1 Data set selection

A comprehensive set of data was chosen to adequately cover the primary phenomena under analysis in this research. This dataset primarily focuses on information dynamics during electoral processes in Spain. It also includes a substantial collection of publications from media outlets or journalists and content from actors identified as distributors of disinformation or deliberately false information for manipulative purposes.

These datasets were selected by current phenomena and the primary information channels associated with them. These channels encompass digital media (including blogs and news portals), micro-blogging services (such as X), and messaging applications (like Telegram). This diverse range of sources enables a comprehensive analysis of the information dynamics across various platforms, offering a holistic view of the digital information landscape during critical political events, such as elections.

The choice of data sources reflects the multifaceted nature of modern information dissemination and consumption, acknowledging the significant role that traditional and new media play in shaping public discourse.

4.1.2 Information broadcasters

In an online social network like X, information is disseminated through various channels. Different users can assume distinct roles based on their personal characteristics, motivations, network position, and engagement timing. A significant portion of social network users are primarily spectators—consumers of content who may interact with others by sharing information or engaging in debate, depending on the context. Likewise, specific events such as electoral campaigns or periods of social unrest can spark interest and activate these users, prompting them to participate by sharing opinions or information. Beyond these particular users and events, we can identify distinct segments of users who take on more active roles. These users enter the network to share information proactively, often with a clear purpose, which is frequently political. We can refer to these users as "information broadcasters," as they actively emit information and are instrumental in initiating the flow of information within the network. This research primarily focuses on three types of information broadcasters, encapsulating the significant social phe-

nomena related to politics on X: political parties, journalists, and accounts that spread disinformation and propaganda. Political parties are defined in detail later in the chapter, while journalists and disinformation actors are discussed in the following sections.

Disinformation actors

Disinformation actors are highly active users whose primary activity on social networks involves disseminating disinformation and propaganda. This type of user is particularly prevalent on the social network X. These users can come from various backgrounds, including anonymous accounts, semi-automated bots, alternative journalists, or public commentators and analysts, often referred to as "pundits." Despite their diverse origins and taxonomy, they exhibit a common pattern of activity: a focus on causes or issues of high interest, often involving social polarization or unrest and the active spread of manipulated information and propaganda related to these topics. In recent years, many of these users on X have been involved in promoting disinformation or propaganda favorable to autocratic regimes seen as alternatives to the Western Nations[1], such as Russia, Cuba, Venezuela, Iran, and, to a lesser extent, China—particularly concerning the origins of COVID-19. While there are also disinformation actors supporting U.S. interests, this thesis focuses on the aforementioned groups due to their significant role in attacking Western democratic systems and their prominent influence on various international political phenomena.

Journalists

Journalists, supported by their profession and their respective media outlets, are the users with the highest professional and ethical credibility on the social network X. They are generally presumed to be reliable and truthful in disseminating information. Due to their professional activities, journalists actively share current news on the platform and engage with the various events being discussed.

Although they are not entirely free from sharing ideologically biased information, their professional standards tend to steer them away from spreading disinformation or outright fake news. This makes them a valuable contrast to disinformation actors, enabling an effective and scientifically sound comparison. As detailed in the following sections, this thesis focuses on a dataset of journalists primarily covering political and social events in Spain. In studying disinformation actors, political figures, and journalists, we employed the snowball sampling technique and tracked all their activities over extended periods.

The specific details regarding our data mining process are outlined below.

4.1.3 Political processes

Online social networks have profoundly impacted how we obtain current information, mainly how we participate in politics. These platforms are serving as dynamic platforms for public discourse and civic engagement.

To accurately study the dynamics of political phenomena on online social networks, a dataset encompassing the major Spanish electoral processes from 2011 to 2019 has been constructed. This period provides a rich context for analyzing evolving political communication and engagement patterns on these platforms.

[1]The ones belonging to international organizations such as the EU or NATO characterized by liberal economies and democratic regimes

The focus on political processes in Spain primarily stems from the fact that this research was conducted in Spain. Additionally, according to PEW Research, Spain's status as a democratic Western nation, a member of the European Union (EU), World Trade Organization (WTO), and North Atlantic Treaty Organization (NATO), with one of the highest rates of online social network usage and particularly notable for its degree of political polarization, makes it an ideal case study for analyzing these phenomena. This context offers a unique opportunity to explore the complex relationship between social media engagement and political dynamics in a contemporary democratic society.

The analysis covers municipal and national electoral processes in Spain from 2011 to 2019. These nine years have been chosen because they are extensive enough to facilitate the identification of consistent and relevant patterns in political behavior and communication.

It is understood that political parties in contemporary Western democratic systems often exist in a state akin to a "constant campaign." They endeavor to utilize news and real-world events for their political objectives, engaging in continuous communication with the citizenry through networks. This perpetual state of campaigning underscores the dynamic nature of political communication in the digital era. However, it is essential to note that the analysis is focused on specific identifiable points within this continuum, providing a structured approach to examining the evolution of political discourse and strategy within the defined time frame.

In order to develop computational tools for evaluating information dynamics in political contexts, this study focuses on the various phases of electoral periods. These include the pre-campaign, the electoral campaign, the election day, and the subsequent week. In Spain, the electoral campaign corresponds to the two weeks preceding the election, with the week before these two constituting the pre-campaign period. This time frame is critical, as political parties as well as the rest of political actors of different orientations articulate and intensify their communication and interaction strategies with users during these periods. Additionally, the study encompasses the election day and the week following it. This approach allows for an analysis of the discourse surrounding the election results, including the reactions of both the winning and losing parties, as well as the response of the broader user base to these outcomes.

The social network chosen for this study is X/Twitter, selected for its particular significance as a hub of political and social debate in Spain and, broadly, the rest of the world. Twitter's platform offers a rich and dynamic environment for understanding the nuances of political discourse, making it an ideal subject for this research. The platform's widespread use and influence in shaping public opinion and political conversations render it an essential element in the study of modern political communication and information dynamics.

4.1.4 Digital activists

Activists represent a particular type of political user present on online social networks, especially on X. They utilize the platform to disseminate political propaganda in a manner that is often more informal, direct, and even aggressive compared to political parties, which are typically constrained by formal tone, political correctness, and the need to focus on official communications.

These users are not primarily associated with the spread of disinformation, although they may occasionally share it [AZ12]. They do not focus on disseminating news, whether biased or otherwise, as any professional code of ethics does not bind them. They typically do not operate within a formal organizational structure or a media outlet. This lack of institutional ties places them in a compelling position

for study. Digital activists are characterized primarily by their propagation of political propaganda. In doing so, they aim to mobilize large groups of users around a cause in a much more informal manner than the political parties or movements with which they may sympathize. Many of these activists have significant followings, highlighting their substantial impact on the dynamics of information flow within social networks [BAB+21].

Given their cross-cutting nature and the complex dynamics surrounding their motivations, use of social networks, and political affiliations, the role of political activists in this thesis has been studied primarily through qualitative analysis. The qualitative analysis methodology is presented further in Section 4.2. This approach complements quantitative research, which forms the bulk of this work.

4.1.5 Data mining process

The data mining process for this research involved the application of web scraping techniques and the use of Twitter's Academic API for downloading information.

Regarding Twitter, the research focused on downloading data about political processes and users identified as information disseminators. The Academic API was utilized to download and index all publications from selected user groups during specified periods relevant to the research objectives. By downloading all publications from these users, metadata associated with these posts, including interactions, publication dates, and authors, were also obtained. These publications were indexed in a MySQL database to facilitate analysis and research.

For publications related to media outlets, web scraping techniques were employed. The data obtained from these sources were also indexed in a separate MySQL database, allowing for efficient organization and accessibility of information.

Political processes

For the practical purpose of generating the relevant datasets, official accounts of the main political parties in Spain were selected. This selection was based on parties that participated and secured election in each of the studied electoral processes, at both local and municipal levels. The complete list of these accounts is available for reference in 4.17.

In creating the dataset, the Academic API of X was utilized for each party to compile a comprehensive set of all publications made by their respective accounts during the designated periods. These periods include the pre-campaign, campaign, election day, and the week following the election for each specific electoral process. The compiled dataset encompassed not only original tweets from these accounts but also retweets made by them. This comprehensive collection provides a robust foundation for analyzing the communication strategies and engagement patterns of these political entities across different phases of the electoral cycle. By including both original content and retweets, the dataset offers a more complete picture of the parties' online presence and influence, reflecting their engagement and messaging strategies during essential political periods.

Upon creating the initial list of accounts, a snowball sampling method was employed to expand the dataset. This technique was instrumental in capturing the immediate environment of users interacting with these political parties, thereby enabling a comprehensive capture of the information dynamics surrounding an electoral process. The complete algorithm used for generating this dataset, detailing the methodology and the steps followed, can be found in Algorithm 1.

Algorithm 1: Snowball Political Dataset Generation from X / Twitter.

Data: X accounts of main political organizations: $accounts$
Election day: $election\_day$
Number of publications: 8
Number of recursions: 2
Result: List of users influenced by political publications

$common\_list \leftarrow$ **empty list**;
$start\_date \leftarrow election\_day - 3$ **weeks**;
$end\_date \leftarrow election\_day + 1$ **week**;
for $account$ ***in*** $accounts$ **do**
  $publications \leftarrow$ **download all publications between** $start\_date$ **and** end_date **from** $account$;
  for $i \leftarrow 1$ ***to*** *number of publications* **do**
    $publication \leftarrow$ **randomly select from** $publications$ **with at least one repost**;
    $reposting\_user \leftarrow$ **randomly select one user who reposted** $publication$;
    $common\_list$.append($reposting\_user$);
  end
end
$iteration \leftarrow 0$;
while $iteration <$ *number of recursions* **do**
  $newly\_retrieved \leftarrow$ **copy of** $common\_list$;
  for $user$ ***in*** $newly\_retrieved$ **do**
    for $i \leftarrow 1$ ***to*** *number of publications* **do**
      $publication \leftarrow$ **randomly select from user's publications with at least one repost**;
      $reposting\_user \leftarrow$ **randomly select one user who reposted** $publication$;
      $common\_list$.append($reposting\_user$);
    end
  end
  $iteration \leftarrow iteration + 1$;
end
return $common\_list$;

The resultant dataset was systematically stored in a MySQL database following structure, see Table 4.1. This choice of database system facilitates relational analysis, allowing efficient organization, retrieval, and examination of data. The use of a relational database like MySQL is particularly beneficial for handling large datasets with complex relationships, as it provides robust data management capabilities and supports advanced querying and analysis techniques. This structured approach to data storage and management is essential for the thorough and systematic analysis required in this research.

As a summary, the data gathered through the sampling process are concisely presented in the Tables 4.2 and 4.3. These tables provide an organized and accessible overview of the collected data, categorizing them based on different criteria relevant to the research.

Information broadcasters

Constructing our data set commenced with identifying "disinformation actors" – accounts that consistently disseminate misleading narratives or counterfeit news. Given the challenging nature of accurately labeling an account as a disinformation agent, we employed the following flexible approach (depicted in Figure 4.1), in a thorough and consistent way to improve its reproducibility. Initially, we consulted verified databases of fictitious news and domain names affiliated with disinformation spreading, courtesy of international organizations such as The European Commission (see the first block in the diagram flow, using [Wan17, SMW$^+$18, Cha19, SV16, Zim17, DCFG21] as databases and [Tas22, ML22, All22] for domain names). Subsequently, we sought out those Spanish accounts (through the use of the 'lang:es' parameter in our search queries) that demonstrated the highest interaction levels with such dubious content. Our decision to focus on Spanish accounts was twofold: it capitalized on the authors' proficiency in the Spanish language and the Spanish political landscape. This, in particular, addressed a significant

Table 4.1: Database description of political process data on Twitter.

Database Table	Description
account_category	Table used to categorize accounts based on predefined types.
annotation	Defines the type of entity and the specific entity that a post refers to.
annotation_prediction_tweet	Represents the relationship between a tweet and its annotations.
domain	Associated domain of a URL mentioned in the tweet.
emotion	Predictions of emotions associated with a tweet, such as Sadness, Joy, Fear, Anger, and Surprise.
hashtag	List of unique hashtags used in tweets.
hashtag_tweet	Relationship between a hashtag and the tweet it appears in.
hate_speech	Classification of a tweet as Hate Speech or non-Hate Speech.
liwc_2007	Classification of a tweet based on the dimensions of Linguistic Inquiry and Word Count (LIWC) in Spanish.
medias	Associated media elements with the tweets.
mention	Mentions of a user in a tweet.
quoted	Records when a tweet is quoted by another.
reply	Represents a reply to a tweet by another tweet.
retweet	Relationship indicating a retweet between two tweets.
sentiment	Represents the sentiment of the tweet, classified as positive or negative.
tweet	Contains the text and basic metadata of the tweet.
tweet_metrics	Metrics of the tweet including number of likes, retweets, replies, and quotes.
url	Unique URL mentioned in the tweet.
url_media	Relationship between a media element (domain) and URL.
url_tweet	Relationship between a tweet and a URL contained within it.
user	Author of the tweet.
user_info	Information about the user including number of followers, following, description, full name, URL, and location.
user_seed	Indicates if the user is a seed (seed node of the snowball).

Table 4.2: Dataset "General Processes" of Spanish general electoral processes from 2011 to 2019.

Electoral Process	November 2019	April 2019	2016	2015	2011
Users	873	715	759	688	611
Posts	527093	423638	616427	465967	317506
Negative Posts	118951	83708	92881	74338	55754
Nodes	872	700	756	614	572
Edges	24460	18524	23371	12420	9369
Degree	56.10	52.93	61.83	67.29	32.76
Avg Retweets	361.11	247.64	74.19	45.14	16.09
Avg replies	3.77	2.49	0.66	0.73	0.80
Avg likes	38.79	26.42	5.39	3.24	1.15
Avg quotes	1.29	0.85	0.19	0.09	0.00

gap in research on Spanish language disinformation.

Our data set generation then incorporated an additional check. We undertook a qualitative review process to mitigate the potential for false positives. The manual filtering process aimed to refine the

Table 4.3: Dataset "Local Processes" of Spanish local electoral processes from 2011 to 2019.

Electoral Process	2019	2015
Users	708	722
Posts	398689	535306
Negative Posts	71816	76030
Nodes	698	719
Edges	16550	19397
Degree	47.42	53.96
Avg Retweets	184.03	44.59
Avg replies	1.80	0.68
Avg likes	20.64	3.58
Avg quotes	0.67	0.00

data set, whereby we favored accounts demonstrating frequent interactions with events or narratives within Spain. This was specified to keep the focus of our research on this regional phenomena, while also allowing us to serve as filtering *experts*, considering our experience and familiarity with the events. Therefore, the creation of our list of accounts associated with disinformation resulted from a multi-step process, as outlined above. This meticulous approach allowed for the compilation of a robust data set in its representation of disinformation activity on Twitter, allowing us to analyze their behavioral patterns in depth. The whole process is depicted in the diagram flow presented in Figure 4.1.

The number of unique accounts identified before the manual inspection was 513. Then, a team of three volunteer students with a background in political science, along with the help of the authors, conducted another qualitative assessment and manually extracted 275 unique disinformation accounts, whose IDs are listed in Table 4.15 (corresponding to the last two blocks in the diagram flow). The process involved selecting the most active and consistent accounts related to disinformation. This was a conscious decision, since these highly active accounts are key in the spread of disinformation, as their primary aim is to achieve maximum possible impact, due to their significantly higher potential for influencing and impacting online conversations. In order to assist the validation of the selected accounts as disinformation actors, they were checked using the "misinfo.me" [2] online service, being all of them flagged as mainly disinformation sharers.

Upon establishing the 275 disinformation-related accounts, we created a commensurate sample size for the comparison group, aiming to identify them as legitimate purveyors of information or more specifically, journalists. We assembled a pool of principal digital media outlets within Spain to construct this sample. This list was substantiated by cross-referencing numerous rankings – such as [OJD22][3] and [Sta22] – to ensure the inclusion of prominent, mainstream outlets. From these sources, we identified individual journalists frequently engaged in public discourse, particularly in socially relevant areas such as politics and society. Indeed, the specific selection of journalists who cover politics and society was not arbitrary. Propaganda and disinformation typically orbit around the themes of politics and societal

[2]Misinfo.me: A platform for analyzing and monitoring the dynamics of misinformation within digital ecosystems. It integrates network visualization tools, data mining, and advanced detection techniques to deliver real-time insights into disinformation campaigns and manipulative narratives on social media.

[3]OJD, `https://www.ojd.es`, (from Spanish *Oficina de Justificación de la Difusión*, in English *Audit Bureaux of Circulations*) is the Spanish organization that provides, among others, services of control and issuing of dissemination reports as well as data consultation figures via the Internet. It belongs to the International Federation of Audit Bureaux of Circulations (IFABC), `http://www.ifabc.org`.

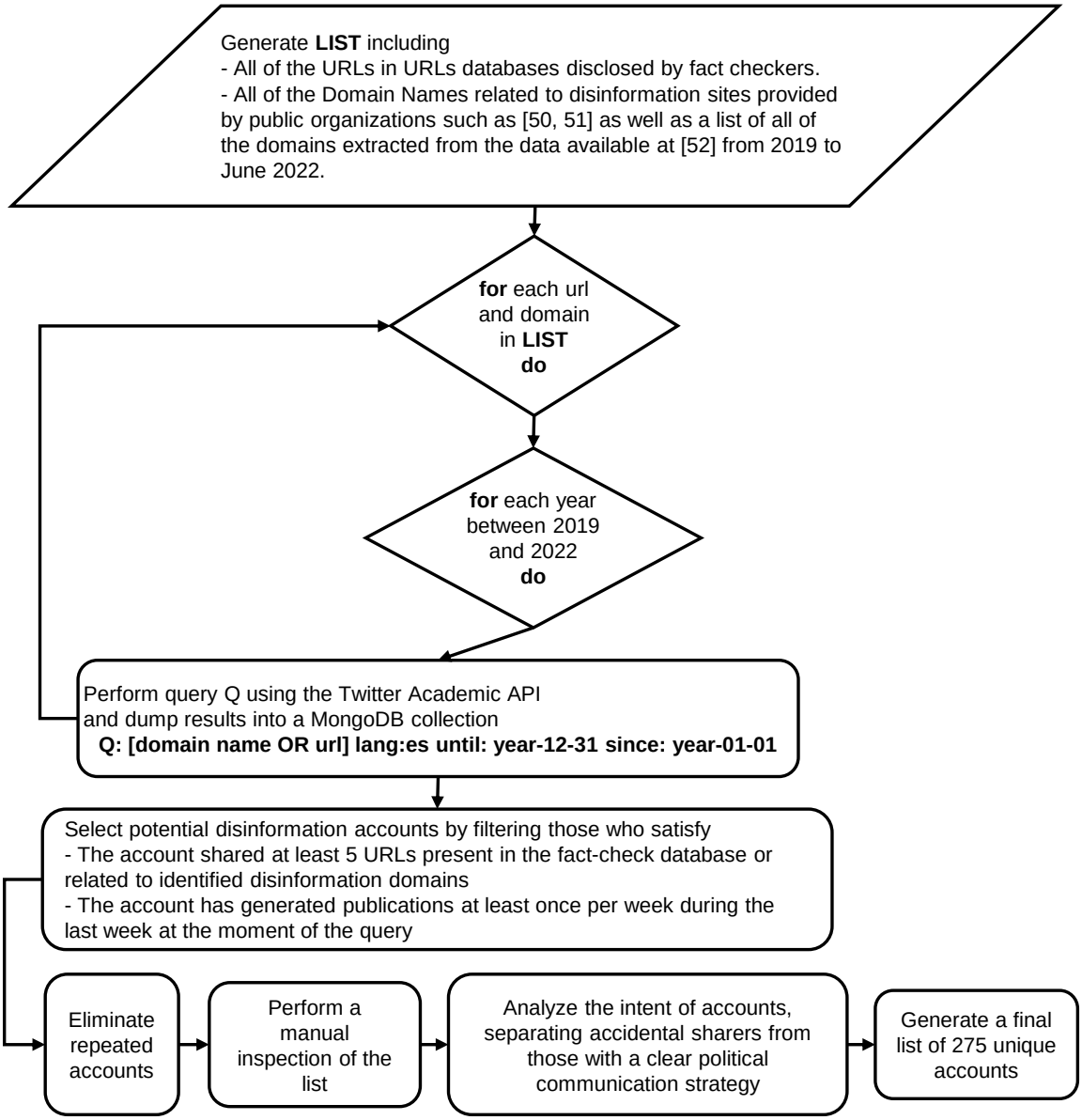


Figure 4.1: Diagram flow of the process for generating a data set of X disinformation accounts.

issues, with these subjects often being the prime targets of such misleading campaigns [Fal15, Ruo21]. Due to their contentious nature and potential for social impact, these topics are prime vehicles for the proliferation of disinformation. Furthermore, such focus areas frequently serve as battlegrounds for public opinion, making them fertile grounds for disinformation actors to exploit.

We further distilled our selection based on activity level from this pool of journalists. The accounts exhibiting the highest degree of interaction were selected for inclusion in our data set. This approach – as it was done for disinformation actors – ensures that our analysis is relevant and focused on entities with the greatest potential to influence the online discourse. This selection method, outlined as a diagram flow in Figure 4.2, provided a balanced, representative sample for studying the behavior of disinformation networks on X in contrast to their legitimate counterparts. A final list of 275 accounts listed in Table 4.16 was generated to be comparable with those obtained through the previous process.

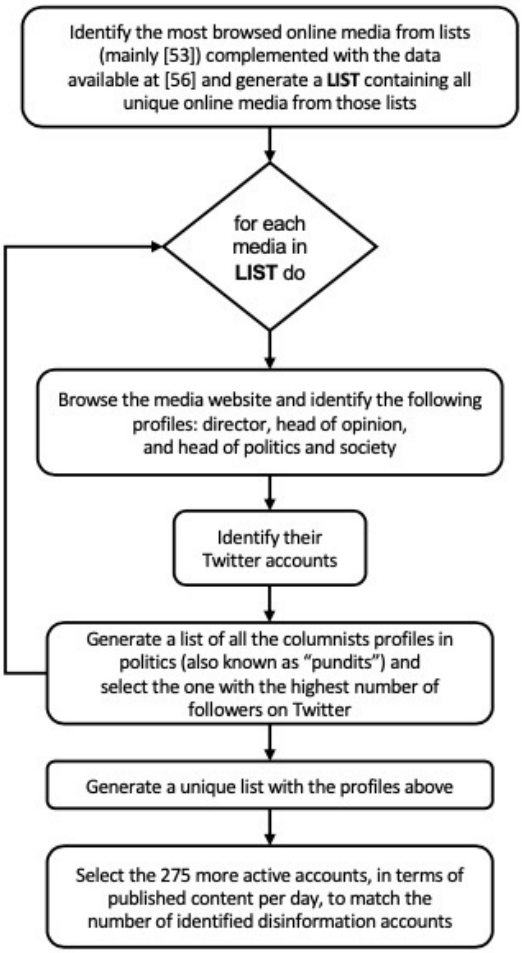


Figure 4.2: Diagram flow of the process for identifying influential media-related (journalists) X accounts.

Table 4.4: Dataset "Information disseminators" Properties of the complete networks (using all data from the entire 2019-2022 period), with 275 users in both Journalists and Disinformation Actors categories.

Property	Journalists	Disinformation Actors
Users	275	275
Posts	3906047	7194766
Negative Posts	118951	513566
Nodes	953	815
Edges	30235	20180
Average Degree	60.15	50.08
Avg Retweets	361.11	1867.52

4.1.6 Data model

The quantitative information in this thesis has been primarily collected from large datasets of posts on the social network X, along with their associated metadata. All this data has been stored in a MySQL relational database to facilitate recurring queries across the various experiments in the thesis. This model has enabled quick information retrieval, result storage, and the execution of multiple tests. Here we detail in Tables 4.5, 4.6, 4.7, 4.8, 4.9, 4.10, 4.11, and 4.12 the employed data model.

Table 4.5: Tweet table.

Name	Type	Key/Index
id	int	Primary Key
tweet_id	varchar(200)	Primary Key
source	varchar(200)	Index
reply_settings	varchar(200)	Index
possibly_sensitive	tinyint(1)	Index
author_id	int	Primary Key
created_at	datetime	Index
lang	varchar(4)	Index
conversation_id	varchar(200)	Primary Key
text	varchar(900)	None
start_time	datetime	Index
end_time	datetime	Index
trending_topic	varchar(200)	Primary Key
campaign	varchar(200)	Index
in_reply_to_user_id	varchar(250)	Index

Table 4.6: Retweet table.

Name	Type	Key/Index
id	int	Primary Key
tweet_id	int	Index
retweeted	varchar(250)	Index
action_date	datetime	None

Table 4.7: Sentiment table.

Name	Type	Key/Index
id	int	Primary Key
tweet_id	int	Index
positive	double	None
negative	double	None

Table 4.8: Hashtag table.

Name	Type	Key/Index
id	int	Primary Key
hashtag	varchar(250)	Index

Table 4.9: Conversation domain table.

Name	Type	Key/Index
id	int	Primary Key
annotation_id	varchar(400)	None
name	varchar(250)	Index
description	text	None
category	varchar(250)	Index

Table 4.10: URL table.

Name	Type	Key/Index
id	int	Primary Key
url	varchar(400)	None
expanded_url	text	None
domain	varchar(300)	None
title	text	None
description	text	None
unwound_url	text	None
disinformation_url	tinyint(1)	None
domain_has_shared_fakenews	tinyint(1)	None
fact_checker	varchar(900)	None
label	text	None
claim	text	None

Table 4.11: User table.

Name	Type	Key/Index
id	int	Primary Key
user_id	varchar(400)	Index
created_at	datetime	None

Table 4.12: User information table.

Name	Type	Key/Index
id	int	Primary Key
screen_name	varchar(200)	Index
name	varchar(200)	None
user_id	int	Index
description	text	None
followers_num	int	None
following_num	int	None
url	varchar(400)	None
location	varchar(200)	None

4.2 Qualitative Data

Digital activists

Participant identification:

To identify profiles that would potentially fill into this new category of activism as articulated in Definition 1 (which will be properly explained in Chapter 5), we first generated a list of seed user profiles on Twitter, representing the leading social network related to political debate. We identified the main political think tanks and associations operating in Spain to generate this initial list. On the right-wing of the political spectrum, we identified organizations like Students for Liberty, Instituto Juan de Mariana, or Fundación Danaes, and on the left-wing organizations such as Instituto25m, Rebeldiajoven, and Sindicato de Estudiantes. Then we complemented the list of seed users by browsing through well-known political X accounts that usually support the ideas or are linked to political organizations or mass media: @clubdeviernes, @CapitanBitcoin, @EugeniodOrs_, or @FrayJosepho on the right-wing of the spectrum, and @spanishrevorg, @kaosenlarednet, @Shine_McShine, or @protestona1 on the left-wing of the spectrum. Finally, the list was also complemented with profiles found at portals such as batallacultural.com, which provides a curated list of right-wing cyber-activist along with their social networks, and redrepublica.es, which provides a list of left-wing anti-monarchy hashtags that can be used to detect X users on the left-wing.

We then manually applied the snowball sampling technique: browsing through all the initial profiles and identifying those recommended profiles following the X contact recommendation system. We applied this process in a recursive manner, limited to three iterations. For each of the retrieved profiles, we manually tried to identify if they were present on other social media platforms by manually browsing through their public URL and searching for the same username on online search engines.

Once the final list was generated, we manually reviewed all the profiles and selected those that comply with the following conditions:

- *Complies with the Definition 1*
- *At least one year of continued activity* (to ensure their behavior was consistent in the long term)
- *Focused almost entirely on topics related to political or social issues* (to avoid heterogeneous or multi-profile accounts)
- *More than 1,000 followers on any of their social network profiles* (as a strong signal of influence[4])
- *An activity pattern of at least 5 publications a week across all of their social network profiles* (to ensure a high activity level, in this case, a minimum of one per weekday)

The first condition was assessed in a qualitative manner, and the other ones in a quantitative manner. The last step involved a manual revision of the most prominent profiles to identify potentially interesting relations, communities, and hierarchies amongst the activists, thus allowing us to identify the most potentially relevant cyber-activists to be interviewed. A total amount of 17 participants, 13 males and 4 females, were interviewed during the period of one year.

[4]This is in line with literature such as [BHMW11] where users with 1.8K followers were predicted to have the largest total influence.

Screening participants:

We then contacted the selected individuals (who, in our case, are activists). The approach was made by a direct message over social media or through their provided e-mail address. The interviews were conducted in-person in Madrid (Spain) or online using the ZOOM app according to the geographical and temporal availability of the participants. The interviews were audio recorded and transcribed verbatim. Verbal consent to record, transcript, and research use was obtained from all the participants before starting the recorded interviews. The interviews ranged from 30 to 180 minutes, with the average interview lasting 40 minutes.

Interview format:

The interview structure was composed of three blocks. In the first block, we asked the participants about their main motivations: how they formed their ideologies, the role of the Internet in their ideology formation process, and their transition into cyber-activism. The second block involved questions related to their tools and methodologies. We mainly asked about the role of digital technology in their activism, the set of tools and social network sites they used to employ, and the specific use case for each. We also asked about the methods, techniques, and strategies they use to promote their narratives and perform their activism online. We focused on activist recruitment, action coordination, and content viralization. In the last block, we asked questions about their relationship with traditional forms of communication and political participation, i.e., political parties and mass media, and their relationship with opposite forms of activism. The structure of the interviews was designed this way to allow one to address the RQs presented, particularly RG1.RQ1 (Which are the main information disseminators in the political context on micro-blogging social networks and how they develop their activity?) and RG1.RQ2 (What motivates political active users to perform online activism and promote narratives?).

Demographic data:

At the end of the interviews, a survey was used to collect features from the participants, such as their age, gender, declared ideology, educational attainment, and occupation.

Table 4.13: Demographic data of the interviewed individuals.

DEMOGRAPHIC DATA		
Mean age:		27.4
GENDER	**N**	**%**
Male	13	76.47
Female	4	23.53
IDEOLOGY		
Self identify as communist	3	17.65
Self identify as socialist	4	23.53
Self identify as conservative	4	23.53
Self identify as libertarian	3	17.65
Self identify as anarchist	3	17.65
EDUCATION		
High school diploma	4	23.53
Vocational school	1	5.88
University degree	10	58.83
PhD	2	11.76
OCCUPATION		
Student	7	41.18
Freelancer	4	23.53
Engineer	1	5.88
Public servant	2	11.76
Journalist	3	17.65
Political analyst	1	5.88

Table 4.14: Social media presence of the interviewees.

SOCIAL MEDIA USE		
	N	**%**
Social network profiles		
Facebook	6	35.30
Instagram	10	59.82
Twiter	17	100
YouTube	8	47.06
Twitch	6	35.30
TikTok	3	17.64
Telegram channels	4	23.53
Tools and communities		
Telegram groups	4	23.53
WhatsApp groups	12	70.59
Patreon profiles	5	29.41
Paypal profiles	3	17.64
Discord servers	4	23.53
Personal websites	2	11.76
Linktree profiles	7	41.17

Table 4.15: Ids of accounts used in our study as disinformation actors.

Id	Id	Id	Id	Id
1000025123263524864	1528872169	253020963	3250842464	535153760
1003133438	1533855511	2569400510	330361451	53789862
100731315	1536167761	257859379	333033292	54233321
1009550909540585473	1561703076	259799108	333348576	55279448
102454320	1613828468	259897306	333936316	555521701
1031419921	1617884742	2605337039	3366058611	558462908
1040334830	1623272011	2609524507	3430772032	564138118
1064953765814509569	163141341	262851272	343933160	590421119
106891797	164110029	266012628	345257174	597712811
107177786	165104029	2667821193	3468902956	59894497
1081168493091962880	165127264	266838221	347300732	601657931
108744588	16799023	268035270	348538097	606623283
1089985800069095426	1688470074	2693017699	357590912	612411163
1104344103179960320	1707867426	2716353140	363421116	617894888
1106569081854066690	176058394	2732715937	36674807	700503810
110922804	17636635	273924214	367070806	708482255
1110891412508340224	178852637	2755279044	369846834	714195188667846657
1112478389309505536	183661695	2766498805	37835750	714556579
1118059060098629632	183786438	278248787	381657036	720743812562337798
112134350	185143177	280081621	390387588	72732893
112170559	1884163777	2809357258	391344883	736185894089199616
1121998616	18856867	2824259531	393476699	745339783
112747809	1896481891	282675582	394229561	755494698584801280
113035227	1923495216	2827483187	399275188	761154976076926977
1133334569577586688	193095342	283409352	4035057615	762405116
114558569	193096110	285255977	407754987	762903092983541761
114741363	1931893196	2858434521	411577733	763100287028568064
1150056069022007298	195446876	287786986	411647930	765599356980498432
1154527447766962176	19599446	289894237	413277087	766221303632240640
115660898	199566583	2919036392	414962189	769562616003960832
116831511	200568348	2922924261	415022746	803388691477630976
1179525037	201517097	2932115764	416154050	804748838330335234
1194010389186527233	201957241	2965135588	416876488	810200597685272576
119497599	203262579	2982700905	41880514	820497732
1199191479094304768	203555695	298993329	425924139	822016688749154305
1210905474754695168	207208127	299661475	435346412	826044679179362304
123975474	21263335	301045311	45013575	840631711427891200
1252255963	214731619	3022877042	452985859	84427144
1281521971	2242909302	303848470	461900216	845571660090671104
1283507407	229598421	3040732982	465085203	85119380
130376756	2333901440	3040948607	475202064	851492096674541569
130452219	2365896248	305514503	4826563611	852269288
1311971648	2372314050	307558964	4831408433	857303965
131795521	2382387620	3079813761	48351615	862585086050533380
13346352	2394020821	309341660	4838961	867818602791018496
135368243	2401859508	3104949454	48668581	877113807461646336
1355594084	2413234485	3106771385	488097570	881197285769654272
1357033094	2425563233	3131419456	488543082	891599857630236672
138726004	2435331090	3133111667	49016599	893474107
1392054620	244077566	314429644	49616273	898740373
139903735	247379224	3171668783	502092248	90432924
141027991	247888588	3208050838	505731001	907246319781195776
145336121	2511075531	3214613968	51280043	909465013370413056
14575708	251290516	32169306	52422182	922357287481757697
1474986842	252016342	3239745664	532490808	923106269761851392

Table 4.16: Ids of accounts used in our study as journalists.

Id	Id	Id	Id	Id
351566384	317226440	67383910	105507745	3359284941
210132467	121366287	78863928	275674102	22748103
2942519806	37062760	184831017	6503172	33313641
14932200	268875728	149635747	698104968105095170	2413025666
116908364	268429272	276480123	118032930	381059099
115793824	228483751	215815774	18932906	392670224
224589305	114037455	242606835	89784280	216755042
722817261795287041	245863642	796684728	171962889	928718891181838336
103841173	263780425	899764291385077760	402035518	1032383648
482053121	196623028	18627726	544781860	361497515
54235496	28550047	865821732770373633	174726190	20164993
843937475068346373	618952760	257962682	822000025	134917034
1082598510	270307259	270607088	283930140	875525911
239765900	392177110	231424061	3306429471	41562449
107153756	8076532	161237361	16694719	551405439
85384885	341988494	286949912	151513481	119404032
189102085	188699808	143455500	246764832	227977305
96639908	418123217	384893636	44336530	486855644
26557207	2575293810	316706708	1413746881	46061866
102977300	236421131	19232900	107759816	320863192
292464252	429796168	278205448	426874087	322003135
33698078	353756954	94144199	870518544	58788205
18944456	255652146	367308015	303131300	155242359
65369125	1015573542	769919	218844578	176297919
559055487	226196017	156630555	342171657	618166944
288881933	3131004953	84186668	228687267	252305989
82863268	159979641	731573	840592769320116224	139371136
301306806	47936941	505412617	879011415922704389	194543506
464057783	29491384	1239229933	268234381	16947439
374737533	139767585	263806815	114235426	14600838
583625672	6794952	106220868	46296077	3874812255
95232591	235211719	407913953	3168171	225187854
522523887	780183727318106113	185985009	83808453	220693082
368859006	94026873	154925267	250092838	264816224
14831098	1701969248	912746615986900994	210913028	601338508
129636906	2196246424	44651546	87815477	219804554
58487243	87768187	371830459	26211178	12822592
205457327	775988391523586048	1660775364	798904559406120960	538978461
280172465	313886137	112796335	135275648	81379216
65609667	149831017	845237932717985793	345371701	19339140
268748579	21121637	127550968	1068208362	401371220
243569143	59087132	23675375	489342588	143422938
11120292	396059470	309705905	144810157	92147974
50598703	161629977	543731127	832986477197983753	266563663
279541793	1222673754	377680686	2781048212	704022213792546816
14371600	88146816	20637082	1186799853378121728	1080173945440141312
562931470	13937642	100502343	60959278	1095068646
16276054	287662628	2345139698	309097402	245847198
224202948	38720717	1387925532	58489786	247031000
15056194	301949958	229259339	191007783	335223244
3245066614	158730655	295854911	16837276	270916490
489848723	87748292	216056423	271519821	1553130476
195909630	106467821	20388141	209239240	17897369
557099769	296763063	607923199	1144954350437113856	531442414
251820322	92377922	428486667	277416367	1175479403868016642

Table 4.17: Ids of accounts used in our study as political party accounts.

Id	Id	Id	Id	Id
1095238556530995200	84053338	1654997503	498722246	144865638
244627171	71341401	20585956	149064381	516087333
1171342846135275521	558982706	3010640649	17618056	296456619
3058976559	217819308	1863514303	64688134	16120227
19028805	1145723037347733506	1234062995258564609	476874364	414892667
1903891723	1508171319434747908	2848435451	262597648	3351004205
2785536912	262163456	794439428185976832	2351613030	3607648335
89543500	393562501	4360802391	77963344	100332112
148867925	1261633938	155820539	59749870	15133318
1635846486	71523791	4286657254	2240681125	38458062
300120714	242402914	1471889304	253788855	1146013092335239169
3161302295	288159350	275661480	13494262	2375743544
3226151715	715982930561122305	10242562	404265706	3067647501
2373783606	983315580	2303221610	279030847	1369949835730817026
1267843368	283943227	2319411390	130876768	2599889528
3151126193	145969179	2872270425	2517663322	1237315908558872577
3005389661	1111096700	2288138575	18979745	1136340341311848453
1490231790	450488947	3005765446	37743703	2201623465
2951865143	392730219	3054128127	1121870430	848650907499859968
905007924132663298	1321037714	3041105807	69352638	3030322627
2361481177	1178807533	3045358571	1124320844300394502	1435170175721123841
3019929155	85407289	3044851768	373931682	1176094393780117507
2398071896	4185234041	2324417802	140020905	1075722904761974789
2423662292	332305283	2291383507	98101353	2373804017
3032171949	136359280	2978479203	90848443	2308987453
2615685211	334731401	2686076408	510203470	1289319994627678213
1400848982	392870735	3055001794	72494826	917833712
2827819039	732986756572844032	3055168421	14273735	1227648572516130818
2342348916	493507751	2299803774	542494722	717771415945146369
2866276162	172129475	2855434408	927551672	4375928739
827160231536844804	425783717	2967547407	36335926	1085590459714539520
3152398361	725615203	2523186290	50982086	789846117563428865
2591051161	14824411	2562977478	276883917	2355076350
3363932933	748256028786032640	2301274694	93929450	1381619021502644225
3165379271	259409783	2371789428	40900410	854017534865092608
405180461	2379309613	2416100726	399954373	378703097
2398162506	1173373987726008321	2300409494	153351659	
3020408320	1046725508615086081	2329184163	13169022	
91796121	1138929492288819200	2308653811	18097083	
83784273	1354186524934479873	2339427847	216155757	
601887686	1003771289092481025	2377043383	17804233	
16579721	1105212399441899521	20509689	33245408	
368765699	155544328	132613373	166767045	
566258312	1011486391	493960916	33116044	
3906310100	124568057	110250944	1471119488158425091	
4909929401	1563636218	4354211962	47975115	
3008074877	212198795	3186884645	165415384	
494478539	3075990394	1154120960792027137	198847800	

4.3 Research Methodology

4.3.1 Network analysis

The central concept around which online social networks revolve is precisely the concept of a *network*. It is from their network structure that ONS has achieved success; this network structure is what gives them meaning. The network is more than just the sum of all its nodes; the networked structure allows the emergence of collective phenomena through its users. This interconnected web of relationships and interactions among users creates a dynamic environment where new behaviors and trends can emerge, leading to ONS's significant influence and reach. Therefore, network-level analysis has been fundamental in the research presented in this thesis to understand these underlying phenomena. In this regard, the primary method of investigation is social network analysis framed within the field of network science.

Network science is an interdisciplinary academic field that studies complex networks such as the ones comprised by the elements of telecommunications, computers, biological networks, cognitive and semantic and social, among others. It seeks to understand the principles governing these interconnected systems' structure, dynamics, and behaviors. Network analysis, particularly within network science, refers to the systematic study of networks by examining the nodes (individual actors, people, or things within the network) and the links (relationships or interactions) that connect them. It involves using quantitative and computational methods to analyze the structure and dynamics of networks, uncovering patterns and insights that are not evident when looking at individual elements in isolation. This approach allows researchers to explore how the configuration and properties of a network influence the flow of information, the spread of behaviors, and the formation of groups within the network.

Network science and network analysis thus provide a solid foundation for analyzing these phenomena, especially concerning information dynamics. In ONS, information fundamentally flows through the network's links. Understanding how information propagates, how influence spreads, and how collective behaviors emerge requires a deep dive into the network's structure and the intricate web of connections that define it.

The network construction techniques employed in this thesis have primarily been:

- The modeling of relationships in the X social network as a graph, where nodes represent users and edges represent the retweet relationship between users. Specifically, if user A has retweeted user B at some point within the time window, an edge is formed. This approach has allowed us to identify communities using methods such as the Louvain algorithm [CKS14], measure information flows through metrics like efficiency, assess the density of these networks to evaluate their robustness, and identify the main leaders in the conversation based on their central position within the network.

- The modeling of author-media and media-media relationships in media networks. Through this modeling, we have generated graphs where the relationships between an author and a media outlet are formed by the author's contributions to that outlet. Similarly, media-media relationships have been modeled based on the number of shared authors. This has enabled us to evaluate the proximity between media outlets and the processes of information diffusion among them through the influence of their authors.

Regarding the analysis techniques employed, these have mainly consisted of:

- Centrality measure analysis: This involves the evaluation of key centrality metrics such as degree centrality, betweenness centrality, and closeness centrality. These metrics help to assess the impact of different types of users on the network, compare networks, and serve as essential components in the development of algorithms and solutions.

- User-level influence analysis: This utilizes algorithms like PageRank [BGS05, BP98] to identify relevant users in various processes, aiming to pinpoint individuals who hold significant influence within the network.

- Information cascade influence analysis: In this context, state-of-the-art algorithms have been implemented to analyze their effectiveness in understanding the dynamics of information in political processes. These have been compared with proprietary solutions that have demonstrated greater robustness.

The specific techniques are detailed at the beginning of each chapter to facilitate the follow-up of the thesis.

4.3.2 Natural Language Processing

The defining characteristic of online social networks is their network structure. Another critical element is the content, specifically the publications made by their users. Thus, the text through which users share content and communicate is central to information dynamics in social networks.

Natural Language Processing (NLP) has been employed to analyze this case. NLP refers to the branch of artificial intelligence that focuses on enabling computers to understand, interpret, and manipulate human language in a valuable way [ID10].

In this particular study, techniques of Natural Language Processing have been primarily based on Latent Dirichlet Allocation (LDA) [BNJ03] for detecting themes in media content and social network publications. Since Twitter's Academic API already offers data categorized by themes and entities, LDA processing has been used with the thematic analysis of social network conversations.

Deep Learning techniques have also been applied for sentiment analysis based on the BERT [SSRK22] neural network architecture. This analysis detects the emotions associated with publications, discerning positive or negative sentiments, and tendencies towards hate speech.

Given the paramount importance of discourse in social networks in shaping conversations, this research has supplemented text analysis with an exhaustive discourse analysis using the LIWC technology [PCI+07]. This tool has enabled the analysis of the discourse style of various political organizations and users involved in these conversations, providing deeper insights into the linguistic and psychological underpinnings of the political dialogue on social networks.

4.3.3 Statistical methods

To ensure the utmost robustness of the analysis, basic statistical methods have been employed to validate the conclusions drawn from the study.

Specifically, the T-test, Kolmogorov-Smirnov, and Whitney U tests have been utilized. These statistical tests are instrumental in confirming the validity of the analyses, particularly in defining and contrasting networks of disinformation with networks of legitimate information.

The T-test is used to determine if there are significant differences between the means of two groups. The formula for the T-test is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{4.1}$$

where $\bar{X}_1$ and $\bar{X}_2$ are the sample means, s_1^2 and s_2^2 are the sample variances, and n_1 and n_2 are the sample sizes of the two groups.

The Kolmogorov-Smirnov test is a non-parametric test that assesses the equality of continuous, one-dimensional probability distributions. The test statistic D is defined as:

$$D = \sup_x |F_1(x) - F_2(x)| \tag{4.2}$$

where $F_1(x)$ and $F_2(x)$ are the empirical distribution functions of the two samples.

The Whitney U test, another non-parametric test, compares two independent samples to ascertain whether they come from the same distribution. The formula for the Whitney U test is:

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 \tag{4.3}$$

where R_1 is the sum of the ranks for the first sample, and n_1 and n_2 are the sample sizes of each group.

Applying these statistical methods provides a foundation of empirical rigor to the research, ensuring that the conclusions regarding the nature and dynamics of various information networks on social media are grounded in statistically validated analysis.

4.3.4 Information Theory

Despite various algorithmic approaches being tested and applied in constructing the analytical tools and solutions in this thesis, the bulk of the theoretical framework has been based on the application of the principles of information theory.

Information theory is a branch of applied mathematics and electrical engineering involving the quantification of information. Historically, it was introduced by Claude Shannon in his seminal 1948 paper, *"A Mathematical Theory of Communication"*, where he sought to address the fundamental limits on signal processing operations such as data compression and reliable data transmission over noisy channels [Sha48]. Shannon's theory laid the groundwork for digital communication and storage systems, transforming telecommunications and influencing numerous fields beyond its original scope.

The core of information theory revolves around entropy, a measure of uncertainty or randomness. Shannon's entropy, denoted as $H(X)$, for a discrete random variable X with possible outcomes $x_1, x_2, \ldots, x_n$ and corresponding probabilities $p(x_1), p(x_2), \ldots, p(x_n)$, is defined by the formula:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \tag{4.4}$$

This formula quantifies the average unpredictability in a set of possible outcomes. The logarithm is usually taken at base 2, and the unit of entropy is the bit. Entropy is a fundamental limit on the best possible lossless compression of any communication, signifying the minimum number of bits needed to encode a string of symbols based on their probabilities.

Another essential concept is cross-entropy, which measures the difference between two probability distributions, P and Q. For two probability distributions over the same underlying set of events, the cross-entropy from Q to P is defined as:

$$H(P, Q) = -\sum_{i=1}^{n} p(x_i) \log q(x_i) \tag{4.5}$$

Cross-entropy is used to quantify the inefficiency of assuming that the distribution Q is the proper distribution when the actual distribution is P. It is instrumental in machine learning for training classification models, where it is used as a loss function to minimize the difference between the predicted probability distribution and the proper distribution of the data.

While originating in telecommunications, information theory's principles have transcended into various domains, including modeling information dynamics in online social networks. These networks are intricate systems where information dissemination, influence, and user behavior can be rigorously analyzed using information-theoretic measures. For example, entropy can be used to measure the diversity of content being shared within the network. At the same time, mutual information can help quantify the amount of information shared between different parts of the network, shedding light on the influence one user or group has over another.

Information theory offers tools to understand and model information flow and dynamics in many modern contexts, including online social networks. In this thesis, information theory has been employed to evaluate the influence of communities, media, and individual actors. The information theory approach has been beneficial for measuring influence relationships beyond the network defined by explicit content sharing. Through information theory, it has been possible to study certain elements' influence on others through the information shared within short time windows.

4.3.5 Simulation

While the earlier parts of this thesis (Parts 2 to 4) takes a much more descriptive and analytical role, focusing on defining the main elements responsible for generating the dynamics of information in social networks as well as developing analytical tools in this regard, the fourth part of this thesis has sought both to explore the role of automatic information filtering systems based on artificial intelligence in catalyzing or modulating these social phenomena, and to explore effective algorithmic intervention strategies to move the network towards a healthier state for its users.

Thus, content recommendation systems in online social networks have been studied through their leading families, mentioned earlier in this thesis in the introductory section. Due to the inherent complexity of such large platforms, the analysis in this part has been based on simulation. Given the inherent complexity of the designs of significant recommendation systems, it is practically impossible to develop a mathematical-analytical solution regarding their role in information dynamics due to the excess of variables to control in each case. Simulation, therefore, emerges as the best solution, allowing us to define algorithmic intervention strategies in the form of new recommendation algorithms capable of addressing the studied risk phenomena [Eks21].

The simulation techniques employed involved partitioning the various datasets corresponding to the phenomena to be studied into temporal windows, generating a graph for each. Starting from the first selected time window data as the base, recommendations were generated in each case using the posts

created by the same set of users in each time window. The various recommendation systems were used to generate the sets of recommendations in each case, and their accuracy was verified by contrasting with the actual user activity in each temporal window. Recommendations were always for posts. Accuracy and the links in the network were measured by retweet actions for their simplicity and ease of comparison. A simple model was followed for user behavior toward recommended posts, assuming that the user would receive four posts with a maximum sharing probability for the first post and a 25 percent decay for subsequent posts.

The recommended approaches tested were content-based, collaborative filtering, reinforcement learning (bandit), and deep learning (deep neural network). The specific details of each approach are detailed in the fourth part of this thesis.

4.3.6 Qualitative methods

Qualitative techniques have been employed in this study to complement the quantitative methods, which constitute its central axis. Given the phenomenon's complexity under investigation and its strong interconnection with social sciences, qualitative techniques have broadened the analytical perspective, ensuring a comprehensive understanding of the studied phenomena.

In-depth semi-structured interviews have been utilized to gain firsthand insights from one of the key players in this analysis: content creators on the internet, particularly those focused on politics. Semi-structured interviews are a flexible, guided form where the interviewer follows a pre-determined set of questions but is free to explore topics that arise spontaneously during the conversation. This approach allows for a detailed and nuanced understanding of the interviewees' perspectives and experiences.

Additionally, the study has been enriched with critical media analysis, a technique used to examine and interpret media texts and their context. This method is advantageous in identifying the framing techniques and biases employed by media outlets, especially in contexts of political propaganda. Critical media analysis involves systematically evaluating media content, focusing on what is said and what is left unsaid, and how the presentation of information may shape public perception and understanding. This qualitative approach complements the quantitative data, offering more profound insights into the strategies and impacts of media in political discourse.

Part II

Online Political Ecosystems

"Cyberspace. A consensual hallucination experienced daily by billions of legitimate operators, in every nation, by children being taught mathematical concepts... A graphic representation of data abstracted from banks of every computer in the human system. Unthinkable complexity. Lines of light ranged in the nonspace of the mind, clusters and constellations of data. Like city lights, receding..."
— William Gibson, *Neuromancer*

Chapter 5

Digital Activism

5.1 Introduction

Digital information ecosystems in the political context of online social networks involve various actors, including political parties, users, activists, journalists, and digital media, all contributing to dynamic information flows. Activists, traditionally operating on the system's margins to challenge the *status quo*, have transitioned from offline to online spheres, leveraging digital technologies to amplify their reach and influence. Historically, activism addressed underrepresented issues through both conventional methods (e.g., press campaigns, signature collections) and direct actions (e.g., wall paintings). With the advent of the internet, activism evolved into digital-activism or cyber-activism, characterized by lower participation costs, global coordination, and online protest actions. This shift led to the creation of alternative communication spheres and a rise in grassroots activists, though often at the expense of organizational cohesion. This evolution highlights the profound role of technology in shaping political participation and the emergence of new activist forms.

This chapter explores the evolution of digital activism, focusing on the emergence of the influencer-activist as a key figure within political ecosystems on social media. Building on the broader research goals of the thesis, this chapter contributes by characterizing the influencer-activist and examining their behaviors, motivations, tools, and relationships with other actors. This analysis provides insight into how these individuals influence online political discourse and how their actions integrate with and reshape traditional political structures.

The chapter is organized as follows: Section 5.2 introduces the concept of digital activism, tracing its historical evolution and examining how new technologies have transformed traditional activism into digital and cyber-activism. Section 5.3 provides a theoretical framework for understanding the influencer-activist, detailing their unique attributes and distinguishing them from other activist forms. Section 5.4 outlines the methodology used in this study, including the participant selection criteria and the data collection process through interviews. Section 5.5 presents the results of the study, focusing on the influencer-activists' ideology formation, activism strategies, tools, and relationships with other actors such as media, political parties, and opposing activists. Section 5.6 discusses the implications of this chapters findings, relating them to the broader objectives of the thesis and situating the influencer-activist within the political information ecosystem. Finally, as conclusions, Section 5.7 reflects on the insights gained and suggests future research directions to further explore the dynamics of digital activism and its

impact on political ecosystems.

Research questions

Our aims in this chapter are summarized in the following research questions, that contribute to better understand the main actors involved in online political processes developing in social networks.

- **RG1:** Develop a valid definition of a political information ecosystem for this study, involving all of its main elements and thus allowing us to study their information dynamics.
  - **RG1.RQ1:** Which are the main information disseminators in the political context on micro-blogging social networks and how they develop their activity?
  - **RG1.RQ2:** What motivates political active users to perform online activism and promote narratives?
  - **RG1.RQ3:** How do this main information disseminators interact between each other during political processes?

5.2 Background

Cyber-activism, also known as digital-activism, refers to the use of digital technologies by an individual or a group to achieve social change. In this new form of activism, that develops as a consequence of the adoption of new communication technologies by activist organizations [MA03], social change is achieved by means of efficient coordination, an expansion of reach limits, and a rapid diffusion of new information [Lyn11]. This kind of activist emerged along with the first Internet connections in the 90s. It went through a set of different stages or waves as communication technologies evolved on par with social processes all over the world. Karatzogianni explores in [Kar15] four stages of cyber-activism.

The first wave began in 1994 with the *zapatista* movement in Mexico and the first anti-globalization protests. During this wave, the key technological factors were telephone networks, the first Internet connections for effective coordination, as well as the first portable photo and video devices that allowed to start recording footage during protests and claims for later sharing. The next stage comes between 2001 and 2007 with the *anti-war in Iraq* protests. In this wave, we find the first consolidated forums and blog networks to be a decisive means for generating counter-narratives and effectively coordinating and mobilizing activist organizations. The third stage started in 2007 and goes hand in hand with the economic growth in countries like Brazil, Russia (and its area of influence), India, and China (commonly known as BRICS). Blog networks and forums were decisive factors in organizing complaints and protests against corruption and authoritarianism. They played a role in protests such as the color revolutions in Eastern Europe. The fourth wave appeared in western and Arab societies between 2010 and 2013 and can be associated with phenomena like Wikileaks, the Spanish 15M protests, the worldwide *Occupy movement*, and the Arab spring [Cas12]. Online social networks such as Twitter, YouTube, and Facebook started gaining widespread traction around 2009-2010. They caught the eye of many activist organizations and individuals, who saw great potential for coordination in them. This fourth stage of cyber-activism is characterized by the heavy use of social networks for tasks such as narrative diffusion, protest coordination, and rapid and, sometimes, personalized news distribution (e.g., news based on the city where the protest is happening). Precisely due to this massive use of online social networks, these protests experienced

tremendous success during their first days due to their possibilities in recruiting participants through social media, although they failed to consolidate in structured and cohesive organizations [Tuf17].

The Internet was the origin of most protests related to this fourth wave. Some forms of protest that appeared here did not even move offline[1]; that is the case of hashtag-protests (that occur when a large set of cyber-activists create and promote a hashtag to perform a claim on social media) or change.org[2] campaigns. Other forms of protest started online with an online target as a goal, such as the case of cyber-attacks like DDoS[3] carried out by so-called hacktivists [Kar21].

The last stage mentioned by Karatzogianni involved a step rarely seen in the other waves. Protests such as the *Occupy movement*, the Spanish 15M, and the ones related to Wikileaks and the Snowden affair significantly impacted the political agendas. Traditional political parties started paying attention to people's claims both on the street and, particularly, online. Some prominent activist figures that gained relevance through their social connections and "virality" online used their newly gained social capital and jumped straight into mainstream politics. By carefully looking at the evolution of the political landscape in western societies after 2014, we have reasons to define a new wave of cyber-activism that started after 2014 and is mainly characterized by events like the 2016 elections in the USA or Brexit, or more recently the COVID-19 outbreak [Sch20]. Also, by new forms of populism offline but online, from populist pundits to Internet trolls [VdBH20, FI20].

In this fifth wave of cyber-activism, the figure of the activist evolved from the 15M / *Occupy movement* activists that were mostly left-wing or even not interested in specific political ideologies, used social networks mainly as a tool to call to action on offline protests or to protest, and lacked a structured network or a clear agenda. That led to a new kind of cyber-activist less reliant on hierarchical organizations, less focused on street protests and offline activity, well aware of the fragility of online environments, more alienated with established political ideas, both left and right-wing, and, most of all, with a clear goal of generating a change in society by pushing a particular ideology. The cyber-activists belonging to this fifth wave were frequently less integrated with structured activist organizations and resembled the figure of the so-called "influencer" [vDD21], that is, an individual that "speaks the language" of the younger generations [DD19b], has a large follower base that posts content related to a particular field (e.g., sports, music, tech), and has a great potential to influence to her audience by recommending them products, services or, in this case, political ideas or even candidates [WHD20].

To our knowledge, though this is a critical and exciting figure that plays a essential role in modern-day political communication and online social networks and protest dynamics, this new influencer-type cyber-activist or influencer-activist is still an under-researched topic. Many studies focused on comprehending new forms of political communication in online social networks are either quantitative and based on big data sets of X only, or very focused on a particular case study. Trying to understand such a complex phenomenon by solely using quantitative methods like social network analysis techniques without a previous deep qualitative study may lead to incorrect or incomplete interpretations. In the following sections, we shed light on this aspect qualitatively by studying how the influencer-activists form their ideology, motivations, tools, and techniques, as well as their relationships with traditional mass media, political parties, and other individual or organization activists.

[1] In this chapter, we refer as offline to the real world, outside Internet and social networks, in line with conventional activism

[2] change.org is an online platform that allows individuals to create and promote claims by signature gathering.

[3] Distributed denial of service attacks aimed at blocking the access of an online service by collapsing its bandwidth.

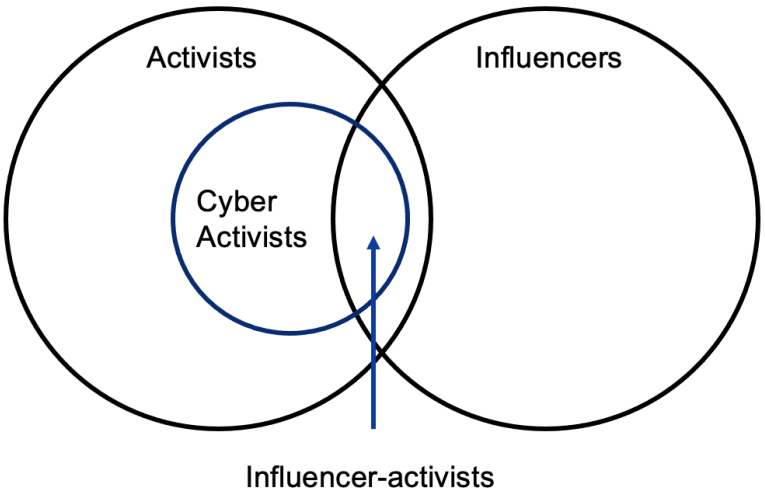


Figure 5.1: Conceptual visualization of influencer-activists.

5.3 Definition

The influencer-activist is a complex figure that moves in a network of online social networks, communication platforms, and overall digital tools. We propose the following definition:

Definition 1 (influencer-activist)**.** *An influencer activist is an activist who is prominent on social media, not an affiliated member of any political party nor a working member of any conventional mass media in an exclusive manner. She focuses most of her activism on online social media, is highly conscientious and very vocal about her political views and ideas, and performs campaigns individually or collectively to push certain topics into the political agenda. To achieve that, she uses a wide variety of tools and strategies online, offline, or hybrid.*

According to our definition (represented in Figure 5.1), influencer-activists share common elements with cyber-activists in their preferences for OSNs and forums, as well as with social media influencers regarding their wide reach and will to promote their online characters as personal trademarks. However, in contrast with classical cyber-activists, they are focused on political campaigns and/or influences their social visibility may achieve.

5.4 Specific Methodology

The data collection methodology used for this study, which focuses on identifying and analyzing political cyber-activists, is detailed comprehensively in Chapter 4. To summarize, a list of seed user profiles was generated through an initial identification of political think tanks, associations, and prominent X users on both sides of the political spectrum in Spain. The snowball sampling technique was employed to expand this list, and several criteria—such as sustained activity, political focus, and follower count—were used to filter the profiles. Manual review and snowball sampling were conducted iteratively, and the final sample included 17 participants who were subsequently interviewed.

These interviews were conducted in 2021 from January to October. The demographic features of the interviewed individuals can be found in Table 4.13. The interviews were conducted with Spanish individuals in the Spanish language who complied with the aforementioned Definition 1, which is of

particular interest due to the relevance of cyber-activism in the country and many mass protest events that happened in the recent history of Spain (around 2011) such as the 15M protests and the "indignados" movement, where social media, especially Twitter, was a fundamental coordination and execution mean [PLCA14], along with the rise of populist parties [Vam20, SR18, SBDMP16] and new forms of political communication from both left and right-wing organizations in the Spanish Twitter-sphere [Pay19].

The interviewed individuals used a variety of social networks that are summarized in Table 4.14. For a full explanation of the data collection and participant screening process, please refer to Chapter 4.

5.5 Results

In the next sections, we summarize the responses collected from the activists interviewed to address the research questions previously presented. We start with their ideology formation process and how their introduction to activism was, followed by an analysis of their social media strategies, including their publication patterns, together with the tools, methodologies, techniques, and strategies used; we finish with their relationship with opposite activists, the press, and political parties. For this, we manually analyzed the collected responses, since the reduced number of interviewees allowed for it. In the future, we aim to use more scalable solutions such as questionnaires and affinity diagramming [Luc15].

5.5.1 Ideology formation

The first research question aims to understand how this new kind of cyber-activist develops her ideology. Historically, activists developed their ideology through personal offline experiences, such as living in poverty or any form of discrimination. In most cases, access to knowledge allowed them to form a set of opinions and to see the world through a particular ideology [Tar97, Joy10]. This first research question involved studying the ideology formation process of this new type of activist by carefully examining the role of the Internet and online social networks, particularly during their introduction into activism, to understand their importance or lack thereof, and answer RG1.RQ1 and RG1.RQ2.

The majority of cyber-activists (90%) we spoke with started getting interested in politics at an early age, most of them (60%) during high school or their first college year. Most of them (80%) stated that they got introduced to their ideologies by a close friend or relative, which resembles the traditional introduction to activism [dPD16]. They then proceeded to reinforce their ideological points of view by browsing through content on online news sites such as hyper-partisan media, forums, and (mostly) YouTube channels and X accounts. Some of them (30%) put more focus on their personal experiences through the role of a close person; a friend or a colleague is a recurrent example.

Other relevant aspects presented by the interviewed activists come in the form of a natural curiosity toward political topics and ideas, as well as a desire to get involved in social matters. After being introduced to their political ideologies in a conventional or offline manner, all the interviewees stated they continued to develop their political ideologies in an online manner, mostly and occasionally by sharing book titles or websites and discussing ideas offline with their friends. When it comes to their online activity, some pointed to online forums and hyper-partisan media. However, the majority signaled Google searches and OSNs (such as Twitter, YouTube, and Facebook groups) as their primary way of getting political content. When asked about the reasons for that, they pointed to the ease of use and the facilities for getting new related content recommended by social media recommendation engines.

When asked if they used to search or even just occasionally saw actively content related to opposite political views, some of them (35%) answered affirmatively and some negatively. Those who did not search for information related to contrary points of view generally blamed the social media algorithms. Others just said straight that they were not interested in contrary points of view.

5.5.2 Introduction to activism

Traditionally, activist individuals started activism straight away, along with their ideology formation process aiming towards a change in society's values and moral standards [SEN18], as their activity was embedded in an offline cohesive organization [Tar97]. Though these patterns still correctly explain the experiences and activities of a large number of activists, this new kind of influencer-activist presents specific dynamics such as the fact that their ideology development process may kick-start offline but develops mainly online, with a less structured and cohesive offline network, and connected with other like-minded individuals in a purely virtual or online manner.

Answering RG1.RQ2 we clearly identify that the majority of the individuals (90%) started their activism online when they felt they knew enough and had solid ideological fundamentals about the topic they wanted to talk about to offer some "relevant and useful content" as they said. Their online activism was inspired by other activists they were following on online social networks such as Instagram, TikTok, YouTube, and Facebook, but mostly X and Twitch. They saw those other activist figures as influential and popular figures and thought, "if it worked for them, then it might work for me".

Furthermore, based on the performed interviews, we detect intrinsic and extrinsic motivations when it comes to creating political content online, thus complementing related studies [LKM17]: on the one hand, the interviewed activists use social media to express themselves. This space allows them to be creative and receive feedback (most interviewees, especially the ones who started on video platforms such as YouTube or Twitch, started with their online activism as a hobby). On the other hand, social media also allows them to gain popularity and develop personal social networks that may provide them with professional opportunities later on. We can also detect a specific bandwagon effect when it comes to them joining political ideologies and starting to post content, as they perceive that joining those ideological streams will make them quickly gain attention.

When asked about the particular reasons that moved them towards being politically active online, they appealed to their desire to elevate the morals of a society that is overlooking essential issues such as women's rights, LGTB community rights, the fight against poverty, protection of the environment, racism, economic liberty, religion, or immigration. Some of them (24%), mainly the ones identified with right-wing ideological positions, appealed to the need to fight a "cultural battle" against the "mainstream media" and "the political ruling elites" as they called it, that is, to generate a particular counter-narrative against the one that is being pushed by the leading Spanish mainstream media and the ruling political parties. When asked about that "cultural battle" concept, they pointed at online social networks as their main "weapon."

As the traditional off-line type of activist focuses their activity offline, in a local context, these new influencer-activists are more focused on the online sphere, and sometimes they even push for offline actions. When asked about the core motivations of their activism, most of them contemplated offline actions, such as demonstrations, to push their issues into the political agendas of the ruling parties. They also considered their actions to force some corporations to take a particular action. The vast majority

of them stated that their main goal was to produce a change in the collective conscience; that is, not to incorporate a particular issue in the political agenda of a party but to root a particular set of ideas into the conscience of the average individual in a Gramscian way[4] [Gra94], making them request a particular set of political organizations in the future.

To summarize their aspirations, they try to develop a collective conscience towards an issue by embedding their particular cause into a complete set of ideas (e.g., taxes–liberalism, anti-racism–socialism). They try to perform agenda-setting for their particular causes forcing political organizations to talk about specific issues. Overall, they try to provide alternative interpretations and behavior around a particular issue at the societal level.

5.5.3 Towards professionalization

As a consequence of the aspects described in the previous section, and according to their experiences, the influencer-activists tend to present very stable publication patterns focused on their particular topics at their starts, and then as time passed and their profiles started to grow, they were offered to take part in specific activist networks. This makes us address RG1.RQ2.

Based on the collected answers, this process started entirely online, usually through an introduction and invitation by a direct message on sites such as X or Instagram, followed by an invitation to take part in a closed group on WhatsApp or Telegram. Some of the interviewed activists also mentioned that the invitations to join organized activist groups on WhatsApp or the like eventually led to invitations to join offline events, such as demonstrations, conferences, political organizations, or even writing/speaking in conventional mass-media. Similarly, as the activists grow in popularity, they may start receiving some revenues due to donations or embedded ads on videos, for example. So they start looking at the platforms more carefully, knowing their followers better, and considering the possibility of self-financing to dedicate all their efforts to the cause. Platforms such as YouTube, Twitch, or Patreon are used for that and will be inspected in more detail in the next section. As the activists start receiving revenue or considering so, they start professionalizing their activity. Professionalization comes by a clear differentiation and interrelation of content between platforms, more strict posting routines, an improvement in the quality of multimedia material and the effort to provide more carefully curated material to their audiences.

Other remarkable aspects of this process include moderation in their narratives and occasional conflicts and controversies with other activists who are not receiving revenues. For example, one of the interviewed influencers told us that he used to hold a hard line on Spanish migration policies, which softened once his fan base widened. As one of the interviewees states:

"Since I started receiving money from Patreon and YouTube, I've been criticized by some of my colleagues, who told me that I "sold myself" to the system. I don't like it either, but you know, I think I can reach way more people and push for a greater change this way." Subject 5, communist.

On the other hand, other interviewees talked about their relationship with their audiences and how they push them into a particular discourse:

"Since I use Patreon, I get in touch a lot with my audience, but I think that I live in an echo chamber, they demand a kind of content, and I have to provide it to them. I don't know if I'm convincing them or if

[4] Antonio Gramsci was an Italian political theorist who stated that in order to achieve a change in the political system, one must change the collective conscience by performing militant activism in all spaces of public life.

they are convincing me. I feel that if someday I change my mind on any political idea... that may have a cost, you know." Subject 12, conservative.

Finally, almost all interviewed participants expressed preoccupations about a possible deactivation or, as some of them referred to, "censorship" of their social media accounts, especially the ones with more radical discourses.

The ones who are at the same time involved in some way with conventional organizations also stated that, since they are now "public figures", they have to moderate their activism in order to integrate it into their political organization and effectively maintain the relationship. Some of them (a minority) stated that as they grew in audiences, they had to start appealing to a large and more diverse base of followers, so they had to moderate.

In general terms, most of the activists start from a point where life experiences made them prone to the interpretation of the world through a particular lens, then some friends or relatives (60%) introduce them to an ideological frame where they can use their previous experiences to interpret the world. From that point, a small portion of the activists (15%) chooses to go further and implicate in the promotion of an ideology and or the defense of a particular social cause. Even a smaller portion (10%) of that set will end up making a life out of it.

We summarize their development into activism in Figure 5.2.

5.5.4 Tools and methodologies

The literature acknowledges that new technologies have been the main factor in the rise of this new kind of political activism. In particular, this new influencer type of cyber-activist proves to be proficient in the management of a variety of online social networks and digital tools. They are very aware of their possibilities and prove knowledge about posting strategies and even about the mechanics of some recommendation algorithms. To gain a deeper understanding of how these activists function in online environments, we need to start by knowing their tools. In this section, we aim to confirm these ideas and, hence, answer RG1.RQ2 and RG1.RQ3.

Online platforms The influencer-activists that were interviewed in this research are active in online social networks and digital applications as listed in Table 4.14, each of them having a specific role:

Twitter: Though a bias may exist as Twitter/X was the original social network from where we extracted candidate profiles, this is often the leading social network of choice amongst the activist community, both traditional activists and especially cyber activists [Tuf17], and of course the influencer-activist. All the interviewed people are active on Twitter. In general terms, they use it to reach their audiences with ideological content, to quickly react at social or political events, to perform calls for action in the form of hashtag promotions and the like, to engage in political debates, and as a central point where they can diffuse content created in other platforms, pointing their X followers to follow them on other media as well.

Facebook: The interviewed influencers who have been active online for the most time signal this platform as one of their main entry point into political activism, though they almost do not use it now. When asked, they stated that Facebook's private groups were a key point in developing a sense of belonging as well as a political identity.

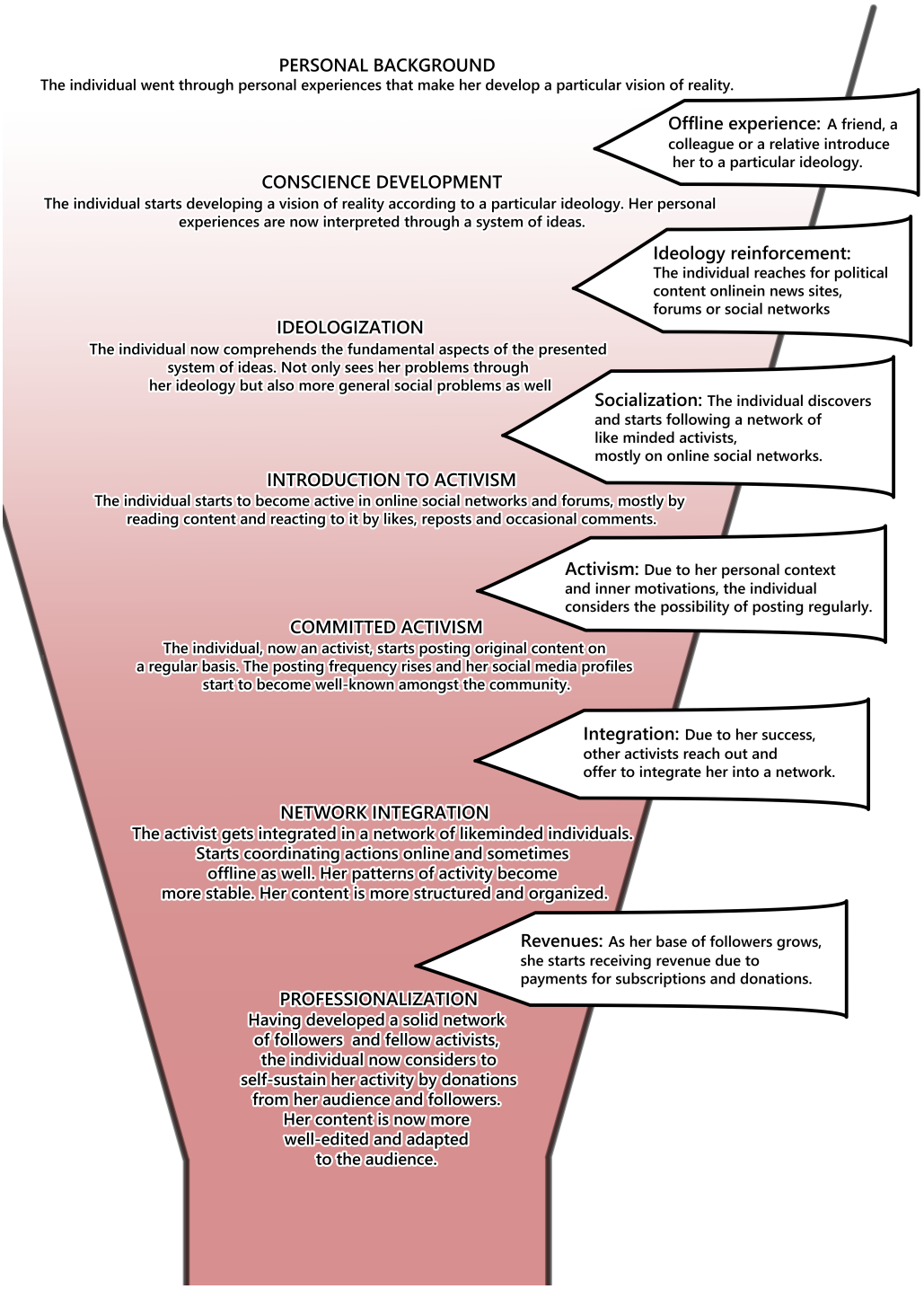


Figure 5.2: Influencer-activism funnel.

TikTok: This network is a novelty for most of them. From what the interviewed activists answered, this is probably one of the less political social networks for now though some of them are starting to use it as a way to mix their personal life with their activism and thus reach wider audiences.

YouTube: Many of the interviewed activists use YouTube as their main communication channel when it comes to promoting their narratives, as this platform allows them to create long and lasting content, develop ideas in detail and report and comment on real-life events quickly. The ones who use YouTube (47.06%) states that they use it when they want to develop and share a solid, informed opinion

on a particular topic. Usually, the videos shared on this platform are promoted later on via X as well as Facebook.

Instagram: All the interviewees who use Instagram (59.82%) use it as a mean to embed their political rhetoric into their daily life by combining the presentation of their daily non-political activities along with their political message in a similar manner as with TikTok through mixing image and video content. As a Christian activist girl said:

"I use Instagram as both a mean to talk about my daily life and a mean to promote my political message. As my message is a bit controversial these days, I like to show that you can be a religious person and have a normal life at the same time, you know, go to pubs on Friday and go to the church on Sunday." - Subject 16, conservative.

Twitch: Twitch is seen as a social platform that is rapidly gaining traction among cyber-activists. They use it as a way to create fast content related to very recent news as well as hot social topics. Most of them often re-use the content created during live streams on Twitch chopping it and uploading it directly to other platforms like X or YouTube expanding the impact of their message. They (35.30%) tend to prefer Twitch to platforms such as YouTube due to the fact that they perceive Twitch to be more transparent, less censoring, as well as more viable in terms of economic interactions such as likes and retweets with their audience. The platform YouNow was also mentioned and is used the same way as Twitch.

Tools Along with their social media profiles of choice, they also presented some digital tools that complement their presence in social networks in their online activism.

WhatsApp: Mostly used for activist-to-activist coordination via private conversations or tight-knit groups. Sometimes, as the individual starts gaining traction on social media, she gets invited to participate in those groups.

Telegram: Telegram is used both as a distribution channel for political narratives via Telegram channels, as well as a tool for activist-to-activist coordination via private groups or perhaps activist-to-followers coordination via public or semi-private groups (invite only) groups.

Discord: It is a complex tool for setting up communities composed by the activist, their inner circle, and part of their audience. The tool allows the creation of diverse chat rooms commonly dedicated to various topics, such as the discussion of a set of political ideas, sharing and commenting on news, or even coordinating some actions. The tool also facilitates the sharing of multimedia files such as videos or PDF files, easily allowing the exchange of political texts and ideas. It also allows diverse user roles, such as content moderators and administrators who, in this case, use to be very active users close to the inner circle of the activist. Sometimes other influencer-activists as well, this way allowing a particular hierarchical structure as well as the formation of collective identities [BS12]. To our knowledge, this tool very often consists of the main gathering platform of the hard-core activist circle, where the most dedicated individuals discuss ideas and actions.

As one of the interviewees (a conservative activist) stated:

"So whenever I find a user who used to be active during my live streams on Twitch, sometimes I invite him to join the Discord server we have here. We used to comment on news and discuss authors, sometimes we stay up till 3 am talking about a single book..." - Subject 4, conservative.

Patreon: It is used as a tool that allows users to get revenues and self-sustain their online activity. Their audiences can subscribe to their content by using it. When subscribing to a content creator, the

activist in our case, starts paying a monthly fee defined by the creator. Subscribers to this service use to get custom or premium content from the activist, such as exclusive videos. Sometimes they have the chance to influence the activity of the account owner by interacting with him via private messages. This is also used by activists as a mean to identify the most politically active followers and recruit them into other private or semi-private groups to get them involved in activism in some manner.

As one of the interviewees states:

"Since I started receiving money from Patreon and YouTube, I've been criticized by some of my colleagues, who told me that I "sold myself" to the system. I don't like it either, but you know, I think I can reach way more people and push for a greater change this way." Subject 5, communist.

On the other hand, other interviewees talked about their relationship with their audiences and how they push them into a particular discourse:

"Since I use Patreon, I get in touch a lot with my audience, but I think that I live in an echo chamber, they demand a kind of content, and I have to provide it to them. I don't know if I'm convincing them or if they are convincing me. I feel that if someday I change my mind on any political idea... that may have a cost, you know." Subject 12, conservative.

PayPal: It is used as a tool for receiving donations. Some activists provide PayPal custom links in their social network profiles and regularly ask for donations to self-sustain their activities.

Linktree aggregates: It is a tool that allows users to generate a custom page containing links to other websites. Some of the interviewed activists use it to list all their social media profiles in one place, to make it easy for new followers to get in touch with their online presence.

In summary, their online presence can be represented as in Figure 5.3, by considering that each social network has its target audience and inherent norms about what should be posted, which is how and why the activists chose them differently. The diagram represents the uses and information flows between each of the different online platforms. Relations between platforms may come in the form of: **promotion and reuse**, where the content generated in one platform is promoted in another one to reach a larger audience; in the form of **comment**, where the content generated by a user, often by a political antagonist, is criticized or mocked by the activist to appeal to her audience; **recruitment** where the activist specifically used the platform to recruit activists that will join her campaigns and help her perform her actions; **finance**, where the activist seeks to obtain revenue from her audience/supporter base; and **action**, where the goal is to start some kind of campaign such as a protest, boycott, slogan promotion, or the like.

Hence, their online presence can be summarized by considering that each social network has its target audience and inherent norms about what should be posted, which is how and why the activists chose them differently.

In this case, as the content created by the activists and their overall activity are created to be shared on online social networks, we observe technology is not the main enabler but the main reason for this type of activism [Lyn11, Cas13] and that social networks are used in a scheme of cooperation and competition as the activists are well aware of the potentialities of each network due to their particular differentiated functions [Cha13]. They integrate each platform into their own custom network and adapt the result to their own political goals, something that conventional activists are not as aware of, due to their lack of IT and social media proficiency and their focus on events outside the Internet.

As we can see, X is the classical tool of choice for many kinds of influencer-activists [SJ16] appears

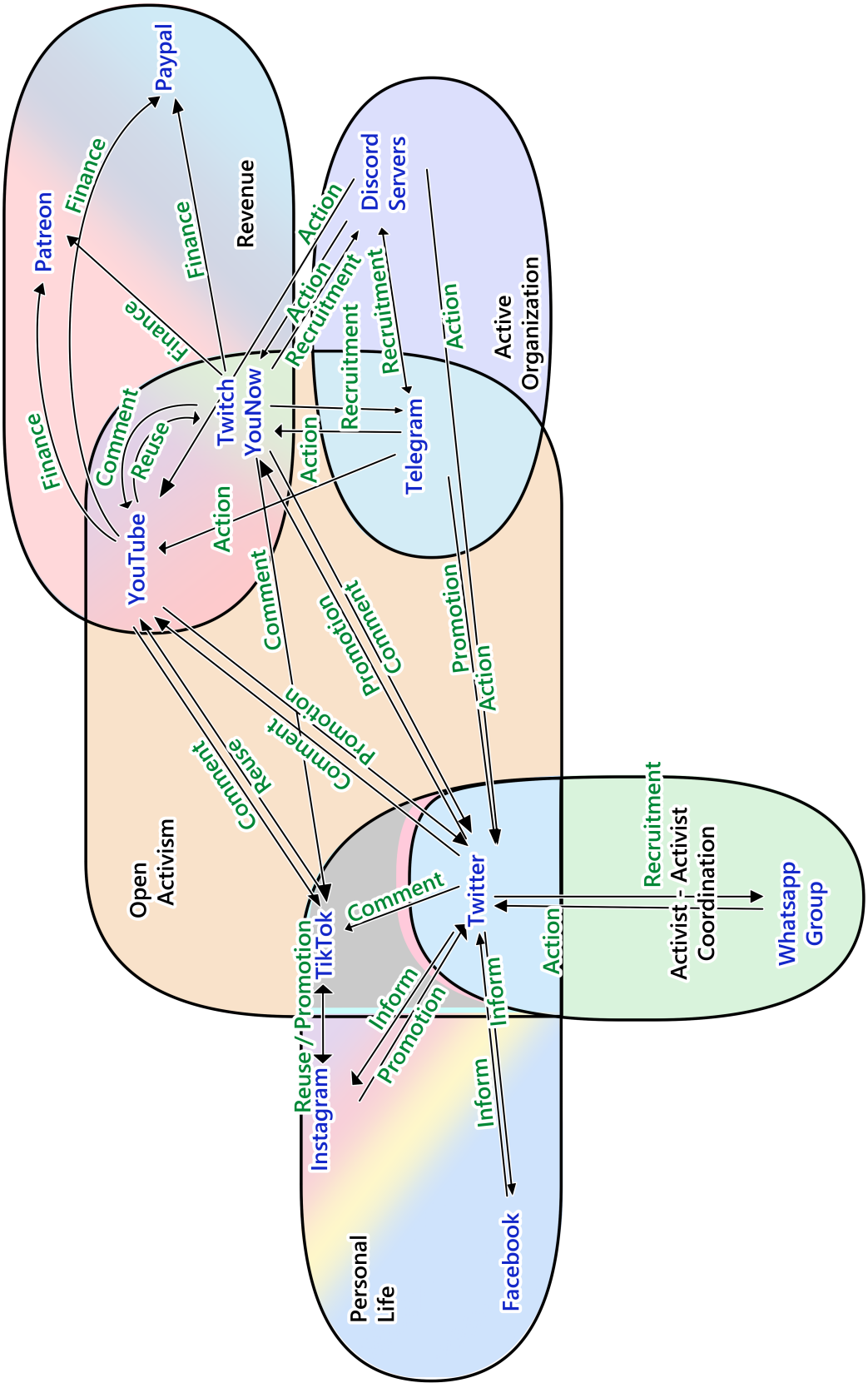


Figure 5.3: Social platforms and tools connections diagram.

to be the central OSN for most of the interviewed activists. They use it to comment on topics of relevance and engage in debates or promote actions, and point their audience to other social networks. We see a clear division of social networks such as Facebook, Instagram, TikTok, or sometimes X that act as platforms less used for sharing political content where the activist fundamentally posts about her personal life. Then, as observed during all the interviews performed, there is a more general set of apps that might function as an open channel of some sort where the individual is active in spreading a particular message by tweeting claims, publishing politically charged videos to educate the audience, complaining about relevant happenings on Twitch, or sometimes even dancing to the partisan beat on TikTok [MSPH20].

Professional activists who are getting revenues from their online activity add a new set of tools to their networks. They usually take a profit from their large base of followers on Twitter, especially on sites such as YouTube or Twitch to ask for revenues in the form of subscriptions to Patreon or donations through PayPal. The last set of tools involves the set-up of custom communities on places like Discord or private or semi-private groups on Telegram that go beyond calls for action and narrative distribution [BS12]. They are often used to recruit and educate a hard-core base of followers that will mostly be prone to follow, if not propose, online and offline actions from trending topics or change.org campaigns to street protests or actions of citizen journalism. Thus, we define five main kinds of relationships the activist can hold with different platforms: i) *inform*, where the activists simply let others know about their profiles or the profiles of other users of interest on other platforms; ii) *comment*, where aside from that, the activist also discusses specific publications to signal their relevance in a positive or negative way; iii) *promotion*, where the user not only informs but actively encourages their follower base to follow and interact with the rest of their profiles; iv) *recruitment*, where the activist identifies, selects, and incorporates parts of their follower base to more tight-knit groups and, finally, v) *action*, where the individual coordinates with other activists, commonly by using closed or semi-private groups to perform online actions not limited to narrative promotion but also ones such as censorship attacks, hashtag hijacking [GB17], or trolling campaigns against their adversaries [KPM19].

5.5.5 Techniques and strategies

Along with the aforementioned tools comes a set of techniques and strategies utilized by these activists in order to make the most of them in spreading their political messages. These will be analyzed now in order to answer RG1.RQ3.

As mentioned earlier, the interviewed individuals were very aware of the potential of online social networks and the Internet to generate new communication spaces and bring politics into every aspect of social life [Pic08]. Some of them (40%) were very conscious that the Internet is one of the few, if not the only space, where they can push their narrative [Cas12]. They also take profit from the ubiquity of electronic communication devices and social interactions through digital means to generate a particular environment where they interpret many aspects of their followers' daily lives through the political lens of their discourses.

When asked about the kind of content they choose to post online, most of the interviewed individuals stated that one of the main aspects of growing a successful profile on social networks is to be very specific, to post regularly, and, if possible, to post in a particular time window.

Another fundamental aspect of their online communication strategies is to be aware of all of the mainstream news, the political happenings, and any form of social controversy to be quick in commenting

on them. Whenever a relevant event happens and is pushed into the news, most of the activists are quick to develop analyses and interpret them through their ideological lenses, where time is critical. During the process, sometimes the actual events that happened and became news can lose their original form to adapt to a particular narrative. In extreme cases, some activists may even deform reality to create false news (also called "fake news") to push for a particular topic and spread propaganda.

Finally, the last publication strategy carried out by individual activists alone involves getting into controversies, debates, or overt aggressive discussions over online social networks. Those happen majorly in networks such as X in the form of replies, quotes, and mentions, on YouTube in the form of criticism over videos, and on Twitch in the form of call-outs and criticism. They signal that when they get into these controversies, they usually experience a growth in interactions such as views and followers.

When they act in a collective manner, that is, they coordinate with other activists, their audience, or both, one can appreciate that their strategies fit mainly into three categories that is financing, recruiting, and political action.

As commented earlier, the influencer type of activists actively explores the possibilities of all the tools they detect to integrate them into their network in the best way. Very often, whenever they feel they have a chance to do so, they would start enabling their online content to be "monetized", that is, to start receiving revenue according to the number of visits they get. YouTube is the central platform used for that, but as some of them point out, it may not be the best platform to do so, as an interviewed activist noted. They usually start on YouTube to grow a base of subscribed followers and point them to other platforms such as Twitch or Patreon.

Some of them (29.1%) complement their activities by asking for donations through Paypal, though Patreon is the main choice for getting revenue for their online activity. Some are worried about the sacrifices they might need to make regarding constancy in publications and the kind of content. The ones that use Patreon amongst the interviewees signaled that they started to use it once they knew they had an excellent base of followers. When asked how they knew about that group of followers, they answered that they noticed a set of users that regularly liked and commented on their tweets and videos, even sending direct messages to them.

Coordination among activists usually happens in private WhatsApp or Telegram groups. Their actions involve hashtag campaigns aimed at making a particular hashtag a trending topic to maximize the visibility of a particular issue. Some examples include asking for a politician's resignation, pushing towards a boycott of a company, or demanding attention from the government for a political issue. Some activists have also been directly involved in pushing hashtags/topics related to political parties during political campaigns though not directly tied to these political organizations. Other aspects of activist-activist coordination involve re-posting content of their colleagues and comrades to promote it and help them grow, as well as recording videos together or participating in live streams on each other channels. During our interviews, some activists pointed us to a particular action carried out on Twitch called *raid*, where a user that is live-streaming content would direct their audience into the other's streamers channel, so the first one can make their audience know the second one and help them grow in popularity.

Coordination also happens between the activists and their audience. According to the opinions and experiences of the interviewed individuals, this usually happens in two forms: in a formal and in an informal manner. Informal coordination usually happens when the activist quotes or signals a social cause, like change.org campaigns or hashtag-protests, and their audience gets the signal and supports

the action [Rhe02, Goo14]. Formal coordination happens when the activist has some experience on the platforms and is well aware of their audience. They would identify relevant active figures among their audience and recruit them into private or semi-private groups in applications such as Telegram groups or Discord communities [Cha13]. Once embedded in those groups, the activist will grow their own network and use it to coordinate offline and online actions better. An interviewed activist admitted that those groups often serve as means to coordinate hostile actions against opposite political activists or any political actor in general with the aim of closing social media accounts, flooding opposite hashtags or trending topics with noise, or just trolling their rivals. When asked about those hostile actions, most of the activists, both left-wing and right-wing expressed preoccupation about them, signaled opposite activists as responsible, and stated that they and their friends used to be the targets.

5.5.6 Relationship with opposite activists

The networked spaces of online communication are characterized by the availability of information and its distribution at a very high speed [PD17]. In the same way, these new spaces are wallless; that is, contrary opinions and points of view can collide at any time. While traditional activists were used to cooperating or colliding mostly on the street in public demonstrations or similar events, the Internet allowed new interactions. During the first waves of cyber-activism as defined by [Bou20], activists would interact with each other through comments in blog-posts or at forums. The nature of those platforms would allow a certain degree of moderation, and the interactions would not be so fast and frequent. Opposite activists would usually prefer to write blog posts criticizing their rivals rather than storm their blog posts with aggressive comments. The technical possibilities of blogs and forums would allow content moderation, so the influencer-activists would mostly be stuck to their own communication spaces. After the rise of online social networks such as Facebook, Twitter, or YouTube, those dynamics changed as the new platforms were open by definition, not owned by any particular group of activists, hardened content moderation, and facilitated fast and anonymous content production [Cas12, Cas10].

While it is clear that an activist of a particular ideology will think that her point of view on the issues of interest is the right one, she can relate to and interact with the rest of the activist community in many ways. To better comprehend the interaction between activists of opposing points of view in the context of online social networks, we consider RG1.RQ2.

Among the interviewed activists, the ones who hold more radical points of view or the ones who are most committed to spreading their messages appear to be very worried about the possibility of getting banned from online media, mainly due to organized actions carried out by their rivals in activism. Almost all of them (90%) were particularly worried about being banned from sites such as YouTube and Twitter.

A group of both left-wing and right-wing interviewed activists acknowledged that some of the activists in their close circle set up and maintain social media groups on platforms such as Telegram, where they coordinate what they call “censorship actions” against opposing activists. They are also aware of their counterparts doing the same thing.

Those actions include massive reports of content posted on the aforementioned platforms arguing copyright or hate speech reasons. Other actions include the rapid gathering of supporters in case there is a hot debate on any of the platforms. In that case, activists themselves or their deputies demand the attention of their followers in online closed groups in WhatsApp or Telegram calling them to act in a particular publication.

5.5.7 Relationship with the conventional press

Starting during the first waves of cyber-activism in the 90s, activists have always looked at traditional media with suspicion and distrust, as they perceived them as agents of the *status quo* closely related to the political elites. The early history of cyber activism has been characterized by the tendency to set up alternative communication spaces by using online technology, with Indymedia [Kid03] as a clear example. These new cyber-activists still see traditional mass-medias as to be sold to the system. However, they tend to adopt a different strategy, assuming possible ideological contradictions and prioritizing the maximization of the spread of their message.

When studying the relationship between this new type of activists and conventional media such as the press, radio, or television networks, we appreciate a two-way relationship.

In general, most of the interviewed (76%) were critical of conventional media, some of them (41%), the ones more related to populist points of view, were very critical accusing them to be "sold to the ruling elites" and having "no potential for pushing societal change" as it is usually seen in this kind of activists [FI20]. Others (50%) stated that "conventional mass media will soon be dead" and that "we are the new thing". They prefer to publish their content directly on the network without any filter, so they are able to talk about their topics of interest that are ignored in the mainstream media.

On the other hand, many of the activists (60%) that were interviewed were collaborating or considering collaborating with conventional media in some way such as participating in interviews, television debates, publishing opinion articles, or political analyses. They integrate their collaborations in those means as another communicative element in their communication network, re-using video clips from their interventions on television on X and YouTube, pointing their followers to their articles in online newspapers and suggesting their readers to follow them on their social media.

When asked about their possible desire to pursue jobs as journalists, in general, they do not see a career in the press or television as a viable option (90% asserted that), as they remark that they would lose their ability to perform critical thinking, and their authenticity, and they will probably work under poor economic conditions, so they are better going by themselves.

Their publication style is generally defined by liquidity and eclecticism, as it is common amongst the ones belonging to the millennial generation [DD19b]. They generally do not tie themselves to a particular newspaper or television, conversely, they rapidly switch between different means of convenience to maximize their visibility and to choose the media that better aligns with their political views. They try to diversify their publications as well as their content on social media among different sites whenever possible to maximize their visibility, avoid platform/media site dependency, and protect their online presence as they fear that one platform may fall in disgrace in the future or their profiles may be prone to censorship or any other form of attack.

5.5.8 Relationship with political parties

Political parties, especially the traditional ones, have been seen [PAF98] as corrupt and useless structures ruled by financial powers and not serving the general interest of the people. That point of view was particularly developed and defended by the *Indignados* in Spain, and by other platforms, worldwide, such as the *Occupy movement* [PLCA14, Cas13]. The Spanish influencer-activists that were interviewed in this research do somehow inherit most of those opinions from the previous wave of activists. Though

their way of thinking is similar in relation to the political parties, they do not deny them totally. For a better comprehension of their relationship with political parties, we address RG1.RQ2.

Even though they do not want to get involved in the internal structure of the party, they consider eventual collaborations that may allow them to keep their ideological independence while trying to incorporate key issues of their activism into the programs of those parties. New political organizations characterized by more populist rhetoric, started after the 15M protest, usually trying to incorporate them into their political communication strategies.

While it is clear that some of the activists that have been interviewed are considering, if not actively seeking, professional opportunities related to their political activities – such as jobs as political analysts in think tanks, campaigners, or lobbyists – they show a clear preference for their own ideological autonomy. They state that if they had to join a political group as a grassroots militant, their discourse would then be controlled and censored by the political organization, thus losing originality and therefore losing traction with their audiences. Most of them prefer the position of the pseudo-independent pundit or even an election candidate that has a clear political position but remains out of the control of the political organization.

5.6 Discussion

As we have presented in this work, the influencer-activist, while it is to some degree similar to the more traditional and well-documented figure of the cyber-activist [MA03], as we discovered from the qualitative interviews done, it has their own set of dynamics and discourses tied to the possibilities of online social networks and the new forms of political participation, in agreement with recent studies [PGC20]. We also detected some general patterns in their behavior such as a strong ideological independence, platform diversification, and inter-related, less-structured relationships with other activists, with a will to professionalize their activism to some degree. But some of the activists may present slightly different visions as well as their own features, thus deeper studies focused on particular kinds of influencer-activists such as left-wing or far-right, for example, are necessary. The variety of tools, techniques, and strategies related to this new kind of activist also requires specific attention, as each of the platforms presents its own dynamics, in particular, different censorship rules, publication timings, and interactions with other users, and is used in a particular way.

More specifically, and as summarized in the methodology section, by answering the research questions considered in this research, we are able to profile and characterize in a particular way this new type of figure. By definition, these influencer-activists are focused on politics (typically engaged at an early age by friends and relatives, see RG1.RQ1) and very active in online social networks, being more proficient than the average user (evidenced by their knowledge and active participation in more than one social network and community, see RG1.RQ2-3). Their motivations are both intrinsic and extrinsic, and related to this dual persona they create and strive to balance in their everyday: creative and influential in political aspects, but being careful to stay popular and do not lose professional opportunities.

Moreover, while the general observation was that their behavior could be summarized as posting regularly but in a particular time window and about a specific content, this should be validated and assessed in the future to better understand these dynamics too, for example, capture which strategies work best, perhaps depending on the target audience or the social network it is being applied to.

Finally, we observed that their interaction with different actors is well-established and carefully planned. First, depending on how radical they are, they may organize against opposite activists what they call censorship actions, especially when their counterparts do the same; this, however, is done carefully to avoid being banned from platforms. Second, while they are critical of conventional press and media, they may use it to their advantage if they believe this could improve their message exposure. Third, their relation with political parties tends to lean towards collaboration, but avoids getting involved in the actual party to keep their voice and autonomy.

With this study, we aim to understand a new figure that arose in the last years in politics and social media. However, in this process, we may expose their inner behavior with enough detail (and by providing explicit technologies and tools) that would make spreading more influencer-activists a trivial task. This may not be harmful by itself unless some actors decide to move this figure into extreme polarities of the political spectrum, as investigated in [GR14] also at a regional level, or in [NO22] in the context of understanding shared misinformation. Another consequence of this study could be the discovery of novel platforms or tools that some actors may not be already using, which could be exploited to gain more followers or increase their influence. Nonetheless, this may easily be already happening, and our study evidences behaviors and procedures that could be automatically detected by neutral institutions. This could even be performed by their own platforms, allowing for tracking (and, subsequently, warning, banning, or censoring) those accounts susceptible to platform misuse.

5.7 Conclusions

The traditional cyber-activist, focused on actions such as online narrative spread on forums, hashtag protests, and actions on social media, leaks, and offline events [MA03, Joy10] has evolved facilitating the rise of a new kind of activists that builds on top of it, combining some of its features with the typical features of the social media influencer figure. This kind of activist is defined by a mastery of social media communication tools and strategies, a strong development of their online personal brand, a well-structured and intertwined presence in a variety of platforms (each one with its own functionality), and a curated scheme of revenue extraction to eventually be able to self sustain their own activity.

This new emergent figure, which coexists with other forms of activism both off and online, prefers not to be directly influenced or associated with conventional political organizations or the press, though they may use them in an occasional and eclectic way to maximize the impact of their messages, extract revenues, and, especially, to push their narrative into the public opinion. The influencer-activists aim at embedding themselves at the center of a networked structure on social media, to influence a large base of followers, and to maximize their possibilities of spreading a message. They still consider offline actions such as protests and the like, though they focus their efforts online and think mid to long-term about their causes, as they generally aim for a change in the conscience of society about the issues they worry about, which will eventually lead to political changes. As this new figure is still in the process of emergence, future work may involve the refinement of the presented definition, deeper analyses of the particular relationships these activists hold with different political organizations and mainstream media, the set of specific tools they incorporate in their social media ecosystems, and their inter-connections along with their specific communication techniques and political actions. Other activist figures may emerge as the interaction of activists, and new technologies widen the possibilities for political participation, they should be documented, understood, and linked to existing kinds of activism.

Chapter 6

Understanding the Dynamics of Online Political Ecosystems

6.1 Introduction

As pointed out in the introduction of this thesis, over recent years, social media platforms have evolved into the contemporary "agora", serving as pivotal hubs for discourse and dissemination of information [MBG+13]. Many of the populace now rely on these platforms for insights into current events, reflecting a transformative shift from traditional media outlets. This evolution underscores online social networks' profound influence and reach in shaping public opinion and informing the masses.

The rise of online social networks has deeply influenced this metamorphosis of political discourse over recent decades. Historically, political conversations were predominantly shaped by mass media outlets and would resonate in homes, cafés, and workplaces. However, in today's digital age, these networks serve a multifaceted role [SLR10]. They act as conduits for news consumption and provide platforms for deliberation on pressing political issues. Furthermore, they have become pivotal channels for spreading and absorbing propaganda. Particularly noteworthy is these platforms' immediacy, enabling real-time discussions and interactions on contemporary matters [SLR10]. This last feature is of particular significance. The capability for real-time self-communication and response enabled by online social networks has revolutionized the communication methods of major political players. Specifically, parties and their candidates have adopted new strategies in the context of electoral processes, further underscoring the transformative impact of these platforms on political discourse and interactions [GSRC+16].

Building on the earlier point, political campaigns, whether at the national or local level, have undeniably evolved to harness the power of social media [SBLS20, VHS13, Ver15]. In contemporary times, it is a given that political parties across major global democracies are leveraging these platforms extensively. They are not merely platforms for sharing information; they have become instrumental in conducting complex political marketing and disseminating propaganda. These strategies are meticulously crafted, with organizations creating and positioning narratives tailored to appeal to and persuade potential voters. This further emphasizes the intertwined relationship between modern politics and online social networks.

X (formerly Twitter) has established itself as the primary platform on which these campaigns are conducted [VHS13, Ver15, Bou20, AS22, Jun16]. As a result of this expansion of political marketing campaigns to online social networks, primarily due to the possibility of occasional real-time anonymous

conversation, new behaviors have emerged: from the proliferation of hate speech and the dissemination of extremist ideas, to the spread of false information and the organization of online harassment campaigns [SNB+21, Jun16, KvS21]. Owing to this, analyzing the information dynamics on these social networks becomes especially relevant. Understanding the types of actors involved in the political conversation and their interrelationship framework is essential to designing strategies to minimize the negative aspects of online political discourse and promote a healthy and constructive online debate.

Following the trajectory of the influence of online social networks on political discourse, this research inquires deeply into the political conversations on platform X during Spain's general electoral processes of 2011, 2016, and April 2019. Central to our investigation is identifying principal user categories actively participating in the discourse. We aim not only to elucidate their defining characteristics, but also to unravel the intricacies of their interrelationships.

The analysis in this chapter of the thesis seeks to provide a comprehensive understanding of the dynamics and nuances that shape the political landscape in the digital age. Though other analyses have explored the role of political parties and political marketing in the context of electoral campaigns in Online Social Networks [SBLS20, VHS13, Ver15, A+13, Jun16, ICMFS12], our analysis is the most recent and complete one to the best of our knowledge. It is also the only one focusing specifically on election campaigns, and the only one that aims to generate a rigorous characterization of the users involved in the public debate around political campaigns. Our analysis is also distinctive from the perspective that it evaluates the role of users other than the political parties or particular politicians, aiming to identify and study the users involved in the conversation in an agnostic way (independent of the topics of a particular conversation), as unsupervised learning is used to identify potential user categories.

Research questions

Our aims in this chapter can be summarized in the following research questions. These specific research questions all belong to the first research goal and will be addressed from a different viewpoint than in the previous chapter:

- **RG1:** Develop a valid definition of a political information ecosystem for this study, involving all of its main elements and thus allowing us to study their information dynamics.
  - **RG1.RQ1:** Which are the main information disseminators in the political context on microblogging social networks and how they develop their activity?
  - **RG1.RQ2:** What motivates political active users to perform online activism and promote narratives?
  - **RG1.RQ3:** How this main information disseminators interact between each other during political processes?

6.2 Background

6.2.1 The role of political organizations in OSNs

Before the ascent of Online Social Networks (OSNs), political organizations played a central, traditional role in shaping and directing political discourse. These entities functioned as the primary gatekeep-

ers of political information, leveraging mass media outlets such as newspapers, television, and radio to disseminate their messages and influence public opinion [Bou20, AS22]. Their power was derived from their capacity to create narratives and their ability to decide which issues would be prioritized in the public domain. Face-to-face interactions, rallies, and community engagements were pivotal tools for garnering support and mobilizing the masses. The reach and influence of these organizations were largely determined by their access to media resources and their grassroots networks. In essence, before the proliferation of OSNs, political organizations operated within a more controlled and centralized communication ecosystem, where information flowed in a predominantly top-down manner [Bou20]

With the dawn of the digital era and the ubiquity of OSNs, political organizations have found themselves amid a transformative communication landscape. This shift needed a profound reevaluation and recalibration of their strategies and communication methods. Firstly, the decentralized nature of OSNs empowered individuals to become not just consumers, but also producers of content. Having recognized this, political entities prioritized direct engagements with their constituents, fostering two-way dialogues rather than the traditional top-down communication. Platforms like Twitter, Facebook, and Instagram became essential tools for politicians to share real-time updates, gauge public sentiment, and rapidly respond to emerging issues [Jun16]. Additionally, the immediacy and virality potential of OSNs meant that political narratives could be disseminated more quickly and broadly than ever before. Harnessing this, organizations started to employ targeted advertising, algorithmic content promotion, and influencer partnerships to amplify their messages and reach specific demographics more effectively.

Moreover, the digital era brought with it the challenge of information overload. To stand out, political entities began crafting more personalized, relatable, and visually appealing content. Multimedia elements, such as videos, infographics, and interactive polls, became their digital communication arsenal staples. However, the adaptation had its challenges. The open nature of OSNs also meant increased scrutiny, rapid dissemination of misinformation, and the rise of echo chambers [CLH+21]. Many political organizations invested in digital literacy campaigns, fact-checking initiatives, and community moderation efforts to navigate this.

Regarding the engagement dynamics between organizations and the public, the digital landscape of online social networks (OSNs) has brought forth a plethora of engagement avenues for political organizations [Bou20]. A predominant method is targeted ads, which enable these entities to pinpoint specific demographics based on interests, location, and online behavior. This precision allows for a more effective allocation of resources and ensures that messages reach the most receptive audiences. Moreover, interactive campaigns have become increasingly popular. By utilizing polls, quizzes, and live Q&A sessions, political organizations can foster active participation and a sense of inclusion among users. Furthermore, collaborations with influencers or notable online figures have proven invaluable. With their substantial followings and credibility, these individuals can amplify political messages, making them resonate more deeply with the digital populace [Bou20].

At the same time, OSNs have revolutionized how politicians and parties interact with the masses. The era when communications were necessitated to pass through conventional media intermediaries has become obsolete. Platforms like X and Facebook have democratized communication, enabling politicians to convey their viewpoints, policies, and reactions without intermediaries [GAS+15]. This direct line of communication fosters a sense of authenticity and transparency, allowing constituents to feel more connected and engaged with their representatives. It also means that politicians can address emerging

issues in real time, ensuring their narratives remain relevant and timely.

One of the most transformative features of OSNs is the immediacy of feedback they offer. Every post, tweet, or update can be instantly met with reactions, comments, and shares. This real-time feedback mechanism is invaluable for politicians. It provides a pulse on public sentiment, allowing them to gauge the effectiveness of their messages, discern prevailing concerns, and identify areas of contention. Such insights empower them to adjust their strategies on the fly, fine-tune their communication, and address issues proactively. In essence, OSNs have created a dynamic feedback loop where politicians and their constituents are in a continuous dialogue, each influencing and being influenced by the other [Jun16, KvS21].

Building on the engagement dynamics and direct communication offered by OSNs, political entities have harnessed these platforms' potential to sculpt public discourse decisively. With the ability to set agendas, they strategically introduce and emphasize topics, steering public attention towards the most pertinent issues. Regular posting, sharing, and engagement allow them to entrench specific narratives, crafting a predominant frame of reference for their audience. Moreover, the agility of OSNs equips them with rapid-response capabilities. When confronted with opposing narratives, these entities can swiftly counter with alternative viewpoints, fact-checks, or even introduce counter-narratives, ensuring their perspective remains dominant.

Simultaneously, the expansive and interconnected nature of OSNs has amplified the reach of political messages. The digital sphere's vastness ensures that a compelling tweet, a thought-provoking video, or a well-designed infographic has the potential to cross geographical, cultural, and linguistic barriers in mere hours. The design ethos of these platforms, emphasizing sharing and interaction, further bolsters the potential for content virality [Jun16, KvS21]. When a message goes viral, its impact is magnified, presenting political entities with opportunities to significantly sway public opinion, rally support, or influence electoral outcomes. This vast reach and potential for virality have redefined political communication in the digital age, ensuring it is not only immediate but also has a far-reaching and powerful resonance.

6.2.2 Conversation dynamics during electoral processes

In today's digital age, a significant portion of political campaigns unfold within the virtual corridors of online social networks, particularly on platforms such as X. These platforms have transformed traditional campaign strategies and introduced novel phenomena intrinsic to the online realm [KvS21]. Instances of viral content, hashtag-driven protests, and real-time debates are becoming increasingly commonplace, directly influencing the trajectory and outcomes of electoral campaigns [KvS21, CRB19]. Such phenomena suggest a complex web of interactions and dynamics within these online spaces. Distinct user categories with unique characteristics and roles emerge and contribute to this complex digital compound.

Understanding these conversation dynamics during electoral processes is paramount. It transcends mere academic interest, holding far-reaching practical implications. A nuanced comprehension of these dynamics can pave the way for more effective political marketing strategies, provide tools to combat the spread of disinformation, and offer insights to counteract extremist or hateful narratives [TJLL18, Sun17, CRF+11, CLH+21, GK20]. Furthermore, it can guide the design of more informed and user-centric online platforms.

6.3 Specific Methodology

To address the research questions presented before, we adopted a multifaceted methodology. At the outset, we gathered extensive information about users, that is their publications and basic profile info, on platform X who engaged in discussions pertinent to the various electoral processes. To distill insights from this data, we employed a combination of unsupervised learning techniques, specifically those related to clustering. Additionally, network analysis techniques were applied to decipher the underlying structures and connections within the discourse. Linguistic analysis methods further enriched our understanding of the content and rhetoric used. To provide a more holistic perspective, we complemented these quantitative techniques with qualitative analysis, ensuring a thorough and nuanced exploration of the topic.

6.3.1 Data set

For this research, we used two primary datasets, as defined in the methodology section of the thesis (Chapter 4), for tables "General Processes" (Table 4.2) and "Local Processes" (Table 4.3), particularly focusing on general elections.

In the context of this research, we analyzed the electoral processes of the Spanish general elections of 2011, 2016, and April 2019, as they are sufficiently spaced out in time to allow us to identify consistent patterns in user behavior and the overall dynamics of the conversation.

While naturally, all electoral processes were taken into account, special attention was given to the information from April 2019, as being more recent, it has more significant potential to elucidate the phenomenon.

6.3.2 Clustering

Once the datasets were generated, our initial goal was to identify the different types of users present within them, allowing us to study their behavior in the context of political discourse subsequently. In order to achieve this, we relied on quantitative and qualitative analyses. From a quantitative perspective, we utilized the unsupervised learning clustering algorithm K-Means [ASI20] to generate these categories. More precisely, the following process was carried out:

1. **Exploratory Data Analysis (EDA)**: Initially, an exploratory data analysis was conducted to gain a deeper understanding of the characteristics of the datasets that could be utilized for the automatic classification of users. Based on this analysis, it was determined that there was a need to employ the aforementioned political datasets with the original features to compute new, more complex and complete features, that could provide useful information for effectively achieving an automatic classification using *clustering* algorithms. Specifically, the following features were computed:

    - **Average daily tweets and retweets**: Considering a time frame (e.g., from 2019-04-05 to 2019-05-05 in the case of the Spanish general elections of April 2019), the number of tweets per day associated with each user was calculated.
    - **Total number of tweets by day of the week**: By considering each day of the week (Monday, Tuesday, Wednesday, etc.), the total number of tweets each user published on each day was calculated, aiming to detect possible patterns in user posts.

- **Total number of tweets by time intervals**: Similarly, four different time intervals were considered (morning, afternoon, evening, and night) and the total number of tweets for each interval was calculated. It is worth noting the hour ranges used to define each of these intervals:
  - **Morning**: from 06:00 to 12:00.
  - **Afternoon**: from 12:00 to 19:00.
  - **Evening**: from 19:00 to 24:00.
  - **Night**: from 24:00 to 6:00.

2. **Percentage of tweets that are retweets, replies, or quotes**: That is, the percentage of posts, within the total tweets of each user, that are retweets, replies, or quotes (three distinct features, measuring each of these aspects respectively).

3. **Percentage of self-retweets**: This feature models the percentage, within the total retweets of each user, that are retweets of a tweet from the same user. In other words, how many retweets are self-retweets (of their own tweets).

4. **Average hashtags and URLs per post**: Specifically, the average number of hashtags each user uses in their posts, and the average number of URLs each user uses in their posts.

5. **Average positive and negative words per post**: That is, the average number of words with positive sentiment and words with negative sentiment contained in each user's posts. The *LIWC 2007* framework [PCI+07] was employed for this, which defines certain categories. This framework associates each word with one or more categories, and it relies on word counts for each category. Thus, the word counts for the "EmoNeg" and "EmoPos" categories allowed for the calculation of the number of negative and positive words, on average, contained in the tweets of each user.

6. **Average times between tweets and retweets**: These two features were also calculated, essentially referencing the "cadence" of each user, i.e., the time they take to post a tweet (first feature) or a retweet (second feature).

7. **Total number of different conversations, tweets, retweets, different users retweeted, quotes, different quoted users, replies, mentions, different mentioned users, used hashtags, different used hashtags, URLs, different used URLs**: Similarly, these other features were calculated to allow for user activity comparison (in magnitude, instead of percentage or average).

8. **Total number of different categories and subcategories (domains) published**: As these two features allow us to differentiate users who only discuss certain topics from those who discuss all conversation topics.

9. **Average success of posts**: The average number of retweets, replies, and likes that each user's posts received was calculated to differentiate *influencers*, i.e., users whose posts go viral.

10. **Profile features**: Moreover, inherent user profile features were added, such as whether or not they are a bot, profile age, number of followers, followees, as well as the follower/followee ratio, which could distinguish famous users.

11. **Average posts for each conversation**: A feature was also calculated to determine the number of posts for each conversation, distinguishing users who "engage" with others on the network from those who do not.

12. **Speech characteristics**: Using the previously mentioned *LIWC 2007* framework, some of its categories were included as features of the user *dataset* on which *clustering* techniques were applied to achieve automatic user classification. Some of the utilized features refer to the use of third-person references, the use of words related to topics like money, work, family, religion, etc.

The computation of the aforementioned features allowed for the creation of a *user dataset* (one per electoral process studied) on which to apply *clustering*. Specifically, the applied process was:

1. **Data Cleaning**: For the creation of the user dataset, users with null or negligible activity during the analyzed time period were not considered. Hence, only users with more than 100 tweets (including retweets, replies, quotes, and original posts) during the analyzed period (31 days) were taken into account. Thus, from the total amount of users in Table 4.2 for a specific electoral process, a small amount was subtracted to avoid considering inactive user in the analysis of the political conversation in Twitter.

2. **Normalization**: Various normalization techniques were used to compare results obtained with different normalizations.

3. **Dimensionality Reduction**: *Principal Component Analysis (PCA)* was applied to reduce the dataset dimensionality to attempt to enhance the performance of the clustering algorithm to be applied. Additionally, using *PCA* was quite suitable for obtaining a visualization of users after applying clustering (as two components can provide a two-dimensional representation that is easily interpretable). An example of a clustering visualization obtained for the *MinMax scaler* technique is shown in Figure 6.1 (similar visualizations were generated for *RobustScaler*, *StandardScaler*, and *MaxAbsScaler*, although the first one did not yield good visualization results, while the last two did not have any significant difference compared to the *MinMaxScaler* shown in the figure). It is worth noting that the use of *PCA* with two components had a relatively low percentage of explained variability, so while the defined categories are perfectly coherent, the presented visualization, though aiding in understanding the phenomenon, might not be entirely accurate.

4. **Clustering**: Finally, various clustering algorithms were used to compare results and understand which achieve the best possible automatic classification. Specifically, algorithms like *Hierarchical and Agglomerative Clustering (HAC)* and *K-Means* were tested. Ultimately, K-Means provided better results, so it was the chosen algorithm for automatic classification. It is worth highlighting that to try and achieve the best possible results, a multi-phase clustering strategy was adopted (specifically, three phases), where in each phase some users were labeled (classified) and others left unlabeled, to maximize the accuracy of such classification based on domain knowledge. Finally, after the three clustering phases, some manual adjustments were made based on a qualitative analysis of the content of the identified categories. This was done to address errors from the automatic classification, resulting in a classification of X users that allowed the analysis whose conclusions are presented in this work.

Thus, five main consistent categories were identified, those were spectators, political parties, politicians, medias, and activists, all detailed in the following sections.

6.3.3 Network analysis

Regarding the categories identified through the conducted clustering, network analysis techniques have been applied to study the interactions between these categories. Precisely, users have been modeled in a network (graph[1]), where each node represents a user and each weighted and directional link represents the action of a user interacting, in a specific manner (retweets, cites, quotes and mentions) with another user, thus effectively capturing the relationships between the users in the best possible way, considering the X's (Twitter) provided environment for user interaction. Thus, we generate a link between A and B with weight N if A has retweeted content N times from B. In our analytical process, particularly when interpreting the generated graphs, we employed the Louvain method [New06]. This approach was specifically chosen for its proficiency in automatically identifying distinct user communities. By leveraging the Louvain method, we aimed to uncover the inherent structures and groupings within the discourse, further enriching our understanding of the dynamics among participants on platform X. By modeling user interactions within a network and applying the community detection method, we have achieved a deeper understanding of how the different user categories group together in political discourse.

6.3.4 Linguistic analysis

Similarly, the discourse of users identified in the various categories has also been analyzed for comparative study. For this purpose, the LIWC software was used [PCI+07], aiming to associate the words in user posts to different categories: by sentiment type, by general theme, by discourse style, whether formal or informal, by the use of verb tenses, or by references to the first, second, or third person.

Similarly, the information provided by the X Academic API service has enabled the identification of the main topics and sub-topics of conversation associated with the downloaded posts. This has been used to assess both the variability and focus on conversation topics by users identified in the various categories.

6.3.5 Qualitative analysis

To round off our research, we incorporated qualitative techniques into our analysis. A pivotal element of this approach was the manual inspection of user activity, complemented by a close examination of specific posts. This hands-on scrutiny was instrumental in delving into the complex information dynamics that encapsulated the electoral processes under study. A particular emphasis was placed on tracing the roots of narratives and charting the information flows—beginning from central content-creating accounts, transitioning through amplifiers, and finally reaching content-consuming ones. While quantitative methods offer invaluable insights, they sometimes grapple with capturing the nuances of abstract relationships, such as the genesis and spread of narratives. This gap underscored the indispensability of our qualitative approach, where manual content inspection illuminated these more elusive dynamics, complementing the insights obtained from the quantitative analysis to capture the full picture of how information flows through OSNs during a Spanish electoral process.

[1] In fact, a multi-directed graph, where different edges are defined for different user-interaction types.

6.4 Results

6.4.1 User sampling per category

In order to enhance the confidence in the clustering conducted, we preset here an analysis on a small subset of users from each category. By comparing values of distinct attributes, our aim is to understand the differences between user categories in online political discourse. While this is carried out for the general electoral process of April 2019, a similar methodology could be applied to the electoral processes of 2016 and 2011 and, in essence, to any other electoral event. Specifically, it is proposed to utilize the following attributes:

- **ScreenName**: This represents the user account name on Twitter, also referred to as the *nickname* or *username*. It uniquely identifies the account under consideration in the comparison.
- **Tweets**: The total count of tweets over the 31 days studied surrounding the electoral process.
- **Followers**: The number of followers associated with the specific account.
- **Distinct URLs**: The total number of unique URLs shared within the posts.
- **Total hashtags**: Cumulative count of hashtags used in the posts, whether they are unique or not.
- **ATT**: This attribute is named after its function – measuring the average time between tweets, abbreviated as ATT (i.e., *Average Time per Tweet*). This average time between any two consecutive tweets from the user is measured in minutes.

6.4.2 User categorization and roles

From the analysis carried out through clustering, we can identify 5 clear categories (explained in the following sections), with the category associated with political parties being the most clearly defined as it can be seen in Figure 6.1, especially for major national parties. To explore into this difference, the weight of the principal components was studied to give them meaning in the domain of OSNs. On the one hand, PC1 (X-Axis) had a high positive weight for features related to user activity, such as the total tweets published across an entire electoral process. In contrast, it had a negative weight for specific LIWC categories, including, for instance, references to supposed third parties. On the other hand, PC2 (Y-Axis) had a high positive weight for other features related to the activity, such as the number of daily tweets or the time-frequency of a user's posts.

Thus, the interpretation of the two-dimensional clustering visualization derived that political parties are characterized by having high activity and more neutral language, positioning them with positive values on the X axis and close to the neutral value on the Y axis, whereas other groups such as spectators, media, politicians, and activists tend to constantly make references to third parties and political subjects, placing them on the left side of the X axis. Activists stand out by having the most extreme values on the Y axis. All this initial information was verified through subsequent analyses described in this thesis.

Table 6.1: Sample of users by category found in the clustering (April 2019).

ScreenName	Tweets	Followers	Distinct URLs	Total hashtags	ATT
Cluster 0: Political Parties					
ppopular	5,051	848,277	2,941	2,823	33.06
PSOE	7,250	850,219	4,888	10,223	25.26
CiudadanosCs	8,382	510,397	5,012	5,593	21.41
vox_es	4,772	497,960	1,914	1,620	33.82
PODEMOS	4,889	1,537,029	2,489	2,132	32.25
IzquierdaUnida	5,251	595,798	2,937	4,150	30.13
Cluster 1: Medias					
el_pais	32,127	8,717,976	23,438	4,310	6.43
la_cerca	16,714	18,096	17,365	18,315	13.14
elmundoes	20,665	4,461,530	18,420	3,669	10.65
larazon_es	23,748	585,258	41,918	19,746	10.22
diariovasco	20,881	223,614	20,834	2,415	10.27
diariosevilla	15,409	223,614	16,155	8,006	14.46
Cluster 2: Politician					
sanchez_castejon	964	1,718,675	836	1,525	182.84
albert_rivera	837	1,204,523	616	494	222.29
pablocasado_	1,493	551,479	996	659	115.83
InesArrimadas	926	687,760	493	206	222.53
cayetanaAT	765	494,177	403	95	205.83
agarzon	1,645	1,172,243	443	334	104.16
Cluster 3: Activists					
algonfdez	45,061	6,371	17,404	10,716	4.83
BarreirAurora	48,432	3,699	10,287	7,891	4.99
ManuelCostoya1	51,573	2,096	13,352	13,588	5.37
IgnotoLugar	59,057	1,525	9,327	10,802	3.44
DoroteaE58	35,642	1,044	4,982	6,283	5.64
esteren3onidos	25,795	1,543	5,302	4,827	7.66
Cluster 4: Spectators					
Miriamferrerdu	951	258	111	548	314.02
VivancosMariano	1,029	1,013	369	389	154.18
maholboy	1,180	701	192	342	201.41
Palmaeixample	490	180	109	590	179.16
GonzatheBoss	1,165	1,162	310	125	158.12
Aliciasotovi	571	23	51	99	161.52

Political organizations The collective of political organizations consists of the official accounts of political parties within the party system, which correspond to approximately 23.8% of the total studied users for the case of the general electoral process of April 2019 (similar scores are obtained for the rest of electoral processes).

The activity level of political organizations is high, exceeding that of individual politicians' accounts. They primarily base their activity on retweets and original tweets. These are generally neither quotes nor replies, indicating that their activity mainly revolves around generating original content and introducing it into the network. Similarly, as detected, political parties also self-amplify their narratives. Almost

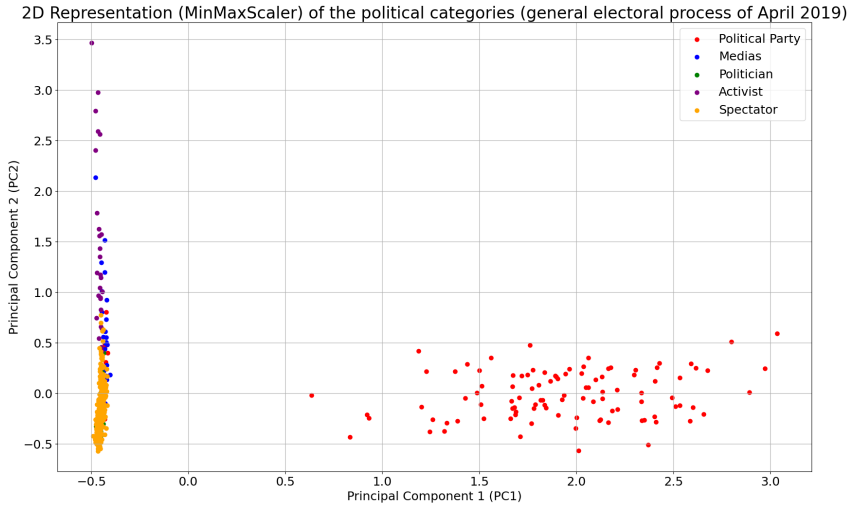


Figure 6.1: Two-dimensional representation of users that participated in the political conversation during the Spanish general electoral process of April 2019.

all of their retweets are dedicated to amplifying their content (category-wise). They complement this activity with numerous mentions, hashtags, and links to websites. They share more than 20 website links per day (most of which are different), doubling the number during the week of the electoral event. During this time, the creation of narratives, supported by media news[2] becomes essential to the parties' objectives. Observing their behavior when using replies, they are primarily used to generate threads as a "storytelling" strategy. In this case, a party will reply to its tweet, creating a content thread to narrate a story, partially due to the length limitation of the tweets.

Regarding the type of links shared, they are predominantly from media outlets, as it is generally the case in the other categories. Our analysis identified more than 20 media outlets consistently shared by different political parties throughout the entire electoral process. While there is a certain variety of media outlets within the studied space (the Spanish political system), we have found that each political party prefers a set of 3 to 5 media outlets deemed ideologically favorable.

Regarding their activity during the electoral processes, it closely mirrors the development of the electoral campaign. It escalates from the pre-campaign phase through the campaign, remains silent on the reflection day, peaks on the election day, and decreases over the subsequent week. Their activity is typically concentrated during working hours on weekdays.

Concerning their speech, we found a remarkable trend by which political parties widely use first-person pronouns in both singular (referring to their candidate) and plural forms (including the reader, i.e., the spectator, the end-user of the created narratives). Additionally, they tend to reference unknown third parties using words such as "They/Them". We also found that the main topics treated by parties tend to change significantly from one electoral process to another, reflecting the society's interests evolution over time, although specific topics such as "economy/money" have been relevant across all the studied electoral processes. Similarly, the verb tense usage, i.e., past, present, and future forms, greatly vary from one process to another, as their usage heavily depends on the narratives, which naturally change over time. It is worth noting that Section 6.4.3 contains figures representing analyses supporting these conclusions.

[2] As they post news whose URLs are used, afterward, by political parties to reinforce their narratives.

Politicians The accounts identified as politicians correspond to the personal accounts of prominent political leaders, who play a role in the electoral process as they seek to be elected and support their party. They correspond to approximately 5.7% of the users that participate in the political conversation on X in the case of the general electoral process of April 2019 (similar results are obtained for the other processes). Their activity patterns are similar to those of political parties. They typically act in great ideological harmony, in a quasi-symbiotic relationship, amplifying the parties' content and having their content amplified by them. These category interrelations are further described in Section 6.4.4.

Regarding their overall activity pattern, it becomes much more propagandist/amplifying during the week of the electoral event, as their number of retweets to already existing content, i.e., content amplification, is significantly greater during this week than during the rest of the electoral process. During that period, they shift from introducing narratives and amplifying them relatively similarly to focusing on content amplification because they take a more propagandist role as the election day approaches.

On the other hand, in terms of how they conduct their activity, they primarily rely on the use of a large number of mentions, along with the sharing of some links and a significant number of hashtags. Both aspects indicate a high degree of coordination when launching communicative campaigns. Their activity is concentrated during working hours on weekdays, but there are peaks on weekends, depending on various political events, including election day.

Regarding the media outlets used, politicians can employ a wide variety of different media, mainly adapting to their local reality. However, some media are more recurrent than others. These media outlets are especially politicized, as we have seen in the case of political parties, and are favorable to the individual's political stance.

On the other hand, their speech is characterized by several aspects. First, they show feelings of discrepancy and exclusion (of opposing opinions) in their posts. Additionally, their speech also represents the "Us versus them" problem, as they tend to use the "I/We" and the "They/Them". This "us versus them" rhetoric is a very well-known factor that appears in the political narrative [Rig97, Mas18] and is closely related to the increase of political polarization in the network [DD19a], as it puts the post reader (probably a spectator in our case) in the situation of deciding which group he or she wants to belong to, understood as opposing groups due to the high usage of exclusion words in their posts (approximately one every four posts contains at least one exclusion-related word) [Mas18]. In terms of topics, politicians enhance the social aspect of the narratives flowing through the network, as the "Social" LIWC category was present, on average, in every single one of their tweets (approximately, an average politician used 1.4 "social" related words per post), thus transforming the narratives created by political parties (as described in Section 6.4.5) to give them a deeper social component. Section 6.4.3 contains figures representing analyses supporting these conclusions.

Medias The media outlets correspond to the accounts of the prominent digital newspapers identified through the carried out clustering. Their primary role within the electoral process revolves around introducing current event information into the network, and they correspond to 5.17% of the total amount of users studied in April 2019 (similar percentages were obtained in the rest of the electoral processes).

Media outlets maintain a very consistent activity throughout the entire electoral process. They are not affected by the peaks present in politicians or parties and have fairly standardized and periodic activity, which increases as the midpoint of the workweek approaches and decreases after that, with minimal activity during weekends, except for election day, where their activity peaks. This day sees the highest

number of tweets, primarily original tweets (mainly news about election day and the near future of society). Naturally, just like political parties, media outlets post mainly during weekday work hours.

Their activity is primarily based on original tweets (not retweets). Within these original tweets, over 90 percent (a figure consistent over time) are neither quotes nor replies, meaning they are new tweets (possibly the news they publish to gain visibility, commenting on a narrative introduced in the network). As expected, their activity is marked by the substantial sharing of links on the network, where the vast majority (over 90 percent) are unique. They also extensively use mentions and hashtags to integrate their news into existing narratives or campaigns on the network. Thus, it is evident that media outlets are neither responders nor amplifiers. Instead, they introduce information to the network, typically with links to their published news, which are subsequently used by political parties to support their narratives.

Similarly, media outlets employ "storytelling" strategies, such as creating "threads" on the network through chained responses, much like political parties. We observe a widespread use of this technique, which has increased over the years. Regarding threads generated by the media, they use them to cover a current event or phenomenon through multiple news items or articles.

Concerning media outlets' speech analysis, they are primarily focused on the "time" category of LIWC, as their main goal is to cover breaking news and events related to the electoral process. They tend to use the "I" (which is the figure representing the media outlet covering news), but also the "You", which is used to make the spectator take part in the covered news or event, potentially enhancing their news' engagement among the readers. Moreover, their posts tend to suggest feelings of discrepancy and negative emotion somewhat (3 out of 10 posts of an average media outlet contain at least one discrepancy and/or negative emotion word). These insights are supported by the analyses described in Section 6.4.3.

Activists Activists are an especially relevant category identified through the clustering performed in this study. This category comprises individual users who are particularly motivated and emotionally tied to a specific "political color", corresponding to approximately 4.98% of the total user population studied for April 2019 (similar to the rest of electoral processes). These users are not political candidates and are especially active in the conversation. Their main objective, as we have observed, is to amplify the narrative of political parties, engage in individual and coordinated propaganda, and generally convey the narrative originated in political organizations to a broad user base through particular dimensions and in a more impassioned tone.

They exhibit increasing activity from the pre-campaign period to the actual campaign, with peaks in activity (mainly retweets) occurring days before the election day. Additionally, there is a significant drop in their activity during the "day of reflection", also known as election silence (which is logical since those they amplify hardly comment at all), and they maintain a similar activity level to the pre-campaign during the post-campaign period.

In general terms, within the conversation, their activity is amplifying/propagandist. About 90% of their tweets (a consistent figure across all the processes studied) are retweets, which shows that their activity is based on amplifying the content of other categories. Thus, activists are the users with the most activity, and this activity is entirely related to narrative amplification in nature. However, even though their role is purely propagandist, we recognize they have many original tweets, which they possibly use to discuss the narratives they are trying to amplify or comment on news from media outlets. Due to the nature of their activity, they use numerous mentions, hashtags (with the majority being repeated, showing their interest in disseminating information in organized campaigns) and links to media news stories.

Similarly to politicians, activists use a variety of media sources, although they focus on massively sharing news from a limited number of media outlets.

Regarding the type of language used, activists display slightly higher levels of hostility and language that incorporates some aspects of informality. This is due to their role in translating more general official narratives to more localized aspects in a combative manner. Similarly, insights are supposed by the analyses described in Section 6.4.3.

Spectators Finally, the spectators constitute the last relevant category identified through clustering. This category comprises many users who mainly consume content, retweeting, and commenting on publications primarily generated by parties/politicians and amplified by activists. More specifically, they correspond to the 60.33% of the total users analyzed for April 2019 (similar percentages were obtained for the other electoral processes), thus being the largest group by far.

Spectators show a similar evolution to the rest in terms of activity patterns but in a more moderate way. During the pre-campaign, their activity is minimal and grows as the campaign progresses. Their activity peaked during the campaign's last week (mainly through retweets and original tweets). It decreases on the day of reflection, rises again on election day, and returns to levels similar to the pre-campaign as the post-campaign passes.

Spectators only comment on what they see on the network, so their activity is based on original tweets expressing their opinions and retweets (amplification) of other publications they agree with. However, while spectators quantitatively contribute little to the political conversation on the network, they are the target of political parties and the flow of information (narratives) in the network conversation. In any case, they are characterized by generally lower activity levels. Their activity is largely based on mentions and hashtags (with more than half being repeated, so spectators tend to publish several posts on the same topic) and news articles. Overall, they have less activity, but using these elements shows their commentator nature (they mention parties and politicians and use hashtags to comment on current events), which no other category has in such a characteristic manner.

Considering their usage of media outlets, they tend to share and/or comment on various links to news from different media outlets, including those related to different political leanings. About their speech, spectators tend to show a discrepancy feeling in most of their tweets, with 7 out of 10 tweets having at least one word related to a discrepancy feeling. They also tend to use "I/We" but not as much as they use "He/She/They" (mainly to refer to the political candidates of the electoral process and their parties), reflecting again that the "us versus them" standpoint is more than spread in the political conversation on the network. Similarly, these insights are supported by the analyses described in Section 6.4.3.

6.4.3 Linguistic patterns and conversation topics

Regardless of the electoral process, all categories have a clear tendency to use neutral language, halfway between formal and informal language. Beyond correct language, we can identify a general dynamic of polarization in the conversation characterized by the "us versus them" approach, as it was already seen in the categories' analyses from political parties to spectators. The usage of "We", especially by political parties and activists, exemplifies the "us versus them" dynamic[3], which is then enhanced by the usage of

[3]Naturally indicating the subsequent polarization of the electoral process.

the pronoun “You”, referring to the idea that the reader (spectator) plays an essential role in what they want to achieve concerning the outcome of the electoral process.

As for the use of verb tenses, this largely depends on the specific characteristics of the particular electoral process. There is no general trend. Instead, its use directly depends on the dominant theme in the electoral process, which also depends on the nature of the introduced narratives.

When considering the hostility in language, no category shows high levels of hostility in their discourse. The exception may be activists, who are characterized by a greater sense of exclusion and the use of swear words. Furthermore, discrepancy is also a constant across different categories, as some tend to disagree in their posts. This is the case for spectators – a trend that remains consistent over time, and political parties – though the tendency is decreasing, possibly because they focus on their narratives without engaging with other narratives outside their ideology.

Considering the LIWC categories, i.e., speech dimensions, treated by each category, they have evolved primarily due to the narratives under discussion. While some dimensions, such as the social aspect, have always been present with varying intensity levels for political parties, politicians, and activists, spectators. Nonetheless, over time, it cannot be asserted that specific dimensions are exclusively tied to a single category of users, except in the case of media outlets and the category related to “time”, due to the nature of their role, i.e., covering breaking news and events. Figure 6.2 displays the results that allow us to derive the above conclusions. Although the figures belong to April 2019, the conclusions remain valid across Spanish electoral processes between 2011 and 2019.

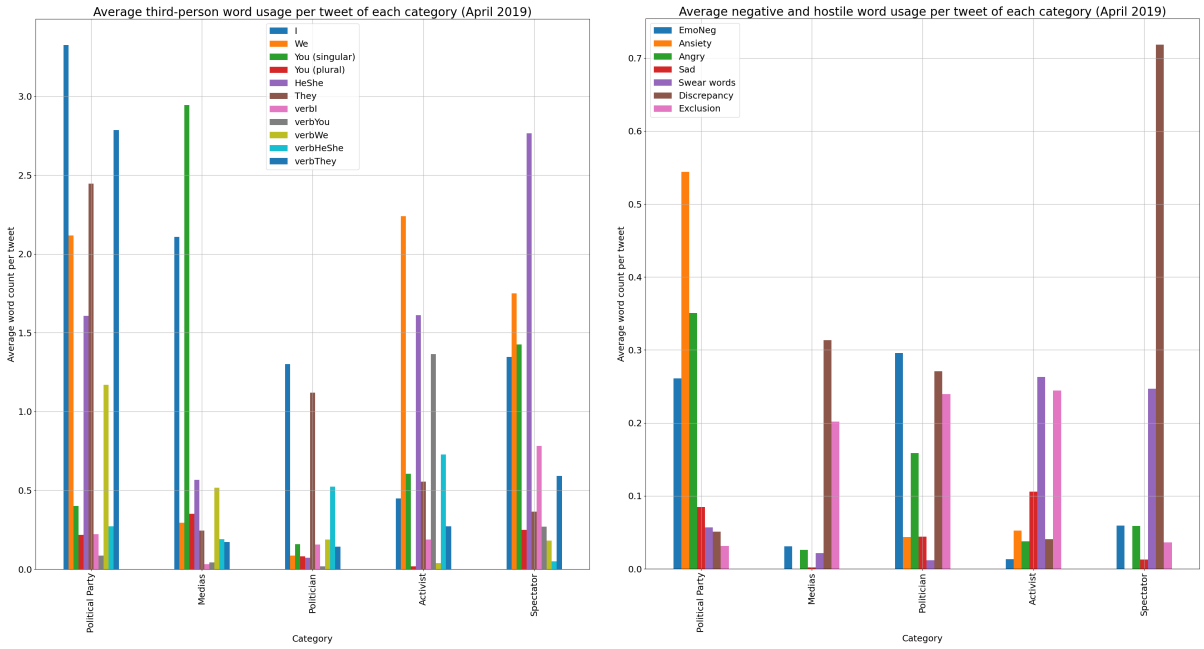


Figure 6.2: Speech analysis per user by category (April 2019).

6.4.4 Activity patterns and interaction dynamics

While media outlets introduce many of the debate topics in political conversation, and political parties generate the narratives that then spread across the network, strictly in quantitative terms, we observe that activists dominate the bulk of the conversation in publications, as it can be seen in Figure 6.3. Their activity is fully aligned with the electoral process: it slightly increases during the pre-campaign period

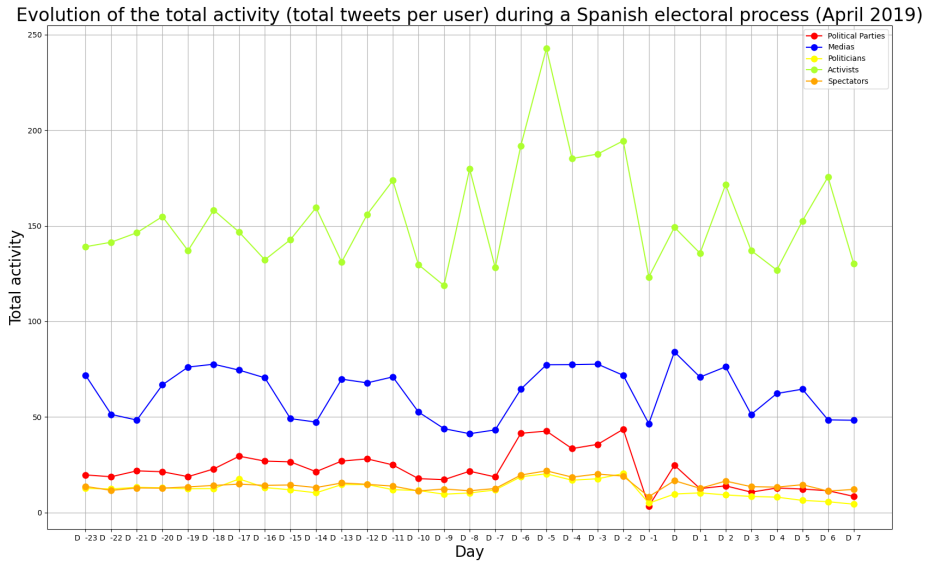


Figure 6.3: Evolution of publications per day per category along the April 2019 electoral process.

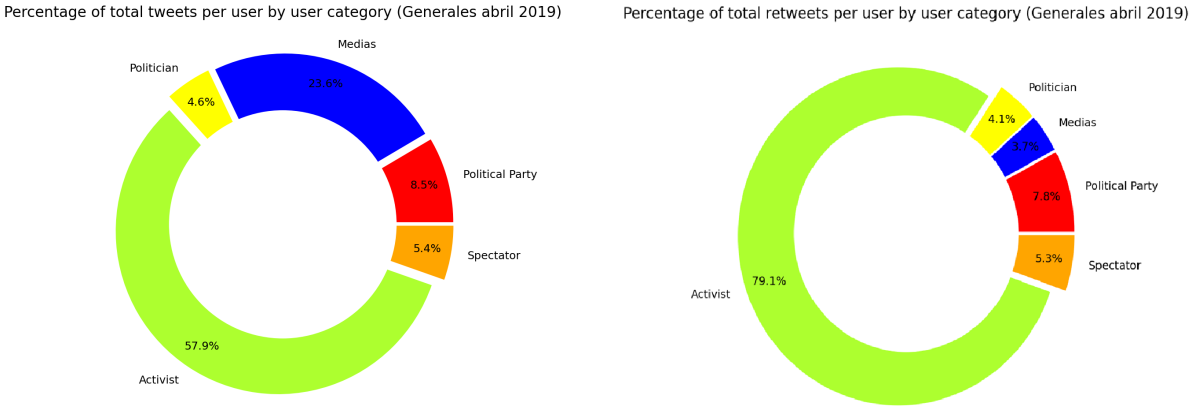


Figure 6.4: Proportion of conversation per category by average of tweets and retweets (April 2019).

up to the campaign, peaks on election day, and becomes particularly relevant in the days immediately following the election, when these users comment on the results either positively or negatively, as the case may be. Likewise, if we look at their activity, we observe that activists primarily play an amplification role, being much more active in retweets than tweets throughout the process.Overall, that can be seen in a graphical way in Figure 6.4.

Political parties show a similar activity pattern to that of activists, as do the accounts of politicians. However, both have a notably lower activity level than the one of the activists. On the other hand, we appreciate how media outlets present a stable pattern corresponding to the usual news cycles. They typically generate content, especially during the week, in parallel with the country's social agenda and events, showing lower activity levels during weekends.

Spectators have low activity levels, focusing on commenting on specific issues through their original tweets while also retweeting (i.e., amplifying) some of the content flowing through the network.

In summary, through the study of categories' interaction dynamics, we observed that, in the political

conversation on the network, everything revolves around political parties, which are the main stakeholders and beneficiaries of said conversation. Therefore, they transmit their narratives in order to, as much as possible, benefit the face of the rest of the categories, including the spectator, who is their ultimate goal. To do so, a complex and complex web of category-interrelation is built to consistently transmit narratives into the network, where the amplification of activists and politicians plays a paramount role, as they make the narratives visible, changing their speech dimensions to those that could be of more interest for the end-users (spectators), enhancing specific dimensions such as the "social" component in the process. Thus, parties and media outlets are the central actors in the conversation, as they are the source of information that impacts the rest of the network, which generally focuses their activity on reacting to their content rather than pumping more content into the network. Both media outlets and parties tend to interact with themselves in terms of parties using links from media outlets, but without retweeting their content, probably to hide a possible "political leaning" from the media outlet concerning the political party. One of the most notable results is the high number of retweets (amplification) that parties give to themselves (as a category), and the same occurs with politicians. The data reflects widespread support among parties or politicians of the same ideology. Thus, regional political parties retweet others and their national variant. The same happens with the leaders or members of these parties, who support each other ("close ranks"). Considering the socio-political reality of Spain, such a natural and evident relationship is faithfully reflected in the data through the analysis performed. Besides, this is as useful as it is evident since it is another way to move and transmit narratives through a social network in a different and complementary way to that of activists.

Similarly, we also observe that these categories consistently group into clearly differentiated political colors. Each political option constitutes a community in the graph (network), as we have seen using the Louvain method, a graphical representation of these communities can be appreciated in Figure 6.5. Each of these communities has users from all categories. That is, it has its reference media outlet, its activists, and a set of loyal spectators. These spectators, in turn, are generally found on the edge of the community, and neighboring political communities might influence them. This is especially relevant because different political parties, politicians, or activists can compete to influence these accounts.

Moreover, we have observed that amplification and self-amplification during the electoral process inevitably lead to a considerable increase in polarization in the electoral process due to the generation of well-differentiated communities that barely exchange information (echo chambers). Such an observation is supported by the obtained results showing that the division into communities becomes clearer when election day approaches, which is a moment of high polarization, partially due to the amplification and self-amplification phenomenon that fosters the generation of well-separated communities where narratives are created and spread through the network. An intuition of this effect is represented in Figure 6.6, where three network representations are shown for different time moments of the Spanish general electoral process of April 2019 (first week of pre-electoral campaign, election day, and first week of post-electoral campaign).

6.4.5 Narrative origination and information flow

Following the studied activity patterns and interrelations between categories, we analyzed the information flow across the network formed during the electoral processes. Although the media do not primarily act as amplifiers and scarcely use retweets, most retweets are directed at political parties. This reveals

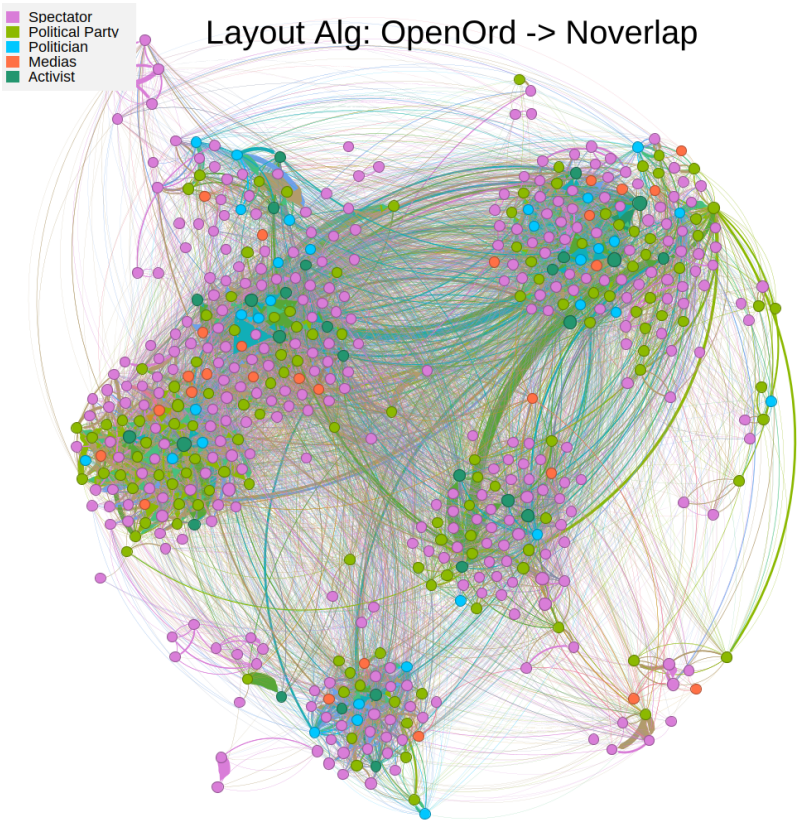


Figure 6.5: Network visualization related to the interactions between user groups, grouped by communities, during the April 2019 election process.

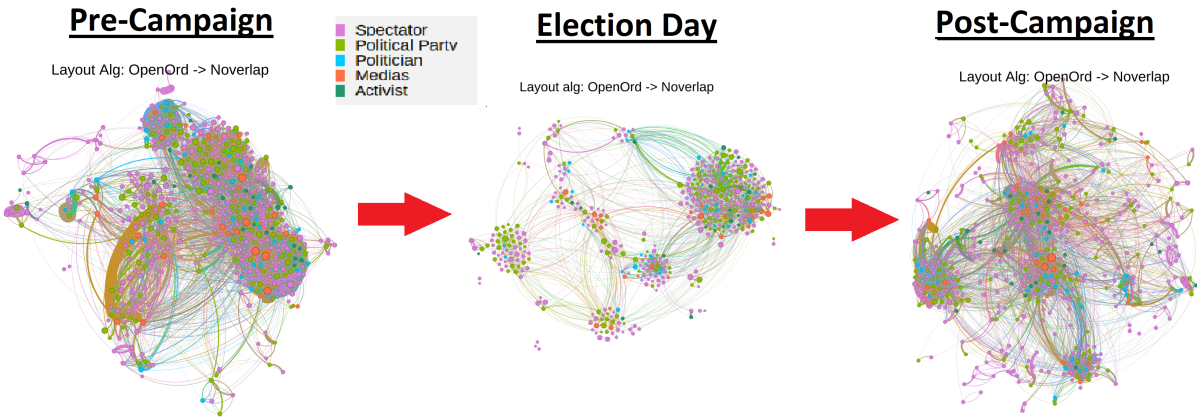


Figure 6.6: Network communities and inter-community interactions in different moments of the electoral process of April 2019.

an interesting relationship: while the media retweet and, somewhat, amplify messages from parties, the relationship is mutual, but not on a large scale (each user in the network representing a political party, be it local, national, or regional, gives on average two retweets to media outlets throughout the entire electoral process).

The fact that this relationship is mutual but without massive feedback is highly relevant, as it supports the identified information flow in the network. Such a flow goes as follows: Media outlets are, evidently, the first to publish news and spread it on the network. These media mention parties and politicians to

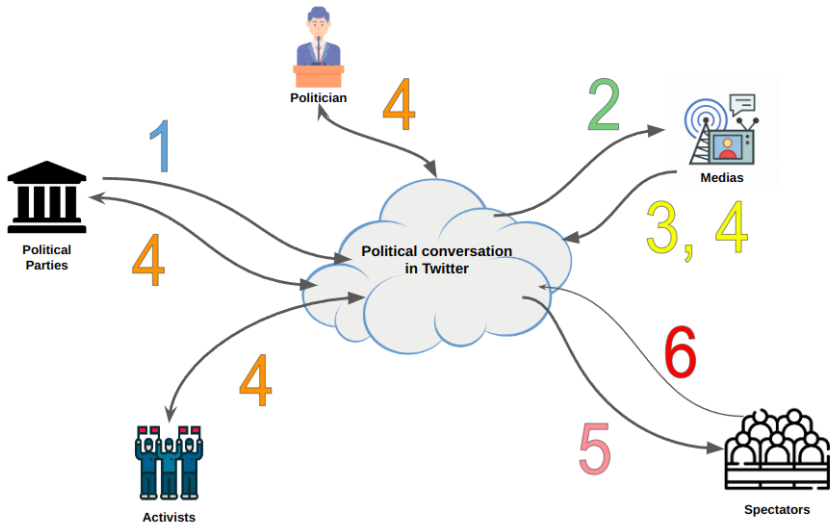


Figure 6.7: Narrative flow between users in the political system.

disseminate the information on social networks. Thus, parties and politicians read the news and create their narratives to spread them. These narratives strongly rely on links to news from these media outlets. However, most of the time, they are not directly supported with a retweet, but they choose to use their tweet (narrative creation) in which the link to the media outlet is included. This justifies the high number of links to media transmitted by parties and politicians. Moreover, thanks to this flow of information, the political party manages to make the media outlet they rely on (even if it is done repeatedly) not be perceived as "politicized" by the viewer since they do not directly retweet most of the time. Instead, they weave their narrative based on a piece of news without the viewer necessarily paying attention to its origin (the media), thus creating consistent narratives on the network. Such an information flow is continued by activists and politicians, who amplify the created narrative, changing the speech dimensions to foster acceptance from the narrative's target: the spectators.

Thus, the observed flow of information during an electoral process within a political context, on a micro-blogging network such as X (Twitter), can be summarized in the following steps (graphically represented in Figure 6.7):

- **Step 1**: Initially, it is the political parties that introduce certain narratives into the political conversation, prompting discussions around them. Logically, these narratives are of their interest, and their objective is to disseminate and maximize their reach. Therefore, it can be genuinely considered that political parties dominate the online political discourse, as they decide on the topics of discussion.

- **Steps 2 and 3**: Subsequently, once the narrative is established, media outlets begin to cover it by producing news articles, which they disseminate across the network with associated links. This step 2 might precede step 1, as the narrative could emerge from public statements, making media outlets the first to "address the topic" on social media. In any scenario, political parties first introduce the narrative, and then the media starts producing news to cover it. These news articles then begin to circulate on the network (step 3). Hence, media outlets play a pivotal role as they provide feedback on the narratives and play a significant part in ensuring their persistence, a concept better understood when considering the subsequent steps. In this context, the increasing use

of the second-person singular ("you") by the media throughout electoral processes contributes to the growth of narratives, as it, in some manner, encourages readers to engage with them.

- **Step 4**: Once the narrative has been crafted and media outlets have initiated their coverage, the narrative expands through various amplification processes:

  - **Political Parties:** They foster the growth of the narrative through diverse interactions. Political parties start by creating posts with links to news from different media outlets to either support the narrative or attempt to discredit it, based on their stance towards it. Additionally, parties aligned with the narrative tend to retweet, meaning they auto-amplify messages related to it. It is vital to underline that when employing news from media sources, political parties tend not to amplify the particular media outlet, but rather utilize the news content itself. The rationale behind this might be to avert conveying to the reader that parties and media are "colluding", hence sidestepping the perception that certain media outlets exclusively favor specific parties due to ideological reasons.
  - **Politicians:** On the other hand, politicians "adopt" the narrative and commence tailoring its tone to exert a more profound impact on the remaining network users. Consequently, they start using the "I" and "They" pronouns, contributing to heightened polarization, as it "compels" the viewer to determine their stance. Moreover, they shift the narrative's tone, providing it with a more social focus, while slightly veering away from themes directly tied to the narrative, such as work, money, religion, etc. Hence, the amplification of narratives by politicians complements the amplification conducted by parties and activists.
  - **Activists:** Concurrently, activists primarily voice the perspectives of political parties, but not exclusively. They also amplify media outlets, politicians, and even spectators, ensuring the rapid dissemination of the narrative across the network, reaching the maximum number of informed users as swiftly as possible. To achieve this, activists amplify the social component of the narrative without sidelining its defining aspects (money, work, religion, etc.), primarily through retweets.
  - **Media Outlets:** Understandably, once the narrative is present on the network, media outlets persist in covering it as extensively as feasible, introducing new links. These links are subsequently shared by political parties, politicians, and activists in their pursuit to expand (or suppress, depending on their objective determined by their ideology) the narrative.
  - **Spectators and the rest of the narrative flow:** Lastly, steps 5 and 6 are outlined in the scheme. These "steps" merely illustrate that, in due course, the narratives introduced and conveyed across the network eventually reach the spectators, who, in turn, comment on them on the platform. It is essential to emphasize that, given the results obtained (low activity from spectators, primarily based on retweets and original tweets), it is understood that spectators mainly focus on discussing the narratives (without necessarily resorting to direct replies) although at a lesser frequency than they receive them. Hence, the "contribution", represented by the arrow in the figure, of the spectators to the network is less than the information they derive from it. Nonetheless, the narrative does not halt there, as steps 4, 5, and 6 recur throughout the entire electoral process until the narrative fades.

An example of this flow is given in this news article, where the left-wing media outlet "El Diario" created an article to cover the exhumation of the Spanish dictator *Francisco Franco*. Once the news article had been introduced in the conversation through a X post, political parties began to create their own narratives around such an event. An example can be found in this X post from the left-wing Spanish political party *PSOE*. Once the narrative had been introduced, the actors with an amplifying role such as politicians and activists further spread the narrative, changing the speech's tone in the process, as it can be seen in right-wing politician *Isabel Diaz Ayuso*'s tweet commenting on such a narrative.

6.5 Discussion

In our analysis, the role of the media emerges as a pivotal force in shaping public discourse, aligning seamlessly with the vast body of literature that underscores its agenda-setting function. This influential role of the media, traditionally seen in various communication mediums, is distinctly mirrored in the digital corridors of online social networks, notably on platforms like X.

Concurrently, political parties manifest as dominant voices in online political dialogues. Their capacity to craft, control, and disseminate narratives positions them at the epicenter of these discussions, underscoring their leadership in shaping the online political landscape.

Furthermore, our findings highlight the nuanced role of activists within this matrix. These individuals adeptly straddle the divide between overarching political narratives and more localized interests. Serving as conduits, they distill and amplify broad political narratives, ensuring these messages resonate with and reach a broader, more diverse audience.

A familiar theme resurfaces after delving deeper into the nature of interactions between different political factions: antagonism. The age-old "Us vs. Them" dichotomy, a staple in political rhetoric, finds new life and vigor in the digital domain. While echoing traditional political discourses, this dynamic is intensified and broadened within the vast and interconnected realm of online platforms.

On the other hand, our analysis has been mainly focused on the communication space in X (Twitter) in Spain. Although we assume that, given the characteristics of the Western media and political system in general, and the Spanish one in particular, the dynamics will be similar, further research is necessary to verify the consistency of the model presented in other political settings. Similarly, this type of research should be replicated in other networks, such as Meta or YouTube, as these may present different user categories or dynamics, complementing the network dynamics presented in this work.

On the other hand, the categories themselves should be scrutinized. Further research should verify the consistency of these categories and their impact on other politically charged processes, such as scandals, demonstrations, and cyber protest campaigns. Similarly, it would be interesting to examine the dynamics of these categories in local electoral processes.

Moreover, further research could also be performed to understand how the presented information flow could impact the generation and transmission of disinformation through the network. In this way, further research could help to clarify how disinformation is injected, amplified, and transmitted through OSNs such as Twitter, potentially identifying user categories that are more likely to introduce and spread disinformation in the network.

6.6 Conclusions

The evolving landscape of political processes on the internet presents a highly intriguing phenomenon in contemporary society. The profound transformation of the political sphere is underscored by the increasing prominence of social media platforms, which have emerged as natural successors to traditional centralized media outlets. This shift has profoundly impacted the mobilization of public opinion, as these digital platforms facilitate rapid and widespread dissemination of information and foster a participatory environment for citizens to engage with political issues. The convergence of technology and politics has ushered in a new era where individuals and grassroots movements can potentially wield considerable influence in shaping the discourse and outcomes of political processes.

In this research, we presented the results of an analysis aimed at understanding the different user types and their roles in the political conversation on a microblog-based OSN, specifically X. To accomplish this, we applied unsupervised machine learning techniques, particularly clustering techniques (specifically K-Means), to various datasets related to Spanish local and general political processes between 2011 and 2019. These datasets were gathered using *snowball sampling*, starting from the accounts of the main political parties in the country.

This analysis also included an in-depth study of the relationships between different user categories to understand further relevant aspects of the political conversation surrounding an electoral process on X. Notably, we found that the quantitative dominance of the conversation, held by activists, does not necessarily translate into actual dominance, as media outlets and political parties play central roles by introducing news and narratives that rapidly spread through the network, reaching user categories with a more amplifying role, such as activists and politicians (particularly during the week of elections). In our research, we found that *storytelling* plays a pivotal role in creating narratives, justifying using X's threads to explain these narratives better.

We also identified a clear information flow that explains the life cycle of narratives, beginning with creating political parties' posts containing website links to news previously shared by media outlets. This flow is then continued by narrative amplification and a shift in speech dimensions, primarily driven by politicians and activists to enhance the social component of the narrative, making it more relatable to the end-users: the spectators. We discovered that spectators tend to have a more commentator role in this conversation, as they create posts around the main topics of discussion in the network, as defined by the narratives that spread across it. This flow was not solely based on quantitative analyses, as qualitative analyses were also performed to confirm its usefulness and veracity further.

While it is known that polarization is inherent in electoral processes, our social network analysis revealed that polarization intensifies as the election day approaches, primarily due to the well-known "us versus them" dynamic, further reinforced by the speech of different categories in their posts (from political parties to spectators).

Furthermore, our conclusions remain valid for various Spanish electoral processes between 2011 and 2019. Future research could explore whether the same conclusions apply to similar political systems in many European countries.

In summary, our research contributes to understanding information dynamics during electoral processes, which could prove helpful in mitigating the polarization and disinformation commonly associated with electoral processes. The results presented in this research could translate into strategies to foster a healthy and open debate around political electoral processes on online social networks.

Part III

Information Dynamics and Recommender Systems in Online Political Social Networks

"Words can kill. Isn't that the point of language, to control people's minds?"
— Revolver Ocelot, *Metal Gear Solid V: The Phantom Pain*

Chapter 7

Modeling Disinformation Networks in Online Political Social Networks

7.1 Introduction

The media, including newspapers, radio, and television, has played for a very long time an instrumental role in shaping societal narratives and influencing public opinion on various subjects, particularly political and social matters. This influence is not merely a reflection of the media's role in information dissemination, but also an indicator of its potential as a tool for power [CM01]. Consequently, many political actors have harnessed this tool to their advantage, utilizing media platforms to propagate their viewpoints and ideologies through highly curated narratives [GTC+20]. This phenomenon is neither new nor transient, as it continues to unfold in the constantly evolving media landscape of the present day [MS72, IK87].

The evolution of traditional media into digital platforms has expanded the reach of these narratives and complexified their dynamics. In this digital age, the line between the producer and consumer of news has blurred, resulting in a significantly more participatory and less controlled environment [LBB+18]. This transformation has paved the way for a paradigm shift in influence dynamics, consequently opening doors to disseminating, not just diverse viewpoints, but also unverified information and disinformation. The accessibility and interactivity of social media platforms, like micro-blogging sites (including Twitter/X, Threads, or Mastodon among others), have made them prime platforms for such activities [ZAB+18].

Indeed, social media platforms have democratized access to information, allowing users to both consume and generate content [TJLL18]. This paradigm shift has resulted in an unprecedented expansion in the volume of information available to the public, contributing to digital media's ascendance over traditional media. X, among other social media networks, has emerged as a significant player in this new era of information exchange, serving as a real-time source of news, opinions, and discourses [BM19]. This evolution not only attests to the dynamic nature of media consumption but also underscores its profound implications for understanding the contemporary digital information ecosystem [BHH18]. In addition, in this digital information era, the internet, particularly social media, allows individuals to tailor their information intake according to their preferences [BAS23]. Users have the autonomy to selectively connect with sources they deem credible, trustful, or align with their perspectives, whether

these sources are legitimate news outlets, individual experts, influencers, or even sources known for propagating unverified or misleading content. This personalized nature of information consumption represents a double-edged sword in the modern media environment [FGR16].

The structure of this and other similar platforms fosters a rapid-fire exchange of information, transcending geographical boundaries and establishing an interconnected global community [VRA18]. While the lack of an editorial filter can enhance the diversity of viewpoints and facilitate the spread of grassroots narratives, it also carries implications for the credibility and veracity of information. The absence of gatekeepers raises questions about the quality of the content shared [XZW23], giving rise to phenomena such as misinformation and disinformation, which have become significant concerns in our contemporary digital information ecosystem [LEC12, SSBW23], in part, because of the difficulty to detect the so-called fake news [JWS+23].

As a consequence, online social media became widely consumed in our societies, which in turn has become the dissemination of organized misinformation increasingly pervasive [ZXWY21]. Misinformation, as defined in Chapter 2, is characterized by the deliberate propagation of incorrect or manipulated information, which is often intended to mislead audiences and influence their perspectives or behaviors. This concept should be distinguished from disinformation (also described in Chapter 2), although the terms are often used interchangeably. While disinformation also involves the deliberate spread of false information, it is typically orchestrated by individuals or organized groups with a calculated intent to deceive, often with political, financial, or societal objectives in mind [VRA18] [MNC22]. These groups can coordinate their (dis)informative action both spontaneously or formally (for example, in the case of nation-state-backed disinformation campaigns). When these accounts consistently act in a coordinated way over time, they constitute disinformation networks, which can be cross-platform, as recently characterized in [NCC22].

Therefore, a disinformation network, particularly in social media like X, is essentially a system of interconnected accounts. These accounts are not just casually connected; they are actively collaborating, either implicitly or explicitly, to disseminate false information or deliberately deceptive narratives. This may occur for various reasons, such as for political gain, to sow social discord, to discredit individuals or organizations, or even to manipulate financial markets among other scenarios [MNC22] [SCV+18]. These malicious actors employ sophisticated strategies to shape narratives and manipulate public opinion. They might present distorted facts, entirely fabricated stories, or decontextualized truths to promote a particular agenda or ideology [SCV+18]. In the digital age, these tactics are not confined to shadowy corners of the internet but are often played out on mainstream social media platforms like X. In these platforms' high-speed, high-volume environment, such content can quickly gain traction, potentially influencing large audiences before corrective measures can be put in place [SCV+18].

Consequently, studying misinformation and disinformation on social media platforms is not merely a niche academic pursuit but a pressing concern with real-world implications. It is a field that requires rigorous analysis to understand the structure, behavior, and impact of these disinformation networks [SCV+18, VRA18, FGR16, LEC12]. Moreover, the design principles and user behavior patterns that make X a fertile ground for misinformation and disinformation are not unique to this platform [SSAW19]. Any micro-blogging or social media platform operating under similar mechanisms and attracting a substantial user base could face the same challenges. Such platforms, too, are susceptible to manipulating their features and algorithms by actors aiming to spread misinformation, thereby perpetuating the cycle.

This reality suggests that the phenomenon of misinformation is not only a concern for the present, but is likely to persist and potentially expand into new digital arenas in the future [GNR19, VRA18, SSAW19].

Therefore, understanding the *modus operandi* of disinformation networks on these platforms, the nature of their interaction with legitimate information sources, and the impact they generate is paramount. Our driving hypothesis in this work is that these networks are characterized by their structure and the dynamics of their interactions. The structure can include elements such as the number and arrangement of nodes (individual user accounts) and edges (connections between accounts, primarily by retweeting as the primary mechanism of content sharing), the presence of clusters or tightly-knit groups, and the overall network density [VRA18, GTC+20]. The network dynamics can include factors such as the speed at which information travels through the network, the frequency and patterns of interaction between accounts, and the evolution of these factors over time [SCV+18]. Through this research, we aim to contribute to that understanding by examining the structural differences and behavioral patterns between disinformation and legitimate sources on X. By doing so, we hope to shed light on the mechanisms of disinformation and provide insights to guide future interventions and policies to combat its spread [SSAW19, GJF+19, BMB20].

Our core research objective lies in understanding the network structure and dynamics characteristic of disinformation networks: sets of user accounts that are interconnected through mutual content sharing (retweeting), and actively engaged in creating, sharing, and promoting disinformation [GTC+20]. In pursuit of this goal, we strive to study the temporal evolution of network properties within disinformation networks during a 3.5-year period (from 2019 to mid-2022), contrasting them with those of networks composed of legitimate information disseminators, both in the context of the Spanish political landscape. Our primary interest lies in determining the efficiency with which information—or rather disinformation—propagates within these disinformation networks.

With this goal in mind, we will contrast disinformation actors against journalists as legitimate sources of information. Unlike anonymous users or those with obscured identities who might engage in the propagation of disinformation, journalists are publicly identifiable entities, which imparts a certain degree of accountability and transparency to their actions on these platforms [MVB20]. They are tethered to the media outlets they represent, which typically uphold strict editorial standards and scrutiny before releasing content [NFN+20]. This adherence to journalistic ethics and the principles of truth, accuracy, objectivity, fairness, and public accountability further distinguishes these professionals from disinformation actors [MVB20]. Furthermore, journalists possess a recognized professional track record, often with a considerable following, influencing public discourse. This visibility and credibility they bring to the platform contrast the often covert, manipulative operations of disinformation actors [MVB20, WD17].

Hence, by comparing and contrasting these two types of accounts – disinformation disseminators and legitimate journalists – this research seeks to uncover their distinctive structural differences, behavioral patterns, and consequent impacts on the X network. Such findings would provide critical insights into the battle against the ongoing disinformation crisis. Moreover, we seek to unravel the factors contributing to forming network structures that facilitate the diffusion of information within disinformation networks. We are especially interested in identifying the conditions under which these disinformation networks manifest increased levels of cohesion and efficiency in their flow of disinformation.

At the outset of this chapter, we explore the structural and behavioral characteristics of disinformation networks on Twitter/X, contrasting them with those of legitimate journalistic networks. Through

a detailed analysis spanning several years, this chapter investigates how these networks differ in terms of connectivity, activity patterns, and information flow. By leveraging network metrics such as density, efficiency, and clustering, we uncover the strategies that enable disinformation networks to disseminate false narratives rapidly and effectively. This chapter contributes directly to the second research goal of this thesis, providing an in-depth analysis of disinformation dynamics and their impact on political information ecosystems. It also aligns with the overarching aim of developing computational tools and insights to address the risks posed by these phenomena in the digital age.

The chapter is organized as follows: Section 7.2 provides an overview of the role of media—traditional and digital—in shaping public opinion, contextualizing the rise of disinformation networks. Section 7.3 examines the structural metrics of these networks, revealing differences in connectivity, clustering, and modularity between disinformation and journalist networks. Section 7.4 investigates the content shared within these networks, including sentiment analysis and thematic focus, to identify behavioral patterns and coordination signals. Section 7.5 synthesizes the findings, discussing their implications for the thesis's research goals and outlining strategies for mitigating the impact of disinformation. Finally, Section 7.6 discusses the chapter's contributions and its limitations, while Section 7.7 concludes the chapter by summarizing its main findings and suggesting avenues for future research.

Research questions

Our aims in this chapter can be summarized in the following research questions which contribute to the overall second research goal:

- **RG2:** Conduct an in-depth analysis of the main risk phenomena in political information ecosystems. Develop a set of tools for their computational analysis.
  - **RG2.RQ1:** How do disinformation networks behave compared to legitimate journalism networks according to network structure?
  - **RG2.RQ2:** How do information content patterns influence the structure of disinformation networks?

7.2 Background

Despite the increasing recognition of the existence and operation of specific social media accounts, particularly on platforms like X, dedicated solely to the propagation of disinformation, the full extent of their impact and functionality still needs to be expanded. Research has confirmed that these accounts, often part of more extensive orchestrated 'influence campaigns', can operate with networks of automated accounts or 'bots', primarily amplifying a particular narrative [ML17].

However, what remains nebulous is the magnitude to which these disinformation actors can outcompete or outmaneuver legitimate actors on these platforms. While it is evident that disinformation campaigns can significantly shape the discourse [SSAW19, PvD21], the precise metrics or mechanisms of their influence vis-à-vis authentic voices have yet to be comprehensively examined. For instance, we lack a complete understanding of their reach, spread, or resonance among the audience compared to legitimate information sources.

Furthermore, the level of coordination within these disinformation networks remains an area requiring more empirical scrutiny [GTC+20]. While a degree of coordination is evident in the concerted distribution of specific narratives, the intricacies of these coordination efforts, such as the command structure, decision-making processes, and synchronization methods, must be thoroughly understood. One approach to address this was recently introduced in [MNC22], where the authors propose a synchronized action framework for detecting automated coordination by constructing and analyzing multi-view networks.

Finally, the consequent effects of these campaigns on shaping public opinion, political attitudes, or behavior are still largely conjectural. While anecdotal evidence and case studies provide insights [BMB20, GJF+19, TCC20], the field still needs robust empirical evidence to quantify the real-world impact of these coordinated disinformation actors.

7.2.1 Micro-blogging networks as tools for political information

Micro-blogging networks, with X as a foremost example, have become pivotal instruments in producing and consuming political information in the contemporary digital environment. These platforms, characterized by real-time updates, concise post lengths, and wide-reaching network structures, are uniquely suited to shaping political discourse and mobilizing public opinion [C+11, JJS12].

Twitter/X, in particular, exhibits several distinct features that make it a potent platform for political information exchange. The platform's real-time nature enables instantaneous reporting and commenting on events, facilitating an active, dynamic political dialogue [C+11, JJS12]. Its broad reach, enabled by network structures that transcend geographical and political boundaries, allows messages to disseminate widely and rapidly. Furthermore, the platform's capacity to accommodate diverse voices, from official political figures and journalists to activists and everyday citizens, fosters a multifaceted and dynamic political discourse.

This potent mix of accessibility, immediacy, and reach gives X significant influence over political information landscapes [JJS12]. However, these same attributes can also be exploited by actors intending to spread disinformation, leading to manipulations of the political discourse and potential distortions in public understanding and opinion. Recognizing the intricacies of these dynamics within micro-blogging networks is a fundamental step towards effectively addressing the challenges of disinformation in our digital societies [HSSP13].

7.2.2 Propaganda and other related concepts

Propaganda refers to the strategic and orchestrated use of information, often biased or misleading, to shape public opinion or behavior toward a particular ideological, political, or commercial objective. It is typically associated with deliberately manipulating facts, ideas, arguments, or even emotional appeals to influence an audience [Ell21, Hen43, Huc16].

In micro-blogging networks like X, propaganda can take on unique characteristics. Given the brevity of content and the real-time nature of these platforms, propaganda is often tailored to be easily digestible and rapidly disseminated [RCM+11]. This can include using sensational or provocative language, visual elements, or hashtags to draw attention and encourage sharing. Moreover, due to the networked structure of these platforms, propaganda can quickly spread beyond its initial audience, reaching and influencing a diverse range of users. Propaganda in these networks is not confined to state actors or organizations; even individuals can become propagators, willingly or otherwise [SSAW19].

Disinformation and misinformation

While often used interchangeably, misinformation and disinformation have distinct implications. Misinformation refers to any incorrect or misleading information, regardless of intent. A user might unknowingly spread misinformation, often due to a genuine mistake or misunderstanding [PELTF23, WD17]. Disinformation, on the other hand, is a subset of misinformation characterized by intent. It refers to the deliberate creation and sharing of false or manipulated information to deceive audiences, often to achieve specific strategic, political, or commercial goals [TJLL18, LEC17, WD17].

In the context of micro-blogging networks, these phenomena become particularly complex. Given the speed at which information spreads on platforms like X, misinformation and disinformation can rapidly reach large audiences. The anonymous or pseudonymous nature of many accounts on these platforms can make it difficult to ascertain the intent behind misleading posts, complicating efforts to distinguish between misinformation and disinformation [AG17]. Additionally, algorithms that prioritize engagement can inadvertently promote misinformation and disinformation, as false or sensational content often elicits strong reactions [LEC17, AG17].

Understanding the nuances between propaganda, misinformation, and disinformation is essential in developing effective strategies to combat these phenomena on micro-blogging networks. Each requires a different approach: counteracting propaganda might foster media literacy and critical thinking; addressing misinformation could entail fact-checking and corrective information, while combating disinformation may necessitate platform-level interventions and policy changes [WD17].

Disinformation and propaganda, while distinct, are closely intertwined. Both are used as tools to influence public opinion, often toward a specific political, ideological, or commercial end. However, they differ primarily in their relationship with truth and intention [HPR21, WD17].

Propaganda may utilize factual and false information, but it is mainly characterized by its use of biased or misleading narratives to promote a particular point of view. Disinformation, conversely, involves the deliberate creation and dissemination of false information intending to deceive [HPR21]. In many cases, disinformation can be a form of propaganda. By creating and spreading false narratives, actors can manipulate public perception and behavior to align with their goals. For instance, a political actor might disseminate disinformation about an opponent's policies or personal life to undermine them and sway public sentiment in their favor [WD17].

Disinformation and political polarization

Disinformation also plays a significant role in political polarization, in particular in the so-called *affective polarization*, which refers to the process where a society's attitudes towards political, ideological, or social issues diverge towards extreme opposing positions [C$^+$11]. Disinformation can exacerbate these divisions by disseminating false narratives, particularly those that play on existing biases, fears, or prejudices. For instance, disinformation that portrays a particular political group as an existential threat to another group can intensify existing animosities, leading to further polarization [C$^+$11, AF18].

Micro-blogging networks like X can amplify these effects due to their structure and algorithms. As users are more likely to interact with content that aligns with their views, platforms may serve them more such content, leading to echo chambers that reinforce and intensify their beliefs [AF18]. Disinformation can thrive in these echo chambers, driving polarization by further entrenching users in their existing viewpoints and making them more susceptible to extreme or divisive narratives [C$^+$11, AF18, T$^+$18].

Therefore, while disinformation is not the sole cause of polarization, it can be a powerful catalyst, leveraging and exacerbating existing divisions for strategic ends [C+11]. Understanding this relationship is essential for developing interventions to counter disinformation and mitigate its impact on societal polarization [C+11, AF18].

7.3 Specific Methodology

7.3.1 Data set

For this study, we employed the dataset "Information disseminators" detailed in the thesis methodology (Section 4.1.5) and composed of two main sub datasets: one containing X accounts linked to verified journalists (described in Table 4.16) and another focusing on disinformation actors (see Table 4.15). This dataset was chosen because disinformation actors disseminate false narratives across a wide variety of events over extended periods, requiring a longitudinal approach to capture their persistent behavior. Such an extended and continuous dataset allows for the identification of stable network properties and the study of how these evolve over time.

On this dataset, we implemented social network analysis techniques to rigorously examine the intrinsic network properties of legitimate information and disinformation nodes. The core objective of this phase was to discern the main difference between these contrasting networks, thereby yielding insights into the probable pathways of information propagation within each of them.

7.3.2 Modeling techniques

In order to analyze the key differences between both networks, we employed content analysis and network analysis techniques. These include the creation of network graphs that depict the complex interplay of interactions between users and provide a visual and quantitative representation of information flow. We divided these graphs into specific temporal segments to capture temporal variations in these dynamics. Complementing this, we dove into the content patterns, using a state-of-the-art deep-learning algorithm for sentiment analysis and observing specific discourse trends. This two-pronged approach enabled us to understand the pathways of information spread and the role shared content plays in these dynamics.

Network generation

We partitioned the activity data gathered for both sets of accounts into temporal segments, which we defined as one week in duration, spanning 2019 to 2022. This decision was grounded in the observation that, on X, news typically has a lifespan of 24 hours, except for certain viral news pieces, especially those concerning social and political issues, that may persist for a more extended period [MSZS14, GRZW21]. In our assessment, a seven-day window aptly captures this fluctuation. Moreover, segregating both networks into associated temporal points allows us to juxtapose their activity patterns over time statistically, as we shall explain in Section 7.3.3.

Therefore, we generated a graph for each temporal segment and data set (i.e., journalists and disinformers). Within these graphs, nodes correspond to the identified accounts, and directed edges symbolize the retweet action from one account (A) to another (B). The weight of these edges equates to the

number of retweets exchanged between accounts during the corresponding time window. This representation offers a quantifiable and visual method of understanding the flow of information and the dynamics within these networks.

Network analysis

After generating graphs corresponding to each temporal window for both data sets, we probed their characteristics utilizing established network metrics. This examination aims to identify the type of network within which information propagates more rapidly and, where feasible, pinpoint contributing factors to this phenomenon. Concurrently, our exploration extended to the patterns of content generation within these networks at each temporal juncture. We sought to comprehend the interplay between the nature and volume of shared content and how it may shape the configuration of the network.

Our analysis was designed to furnish a comparative evaluation of the evolution of both networks over the designated period. Furthermore, it was instrumental in deciphering the dynamics that either accelerate or decelerate the flow of information within the network. Notably, the robustness of the network — its resilience to disruptions or perturbations — was also a focal point of our investigation.

Through this in-depth analysis, we aimed to unveil the complex workings of these networks, offering invaluable insights into the behavior of disinformation networks on X and their subsequent impact on the legitimate journalistic landscape. Next, we summarize the network metrics employed in this work, based on well-known concepts from the area [New10].

Density: The density of a directed and weighted graph is a measure of how many edges are present in the graph compared to the maximum possible number of edges. It quantifies how "connected" the graph is. Given a directed and weighted graph G with N nodes and M edges, the density can be defined as:

$$D(G) = \frac{M}{N(N-1)} \tag{7.1}$$

Average Degree (Weighted): The average degree (also called average degree centrality measure) of a directed and weighted graph is a measure of how many connections, on average, each node has, considering the weights of the edges. The average degree can be calculated separately for in-degree ($\bar{k}_{in}(G)$), that is, the number of edges pointing to the node, and out-degree ($\bar{k}_{out}(G)$), that represents the number of edges starting from the node in a directed graph.

$$\overline{k}_{in}(G) = \frac{1}{N}\sum_{i=1}^{N} k_{in}(i) \tag{7.2}$$

$$\overline{k}_{out}(G) = \frac{1}{N}\sum_{i=1}^{N} k_{out}(i) \tag{7.3}$$

Efficiency: The efficiency of a directed and weighted graph G with N nodes and M edges is a measure of how efficiently information can be transmitted across the network. Efficiency is calculated as the inverse of the average of the shortest paths between all pairs of nodes in the graph.

$$E(G) = \frac{1}{N(N-1)}\sum_{i\neq j\in G}\frac{1}{d(i,j)} \tag{7.4}$$

where $d(i,j)$ computes the shortest path distance between nodes i and j.

These metrics can be used to analyze the structural properties of a graph, providing valuable insights into the connectivity, clustering, community structure, and overall efficiency of the network.

Modularity: The modularity of a directed and weighted graph measures the strength of its community structure. It quantifies the difference between the number of edges within communities and the expected number of edges if the edges were distributed randomly, preserving the nodes' in- and out-degree. Thus, given a directed and weighted graph G with N nodes partitioned into C communities[1], the modularity can be defined as:

$$Q(G) = \frac{1}{C} \sum_{c=1}^{C} \left[e_c - \frac{k_{in}^{(c)} k_{out}^{(c)}}{C} \right] \tag{7.5}$$

Although several approaches could be used to partition the network in communities, in this work we use the Louvain method [BGLL08].

Average Clustering Coefficient: The average clustering coefficient of a directed and weighted graph measures the degree to which nodes in the graph tend to cluster together, considering the weights of the edges. It is the average of the local clustering coefficients of all nodes in the graph.

$$\overline{Cc}(G) = \frac{1}{N} \sum_{i=1}^{N} \frac{2E_i}{k_i(k_i - 1)} \tag{7.6}$$

where $\overline{Cc}(G)$ represents the average clustering coefficient of a graph G. It is calculated by summing up the local clustering coefficients for each node i in the graph (computed as the ratio of twice the number of edges E_i between the neighbors of node i to the product of the (in-)degree k_i of node i and its degree minus one), and dividing by the total number of nodes N. This quantity measures the overall tendency of nodes in G to form clusters.

Average Eigenvector Centrality: The average eigenvector centrality of a directed and weighted graph G with N nodes is a measure of the overall importance or influence of nodes in the network. It takes into account not only the number of connections a node has, but also the importance of the nodes to which it is connected.

$$\overline{EVC}(G) = \frac{1}{N} \sum_{i=1}^{N} EVC(i) \tag{7.7}$$

where $EVC(i)$ is the eigenvector centrality of a node, a measure of the influence of that particular node in a network. It assigns relative scores to all nodes in the network based on the principle that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. Given a directed and weighted graph G with N nodes and adjacency matrix A, the eigenvector centrality $EVC(i)$ of node i can be found as the element i of the eigenvector v corresponding to the largest eigenvalue λ of the adjacency matrix, i.e., $Av = \lambda v$. It should be noted that this metric is only defined for undirected networks; hence, to get around this issue and use it in our directed networks, we adapted it by symmetrizing the graph. Specifically, if account A retweeted account B, or vice versa, we treated this as a non-directed link between A and B for the purposes of calculating Eigenvector centrality.

[1] As groups of tightly inter-connected nodes.

Content analysis

In addition to analyzing the properties of the networks, which remain independent of the specific content shared by each account, we ventured into the examination of the type of content posted within these networks. This deeper dive aimed to discern any possible correlation between the nature of shared content and the evolving structure of the network over time. Specifically, we sought to identify whether certain types of content or attitudes could contribute to or be associated with increased efficiency or density within the network.

Tweet sentiment: In order to achieve this, we started by analyzing the sentiment associated to each publication. We employed a state-of-the-art deep learning-based algorithm [PGL21] for classifying user posts based on their sentiment. The classification divided posts into those displaying predominantly positive sentiment and those with predominantly negative sentiment. From there, we could compute the average number of predominantly negative tweets during a particular period.

$$\text{sentiment}(tweet) = \begin{cases} \text{'negative'} & \text{if neg}(tweet) > \text{pos}(tweet) \\ \text{'positive'} & \text{otherwise} \end{cases} \tag{7.8}$$

References to controversial events per tweet: We also sought to identify references such as words and slogans to significant geopolitical events, such as NATO-related activities or the war in Ukraine. These topics have been a focal point within the Spanish communication space during the period under review and have been substantially covered by actors disseminating disinformation, as indicated by existing research [Gar23, SWB+22]. Besides, in light of the COVID-19 pandemic's significant presence in major disinformation studies, we also included it in our analysis [KJK+20]. These three events, while global in their implications, had direct and significant impacts on Spain: a) Spain hosted the NATO summit in Madrid in 2022, which brought NATO-related discussions and narratives to social media; b) the conflict in Ukraine also held substantial relevance in Spain, both because it was the first major war on European soil since the Balkan conflicts, but also because the Spanish government was divided in their discourses, further increasing the controversy around this event; and c) the impact of COVID-19 on Spain was particularly pronounced, given the country's implementation of a highly restrictive and controversial lockdown policy, making it a central topic of discourse and disinformation within the Spanish X sphere.

In this context, we calculated the average number of references to COVID, NATO, and Ukraine per tweet within the network for each time window studied. In order to do that, we started by calculating the number of references to each of the topics by searching for the words 'COVID', 'NATO', and 'UKRAINE' in each text and then averaging all the occurrences found per tweet.

$$\text{avg\_nato} = \frac{1}{T} \sum_{tweet=1}^{T} \text{references\_nato}(tweet) \tag{7.9}$$

$$\text{avg\_ukraine} = \frac{1}{T} \sum_{tweet=1}^{T} \text{references\_ukraine}(tweet) \tag{7.10}$$

$$\text{avg\_covid} = \frac{1}{T} \sum_{tweet=1}^{T} \text{references\_covid}(tweet) \tag{7.11}$$

where T is the number of tweets in a specific instance of the network.

Table 7.1: Properties of the complete networks (i.e., using all the data from the entire 2019-2022 period), in both cases, with 275 nodes[2]. RTs refers to retweets (edges), and the last four columns correspond to metrics defined in Section 7.3.2 (in all cases, metrics are in the $[0, 1]$ range, where a higher value denotes the network is more efficient, modular, or clustered).

	Tweets	RTs	$E(G)$	$Q(G)$	$\overline{Cc}(G)$	$\overline{EVC}(G)$
Journalists	3,906,047	96,551	0.170	0.743	0.334	0.028
Disinformation	7,194,766	513,566	0.268	0.580	0.502	0.036

URL and hashtags per tweet: In the final phase of our content analysis, we noted the number of URLs and hashtags shared per tweet on average within the network for each studied time window. This allowed us to observe potential patterns or shifts in content sharing behaviors over time.

7.3.3 Experimental settings

To address the research questions considered throughout this work, we have processed the data collected as explained in Section 7.3.1. First, let us recall we create two (sets of) networks by exploiting the retweet action among two subsets of users: those categorized as journalists and those as disinformation actors (see Section 7.3.2). A summary of the overall graphs generated when using all this information is presented in Table 7.1.

Moreover, since the data was collected during 3.5 years (2019-2022[3]), a different network was created for each temporal segment of one week of duration, resulting in 338 different networks. These networks are the ones considered for analysis in this section. In the experiments we present in the following sections, we use this data in several, complementary ways. In some cases, we consider the temporal evolution (time series) of all the network metrics defined in Section 7.3.2. This means that those metrics were computed on each network, and the obtained scores were recorded for every instance of the network throughout the 2019-2022 period (once for each temporal segment). These time series will be considered to assess whether any statistically significant difference exists between the two types of networks (journalists and disinformation), as our aim is to delineate the behavioral patterns distinguishing disinformation networks from journalist networks, with a particular focus on the interaction structures within each network.

For this, we initially applied a Mann-Whitney U test, assuming the null hypothesis (H0) is that *there is no difference between the journalists and disinformation actors groups* for each scenario. This test was used because, after checking the normality of the data using the Shapiro-Wilk test, the results indicated that the data were not normally distributed for both groups. Thus, a non-parametric test was chosen for further analysis.

In conjunction with the Mann-Whitney U test, our methodology also incorporated a one-sample t-test on the weekly differences in average metrics between the two groups. This approach enables an examination of whether the observed weekly mean differences in these metrics are significantly distinct from zero, thus offering insights into the dynamic interplay between the networks over time. The integration of this test complements the distributional analysis provided by the Mann-Whitney U test, shedding

[2]The 275 nodes were selected to provide a statistically significant sample size, ensuring a confidence level of 95% with a margin of error below 6% for the observed network metrics.

[3]In fact, only half year of 2022 was considered, since that was the most recent data available at request time.

light on both the distributional differences and the temporal consistency and significance of these differences. However, since it can only be applied to normally distributed data, we applied a logarithmic transformation before running the test.

Furthermore, to augment the robustness of our findings, we employed the Kolmogorov-Smirnov (KS) test for a granular analysis at the individual user level. This was particularly pivotal for our study of eigenvector centrality and degree centrality. For each user in both the disinformation and journalist networks, we computed the weekly averages of these metrics. The KS test was then applied to these data sets to determine if the distributions of eigenvector centrality and degree centrality values for individual users differed significantly between the two networks. This level of detailed analysis allows us to assert with greater confidence whether the observed patterns in network metrics are indeed reflective of underlying differences in the behavioral dynamics of disinformation actors and journalists.

The integration of the Mann-Whitney U test, the one-sample t-test, and the Kolmogorov-Smirnov test in our methodology provides a comprehensive and scientifically rigorous framework. This multifaceted approach enhances the depth of our analysis, as it examines the contrasting behaviors of disinformation and journalist networks across both aggregate and individual levels over a temporal spectrum.

Finally, the other main method used in our experiments consists of a correlation analysis via scatter plots, where a linear fit of the data is attempted, producing a measure of the goodness-of-fit and the probability that the relationship between the two variables is equal to zero (p-value).

7.4 Results

In this section, we will address the research questions introduced at the beginning of this chapter. In Section 7.4.2 we will analyze the output of the structural network metrics presented before (**RG2.RQ1**), while the output of those metrics related to information flow and propagation will be considered for **RG2.RQ2**. Moreover, for all the considered metrics we will compute significance test statistics to address the statistical significance of the variations in the temporal patterns of activity between disinformation networks and legitimate journalism networks, whereas correlation between the type of content and the structure of the network will be presented in Section 7.4.3 to answer **RG2.RQ2** studying how do the information content patterns influence the structure of the disinformation network. Before that, in the following Section 7.4.1, we introduce an initial analysis on the collected data.

7.4.1 Initial analysis of collected data

In our preliminary analysis, journalists and disinformation actors displayed consistent patterns of reciprocal retweeting. This common behavior of sharing each other's content over time results in the creation of information networks. Within these networks, users receive information from a variety of sources. In turn, any information – news, commentary, slogan, etc. – inserted into a network can circulate within that network via retweets. Given this observed phenomenon, we concluded that building network graphs is the optimal approach to investigate the dynamics of information dispersion within both groups, as done in previous works [SC18].

As Table 7.1 shows, the two analyzed networks evidenced different values of efficiency ($E(G)$), modularity ($Q(G)$), average clustering coefficient ($\overline{Cc}(G)$), and average eigenvector centrality ($\overline{EVC}(G)$), in particular, the disinformation network is more efficient and evidences a higher average clustering

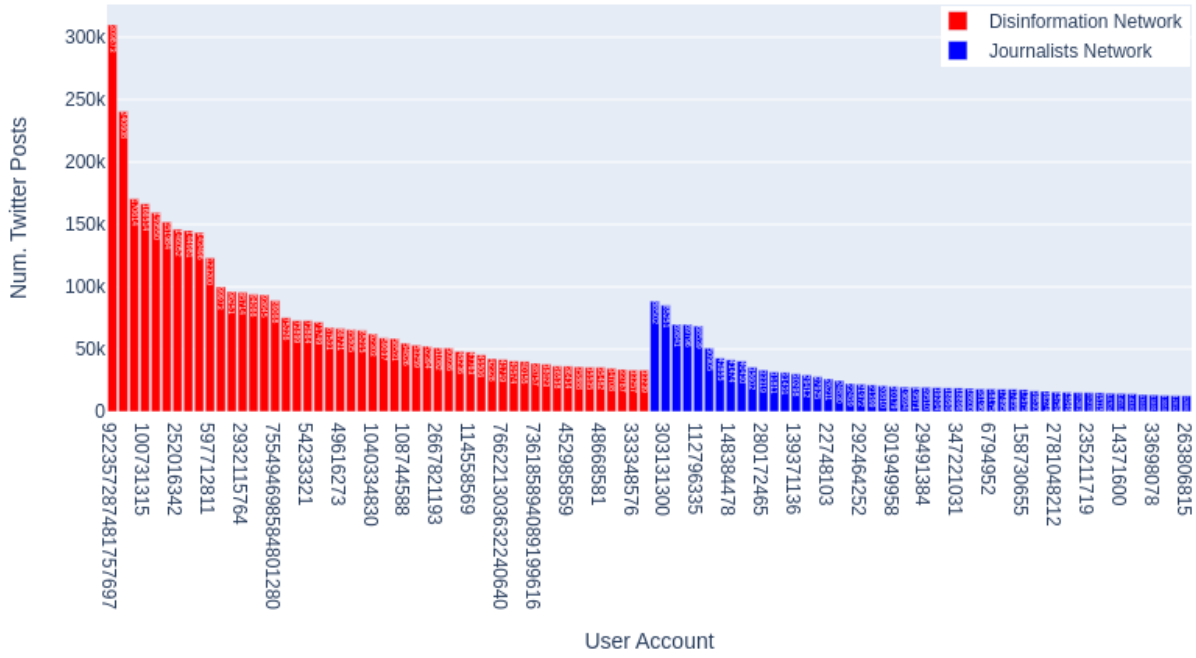


Figure 7.1: Number of publications of the 50 top accounts in each network 2019-2022.

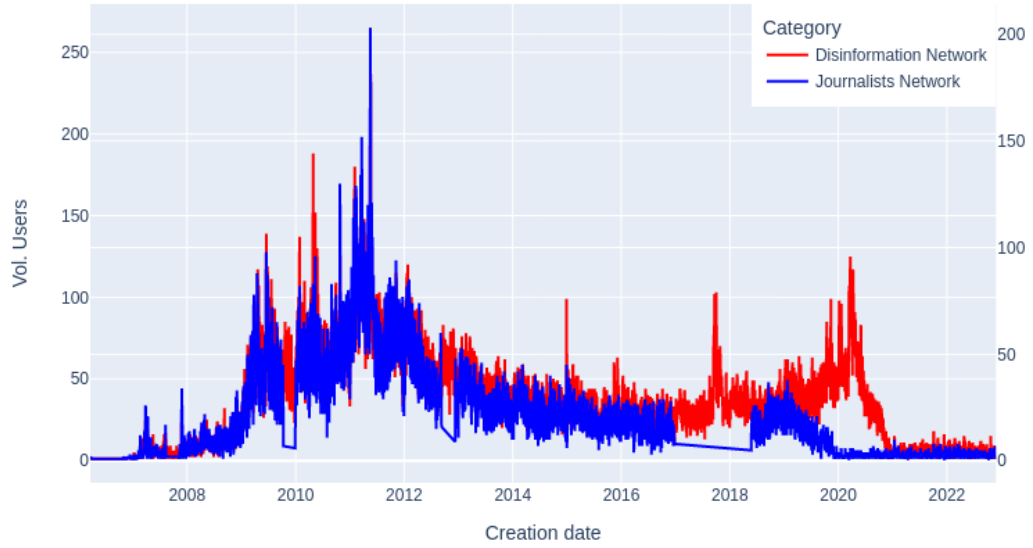
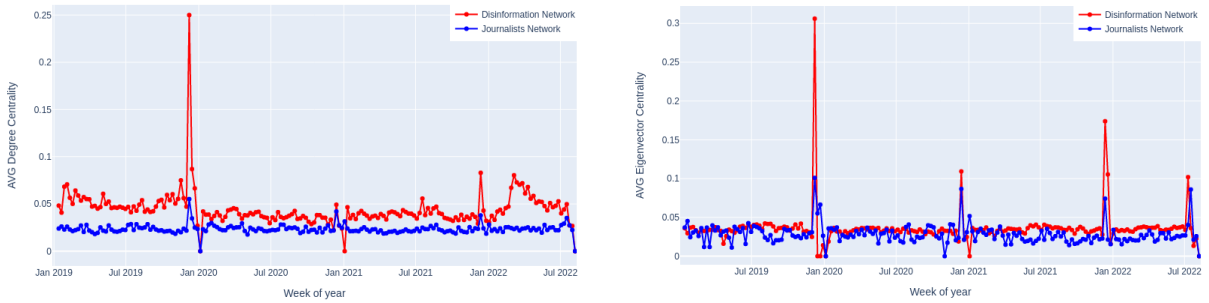


Figure 7.2: Number of accounts created per month in the Disinformation and Journalists networks. Including the accounts mentioned, quoted or retweeted by them.

coefficient, highlighting its internal cohesion. However, since these values are collected for the entire networks, no fine-grained analysis can be performed – something we shall show later in subsequent sections.

Moreover, upon initial observation of account activity throughout the studied period, we note that the disinformation network and the network of legitimate actors demonstrate patterns where few users are responsible for most posts (see Figure 7.1). It is also evident that the disinformation network produced a (total) higher volume of posts over the study period, since the top users of each network produced a remarkably different number of publications: more than 300K for the disinformation network and around 90K for the journalists.



(a) Evolution of average degree centrality.

(b) Evolution of average eigenvector centrality.

Figure 7.3: Evolution of average degree centrality (left) and average eigenvector centrality (right) on the Disinformation and Journalists Networks.

Considering the account creation dates in Figure 7.2, we observe that most accounts associated with legitimate actors were established between 2010 and 2012, coinciding with X's rise in popularity in Spain, but also with an electoral period. In contrast, we noted two periods of substantial account creation within the disinformation network, one in 2018 and another in 2020. Interestingly, the latter coincides with the onset of the COVID-19 pandemic.

7.4.2 Behavior of disinformation networks according to the network structure

In this section, we perform different experiments to understand the behavior of disinformation networks (in comparison with the behavior of journalist networks) by considering the structure of the network derived through the interactions between the nodes in each network. For this, as explained in the methodology, we will contrast and compare those two networks throughout the time dimension, in particular computing the Mann-Whitney U significance test assuming the null hypothesis (H0) is that there is no difference between the journalists and disinformation actors groups.

The results of these tests are presented next, according to the type of metrics being analyzed: first, connectivity and centrality, and later community structure and information flow.

Connectivity and centrality

In this section, we focus on two definitions of centrality: average degree and eigenvector. Figure 7.3 shows the evolution of these metrics throughout the studied period of 2019-2022 for both networks. The results of running the significance tests on these data are:

1. For the average degree (also called average degree centrality, $\bar{k}_{in}(G)$), the Mann-Whitney U test demonstrated a statistically significant difference between the Journalists and Disinformation actors groups (U = 5373.0, p = 3.48e-40). We reject the null hypothesis (H0) for the average degree centrality, indicating that the median degree centrality for both groups is significantly different, being 0.0445 for the disinformation network and 0.0235 for the journalists network.

   Additionally, a one-sample t-test on the weekly differences in average degree centrality revealed a t-statistic of -16.44 with a p-value of 5.01e-38, robustly rejecting the null hypothesis (H0). This indicates that the mean difference in average degree centrality between the two groups is significantly different from zero, with the negative t-statistic suggesting a higher average degree centrality in

the disinformation network compared to the journalists network. Further enhancing our analysis, the Kolmogorov-Smirnov Test was conducted at the individual user level to compare the weekly averages of degree centrality, for each user within both networks over the studied period. This test yielded a KS statistic of 0.78 and a p-value of 2.81e-15, confirming that the distributions of degree centrality values are significantly different between individual users of the disinformation and journalists networks.

These comprehensive findings, which include both network-level and individual user-level analyses, strongly support the conclusion that there are not only significant differences in the distribution of average degree centrality values but also a consistent and notable divergence in the average and median values of this metric over time between the two networks.

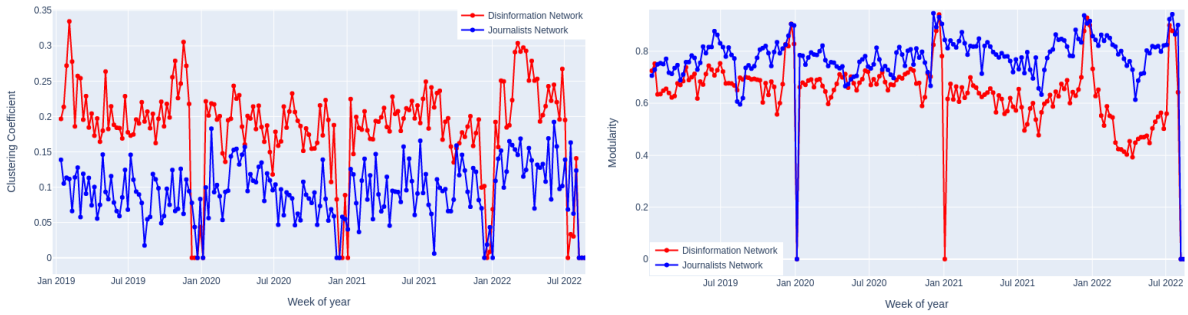
2. For the eigenvector centrality ($\overline{EVC}(G)$), the Mann-Whitney U test indicated a statistically significant difference between the journalists and disinformation actors groups (U = 13348.0, p = 1.36e-11), suggesting distinct patterns in node influence and connectivity. We reject the null hypothesis (H0) for average eigenvector centrality, indicating that the median eigenvector centrality for both groups is significantly different, being 0.0362 for the disinformation network and 0.0284 for the journalists network.

   Further, a one-sample t-test on the weekly differences in average eigenvector centrality yielded a t-statistic of -4.89 with a p-value of 2.21e-06, robustly rejecting the null hypothesis (H0) and indicating a significant mean difference between the groups. The negative t-statistic implies that, on average, the disinformation network exhibits higher eigenvector centrality compared to the journalists network. To deepen our analysis, we again conducted the Kolmogorov-Smirnov Test at the individual user level, comparing the weekly averages of eigenvector centrality for each user within the networks across the studied period. This test resulted in a KS statistic of 0.56 with a p-value of 1.45e-07, confirming that the distributions of eigenvector centrality values are significantly different between individual users of the disinformation and journalists networks.

   These collective findings, encompassing both network-level and individual user-level analyses, strongly support the conclusion that there are not only significant differences in the distribution of eigenvector centrality values but also a consistent and substantial divergence in the average and median values of this metric over time between the two networks.

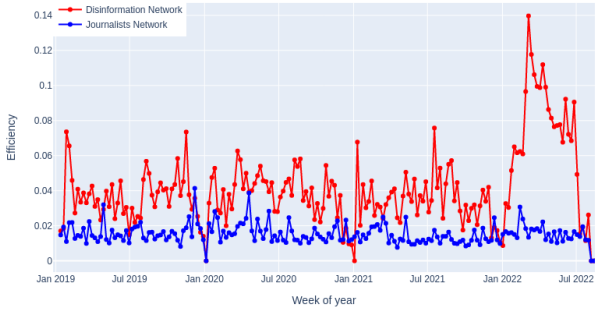
According to the tests, when considering the elements of centrality, the nodes within the disinformation network have displayed a sustained pattern of more connections over time. This attribute feeds into a network structure that fosters and facilitates the rapid spread of information. Furthermore, the higher average eigenvector centrality in the disinformation network reveals that the nodes within these networks are more numerous in their connections and boast superior quality connections. This, in turn, enables a faster distribution of information.

An intriguing aspect revealed by these metrics – in particular, according to EVC – is the potential presence of well-connected ‘conversation leaders’ within these networks, equivalent to webpages with high PageRank, a classical proxy for authority [BP98]. They could be perceived as strategic coordinators or influencers who may help steer the direction of the shared narratives. However, a thorough investigation into this phenomenon would necessitate more detailed research. Identifying and understanding these key actors could be essential for devising strategies to mitigate the influence of disinformation networks.



(a) Evolution of the clustering coefficient.

(b) Evolution of the modularity.



(c) Evolution of the efficiency.

Figure 7.4: Evolution of clustering coefficient (top left), modularity (top right), efficiency (bottom left) on the Disinformation and Journalists Networks.

Community structure and information flow

We now focus on efficiency, modularity, and clustering coefficient network metrics, more associated to how the information flows throughout a given network. Figure 7.4 shows the evolution of these metrics, and the corresponding results when running the tests to contrast the previously defined null hypothesis (there is no difference between the journalists and disinformation actors groups) are:

1. For efficiency ($E(G)$), the Mann-Whitney U test revealed a statistically significant difference between the Journalists and Disinformation actors groups (U = 6959.0, p = 4.80e-33). We reject the null hypothesis (H0) for efficiency, indicating that the mean efficiency for both groups is significantly different, being 0.0416 for the Disinformation network and 0.0154 for the Journalists network.

2. For modularity ($Q(G)$), the Mann-Whitney U test showed a statistically significant difference between the Journalists and Disinformation actors groups (U = 35077.0, p = 5.16e-28). We reject the null hypothesis (H0) for modularity, indicating that the mean modularity for both groups is significantly different, being 0.6451 for the Disinformation network and 0.7807 for the Journalists network.

3. For average clustering coefficient ($\overline{Cc}(G)$), the Mann-Whitney U test revealed a statistically significant difference between the Journalists and Disinformation actors groups (U = 7728.0, p = 6.97e-30). We reject the null hypothesis (H0) for $\overline{Cc}(G)$, indicating that the mean average clustering coefficient for both groups is significantly different, being 0.1871 for the Disinformation network and 0.0956 for the Journalists network.

Drawing from our findings, as proven by the statistical tests, it is clear that information tends to flow more swiftly within the disinformation network. This trend of enhanced efficiency in information propagation has been consistent throughout the study periods from 2019 to 2022. Likewise, the observed disparities in the clustering coefficient and modularity indicate a more fragmented structure over time in the network of legitimate informers and a more cohesive structure among disinformation actors. This suggests that these networks, while vital for disseminating truthful and reliable information, may not be as interconnected or tightly-knit as their disinformation counterparts. Consequently, this could hinder the speed at which accurate information is disseminated within and across these networks.

There are, however, at least two aspects from Figure 7.4 that deserves further explanation. First, the drops in modularity values for both the disinformation and journalist networks, as shown in Figure 7.4b. These are primarily due to distinct periods of reduced activity within these networks. On certain days, the studied accounts, although typically active, exhibited lower levels of engagement. This resulted in smaller networks with fewer retweets, directly impacting the network structure, since modularity, being a measure that hinges on the existence and definition of communities within a network, is sensitive to changes in network size. Second, the substantial spike in efficiency for the disinformation networks in 2022, as indicated in Figure 7.4c, correlates with the significant geopolitical events surrounding the Russian campaign and subsequent large-scale land invasion in Ukraine. During such period, these networks exhibited peaks of activity, likely as a response to the unfolding events. This heightened activity led to increased connectivity and coordination among the accounts within the disinformation network, thus resulting in the observed spike in efficiency.

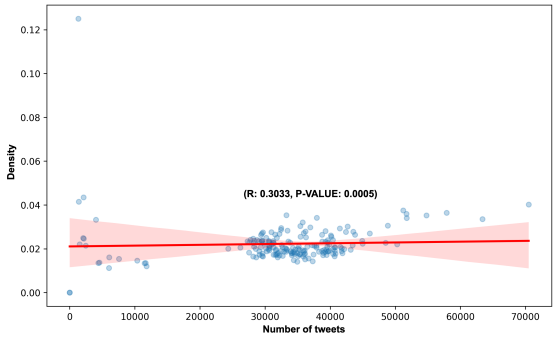
7.4.3 Behavior of disinformation networks according to the network content

Our analysis sought to establish a correlation between the nature of network activity and its structure, emphasizing activities that could hint at coordinated behavior.

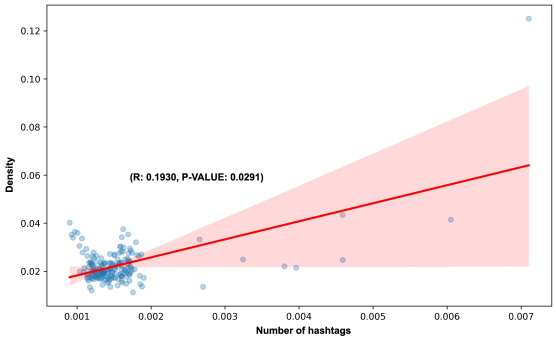
Our analysis revealed significant relationships between network density and various content-related metrics in the disinformation network. As shown in Figure 7.5, panel (a) indicates that higher network density correlates with an increased total number of tweets, suggesting that activity levels contribute to the compactness of these networks. Panel (b) demonstrates a positive association between network density and the number of hashtags per tweet, indicating that disinformation actors may strategically employ hashtags to enhance connectivity. Panels (c) and (d) reveal that specific thematic references, such as NATO and Ukraine, are also associated with increased density, suggesting coordinated messaging around geopolitical topics. Finally, panel (e) shows a notable relationship between density and tweets with negative sentiment, emphasizing the role of emotionally charged content in fostering stronger network cohesion. These findings collectively highlight how content production and thematic focus influence the structural properties of disinformation networks, providing deeper insights into their operation.

For this reason, in this experiment, we analyze the networks at different moments in time and correlate their content characteristics against their density and efficiency to identify specific patterns and behaviors that might indicate coordinated actions or strategic use of thematic narratives.

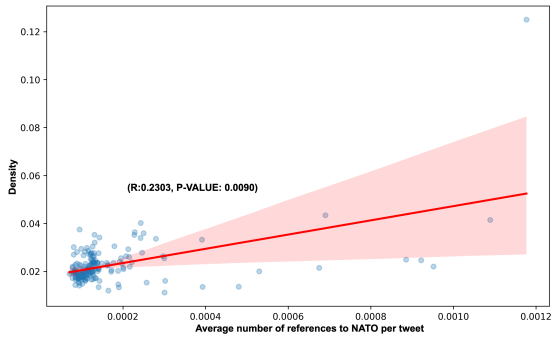
First, in Figure 7.6 we show the journalist and disinformation networks at different moments in time. Each dot in the graph is a node (X account) of a given network, its size is proportional to the number of retweets it made during that period, and an edge exists if a retweet was made between those nodes. The colors represent communities detected by the Louvain method [BGLL08]. Based on these graphs, we



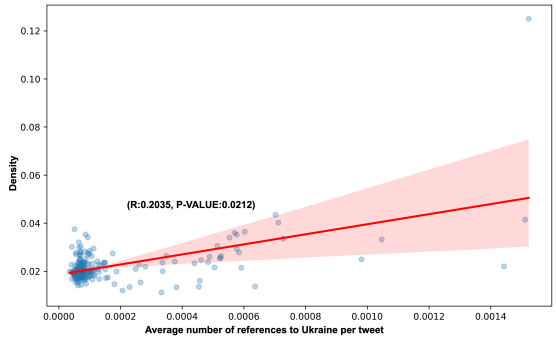
(a) Density VS Total number of tweets.



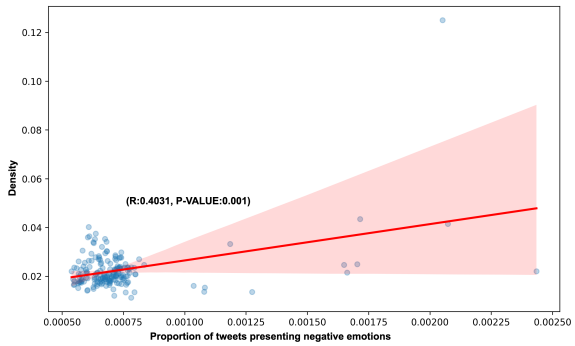
(b) Density VS number of hashtags per tweet.



(c) Density VS NATO references per tweet.



(d) Density VS Ukraine references per tweet.



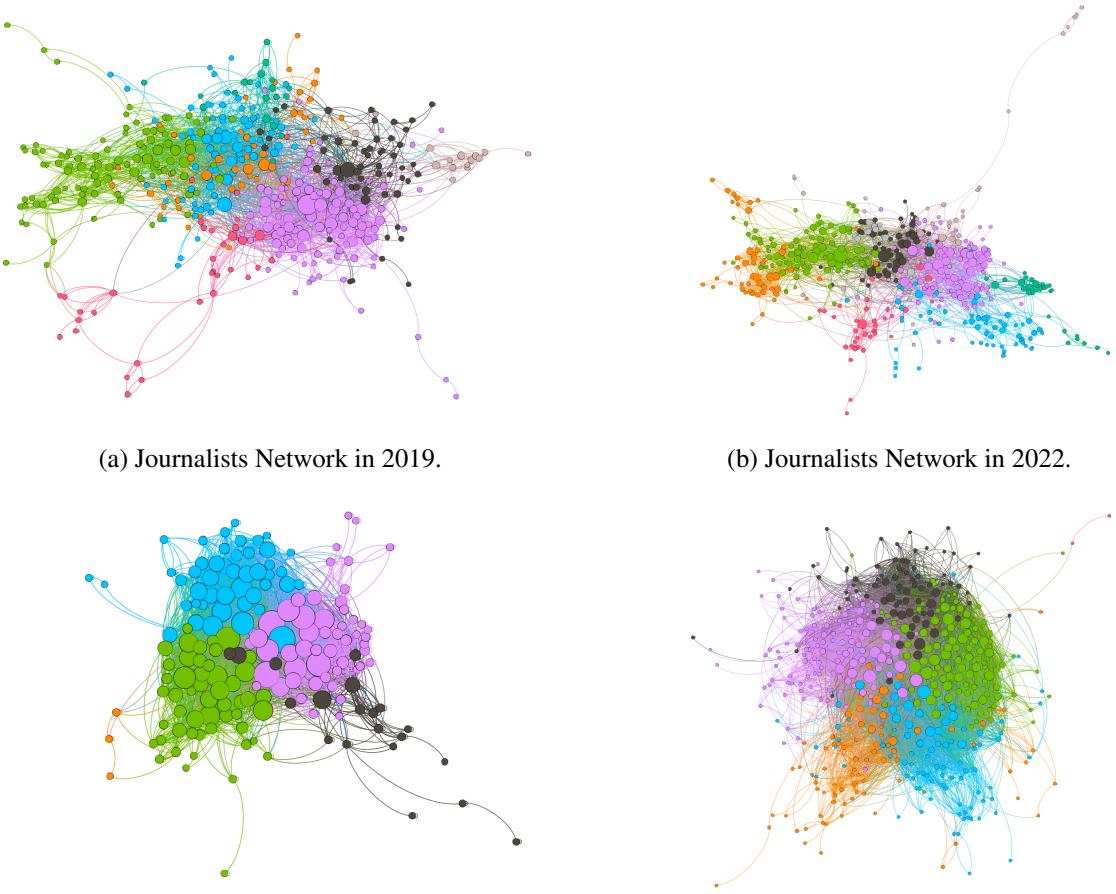
(e) Density VS Number of tweets with negative emotion per tweet.

Figure 7.5: Scatter plots between density and other variables in the disinformation network, including the p-value and the goodness of fit (R) of a linear fit on such data.

observe that the disinformation networks tend to be less spread than the journalists networks. There are also more isolated communities in the journalist case, and the size of their nodes (the number of retweets) is smaller, evidencing their lower rate of interaction with the rest of the network.

Second, we found that the disinformation network density tended to be higher during periods with increased post and hashtag volumes, suggesting that the network becomes denser when it resonates or orchestrates a communication campaign. Some examples of this behavior are shown in Figure 7.5, where the reported metrics are computed weekly on the disinformation network and plotted against their density. While no clear correlations emerged regarding posts related to COVID-19, we observed that the network displayed increased density when its focus on Ukraine or NATO intensified. This could indicate a coordinated effort by state actors or organized groups around these topics.

Similarly, when contrasting the network density against the negative emotion of the information, we



(a) Journalists Network in 2019.

(b) Journalists Network in 2022.

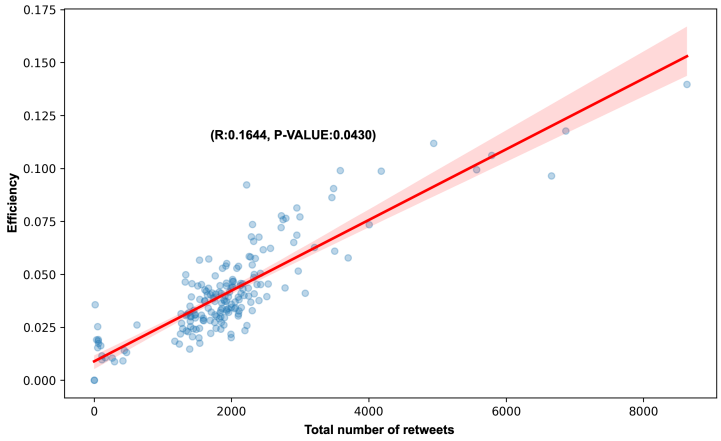
(c) Disinformation Network in 2019.

(d) Disinformation Network in 2022.

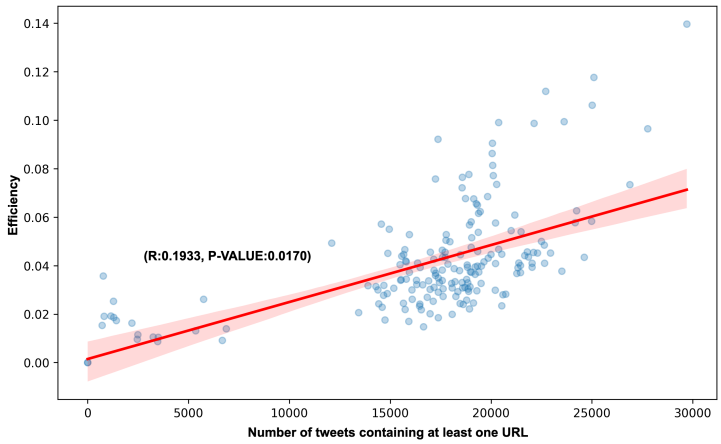
Figure 7.6: Retweet graphs of the Journalists and Disinformation networks captured in 2019 and 2022.

observed periods of increased network density coincided with more negative sentiment. This may imply that these possible campaigns or coordinated actions are deeply emotional, for example, by using more aggressive or strong vocabulary. It is interesting to observe that, among the five variables analyzed with respect to density, this dimension achieved the highest R value, indicating a stronger relation between those variables.

The correlations we discovered were overall weak (except, to some extent, with respect to the tweet sentiment), making it difficult to conclusively establish a cause-effect relationship between the network's shape and the nature of its content. However, these relationships offer intriguing insights, such as those already discussed. Nonetheless, we noted an improvement in efficiency when the total number of retweets within the network was higher and when a more significant proportion of published tweets contained URLs (see Figure 7.7). This suggests that the disinformation network becomes more efficient when it absorbs and disseminates information, demonstrating the remarkable capacity of these networks to facilitate information flow.



(a) Efficiency VS total edge weight (total number of retweets between users in the network).



(b) Efficiency VS number of tweets containing URLs.

Figure 7.7: Scatter plots considering efficiency in the disinformation network.

7.5 Analysis of Results

In the preceding sections, we have examined the defining characteristics of disinformation networks on X and outlined their potential operation strategies within the Spanish communication space. In this section, we will interpret these findings and their implications for understanding and mitigating the impact of disinformation. Furthermore, we will acknowledge the limitations of our study and outline prospective directions for future research in this domain.

7.5.1 Implications for understanding disinformation networks

Firstly, the actors involved in disseminating disinformation are markedly more active and work complexly within their networks. This heightened level of activity, combined with a strong interconnectedness, allows them to amplify their visibility and draw more attention to their narratives.

Moreover, disinformation actors can coordinate their efforts, particularly during specific campaigns or around contentious topics. Our study showcases such coordination in the case of NATO-related narratives. This phenomenon aligns with substantial research and spotlights nation-states as key disinformation actors within online networks, by using, for example, a network of both human-operated and automated accounts (known as 'bots') to disseminate misleading narratives, amplify divisive content,

and create a false impression of grassroots support (or opposition) to specific issues (a tactic known as 'astroturfing') [Ste19]. In addition to these overarching strategies such countries employ, there are also standard tools and techniques used in these disinformation campaigns; conspiracy theories, health misinformation, and the propagation of extreme political narratives are among the most frequently observed [KB18].

A noteworthy pattern we observed is the surge in the activity of disinformation networks during times of social crisis. During these periods, they actively disseminate URLs, especially those linked to disinformation media outlets and sources known to spread fake news. By exploiting social vulnerabilities and heightened emotions during crises, they amplify their influence, reaching a broader audience.

Therefore, given their higher density, increased levels of activity, and unique network structure, disinformation networks on X possess a significant potential to captivate users. Once these users fall into the network, they are more likely to be exposed to and receive false or biased information and propaganda faster than legitimate information. Hence, for those users who already belong (probably inadvertently) to such disinformation networks, the spread and impact of misleading narratives would be exacerbated, in particular, when compared to users belonging to journalists or other neutral actors within the social network.

7.5.2 Strategies for mitigating the impact of disinformation

Our research underscores that disinformation networks distinguish themselves through their notably higher density, specialized structure, and proficient communication flows. Therefore, to effectively mitigate the spread of disinformation, we must devise strategies that target these unique characteristics.

One of the key strategies involves disrupting the complex web of connections within disinformation networks. These networks function effectively because of the interconnectedness of their actors, who continuously reinforce each other's messages through retweets, participation in hashtags linked to disinformation campaigns, or the widespread sharing of fraudulent news. Interrupting this reinforcement chain would undermine the network's efficiency and reach, limiting its impact. Advanced algorithms can be developed to identify and shut down inauthentic accounts, thereby disrupting these networks at their core.

Simultaneously, it is of high importance to analyze network activity patterns to understand and identify coordinated inauthentic behavior. This analysis will provide a foundation for targeted interventions and astroturfing detection mechanisms. Policy recommendations for social media platforms can be developed to enforce stricter measures against coordinated inauthentic behaviors, enhancing the overall integrity of the information ecosystem.

Reinforcing the links within networks disseminating legitimate information is equally important. This strategy increases the spread and visibility of accurate information and provides a counter-narrative to disinformation. Moreover, strengthening these networks would equip users to resist the influence of disinformation networks and help create a more balanced information ecosystem on social media platforms, even though people may struggle to change their beliefs even after finding out that the presented information is incorrect or misleading [GNL13]. Public awareness campaigns and media literacy programs are essential in educating users to recognize and respond to disinformation. Collaboration with independent fact-checkers will aid in quickly debunking false narratives, reducing their spread and impact.

The urgency of these actions is especially pronounced during periods of social unrest, electoral contexts, or events with the potential to disrupt public security significantly. During those volatile times, disinformation networks are often the most active and have the highest potential to cause harm.

In this endeavor, we recognize the pivotal role that recommendation systems play [RRS22]. These systems, which are responsible for content distribution on platforms like X [GGL+13], can be leveraged strategically to minimize the visibility of disinformation. By deprioritizing content from disinformation accounts – especially within the disinformation networks themselves – these systems could weaken the disinformation networks' structure and efficiency. Our results on the impact of disinformation in recommendation algorithms shall be presented in Chapter 9, whereas our proposal to mitigate and alleviate these effects shall be introduced in Chapter 11.

However, it is essential to recognize that content does not exist in isolation – it circulates within networks. Strategies to combat disinformation should focus not just on individual users, but also on the broader network dynamics. Understanding and altering information flow dynamics at the network level enables more effective impediment of disinformation spread and boosts the spread of accurate, reliable information. Furthermore, international cooperation against cross-border disinformation and investment in research on behavioral patterns of disinformation spread will provide a holistic approach to combating this global issue. By incorporating these multifaceted strategies, we aim to create a more resilient and informed digital community, equipped to resist the influence of disinformation networks and foster a balanced information ecosystem on social media platforms.

Finally, in Table 7.2 we summarize the roles and responsibilites that different actors may have in preventing disinformation, by implementing the strategies discussed before. Building on the insights gained from our analysis, it is evident that countering disinformation requires a coordinated effort across multiple actors. Each stakeholder in the information ecosystem plays a unique and complementary role in addressing the structural and behavioral dynamics of disinformation networks. This requires various actions, from technical interventions to policy-making, public education, and global cooperation. Below, we provide an expanded discussion of these roles, categorized by the key stakeholders, with their responsibilities and actions as summarized in Table 7.2.

Social Media Platforms have a critical role as gatekeepers of the digital public sphere. Their ability to identify and remove inauthentic accounts can directly disrupt the interconnected structures that disinformation networks rely on for amplification. Platforms must also refine their algorithms to detect coordinated inauthentic behavior more effectively, leveraging advancements in artificial intelligence and network analysis. Furthermore, these platforms can adjust recommendation systems to prioritize legitimate content and minimize the visibility of disinformation. Reinforcing legitimate news sources through enhanced visibility ensures that users are more likely to encounter accurate information, countering the reach of false narratives.

Governments and Institutions must address disinformation's regulatory and educational dimensions. Promoting media literacy campaigns equips citizens with critical thinking skills necessary to recognize and resist disinformation. Simultaneously, governments can enforce policies targeting coordinated inauthentic behavior and collaborate internationally to tackle cross-border disinformation campaigns. Supporting research initiatives to understand disinformation's spread and impact further enhances institutional preparedness.

Users, as the final recipients of information, play a key role in resisting disinformation. By engaging

Table 7.2: Roles and responsibilities of various actors in preventing disinformation.

Actor	Possible Actions to Prevent Disinformation
Social Media Platforms	- Deleting fake accounts - Breaking disinformation networks - Reinforcing legitimate news sources - Implementing advanced algorithms for detecting inauthentic behavior - Adjusting recommendation systems to deprioritize disinformation
Governments and Institutions	- Promoting media literacy campaigns - Developing and enforcing policies against coordinated inauthentic behavior - Collaborating internationally to tackle cross-border disinformation - Funding research on disinformation spread and its impact
Users	- Engaging in media literacy education - Learning to recognize and respond to disinformation - Using critical thinking to assess the credibility of information - Reporting suspicious or misleading content to platforms
Independent Fact-Checkers and NGOs	- Identifying and debunking disinformation quickly - Collaborating with social media platforms to highlight accurate information - Educating the public about identifying fake news
Researchers and Academics	- Conducting behavioral studies on disinformation spread - Developing new tools and methods to detect and analyze disinformation networks - Collaborating with platforms and governments to provide insights and recommendations
International Bodies and Coalitions	- Facilitating cross-border cooperation in combating disinformation - Establishing global standards and protocols for information integrity - Coordinating efforts among member states to address disinformation challenges

in media literacy education and employing critical thinking, users can better evaluate the credibility of information and avoid amplifying false narratives. Reporting suspicious or misleading content to platforms contributes to disrupting disinformation at its source and helps platforms take timely corrective action.

Independent Fact-Checkers and NGOs operate as essential partners in debunking disinformation and ensuring the availability of accurate information. Their role includes identifying false claims quickly, collaborating with social media platforms to promote legitimate content, and educating the public on recognizing disinformation. By building trust with audiences, these actors contribute to strengthening the overall resilience of the information ecosystem.

Researchers and Academics bring a scientific lens to understanding disinformation networks and their dynamics. Conducting behavioral studies, developing advanced detection tools, and providing recommendations to platforms and governments are central to their contributions. Their work ensures that evidence-based interventions are continuously improved to counter emerging disinformation tactics.

International Bodies and Coalitions play a pivotal role in fostering cross-border cooperation, given the global nature of disinformation. By establishing international standards and coordinating among member states, these bodies can address the challenges of disinformation campaigns that transcend national boundaries. Their efforts provide a unified framework for combating disinformation on a global scale.

7.6 Discussion

While our research provides illuminating insights into the behavior of disinformation networks, it is essential to acknowledge certain limitations within our study. Firstly, the scope of our research was primarily concentrated within the Spanish communication space in Spain. As such, the samples studied, albeit systematically and rigorously collected, are specific to this geographic and cultural context. Extending this research to include other linguistic and cultural contexts could provide a more comprehensive understanding of the global dynamics of disinformation.

Additionally, while X remains a widely used platform for information dissemination, it is only one of many social media platforms, each with its unique dynamics. Hence, the behaviors and patterns observed on X may not completely represent disinformation strategies across all platforms. Furthermore, the rapidly evolving social media landscape at the time of writing this article may introduce new dynamics that could either exacerbate or mitigate the disinformation strategies we have studied.

Our study spans more than three years, a substantial period that provides consistent and relevant results. However, it is essential to consider that some of the accounts studied may have been active before the start of our research period. This prior activity could have influenced the network structure and dynamics we observed, but was not accounted for in our analysis.

Finally, when studying coordination within disinformation networks, it is essential to understand that such coordination can arise spontaneously due to shared interests or ideologies among individuals consuming and producing content. However, there may also be more calculated and organized strategies being managed on other platforms or in offline environments. Distinguishing between these two types of coordination can be challenging, and our study may need to be extended in the future to capture this complexity fully.

Despite these limitations, our research offers valuable insights into disinformation networks' behavior and strategies, contributing to the broader understanding of how false information spreads and how it can be mitigated. Future research should address these limitations, broadening the scope and deepening the understanding of the dynamics at play.

Future research avenues should investigate disinformation networks across various linguistic and cultural communities. English, Russian, Arabic, and Chinese represent significant sectors of the global internet user base, each with its cultural nuances and potential variations in disinformation dynamics. A comparative analysis across such diverse linguistic and cultural backgrounds would undoubtedly enrich our understanding of the global patterns of disinformation and its impact.

Further research could explore deeper into the types of media shared within these networks by developing a more refined taxonomy. This approach could yield insights into political biases and the most successful disinformation narratives, and identify political groups exhibiting higher levels of coordination and efficiency. Understanding the types of narratives that gain traction within these networks could inform more targeted and effective countermeasures.

The replication of our study on other social media platforms, such as micro-blogging or general-purpose networks, is another vital avenue to explore. Given the varying dynamics across different platforms, it would be invaluable to ascertain whether the patterns we observed on X are consistent across other platforms or if each platform presents unique challenges and opportunities in combating disinformation.

While our study focused on journalists as primary disseminators of legitimate information, future

research could incorporate other influential user categories. These could include politicians, influencers, or cyber activists, whose roles in the information dissemination process could significantly impact the spread of (dis)information and the efficacy of countermeasures.

The most critical focus for future research, however, should be developing and implementing strategies designed to disrupt disinformation networks and enhance the efficiency of legitimate information dissemination. Such strategies could range from redesigning recommendation systems to implementing more sophisticated communication campaigns that target specific areas of these harmful networks.

Developing these strategies necessitates an understanding that tackling disinformation is not merely about fact-checking or debunking individual false narratives. Instead, it requires a strategic shift in the information flows within and between these networks. This includes re-engineering algorithms that govern these flows, changing the incentives for sharing information, and creating an environment that fosters critical information consumption among users.

In essence, the battle against disinformation is a contest over the control and direction of information flows. As such, dismantling disinformation networks involves disrupting the existing harmful flows and proactively shaping beneficial ones. By focusing on these two dimensions, we can disrupt these networks and mitigate their impact. This represents a challenging but essential task for researchers, policymakers, and practitioners committed to preserving the integrity of our information ecosystems.

7.7 Conclusions

This research inquires into the structure and behavior of X accounts associated with legitimate journalists and disinformation actors. Our data set spans from 2019 to mid-2022, encompassing various accounts and their activities. The focus of our study is to illuminate how these diverse actors form networks within the X platform and engage in distinct dynamics of content production and sharing. Our findings reveal that disinformation actors form considerably denser networks than journalists create, which underscores clear signs of coordination within their information-sharing dynamics. This characteristic is critical as it indicates a calculated, collective approach to disseminating disinformation, contributing to its pervasive nature on the platform.

Moreover, our analysis utilizes network metrics such as efficiency, which measures the speed at which information propagates within a network. We observe that information within disinformation networks flows considerably faster than networks formed by legitimate journalists, which exhibit a higher degree of fragmentation. This faster propagation of information allows for the rapid and widespread distribution of disinformation, often outpacing the dissemination of corrective or countering information from legitimate sources. This contrast between the behavior of disinformation networks and legitimate information sources offers insights into the challenges of countering disinformation on Twitter/X. The orchestrated network structure and efficient information dissemination within disinformation networks pose significant obstacles to mitigating the impact of misinformation on the platform. By shedding light on these dynamics, our study contributes valuable insights to the ongoing discourse on tackling the disinformation crisis in the digital age.

In summary, disinformation networks demonstrate a unique capacity for adaptability, elevating their density levels during periods of heightened social controversy, such as the war in Ukraine or debates concerning NATO. This heightened activity often correlates with a more negative sentiment within the network, hinting at possible coordinated actions. Though the correlation is not definitive and further

investigation is required, this trend aligns with the discourse typically driven by nation-states around such topics.

This superior efficiency of disinformation networks in communication flow and their adaptive nature underscores the challenges in combating disinformation. Nevertheless, it also points to potential avenues for intervention. For instance, strategies that fracture the efficiency and density of these disinformation networks could be particularly impactful. As such, future research should explore into network activation and coordination mechanisms and expand to include other national and cultural contexts. Another potential intervention in these scenarios may include evidencing reasons or contexts behind specific tweets, as presented recently in [LHW+22], probably not to everyone in the network but depending on the characteristic of the information being sent (number of hashtags or URLs) or based on the sender/receiver.

Considering our findings, recommendation systems on platforms like X could be valuable targets for future intervention research. The potential to influence these systems to disrupt the efficiency of disinformation networks while simultaneously enhancing the efficiency of legitimate networks may be a critical component in the fight against disinformation. In the digital age, such strategic interventions are more critical than ever for preserving the integrity of our information ecosystems.

Chapter 8

A Novel Polarization Metric for Online Political Social Networks

8.1 Introduction

Political polarization is a significant phenomenon in today's interconnected digital age. Defined by deep-rooted ideological divides and emotionally charged beliefs, it finds fertile ground in platforms like X. These platforms, serving as dominant channels for political discourse, reflect societal sentiments and magnify and distort them in many instances [GAS+15]. Challenges like the rapid spread of misinformation, the creation of echo chambers, and algorithm-driven content recommendations further compound the problem [CLH+21]. However, it is essential to recognize that these platforms have a dual nature, given their vast user base and real-time data processing capabilities. While they can amplify divides, they also have the unparalleled potential to bridge gaps, encourage diverse dialogues, and shape global political sentiment [Tuf17].

Electoral campaigns offer an ideal setting to analyze political polarization on social media. During these periods, parties and candidates ramp up their outreach to influence and galvanize followers. The surge in news, commentary, and targeted messaging, combined with the platforms' dynamics, magnifies existing polarization, making campaigns a focal point for studying such divides.

The current state-of-the-art algorithms employed to gauge polarization tend to hinge on three primary approaches: exploring network topology, content analysis of posts, or applying hybrid methods combining both. These techniques, equipped with their strengths, capture distinct facets of the polarization phenomenon. For instance, studying network topology can reveal clusters or echo chambers, while content analysis might shed light on the nature and intensity of polarized rhetoric. However, these algorithms often fail to provide a comprehensive view despite their varying sophistication. One of their main challenges is the need for more adaptation to the particular political landscapes in which they are applied. A one-size-fits-all approach is less effective, especially in multi-party systems, where political dynamics can be more complex and varied. This emphasizes the necessity for developing algorithms and methodologies that are attuned to specific political contexts and can navigate the complexities inherent in different political systems.

In light of these observations, the research in this chapter undertakes several critical tasks. Firstly, we analyze the state of the art, focusing explicitly on the evolving definitions and nuances of political

polarization within the unique ecosystem of online social networks. Building upon this foundation, we shift our lens to a detailed exploration and comparison of existing algorithms, delving into their methodologies, strengths, and weaknesses. Such a comparative study, while highlighting the current landscape, also uncovers gaps and opportunities for innovation. To this end, our research culminates in the proposal of *SPIN* (Social-political Polarization analysis by Information Theory), a novel algorithm rooted in the principles of information theory. This algorithm aims not just to measure but to shed light on the underlying complexities of political polarization in the digital realm, offering scholars and practitioners a new tool to understand and navigate this pressing issue.

At the outset of this chapter, we delve into the concept of political polarization in the context of online social networks, particularly during electoral campaigns. This chapter examines the structural and informational factors contributing to polarization, highlighting the role of social media dynamics, such as echo chambers and algorithmic biases, in amplifying divides. Through an analysis of existing approaches to measuring polarization, this chapter identifies the limitations of current methodologies and introduces SPIN, a novel algorithm that integrates information theory with network analysis. This work provides critical insights into the evolution of polarization dynamics and proposes a computational tool designed to capture and analyze these patterns effectively.

The chapter is organized as follows: Section 8.2 reviews the state-of-the-art approaches for measuring polarization, detailing their methodologies, strengths, and limitations. Section 8.3 presents the specific methodology followed for the approach introduced here. Section 8.4 outlines the SPIN algorithm, describing its theoretical foundation in information theory and its practical implementation for measuring polarization. Section 8.5 presents the results of benchmarking SPIN against existing algorithms using datasets from Spanish electoral campaigns, showcasing its ability to provide nuanced insights into polarization dynamics. Section 8.6 includes a discussion of the obtained results. Finally, Section 8.7 concludes by summarizing the chapter's contributions and suggesting directions for future research on polarization in digital ecosystems.

Research questions

Our aims in this chapter can be summarized in the following research questions included in the second research goal of the thesis:

- **RG2:** Conduct an in-depth analysis of the main risk phenomena in political information ecosystems. Develop a set of tools for their computational analysis.
  - **RG2.RQ3:** How can polarization in a social network in a political context be measured accurately?
  - **RG2.RQ4:** What are the main factors contributing to the emergence of polarization?

8.2 Measuring polarization

8.2.1 Background

Online social networks have become pivotal for political participation and information consumption, transforming platforms like X, Facebook, and Instagram into digital agoras for global civic discourse

[Bou20, AS22]. These platforms democratize information dissemination, enabling users to bypass traditional media gatekeepers and contribute directly to societal conversations. However, this accessibility often amplifies challenges such as misinformation and the formation of echo chambers, driven by personalized algorithms that prioritize emotionally charged content [Sun17, KvS21]. As these networks increasingly shape public opinion, their influence on political dynamics and polarization cannot be overstated [TJLL18, CLH+21].

Political polarization refers to the growing ideological, affective, and behavioral divides between groups, often manifesting as antagonism towards opposing political identities and homophily within like-minded communities [Wag21, BBF21]. Social media intensifies this phenomenon by creating environments that reinforce pre-existing beliefs, fostering ideological silos and heightening intergroup tensions [VBP21, ISL12]. While platforms like X mirror these societal divisions, they also hold potential for mitigating polarization by promoting diverse interactions and curating content to encourage constructive dialogue [SLL21, MC21]. Understanding these dynamics is essential for developing tools to measure and reduce polarization in the digital age.

8.2.2 Approaches to measure polarization

Measuring polarization, also called controversy, is a complex task, primarily because the term can be interpreted in various ways. As we have seen, depending on context and perspective, definitions of polarization may differ, leading to inconsistencies in data and analysis. Accurately quantifying and comparing polarization across different mediums or regions with a universally accepted metric or definition becomes easier. Nevertheless, most existing measuring algorithms agree on a set of basic terms.

Overall, according to the state of the art, polarization on social media is prominently evidenced by the emergence of echo chambers: insular user groups that predominantly revolve around a unified political or ideological perspective. These chambers continuously reinforce similar viewpoints while dissenting opinions are minimized or outright excluded. This homogeneity in thought limits exposure to diverse perspectives and often fosters an environment where members develop adversarial attitudes towards outside communities. The inherent design of many social media platforms, prioritizing engagement and like-minded interactions, inadvertently fuels this phenomenon, leading to further divisions and intensifying feelings of “us versus them”.

However, most of these algorithms capture the phenomenon only partially. Also, they focus on general domain polarization, instead of exploring this phenomenon on the political context.

Indeed, while polarization is often associated with a two-party system, where divides appear distinctly binary, it is not exclusive to such structures. Despite having a broader array of political stances and entities, multi-party systems can also exhibit pronounced polarization. Fragmentation can occur among various parties or ideological groups in these systems, leading to multiple echo chambers. Each group can become insular, intensifying its internal consensus while growing increasingly distant from or antagonistic toward other groups. Hence, polarization is not inherently bipartite or bipolar; it can manifest differently across different political landscapes, underscoring its complexity and the need for nuanced approaches in its study and mitigation.

In the subsequent sections, we present a concise overview of the state-of-the-art algorithms to measure polarization. These topology-based, content-based, and hybrid algorithms offer insights into various facets of the complex landscape of digital polarization. After discussing their dynamics, strengths, and

potential limitations, we introduce our method based on Information Theory. Drawing from the lessons of existing tools and addressing their gaps, this proposed approach aims to provide a more comprehensive and nuanced understanding of polarization dynamics in digital spaces.

Topology based algorithms

Topology-based algorithms hinge primarily on the network structure to discern polarization patterns. By analyzing user connections and interactions, these algorithms map out the network's layout to identify clusters, bridges, or isolated nodes. Such clusters or echo chambers represent groups of like-minded individuals who frequently interact with one another, often reinforcing shared beliefs.

In these topology-based networks such as X, individual users are denoted as nodes, with their interactions, like retweets, likes, replies, and quotes, symbolizing the links or edges that connect them. For instance, if User A retweets User B, this establishes a directional link from A to B. Similarly, a 'like' or 'reply' would form another type of connection. As more interactions accumulate, a more complex web of connections emerges, vividly illustrating the flow of information and the nature of interactions among users. Over time, distinct clusters or communities become apparent, often signifying groups with shared beliefs or interests. Topology-based algorithms can infer the degree and patterns of polarization within the network by analyzing the pattern and density of these links. The granularity of these interactions provides a detailed map of the digital landscape, indicating areas of consensus, contention, and isolation.

Indeed, the topology's structure is insightful, working on the assumption that highly polarized networks exhibit fewer interconnections between differing groups and denser internal connections within like-minded clusters. This pattern reflects the echo chamber effect, where users mainly interact with those sharing similar beliefs, creating clear divisions within the more extensive network.

The main topology-based algorithms are listed below:

- **Boundary Connectivity (GMCK/BC[1])** [GJCK13] [GMGM18]: originally proposed in [GJCK13] a controversy score initially called *P* is defined. Essentially, in such research (and then in [GMGM18]) the authors discuss and propose a network controversy index based on the concepts of "boundary nodes" and "internal nodes". A boundary node is essentially a node connected to at least one other node from its partition and at least one node from the opposite partition. Thus, when calculating the set of boundary nodes, the computation is performed for each pair of partitions in each given direction (not admitting more than two communities for this calculation), as the boundary of partition X with regard to partition Y, denoted as B_{xy}, is different from the boundary of partition Y with regard to partition X, B_{yx}. However, in the proposed algorithm's case, the polarization index only admits two network partitions. The index's intuition is as follows: if the network is polarized, the boundary nodes should have more connections to internal nodes than to nodes from the other network. Following this intuition, the authors propose a controversy index ranging in the interval $[-0.5, 0.5]$. The lower boundary indicates the absence of polarization, while the upper boundary indicates maximum polarization in the considered network. Moreover, GMCK (also known as BC) is a polarization index that was originally devised in [GJCK13]. In that research, the authors designed a polarization detection algorithm based on community boundaries that could not be computed when the boundaries were empty. Such an algorithm, named P score, is virtually identical to the one proposed in [GMGM18] (i.e., GMCK).

[1] In literature, one can find this algorithm named in various ways, primarily as GMCK and BC.

- **Dipole Moment (MBLB)** [GMGM18][MBLB15]: originally proposed in [MBLB15], the *Dipole Moment* is a controversy measure inspired by concepts purely related to physics. Specifically, it is based on the idea that if two communities are polarized against each other (measured with a node-associated label indicating its "ideology", then most labels should be closer to the extremes than the center, as both political communities repel each other. Again, this polarization index can take values in the range $[0, 1]$, where 0 indicates no polarization and 1 indicates a maximum of polarization.

- **Random Walk Controversy (RWC)** [GMGM18]: One of the most fundamental algorithms found in the literature for detecting and quantifying polarization is *RWC*. This algorithm starts with a network, represented as a graph where nodes are users and edges are interactions (these could be retweets, mentions, replies, etc., or even combinations of the above). This network must also be partitioned into exactly two distinct partitions. The basic essence of the algorithm is to perform "random walks" starting from a specific node until reaching a limit (this could be reaching an influential node in the network, a maximum number of steps, etc.). Based on the starting community of the walks (community of the nodes where the N walks begin) and their ending communities, conditional probabilities are calculated, for example, the probability of having started in partition X, if it ended in X, denoted as P_{xx}. Using these probabilities, the authors define a formula to compute a polarization index of the network that takes values in the interval $[0, 1]$, where 0 represents no polarization and 1 represents maximum polarization.

- **Betweenness Centrality Controversy (BCC)** [GMGM18]: Another simple polarization detection and quantification algorithm from the literature is directly related to and inspired by the *Betweenness Centrality* concept referred to the network edges. The algorithm's intuition is relatively straightforward: if the network is polarized, the edges connecting nodes from one partition to another should have a very high betweenness centrality (since many of the shortest paths from a node in partition X to another node in partition Y will necessarily pass through them), contrasting with the low betweenness centrality of edges connecting nodes from the same partition. Thus, the authors propose an easy-to-calculate polarization index, which is intuitively sound and entirely based on the network structure.

- **Embedding Controversy (EC)** [GMGM18]: Similarly, we have another polarization index. It is entirely based on the *Force Atlas 2 (fa2)* algorithm [JVHB14]. This is rooted in the idea that this algorithm maximizes modularity, and polarized networks are characterized by high modularity. This algorithm calculates the embedding of each node (bi-dimensional representation) in the network, to compute distances between nodes of the same partition and between different partitions (a maximum of 2 partitions, like the previous algorithms). Thus, this polarization index can detect and quantify network polarization using values in the interval $[0, 1]$, where 0 implies no polarization and 1 implies maximum polarization.

- **Authoritative Random Walk Controversy (ARWC)** [VPV21]: There are other detection and quantification algorithms for polarization that follow the same concept of *RWC*, but introduce certain complexities to allow the algorithms to better capture the polarization of the network. This is the case with *ARWC*, a variant of the original *RWC*, with the primary distinction being that the walk concludes once it reaches an influential node within the network. Node relevance is determined

based on the degree of each node. According to the authors, each walk should conclude when it arrives at a node that is within the top 15% of nodes by degree, in either partition. Probabilities are similarly computed, and a polarization index is returned. Interestingly, the only distinction between *ARWC* and the original *RWC* is the one described above, as the mathematical formula defining the polarization index remains identical in both instances. Once again, the polarization index values lie within the interval $[0, 1]$.

- **Displacement Random Walk Controversy (DRWC)** [VPV21]: Alongside *ARWC*, the authors also propose another network controversy measure (polarization index) with further deviations from the original *RWC*. Specifically, *DRWC* is a variant where a fixed length is considered for all random walks, and instead of considering the end community of the walk, it takes into account the number of community changes that have occurred during the walk. The intuition behind the algorithm is straightforward: if there are few community changes during the walk, it suggests a high level of controversy, keeping the polarization index close to 1. However, if many community changes occur during the random walks, it indicates low polarization between the communities (since they are well interconnected). Once again, the resulting value from the algorithm lies in the interval $[0, 1]$.

- **ERIS** [GGLC22]: Similar to already existing algorithms, such as *GMCK/BC*, the authors propose an algorithm, called ERIS, to carry out polarization detection on a network using the concept of boundary nodes. This algorithm allows to reliably capture the polarization of a network by pairs of communities. Essentially, the algorithm accepts having a network partitioned in more than two communities, and for each pair of communities (in each direction) porosity and antagonism values are computed. On the one hand, porosity refers to the influence that a community can have over another, on the other hand, antagonism refers to how opposite a community looks from the point of view of a different community. Thus, this algorithm allows an in-depth analysis of the polarization of a network, based on the analysis of the antagonism and porosity matrices.

As we see, while topology-based algorithms provide insight into the structural patterns of networks, they overlook the content of posts, a vital component in assessing polarization. The very essence of a post, especially when laden with negative sentiments or direct opposition to contrasting views, amplifies polarization. It is not just about how clustered or isolated networks are, but also about the intensity of sentiments within those clusters. Real polarization is underscored when isolated networks echo shared beliefs and hostility towards differing perspectives.

Content-based algorithms

Content-based algorithms focus on analyzing the text within posts to gauge polarization. By examining language use, sentiment, and thematic content, these algorithms can discern the tone, intensity, and nature of the discussions, allowing for a deeper understanding of underlying beliefs and attitudes. Such algorithms can detect patterns of extreme views, recurring divisive topics, and the frequency of negative or adversarial language, offering a nuanced picture of polarization beyond mere network structures.

The main content-based algorithms are listed below:

- **Purely NLP-based algorithms for polarization detection**: besides the aforementioned algorithm, in the literature, it is possible to find many polarization detection algorithms purely based on NLP, such as [DEB96, CLDB18, Mat17, WMZK18]. All these algorithms apply NLP concepts to extract information from the content of the posts (they could be tweets, for instance) with which to obtain a polarization index. However, the index does not really consider interesting aspects such as the topology of the network or the temporality of the posts. In fact, these algorithms do not take into consideration the interaction between users (as that depends on the network topology) during the calculation, thus it is acceptable to say that they are limited.

- In [YWLD17] the authors propose a content-based measurement of the polarization of a network called *Normalized Cut (NC)*. Such a measurement proved to work significantly well when considering, in a separate manner, the tweets of the network by their sentiment. Thus, when considering separately positive, neutral, and negative tweets, the polarization measure proved to reliably capture the polarization of the network, in the use cases where the authors applied it.

- In [KM22], the authors propose a Deep Learning based approach to carry out ideology detection and polarization detection using the sentiment analysis from tweets, in the context of the Covid-19 pandemic.

Indeed, while content-based algorithms excel in deciphering textual nuances indicating polarization, their neglect of the network's structure is a limitation. Without considering how users connect and form clusters, these algorithms may overlook inter-community polarization dynamics. They primarily offer insights into the overarching sentiment or the general polarization levels related to specific topics, rather than the interactions and divisions between distinct user groups or communities. As a result, they may miss the subtleties of how polarization manifests and propagates across different segments of the network.

Moreover, many of the studied content-based algorithms lean heavily on Machine Learning (ML) and Deep Learning techniques for classifying and understanding labelled posts. While these methods can offer high accuracy, they come with inherent challenges. The dependency on labelled data means that as the digital landscape and discourse evolve, there's a consistent need to train and retrain these models to maintain their accuracy. Moreover, Deep Learning models, especially, often operate as "black boxes," making it challenging to discern how they arrive at specific classifications. This lack of transparency, known as the explain-ability problem, can hinder the broader acceptance and trust in these algorithms, especially in contexts where understanding the reasoning behind classifications is essential [BSKNS22].

Precisely, the inherent complexities of these ML and Deep Learning-based algorithms, combined with their need for continuous retraining and their limited explain-ability, have made them more challenging to deploy effectively. These hurdles can impede widespread adoption, especially in contexts where stakeholders value transparency, understand-ability, and adaptability in the tools they utilize.

Hybrid algorithms

Hybrid, or mixed algorithms, meld the strengths of both approaches: they incorporate the structural insights derived from network topology with the nuanced content analysis of posts. By doing so, they aim to provide a more holistic view of polarization, capturing both the overarching patterns of connectivity and the underlying sentiments and discourses prevalent within the network. This integration allows for

a more comprehensive understanding of the multi-dimensional facets of polarization in digital spaces. Although scarce, some of the hybrid algorithms that we can find in the literature are described next:

- **Biased Random Walk (BRW)** [ENT+20]: In the literature, it is possible to find many variations of the original RWC polarization measurement. One of them is BRW, which offers a hybrid approach to detect and quantify the polarization of a network. The main idea behind this algorithm is to give an "initial energy" for the random walk, assigning a "loss energy" to each node in the network. Thus, when a walk is performed from a specific origin node, the initial energy is given to the node. Then, the initial energy is decreased as the random walk progresses, due to the loss of energy applied at each movement from node to node. These energies can be computed in different ways, including the use of the content of the network for their calculation, hence leading to the creation of a hybrid polarization detection algorithm.

- **Diffpool based approach for polarization detection** [BAB+21]: As opposed to other polarization detection algorithms that rely on Deep Learning, Diffpool is an algorithm that combines both the structure of the network and its content, hence being a hybrid algorithm, to measure the polarization of the network. Using Diffpool-based pooling layers, the input network is then processed into a more coarsened network (in one or more steps), finally obtaining a measurement of its polarization. However, this algorithm does have the same weakness as the aforementioned Deep Learning algorithms, as it requires intensive training and the resulting model is often a "black box" with reduced explainability on the final outcome (the polarization of the network), thus limiting deeper analysis to be performed with the algorithm.

- **Multi-Opinion based method for controversy detection** [SIK22]: Given that many algorithms (polarization indices) are constrained by the number of communities they accommodate (many of them only allow two), some seek to overcome this limitation. Such is the case with the polarization detection and quantification framework proposed by the authors in this research. Specifically, they build on the idea of creating, on one hand, a network partitioning (which could be achieved using algorithms like *METIS* [KK97] or Louvain [BGLL08]) and, on the other hand, another semantic-level partitioning (for instance, this could represent user ideologies in a political context). This algorithm then weights these partitionings, recalculating the edge weights (an initial weighted graph is required) to produce a polarization index for the network within the range $[0, 1]$, where 0 indicates no polarization, and 1 denotes maximum polarization.

- **Generalized Euclidean (GE)** [HDC23]: another hybrid algorithm used for polarization detection and quantification is based on the usage of a measure derived from Euclidean distance called *Generalized Euclidean Distance*. More specifically, this algorithm offers a polarization measurement based on the distance between nodes in different partitions. As a result, the measurement is maximized when the distance between nodes of different partitions is maximized, which intuitively leads to a polarized network.

However, while hybrid algorithms offer a more encompassing approach, they predominantly focus on the sentiment and directionality of posts. This means they primarily discern positive or negative sentiments and determine whether the content is leaning towards or against specific viewpoints or topics. Consequently, they might overlook other complex aspects like sarcasm, cultural nuances, or deeper

thematic contexts that can provide a richer understanding of polarization dynamics. This limitation underscores the challenge of balancing breadth with depth in algorithmic analysis.

Indeed, a notable limitation of current hybrid algorithms is their lack of consideration for temporal patterns. This means they might overlook the evolution of discussions, sentiments, and network structures over time. Understanding how and when polarization intensifies, ebbs, or shifts, especially in response to real-world events or online triggers, is essential. Without this temporal dimension, we miss out on the dynamics of polarization, potentially leading to static or outdated interpretations of the digital landscape.

The vast majority of controversy/polarization detection algorithms available to date rely on the use of lattice structures that store information related to a specific moment. As such, few current approaches allow for the detection of controversy using more than a static snapshot of a network at a given point in time. Furthermore, many of the proposed algorithms are limited to studying controversy between two groups or communities, precluding the possibility of considering more communities.

Similarly, these algorithms do not consider coherence in discourse. Polarization will be pronounced when each network maintains its own narrative, especially if it's negative towards the other network. Polarization won't be as pronounced when, despite the presence of well-defined communities, there exists a plurality or richness in the discourse within each one of them. Likewise, all of the algorithms studied do not clearly adapt to the political context. They operate generally on a graph, regardless of its theme.

While current methodologies offer insights into the phenomenon of polarization, they capture only a fragment of its complex reality. This particularly resonates with their omission of essential elements like the temporality of information flows, the internal consistency within communities, and how these communities intertwine with the broader network's structure. A glaring oversight is the lack of contextual awareness, especially recognizing that polarization is often magnified within specific political contexts. Our subsequent work aims to bridge these gaps. We introduce an innovative approach encompassing these overlooked dimensions, providing a more holistic view of polarization dynamics. To this end, we have employed the curated specialized dataset described in Table 4.1 and crafted a new algorithm, benchmarking its performance against existing methods. We eagerly present our findings and the specifics of our unique solution in the ensuing sections.

8.3 Specific methodology

8.3.1 Data set

A meaningful evaluation of our algorithm requires datasets rich in political discourse, allowing for an analysis of its ability to detect and measure polarization effectively. Electoral processes, with their distinct phases—pre-campaign, campaign, voting day, and post-campaign—provide ideal conditions for this. Each phase presents unique dynamics and discourse patterns, enhancing the robustness of our algorithm's evaluation within real-world political landscapes.

Notably, polarization typically intensifies through the campaign period, peaking on election day. The preceding "election silence" period, where campaign activity legally halts, provides a quieter moment for analysis. Using data from Spanish electoral processes on X (Twitter), we captured these phases to chronicle the dynamics of polarization.

The dataset was generated starting from the main accounts of the parties that secured representation in each general electoral process. We then applied snowball sampling with three levels of depth. This involved collecting all the posts from the recursively identified accounts during the 15 days leading up to the electoral date, on the day itself, and the subsequent 7 days.

Thus, as we did in Chapter 6, we used two primary datasets for this analysis, previously defined in the methodology chapter (Chapter 2), for tables "General Processes" (see Table 4.2) and "Local Processes" (Table 4.3).

8.3.2 Evaluation and comparative analysis of algorithms

In order to correctly assess any polarization measurement algorithm, we must start from a set of assumptions about political polarization throughout an electoral process. Based on these, we can determine if existing or proposed solutions are capable of accurately capturing these assumptions. Similarly, the proposed algorithm should meet basic conditions related to stability, explainability, and efficiency.

These patterns are based on the following conditions, matching the actual consensus around political polarization [HAR21, GOBG23, HKP17, OCLB19]:

- Polarization increases steadily from the pre-campaign until the day of reflection.
- On the day of reflection, it drops, reaching local minimum levels.
- On election day and the day after, it rises, reaching a global maximum.
- Subsequently, it progressively decreases after the election, reaching levels lower than during the campaign and pre-campaign.

Similarly, when defining an index, the following conditions should be taken into account:

- The index should range from 0 to 1, for full understanding and explain-ability.
- The higher the polarization, the closer the index value should be to 1.
- The polarization values generated by the index must follow a stable pattern according to the previously defined conditions, without presenting large daily oscillations or abnormally high peaks.
- The underlying logic of the process to generate the index must be easily explained and understood.
- The polarization index of the algorithm must be calculated in a reasonable time.

8.3.3 Information Theory based approach

Entropy is one of the foundational concepts in information theory and was proposed by Claude Shannon in 1948 [Sha48]. The intuition behind this concept is to measure the average information expected when observing the value of a variable based on its probability distribution. From the original concept of entropy, another known as cross-entropy is proposed, which measures the difference in probability distributions between two given users. This entropy can be calculated, especially in the context of social networks, based on the temporal moment. Thus, the idea is to see how much information is expected to be seen from one user, based on what another has generated (time-synchronized cross-entropy) [SSRM22].

By scaling this concept at the community level, we are capable, thanks to information theory, of modeling the flow of information between communities using the entropy concept, which is the basis of our algorithm: *Social-political Polarization analysis by INformation theory (SPIN)*.

This intuition over which we built our algorithm was devised in [SSRM22], where the authors proposed the application of entropy measurements to reliably quantify the influence between the communities that coexist in social networks. The basic intuition is that, if a "target community" re-uses part of the information that was generated by a "source community", then there was a clear influence of the "source community" on the "target community". In fact, this influence relationship is time-bounded: the "source community" was able to influence the "target community", because they generated the information before, and then it reached the target community, which further expanded a part of it. This intuition matches perfectly the aforementioned concept of *time-syncronized cross-entropy*. Furthermore, this intuition has already been used in research, for instance, in [SWB+22] where these concepts were used to analyze the influence of bots, activists, and disinformation users in online social networks, also in the context of political campaigns.

8.3.4 Methods for the estimation of information flow

In the literature, it is possible to find several entropy measurements to estimate the information flow between the users of a network. In [SSRM22], the authors show how using the entropy rate estimator (h), originally proposed in [KASW98] , some other entropy measurements could be derived. The entropy rate is defined next:

$$h = \frac{N \log_2 N}{\sum_{i=0}^{N} \Lambda_i} \tag{8.1}$$

where N refers to the length of the text whose entropy is being calculated, and Λ_i refers to the length of the longest non-contiguous subsequence from the target text that appears, previously (in the previous i symbols), as a contiguous subsequence in the text.

From that basic information flow estimator, the authors propose the time-synchronized cross-entropy metric [SSRM22], which leverages the usage of entropy to calculate the information flow between two users, taking into account the dynamics of the conversation, which can be considered useful in the context of social networks. Such a measurement is mathematically defined below:

$$h(T||S) = \frac{N_T \log_2 N_S}{\sum_{i=1}^{N_T} \Lambda_i(T|S_{\leq t(T_i)})} \tag{8.2}$$

In this expression:

- $h(T||S)$ represents the **cross-entropy of the target (T) with respect to the source (S)**, indicating how much uncertainty exists in T given the information from S.
- $T|S$ in the denominator refers to T **conditioned on** S, which reflects the relationship between the target and source texts, incorporating the temporal dynamics of their interaction.

Here, T refers to the target (e.g., a target user), S refers to the source (e.g., a source user), N denotes the length of the text (where the sub-index specifies whether it pertains to T or S), and Λ_i represents

the longest sub-sequence in the target text that appears contiguously in the source text. This metric also considers the timing of the posts if applied to information flow in social networks.

Although the time-synchronized cross-entropy is already a valid metric for measuring information flow, it was decided to measure the information flowing through a network using an entropy metric derived from it, called *Neighbor Normalized Information Flow (NNIF)*, also defined in [SSRM22]. The reason behind the decision is simple: based on the benchmark carried out by the authors in [SSRM22], NNIF is able to measure a network's information flow in a much more reliable way than the time-synchronized cross-entropy, and other entropy measurements derived from it. In fact, such an entropy estimator has already been used in the literature (as it was aforementioned) achieving good results [SWB$^+$22]. As a result, it was decided to use NNIF as the metric to estimate the information flow in a network, when calculating our polarization index, SPIN. Formally, the NNIF metric is defined next [SSRM22]:

$$NNIF = \frac{h(T||S)}{\sum_X h(T||X)} - \frac{h(S||T)}{\sum_X h(S||X)} \tag{8.3}$$

Where $h(\alpha||\beta)$ refers to the time-synchronized cross-entropy between a target α and a source β, T refers to the target node, S refers to the source node and X refers to any node in the local neighborhood.

8.4 The SPIN Algorithm

The proposed polarization index is based on the idea that we can borrow the fundamental concepts from Information Theory to measure the flow of information between communities of different characteristics (in the context of social network polarization in politics, we refer to communities based on ideological positioning, associated with clearly identified political organizations), primarily through the concept of entropy (and the metrics, defined in the present literature, to estimate it). Given that entropy indicates the flow of information between two users, we can define an algorithm utilizing it to detect polarization. Our proposal goes as follows:

1. **Calculation of intra-community entropy**: Considering those nodes (users) that are somehow related, we can calculate the flow of information between them, estimating the amount of information in the network that flows within each specific community.

2. **Calculation of inter-community entropy**: By considering the connections between nodes (users) from different communities, through the calculation of entropy, we can quantify the amount of information that flows between communities.

However, these concepts alone would not allow us to calculate a polarization index. For this reason, we also employ the concept of negative entropy, referring to the measurement of the amount of negative, as referred to hostile, information being transmitted through the network.

Since negative entropy allows the measurement of the flow of negative information generated in the network, the intuition behind the usage and calculation of negative entropies is simple: the more negative information transmitted through a network, the more polarized it should be; hence, measuring negative entropy becomes the cornerstone upon which *SPIN* operates.

Specifically, for calculating negative entropies, the *LIWC 2007* framework [PCI$^+$07] was used. This framework allows, through text analysis, to obtain the count of words from each specific category present

in a particular text, for instance, in a tweet. These categories correspond to time, money, food, exclusion, inclusion, family, people... Moreover, other categories of particular importance for calculating negative entropy are associated with sentiment analysis: number of words with positive emotion, negative emotion, number of words expressing sadness, anxiety, etc. Beyond all these categories, the framework also allows characterizing users' speech through other categories like the number of personal pronouns (in each personal form) and number of verbs (in different personal forms), allowing, for instance, to analyze if certain users consistently refer to supposed third parties.

Considering the above, it was decided to use this framework to count the words of each category, filtering those tweets with any negative sentiment (not necessarily having a value for the category $EmoNeg > 0$, as more complex conditions could be used involving other categories related to speech, or certain sensitive topics like money or work). Thus, by filtering for negative tweets, the negative entropies are calculated, which provide information on how much negative information is truly flowing through the network.

Thus, the essence of this algorithm is to consider that a network is polarized when polarization exists:

1. **Between different communities**: The flow of negative information between different communities is naturally a clear indicator of the existence of polarization in the network.

2. **Within each community**: Nevertheless, there is also the possibility that communities are highly polarized because negative information (or information against other communities) circulates within each of them, without this information necessarily flowing to other communities. To model this possibility, it is essential to also consider the flow of negative information within each community, and not just between communities.

Thus, as other existing algorithms do, the polarization index proposed in this document is based on a network structure representing the network on which polarization analysis (quantification) will be carried out. The network input to the proposed algorithm must be partitioned so that nodes with similar characteristics (ideologies, in the case of using polarization detection in social networks in politics) are located in the same partition and different partitions from nodes with different characteristics.

Additionally, since the algorithm employs the approach of information theory, it requires (in the case study of polarization in social networks) the content of the nodes' (users') posts, as well as the temporal moment they are published (timestamps) because with this information the entropy metric is calculated, which aims to estimate the flow of information in the network.

8.4.1 Network representation

In our proposed algorithm, SPIN, the representation of Online Social Networks, such as X, is done through a multi-directed graph whose nodes represent users and whose edges represent connections between users. Those edges must, in fact, be weighted, as the edge weight represents the frequency of the interactions between users. The intuition behind this representation is quite simple: two users (nodes) could have one or more connections with each other in any direction, and these connections are related to user interactions that could have occurred multiple, potentially many, times due to the natural interactions and information exchange that characterizes micro-blogging Online Social Networks, such

as X. However, as the representation involves a (multi)directed graph, it is fundamental to define both the possible interactions between users, as well as the direction of those interactions[2]:

- **Retweets**: The connection (edge) between the user (user A) that creates the retweet and the retweeted user (user B) must go from the retweeted user to the user that creates the retweet, i.e., from user B to user A. The intuition behind this is simple: the information in the network flowed from user B (who created the original post) to user A (who read the post and retweeted it).
- **Replies**: The connection between the user (user A) that replies to the post created by another user (user B) must go in the direction of the replied user to the user that created the reply, following the same direction as the information flow in the network, i.e., from user B to user A.
- **Quotes**: Similarly to the replies, the connection between the user that creates the quote (user A) and the user whose post is quoted (user B) must go in the direction of the quoted user to the user that created the quote, i.e., from user B to user A.
- **Mentions**: In the case in which a user (user A) creates a post in which it mentions another user (user B), the edge must go in the direction of the user that creates the mention (user A) to the mentioned user (user B), as the intention of user A is to make user B read the post, thus information flows in that direction.

In the case of SPIN, it is convenient to consider the previous four possible connections between users instead of simply considering retweet networks, follow networks, mention networks, or hashtag networks separately, as it has already been done previously in the literature [GMGM18, SIK22], because by considering those four possible connections (simultaneously), it is possible to represent a bigger portion of the information flowing through a network, which is what our algorithm, SPIN, tries to detect and quantify, thus a correct representation of the network is completely key to the correct functioning of SPIN.

8.4.2 Community detection

In order to apply the polarization algorithm as detailed in the document, a pre-partitioned network is required.

Since SPIN is mainly focused on socio-political analysis, considering aspects such as the ideology of the users is fundamental, as it allows the partition of the graph into well-separated communities from a domain knowledge point of view. For instance, a user with a left-leaning political view should not be in the same partition as a user with right right-leaning political view. Consequently, the usage of other graph partitioning algorithms, such as METIS [KK97] or Louvain [BGLL08], might not be as accurate (from a domain knowledge point of view) as SPIN requires because those methods are typically based on maximizing network modularity [New06], without considering important domain knowledge information, such as the political leaning of the users.

Thus, in order to generate a set of partitions, more adequate to the study of the political conversation in Online Social Networks such as X, we propose a label propagation approach (*label propagation*). More specifically, our label-propagation algorithm has several prerequisites:

[2]X (Twitter) interactions are defined, although similar interactions could be defined for other social networks.

1. **Number of communities**: Firstly, it should be able to support the division into 2 or more communities, as its application in studying polarization, as obvious as it seems, may require more than two partitions in the conversation graph. For instance, when considering political data, it can be inferred that there are multiple communities or partitions of users, and not necessarily just two; there could be many more. At least as many as the number of political parties in the system.

2. **Correct partitions from a semantic (domain knowledge) point of view**: Moreover, the algorithm should be able to encompass, according to a certain semantic component, nodes that are similar within the same partition, and in a different partition compared to those that are semantically different. This is starkly different from other existing partitioning algorithms, such as *METIS* [KK97] or *Louvain* [BGLL08], where the graph's structure is used to create the partition. Additionally, it should also utilize the structural aspects of the network (e.g., connections) to assign communities in the best possible way, but always respecting the meaning of the partitions. For instance, in a political use-case, nodes with different ideologies should not be included in the same partition, even if that made sense from the topological point of view of the network.

With this in mind, a *Label Propagation*-based approach was chosen, allowing nodes to "influence" their neighboring nodes until converging on a final partition. The community generation algorithm takes a weighted directed (potentially multi-directed) graph as its main parameter. In the graph, each node represents a user, each link represents a retweet relation (A retweeted B). The weight of the link corresponds to the number of retweets from one node to another.

This algorithm hinges on the idea that every node must have an associated community vector, with as many communities as specified in the input. Each position of the vector corresponds to a community, and the value of each position indicates the node's degree of membership in that community (a value normalized between zero and one, where zero indicates no membership and one indicates maximum membership). Furthermore, there are two different types of nodes:

1. **Labeled nodes**: On one hand, the algorithm requires as input a set of already labeled "seed" nodes. That is, an input node-vector dictionary is needed, where for each node a community vector to which it belongs is stored. Typically, this vector is expected to be full of zeros, except for one specific community where its value should be one (the community to which this seed node belongs). However, this greatly depends on the specific use-case where the algorithm is to be applied, so different values for the node membership vectors could be useful depending on the situation. These nodes are unique in that their labels will never change since they provide the initial information, the only one considered as the ground truth (*ground truth*).

2. **Unlabeled nodes**: On the other hand, there are nodes that are not labeled (the majority of them), and their community must be "inferred" by the algorithm. These nodes receive a label update[3] in each iteration of the label propagation algorithm based on their neighboring nodes' labels. This update is done in a random order to prevent potential biases). Initially, unlabeled nodes have an equitable membership value for each community, meaning that before its first iteration, they belong equally to all communities (because their community membership is unknown).

The employed propagation algorithm is defined as follows:

[3]Note that the concept of "label" in this context refers to each node's community vector.

1. **Label initialization**: Initially, labels are set using the already labeled nodes, specifying that for the unlabeled nodes, the membership vector has a value of $\frac{1}{N}$ for each community, where *N* is the number of communities the network is being divided into.

2. **Iterations**: As the underlying label propagation algorithm on which this partitioning algorithm is based is iterative, each iteration updates the labels of each node as follows:

    a) **Random ordering of unlabeled nodes**: First, the unlabeled nodes are randomly ordered (to prevent biases, as mentioned earlier).

    b) **Updating each unlabeled node**: For each unlabeled node, a new membership vector is considered with null (zero) membership values for each community, and:

    - **Connections to the node**: For every connection to the node that is being updated, each community membership value (from the neighboring node's membership vector) is multiplied by the connection's weight (frequency of that interaction between the users). The resulting vector from multiplying the weight by the neighboring node's membership vector is added to the new membership vector created for this node in this iteration (which was initially a zero vector). Intuitively, this allows the algorithm to group nodes in the same partition that have more connections, and connections with higher intensity, with nodes of that partition.
    - **Connections from the node**: Similarly, for every connection from the node being updated, the neighboring node's membership vector is multiplied by the connection's weight, and the resulting vector is added to the node's new membership vector. Thus, in-neighbors and out-neighbors are taken into consideration for the update of the nodes' labels.
    - **Normalization**: After considering all connections in updating the specific node, the resulting vector must be divided by the sum of all the node's connection weights (for both in-edges and out-edges), so that the membership values for each community are always normalized in the range $[0, 1]$.
    - **Key aspects related to connections between nodes**: There are a few aspects that should not go unnoticed when considering the connections between nodes in the network:
        - **Connection types**: As it was aforementioned, the input graph could be a multi-directed graph, thus allowing a pair of users to have multiple connections of different types: retweets, replies, quotes, and mentions (considering X's possible interactions between users).
        - **Filtering edges**: Using a multi-directed graph is quite convenient for the good performance of *SPIN*. However, having multiple edges (of different types) between the same nodes might not be as convenient for the label propagation algorithm, mainly because the label propagation must be done based on endorsement. As a result, for the label propagation algorithm, only retweet edges should be considered, because only retweets suggest that there is full agreement on a specific post (tweet), i.e., mention, reply, and quote edges might not suggest actual endorsement between users and could lead to wrong label propagation between users.

- **Weighted connections**: Additionally, it must be noted the edges must be weighted. Such a weight corresponds to the frequency of the interaction between two users, i.e., if a user has a retweet connection to another user, such a connection will have a weight based on the number of retweets between the users in the corresponding direction. This is quite a relevant aspect of the label propagation algorithm, as the labels of the neighbors are weighted based on the connection weight with those neighbors.

c) **Stopping condition**: As this algorithm is iterative, when the desired maximum number of iterations is reached, the algorithm stops. Thus, the labels from the maximum iteration are each node's final labels, i.e., the final membership vectors indicating the degree to which each node belongs to each community. Likewise, a stopping condition based on a minimum value of membership changes from one iteration to the next could be implemented.

The associated pseudocode is reflected in Algorithm 2.

Indeed, the label propagation step of the previous algorithm only generates a list of potential members for the communities provided. Therefore, another step is required to determine which communities a node belongs to, based on a minimum community membership threshold that must be also provided as an argument to this algorithm (τ). If a node has a value, for one or more communities, higher than the threshold, such a node will belong to all those communities meeting the condition.

However, it could happen that a node does not have any community membership value higher than the minimum community membership threshold (τ). As a result, that node would not belong to any partition. Since that behavior is normally not desired to carry out Social Network Analysis, and, more specifically, polarization detection and quantization (because it would be desired to have all the nodes assigned to a partition, to study the polarization between them considering all the members of the network, and not only a subset of them), a boolean argument called $create\_others$ can be specified in our community generation algorithm. Considering this parameter, two situations can arise:

- **create_others = True**: the node will be assigned to partition $N + 1$, which is the "other" partition (considering the aforementioned 1-origin indexing).
- **create_others = False**: The node will not be assigned to any partition, so its associated partition list will be an empty list. This is essentially useless for most algorithms that require partitioning, so it is recommended to use the "other" partition to prevent nodes without an associated partition.

Thus, this $create\_others$ parameter simply allows the creation of a new community containing all those nodes that could not be assigned to any other partition, as they did not have values (in their final community vector, i.e., their final label) greater or equal than τ for any community. Such a community is not considered within the N communities specified, thus if N partitions are to be created and the $create\_others$ parameter is enabled, there will be a partition $N + 1$ holding nodes that do not belong to any of the N partitions.

8.4.3 Intra-community entropy

Considering the implemented entropy metrics (outlined in Section 8.3.3), and taking into account the general idea of the algorithm (described in Section 8.4), there arose a need to design and implement an

Algorithm 2: Community generation based on label propagation.

Data: Directed graph: $network$
Initial dictionary of labelled nodes: $labelled\_nodes$
Number of iterations: max_iter
Number of communities: N
Community membership threshold: τ
Creation of the "others" community: $create\_others$

Result: Partition assignment to each node

$final\_labels \leftarrow$ **copy of** $labelled\_nodes$;
$unlabelled\_nodes \leftarrow$ **set difference of** $network.nodes$ **and** $keys(labelled\_nodes)$;
// Initialize the labels of the unlabelled nodes to $\frac{1}{N}$
for $node$ ***in*** $unlabelled\_nodes$ **do**
    $final\_labels[node] \leftarrow vector(N, \frac{1}{N})$; // vector of N $\frac{1}{N}$ values
end
for $iteration \leftarrow 0$ ***to*** $max\_iteration - 1$ **do**
    $unlabelled\_nodes \leftarrow shuffle(unlabelled\_nodes)$;
    for $node$ ***in*** $unlabelled\_nodes$ **do**
        $new\_label \leftarrow vector(N, 0)$; // vector of N zeros
        $weight\_sum \leftarrow 0$;
        for $(neighbor, node, weight)$ ***in*** $network.in\_edges(node)$ **do**
            $new\_label \leftarrow new\_label + weight \cdot final\_labels[neighbor]$;
            $weight\_sum \leftarrow weight\_sum + weight$;
        end
        for $(node, neighbor, weight)$ ***in*** $network.out\_edges(node)$ **do**
            $new\_label \leftarrow new\_label + weight \cdot final\_labels[neighbor]$;
            $weight\_sum \leftarrow weight\_sum + weight$;
        end
        $new\_label \leftarrow \frac{new\_label}{weight\_sum}$;
        $final\_labels[node] \leftarrow new\_label$;
    end
end
// Assigning the list of partitions of each node
$node\_partitions \leftarrow \{\}$;
for *node* ***in*** *network* **do**
    $list\_partitions \leftarrow$ **get** communities of $final\_labels[node]$ with value $\geq \tau$;
    if $create\_others$ ***and*** $list\_partitions$ ***is*** $empty$ **then**
        $node\_partitions[node] \leftarrow [N + 1]$;
    else
        $node\_partitions \leftarrow list\_partitions$;
    end
end
return $node\_partitions$;

algorithm to calculate the information flow within each community of nodes (users). Broadly speaking, the intra-community entropy calculation algorithm is defined as follows:

1. **Communities**: Each community is considered individually.
2. **Connections between nodes**: For each pair of connected nodes (in any direction), the entropy (NNIF) 8.3

between them is computed. This pairwise operation incurs the highest computational cost of the entire algorithm (due to the intensive use of NNIF). Note that the entropy between two nodes should never be calculated twice. If two nodes interact multiple times, the entropy is only computed once; otherwise, it would skew the actual amount of information flowing through the network. It must be noted that, in order to model the information flow in the best possible way, the connections between nodes must include retweet, mention, reply, and quote edges (considering X's interactions between users) because all those interactions between users involve the exchange of information between users, either implicitly or explicitly, as it was explained in Section 8.4.1.

3. **Intra-entropy of the partition**: Once all entropies related to connected nodes have been computed, the absolute average (regardless of direction) of the entropy (calculated with the NNIF metric) values is determined. This provides an average of the amount of information flowing within each partition, which is returned as the result. Additionally, the calculated entropies for each pair of nodes are also computed, as they may be useful in other contexts. For instance, they could highlight the most influential node (user) in terms of information flow within each community, which could be useful in the context of a socio-political analysis.

In order to execute this algorithm, it is essential that NNIF can be run, and this metric, in turn, requires the posts (and their timestamps) for both the source node and the target node, as well as the posts (and their timestamps) for nodes in the local neighborhood. For the case of SPIN, this information can be modeled through an attributed network, where a node attribute can hold this information so that the NNIF metric to be computed. Thus, the pseudocode of the algorithm can be found in Algorithm 3.

Algorithm 3: Intra-Community Entropy Calculation.

Data: Multi-directed and attribute-based graph: $network$
Partition dictionary: $partition\_dict$
Result: Average intra-entropy per community: $partition\_cross\_entropy$
Entropy of each pair of nodes connected by community: $cross\_entropies$

$partition\_cross\_entropy \leftarrow \{\}$;
$cross\_entropies \leftarrow \{\}$;
for $comm$ ***in keys of*** $partition\_dict$ **do**
  $current\_partition\_entropy = \{\}$;
  for $node1\_pointer \leftarrow 0$ ***to number of*** $comm$*'s* $nodes - 1$ **do**
    for $node2\_pointer \leftarrow node1\_pointer + 1$ ***to number of*** $comm$*'s* $nodes - 1$ **do**
      $node1 \leftarrow node$(node1_pointer);
      $node2 \leftarrow node$(node2_pointer);
      if $network$*.has_edge(*$node1$*,* $node2$*)* **then**
        $current\_partition\_entropy[(node1, node2)] \leftarrow NNIF(node1, node2)$;
      end
    end
  end
  $cross\_entropies[comm] \leftarrow current\_partition\_entropy$;
  $partition\_cross\_entropy[comm] \leftarrow avg(abs(values$ **of** cross_entropies[comm]$))$;
end
return $partition\_cross\_entropy$, $cross\_entropies$;

8.4.4 Inter-community entropy

On the other hand, the algorithm used for the calculation of inter-community entropy is as follows:

1. **Communities**: Each pair of communities is considered individually. Each pair of communities is considered only once, as considering them more than once would double the real flow of information between communities.

2. **Essence of the Algorithm**: The goal of the algorithm is to determine the flow of information from one community to another community and vice versa. For this, a matrix M of dimension $C \times C$ is used, where C is the number of communities being handled in a specific use case. In this matrix M:

   - **Main Diagonal**: It should be completely ignored for calculations, as it does not refer to the flow of information between different communities (it is not inter-community entropy).
   - **Non-symmetry**: Moreover, the matrix is non-symmetric, and this has an explanation: for two communities A and B, two separate information flows are calculated, one from A to B ($X \rightarrow Y$) and one from B to A ($Y \rightarrow X$). Thus, when calculating entropies between nodes of these communities, the direction of the connection and the sign of the calculated entropy should be considered:
     - **Edge from A to B**: If the edge goes from community A to community B, there are two possible cases. If the entropy is positive, the information flow should be considered (added) for the position corresponding to that direction ($M[A, B]$). However, if it is negative, it should be considered for the symmetric position ($M[B, A]$). In this way, only positive entropy is considered, as based on the direction, it is used to calculate one matrix position or its symmetric position.
     - **Edge from B to A**: If the edge goes from community B to community A, the same principle applies. If the calculated entropy is positive, it is used in the calculation of $M[B, A]$. Otherwise, it's used in the calculation of $M[A, B]$.
     - **Information Flow Division**: This division of the total information flow between A and B (which is split into $M[A, B]$ and $M[B, A]$) should be "normalized" according to the total number of entropies calculated between communities A and B. Thus, the sum of entropies for each matrix cell corresponding to the flow of information between communities ($M[A, B]$ and $M[B, A]$) should be divided by the total number of entropies calculated, both from A to B and vice versa. The result is an accurate division of information flow in two different matrix positions, indicating the direction of the information flow between communities.
   - **Return Value**: The return value of the algorithm is the matrix obtained after considering all community pairs, performing the sums of information flow (divided into different matrix positions based on the direction indicated by the entropy sign), and normalizing these sums considering the entropies calculated for each pair of communities. This matrix represents the amount of information flowing from the community represented by the rows to the community represented by the columns. Thus, position $M[A, B]$ represents how much information

flows from community A to community B. Such a matrix is quite relevant in a political context, as the one proposed to benchmark our algorithm because it shows the negative influence of each community over the others, which could significantly help to carry out an in-depth socio-political analysis in a reliable manner, as it is based on the information that flows across the different communities that can appear in a social network.

The pseudo-code associated with the inter-community entropy calculation algorithm is defined in Algorithm 4.

Algorithm 4: Calculation of inter-community entropy.

Data: Multi-directed and attribute-based graph: $network$
Partition dictionary: $partition\_dict$
Result: Inter-community entropy matrix

$num\_comm \leftarrow length$ **of** $keys$ **of** $partition\_dict$;
$inter\_comm\_entropy \leftarrow matrix(num\_comm, num\_comm)$;
for $comm1 \leftarrow 0$ ***to*** $num\_comm - 1$ **do**
  for $comm2 \leftarrow comm1 + 1$ ***to*** $num\_comm - 1$ **do**
    $entropy\_c1\_c2 \leftarrow \{\}$;
    $entropy\_c2\_c1 \leftarrow \{\}$;
    $pairwise \leftarrow 0$
    for $node1$ ***in*** $comm1$*'s* $nodes$ **do**
      for $node2$ ***in*** $comm2$*'s* $nodes$ **do**
        if $network$.***has_edge(***$node1$*, *$node2$***)** **then**
          $entropy \leftarrow NNIF(node1, node2)$;
          if $entropy \geq 0$ **then**
            $entropy\_c1\_c2[(node1, node2)] \leftarrow entropy$;
          else
            $entropy\_c2\_c1[(node1, node2)] \leftarrow -entropy$;
          end
          $pairwise \leftarrow pairwise + 1$;
        end
      end
    end
    $inter\_comm\_entropy[comm1, comm2] \leftarrow \frac{sum(values(entropy\_c1\_c2))}{pairwise}$;
    $inter\_comm\_entropy[comm2, comm1] \leftarrow \frac{sum(values(entropy\_c2\_c1))}{pairwise}$;
  end
end
return $inter\_comm\_entropy$;

8.4.5 Calculation of negative entropies

Another critical aspect of this algorithm is the aforementioned concept of negative entropy, which leverages the usage of entropy measurements, such as NNIF, for the analysis and quantification of negative (hostile) information flowing through the network.

Before calculating the negative intra-community and inter-community entropies, it is first required to filter the tweets keeping only the negative tweets, that is, tweets with potentially hostile information, as they clearly contribute to the polarization of the network. To do so, the aforementioned LIWC 2007

framework was used to count the number of words of each category per tweet, and then these category counts were used to filter the original tweets, keeping just a subset of them. The tweets can be filtered by one or several categories attending any desired criteria, however, for the SPIN benchmark we filtered tweets having the category $EmoNeg \geq 1$ (EmoNeg refers to the number of words with negative emotion in the input text, in this case, a tweet).

Then, regarding the calculation of intra-community and inter-community negative entropies, the same algorithms as those defined in Sections 8.4.3 and 8.4.4 were used, because the *NNIF* metric requires the posts, together with their timestamp, published between the two nodes of the network whose (NNIF) cross-entropy is being calculated (as well as the posts and their timestamps for the nodes in the local neighborhood, which is the remarkable feature of this entropy metric). In the case of the SPIN algorithm, this information can be easily represented through an attributed network with the following attributes:

- **Timestamp-Tweet Dictionary**[4]: A dictionary whose keys are the timestamps of the publications (from each user), and whose values are each of the publications associated with those timestamps.
- **Negative Timestamp-Tweet Dictionary**: This attribute is exactly like the previous one, but only considers negative (hostile) posts. Note that the SPIN polarization index may vary depending on how a negative tweet is defined. In our case, we use the LIWC 2007 framework, applying filters on the categories obtained through text analysis. For benchmarking our algorithm, we defined the set of negative tweets as tweets that meet the following condition $Emoneg \geq 1$, being EmoNeg the number of words with negative emotion, detected through textual analysis by LIWC 2007.

Thus, an optional parameter not specified (for simplicity reasons) in the previous algorithms is required to calculate the entropies. Such a parameter is related to the network attribute that should be used to calculate the entropies. If the attribute contains the timestamp-tweet dictionary (without any filter), the total intra-community (or inter-community) entropy will be calculated depending on the algorithm that is to be executed. However, if the attribute containing the timestamp-tweet dictionary (filtered by negative tweets) is specified, the negative intra-community (or inter-community) entropy will be calculated, using the exact same algorithms.

8.4.6 Polarization score

Finally, considering the calculation of intra-community entropies (see Section 8.4.3), the calculation of inter-community entropies (see Section 8.4.4), and their counterparts in terms of negative entropy (see Section 8.4.5), the network polarization index is defined as follows:

1. **Ratio of Negative Intra-community Information Flow**: Firstly, it is necessary to calculate the percentage of negative information being sent within each community. This is important because it allows us to model the situation where two communities are polarized due to the exchange of negative information (possibly about the other community), even if there are no direct connections that allow us to deduce this. This ratio is calculated as follows:

[4]Timestamp-Tweet considers the X (Twitter) case study, but it could be any post in general.

a) **Intra-community Entropy**: Firstly, it is essential to understand that we start from the partition-entropy intra dictionary (and also the negative), which specify the average intra-community entropy of each total and negative partition respectively (obtained from the first of the two outputs of the intra-community entropy calculation algorithm, defined in Section 8.4.3).

b) **Negative Intra-community Entropy Ratios**: In order to calculate each of the negative information flow ratios within each community, each partition is individually considered, and the negative information flow in that community is divided by the total information flow of that community.

c) **Ratios Aggregation**: Finally, an average of the negative intra-community entropy ratios from all communities is taken to obtain the average negative information flowing within the communities. This result is returned as the negative intra-community information flow ratio.

2. **Ratio of Negative Inter-community Information Flow**: Similarly, it is important to calculate the percentage of negative information exchanged between communities and not just within them, as if the percentage is very high, there will undoubtedly be polarization between the communities. To calculate this negative inter-community information ratio, proceed as follows:

   a) **Inter-community Entropy**: Firstly, you must start from the matrices of total and negative inter-community entropy.

   b) **Negative Inter-community Entropy Ratios**: For each position in the matrix that does not belong to the main diagonal, the value found in the negative inter-community entropy matrix is divided by the value found in the total inter-community entropy matrix. Resulting in $C \times C - C$ negative information flow ratios between communities (they are $C \times C - C$ because the main diagonal is ignored, as mentioned previously).

   c) **Ratios Aggregation**: Finally, the average of the ratios calculated in the previous step is taken by summing the ratios and dividing by $C \times C - C$, which is the number of aggregated ratios. This result is returned as the negative inter-community information flow ratio.

3. **Polarization Index**: Finally, considering the above ratios, the polarization index is defined as the weighted average of both ratios, through two coefficients α and β. The former determines how important the negative information flow is within the communities, while the latter models how important the negative information flow is between the communities. By default, they both take the same value $\alpha = \beta = 0.5$. However, the coefficients can be freely adjusted to properly fit each study case, as long as their sum equals 1, i.e., $\alpha + \beta = 1$.

The pseudocode for calculating the polarization index is defined in Algorithm 5.

Algorithm 5: Calculation of the polarization index.

Data: List of communities: $communities$
Intra-community entropy dictionary: $partition\_intra$
Negative intra-community entropy dictionary: $partition\_neg\_intra$
Inter-community entropy matrix: $inter\_matrix$
Negative inter-community entropy matrix: neg_inter_matrix
Adjustment coefficients α and β: α, β

Result: Network polarization index

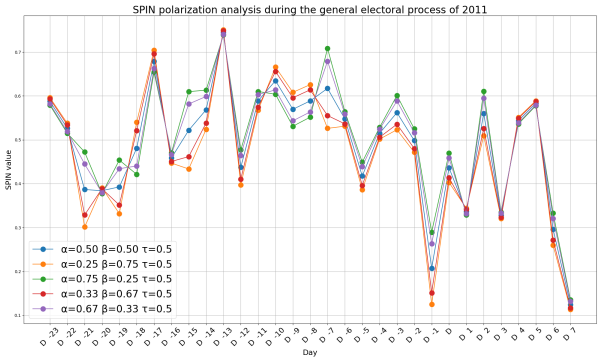
$intra\_neg\_ratio \leftarrow 0$;
$inter\_neg\_ratio \leftarrow 0$;
for $comm$ ***in*** $communities$ **do**
  $intra\_neg\_ratio \leftarrow intra\_neg\_ratio + \frac{partition\_neg\_intra[comm]}{partition\_intra[comm]}$
end
$intra\_neg\_ratio \leftarrow \frac{intra\_neg\_ratio}{length(communities)}$
for $comm1$ ***in*** $communities$ **do**
  for $comm2$ ***in*** $communities$ **do**
    if $comm1$ ***not equal*** $comm2$ **then**
      $inter\_neg\_ratio \leftarrow inter\_neg\_ratio + \frac{neg\_inter\_matrix[comm1,comm2]}{inter\_matrix[comm1,comm2]}$
    end
  end
end
$inter\_neg\_ratio \leftarrow \frac{inter\_neg\_ratio}{length(communities)\cdot length(communities)-length(communities)}$
return $\alpha \cdot intra\_neg\_ratio + \beta \cdot inter\_neg\_ratio$;

8.5 Comparative Evaluation of Algorithms

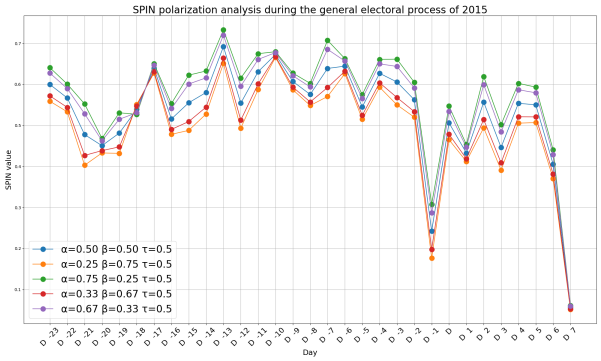
In order to benchmark the utility of our algorithm, we applied the algorithm defined in Alg. 1 to obtain datasets related to each of the Spanish electoral processes from 2011 to 2019, including both general and local electoral processes. Thus, we propose the usage of the political scenario of Spain to benchmark SPIN, as the country is characterized by a complex political scenario where the society is represented by several political parties (not only two), with different political leanings. We consider that using such a complex political scenario could show the true potential of SPIN to carry out reasonable and precise polarization detection and quantification, at the same time that it allows an in-depth socio-political analysis thanks to the high explainability of its results.

First, we compare the performance of SPIN when using different hyperparameters, obtaining the results described in Fig. 8.1. Based on this, we observe the SPIN algorithm adeptly captures the polarization dynamics outlined in the initial axioms. We see a subtle and incremental rise in polarization from the pre-election phase leading up to the electoral campaign. This is followed by a decrease on the day before the elections, known as the "day of reflection," due to the inactivity of politicians and parties. There is a notable surge during the election event, which then gradually tapers off post-election.

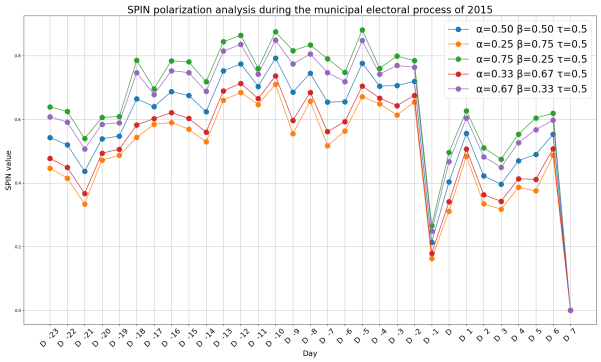
Both the rise in polarization during the campaign and, notably, its decline during the "day of reflection" serve as strong indicators of the role political parties and candidates play in the escalation of political polarization. Similarly, we also observe that polarization peaks on the days when electoral debates are held on the country's main television channel (RTVE), which take place 5 days prior to the general elections.



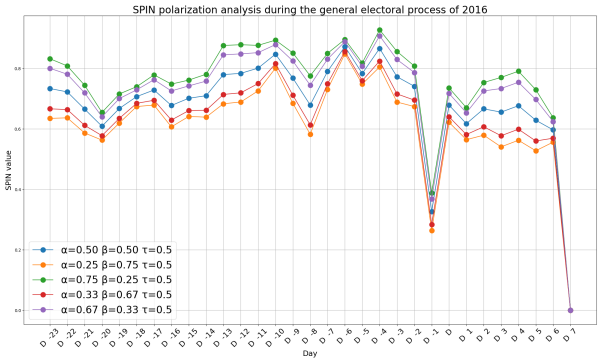
(a) Spanish elections of 2011 (general).



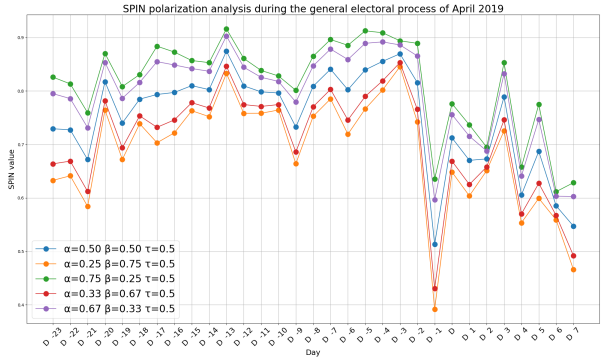
(b) Spanish elections of 2015 (general).



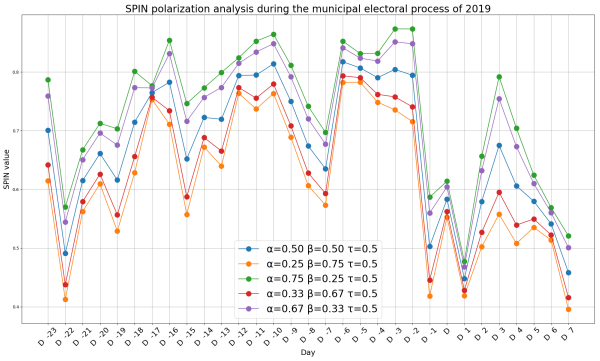
(c) Spanish elections of 2015 (local).



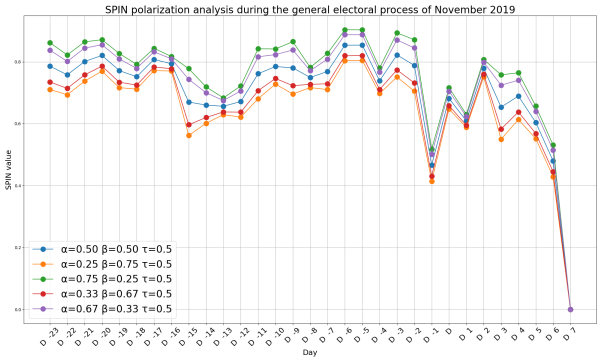
(d) Spanish elections of 2016 (general).



(e) Spanish elections of April 2019 (general).



(f) Spanish elections of 2019 (local).



(g) Spanish elections of November 2019 (general).

Figure 8.1: Comparison of SPIN performance with different parametrizations across Spanish electoral processes from 2011 to 2019.

Regarding the evolution of polarization, we broadly observe that polarization has increased over time, displaying consistently higher and more sustained patterns throughout the entire process. This is

especially evident during the general electoral processes, with less polarization and greater variability during local processes. This is likely due to the diversity of cities and options, as well as the adoption of decentralized communication strategies by political parties.

However, not only did we evaluate the performance of SPIN by testing its performance with different hyperparameters, but we also compared its polarization index with the polarization index obtained from other algorithms of the literature, so as to determine whether, for the specific context of the use-case (political scenario), the algorithm proves to work better or worse than already existing algorithms. The comparison results are described in Fig. 8.2.

When comparing the SPIN algorithm with the rest of the studied algorithms, we find that almost all of them fail to capture polarization according to the defined axioms, particularly during the "reflection periods" and the post-election week.

Similarly, a majority of the studied algorithms exhibit particularly high and sustained values over time (*RWC, MBLB, Multi-Opinion) or notably low values (ERIS, GMCK, P).

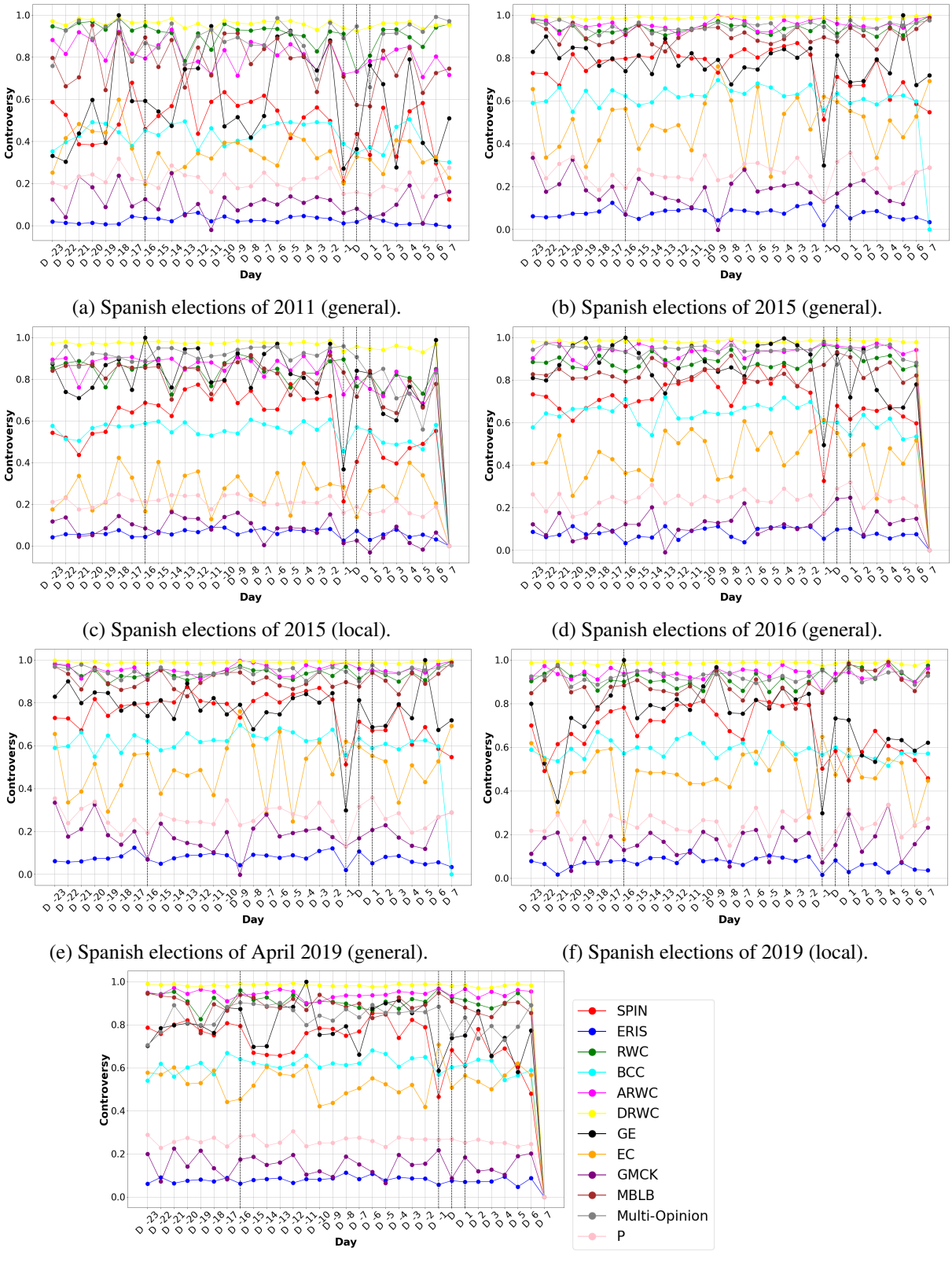
Among the most similar algorithms, capable of showing centered values throughout the entire spectrum (neither too high nor too low), we find our proposal alongside GE, EC, and BCC. However, both EC and BCC particularly fail to capture the post-election week and the "reflection day." They also display unstable patterns that fluctuate throughout the process. While GE comes close to meeting the proposed axioms, its oscillation complicates its application for analysis.

This trend is consistent across all data sets.

Additionally, we used our algorithm, SPIN, to gain a general understanding on the evolution of the (political) polarization in Online Social Networks, in this case, using X (Twitter) as the source of data. This study was divided into two separate analyses: the evolution of polarization across general electoral processes (see Fig. 8.3), and the evolution of polarization across local electoral processes (see Fig. 8.4), as this approach also allowed to compare the polarization between the two types of electoral processes. The conclusions that were derived from these results can be described below:

Additionally, we seek to study whether SPIN is capable of identifying the average polarization evolution across different Spanish electoral processes, considering both general and local electoral processes. Such a study is described in Fig. 8.5

As we have observed on an individual election level, the polarization surrounding political discourse on X (Twitter) in Spain has experienced an increasing trend from 2011 to 2019. This trend has been particularly pronounced in national politics as opposed to local politics.



(a) Spanish elections of 2011 (general).

(b) Spanish elections of 2015 (general).

(c) Spanish elections of 2015 (local).

(d) Spanish elections of 2016 (general).

(e) Spanish elections of April 2019 (general).

(f) Spanish elections of 2019 (local).

(g) Spanish elections of November 2019 (general).

(h) List of compared algorithms.

Figure 8.2: Comparison of SPIN performance with different parametrizations across Spanish electoral processes from 2011 to 2019.

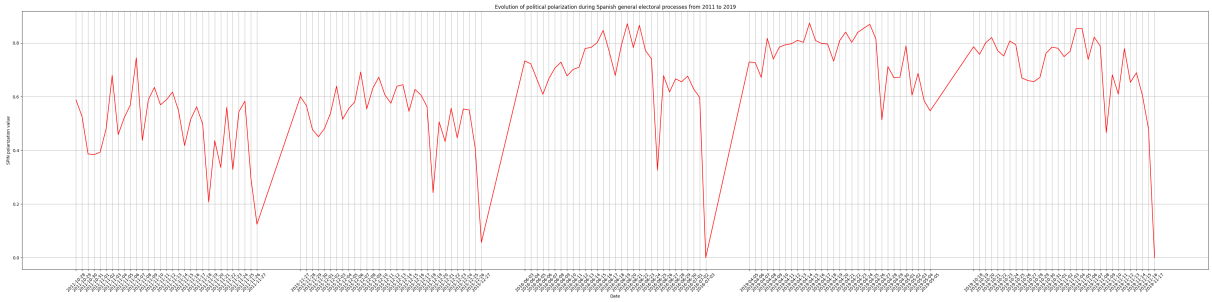


Figure 8.3: Evolution of polarization during Spanish general electoral processes from 2011 to 2019.

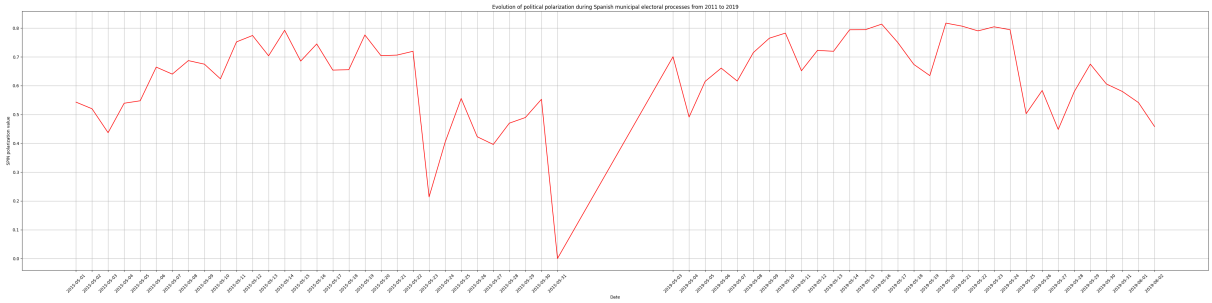


Figure 8.4: Evolution of polarization during Spanish local electoral processes from 2011 to 2019.

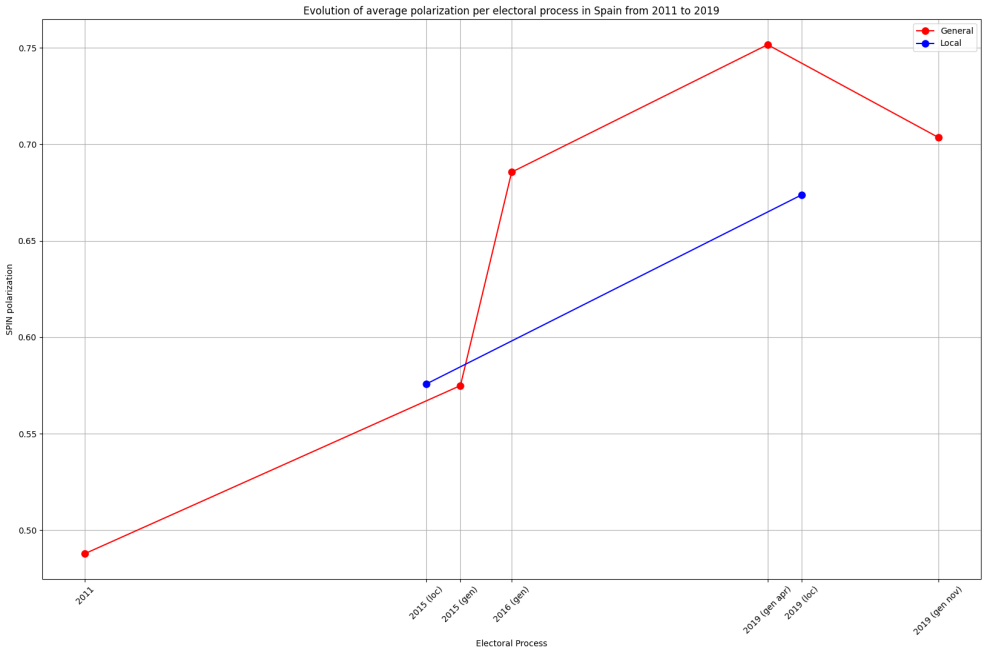


Figure 8.5: Average polarization per electoral process in Spanish electoral processes from 2011 to 2019.

8.6 Discussion

The main challenges of the algorithm are primarily found in efficiency and accuracy. Computational efficiency is a concern. There are several factors that can impact the efficiency of the algorithm. While the algorithm has proven its capability and utility in analyzing multiple electoral processes in Spain, it is clear that for especially large data sets (graphs), the process could take an excessive amount of time. However, this is a challenge common to almost all of the algorithms studied.

Regarding this, in our particular case if the communities are large, the number of cross entropies to be calculated increases significantly (pairwise based on connections between nodes), which can notably increase the execution time.

Similarly, nodes with many publications also affect the efficiency of the algorithm, since entropy takes longer to calculate in these cases.

The accuracy of the partitioning can also influence the result. The proposed algorithm is heavily dependent on the network partitioning algorithm. If a good partitioning is used, the results will be accurate. However, if the partitioning is not appropriate (semantically, not structurally), then the results can deviate significantly from reality. Specifically, in the case of the label propagation algorithm, if the "seed" nodes are not selected in a way that truly represents the political system, the effectiveness of the algorithm will not be accurate.

8.7 Conclusions

Answering the first research question addressed in this chapter, that is, RG2.RQ3, the SPIN algorithm was developed and successfully tested. The uniqueness of the SPIN algorithm lies in its foundation in information theory, offering a fresh perspective on measuring polarization. This approach allows it to encapsulate the core tenets of polarization as articulated by predominant definitions in the field. While other algorithms might offer insights based on surface interactions or apparent divides, SPIN inquires into the inherent informational structures and patterns. This deep-rooted analysis ensures a more nuanced and precise understanding of polarization, making SPIN stand out from the other approaches.

A standout feature is its incorporation of the temporal dimension, tracking the evolution and flow of information over time. This temporal consideration is essential as polarization is not static; it evolves, intensifies, or diminishes in response to real-time events and discourses. Additionally, SPIN emphasizes coherence in content, ensuring that consistent themes and narratives within a community are recognized and factored into the analysis. Considering the community's structural intricacies, it keeps sight of the network's topology. By weaving together the temporal, content coherence, and structural aspects, SPIN offers a comprehensive and nuanced insight into the multifaceted nature of polarization.

Consequently, our algorithm accurately assesses the polarization phenomenon, can produce results within a reasonable time frame, and the outcomes it generates align with the postulated axioms regarding what polarization should be during an electoral campaign.

All in all, our algorithm takes into consideration both the topology of the network, as well as the information flowing through it, hence utilizing as many resources as possible to carry out polarization detection and quantification. Indeed, other hybrid algorithms have already been proposed and used for the same purpose, nevertheless, none of these mixes a hybrid approach with the analysis of the dynamics of the network, i.e., considering the information that is flowing through the network at every moment

of time, which could have a significant influence on the other posts and information exchanges between the users in the network. As a result, our algorithm, SPIN, blends the best of both Information Theory and Social Network Analysis to carry out precise polarization detection and quantification, thanks to its support to multiple communities and its ability to model the relationships between them, confirming that polarization is rooted in the structure of the network as well as the information flowing through it, answering RG2.RQ4.

Based on the benchmark results aforementioned, we consider SPIN to be a useful contribution to the socio-political analysis, as it was able to model the polarization along different Spanish electoral processes, from 2011 to 2019, in a complex political scenario, where old political parties are still supported by a vast amount of people (PP and PSOE), at the same time that new political parties are emerging and gaining more and more traction in the political arena (such as Ciudadanos, Vox or Podemos), and divides between supporters of the different political parties tend to broaden more and more. Indeed, the high degree of explainability of the proposed polarization index makes the SPIN algorithm a suitable option, not only for polarization detection and quantification, but also for gaining a deep understanding of the socio-political elements that most contribute to the polarization of the network to potentially prevent it, as polarization detection and quantification can be used as the tools of today's society for closing divides between the different communities that we can find in social networks, which are no more than a digital reflect of our society.

Prospective research trajectories for the SPIN algorithm present a plethora of opportunities. There is potential in refining and optimizing its mechanics for even more accurate results. Venturing beyond the Spanish-language datasets, testing its applicability across diverse linguistic and cultural contexts can further validate its universal relevance. Moreover, while its current focus is on political polarization, SPIN's underlying principles hold promise for broader applications. Areas like marketing can benefit from understanding consumer polarities, preferences, or brand loyalties. Similarly, in misinformation, the algorithm can be pivotal in discerning echo chambers, biased information flows, and the intensity of misleading narratives. Such expansive applications could position SPIN as a versatile tool for various analytical challenges.

Indeed, the implications of the SPIN algorithm extend beyond mere measurement. One of its most profound utilities lies in its potential to inform and shape strategies to counter polarization. By offering a detailed insight into the complex webs of polarization – including its temporal evolution, content coherence, and community structure – SPIN provides policymakers, platform developers, and community leaders with a granular understanding of where and how divisions occur. With this knowledge, targeted interventions such as developing custom recommendation strategies can be formulated to bridge divides, foster understanding, and promote more cohesive dialogues. The SPIN algorithm is not just a diagnostic tool; it is a foundation upon which effective solutions to the challenges of polarization can be built.

Chapter 9

The Role of Recommender Systems in the Formation of Disinformation Networks

9.1 Introduction

We know the rise of social media as a primary information source has shifted from traditional, centralized media to a decentralized, user-driven model, empowering widespread content creation and consumption. This shift, while democratizing information, has bypassed traditional editorial oversight, relying on algorithmic recommendations that often create echo chambers and spread misinformation. This chapter of the thesis evaluates the role of various recommendation algorithms (user-based, content-based, deep learning, and reinforcement learning) in disinformation spread, using a three-year dataset of disinformation and legitimate information networks. By simulating content recommendations, we assess how these algorithms impact disinformation dynamics and propose models to reduce misinformation while maintaining recommendation accuracy, fostering healthier information ecosystems.

At the outset of this chapter, we investigate the influence of recommendation algorithms on the proliferation of disinformation networks within online social media platforms. The chapter evaluates the extent to which various recommendation systems—including user-based, content-based, deep learning, and reinforcement learning approaches—contribute to the generation and consolidation of disinformation networks. Leveraging simulations performed on a large dataset spanning three years, we identify patterns that differentiate these algorithms in their potential to amplify disinformation. By analyzing key network properties and content dynamics, this chapter directly addresses the second part of the third research goal of this thesis: proposing mitigation strategies for recommendation systems that exacerbate disinformation. It also supports the overarching aim of this research by advancing our understanding of the mechanisms behind disinformation propagation in digital ecosystems.

The chapter is organized as follows: Section 9.3 outlines the methodologies used for data collection, simulation, and network analysis, detailing the selection of algorithms and their respective properties. Section 9.4 presents the results, exploring how different recommendation algorithms impact the structural and behavioral characteristics of disinformation networks. Section 9.5 provides a comparative discussion, analyzing the interplay between recommendation system accuracy and disinformation network

formation. Finally, Section 9.6 concludes with a summary of findings, emphasizing the need for more nuanced and responsible algorithmic design to mitigate disinformation in the digital realm.

Research questions

Our aims in this chapter are summarized in the following research questions, as a starting point to cover the third research goal of the thesis:

- **RG3:** Analyze the role or contribution of recommendation systems in promoting these phenomena. Propose mitigation strategies through these systems.
  - **RG3.RQ1:** Which approach or type of recommendation system contributes most to the formation of disinformation in the form of networks?

9.2 Background

Social media platforms are increasingly supplanting traditional media as the primary source of information, offering unparalleled immediacy and accessibility [HRC11, RT14, YK01]. Unlike traditional media's centralized and editorially curated nature, social media democratizes content creation and dissemination, enabling users to share information instantaneously without editorial oversight [JVS00, YK01]. This shift is largely driven by algorithmic recommendation systems, previously described in Section 2.16, which analyze user behavior to curate personalized feeds [ASIM18]. However, these systems can inadvertently foster echo chambers, reinforcing ideological silos and amplifying misinformation through selective exposure [VdBGH22, Par11].

Echo chambers, characterized by dense networks of ideologically similar users, amplify politically charged misinformation while limiting exposure to diverse perspectives [VPV21, Par11], as described in Section 2.15. Recommendation algorithms (RAs), such as user-based, content-based, deep learning, and reinforcement learning, further shape these dynamics by influencing content visibility and engagement [CLDB18, JSH+21, ZYST19]. This study investigates the role of these algorithms in promoting or mitigating disinformation by simulating their impact on network structures and content dissemination. By analyzing how these methods influence the formation and amplification of disinformation networks, we aim to develop strategies that balance personalized recommendations with the reduction of misinformation spread, fostering healthier information ecosystems.

9.3 Specific Methodology

In this section, we describe the methodology that we used to carry out our research. First, we describe the data collection and utilization process in Section 9.3.1. Then, we perform a literature review to discuss the existing recommendation algorithms that have been widely applied in the context of Online Social Networks. Indeed, Section 9.3.2 describes the existing approaches to carry out content recommendation in Online Social Networks, and it also describes the RA selection strategy that we used for our research. Additionally, in Section 9.3.4 we describe the methods and techniques that we used to carry out simulations with the selected RAs, together with the techniques that we used for the correct analysis of their results. Previously, in Section 9.3.3, we discuss the used approach for generating the networks with

which our study was performed. Last, we leverage the use of different network analysis techniques in Section 9.3.4 to rigorously answer the proposed research questions.

9.3.1 Data set

For this study, we employed the dataset "Information disseminators" detailed in the thesis methodology (Section 4.1.5) and composed of two main sub datasets: one containing X accounts linked to verified journalists (described in Table 4.16) and another focusing on disinformation actors (see Table 4.15) These datasets were previously used in Chapter 7 to analyze the behavior of disinformation networks and better understand their network structure and content.

In this research, we are mostly interested in the study of recommendation algorithms and their impact on the generation of disinformation networks. For that reason, we crafted a single dataset containing information regarding both the journalist and disinformation accounts, that was stored in the aforementioned datasets (journalist and disinformation datasets).

In order to craft a correct dataset, we carefully mixed both datasets, avoiding user duplication (as one user could be both a journalist and a disinformation actor) while linking their posts to the same (unique and correct) user, and adding extra pieces of information to each user to understand whether they appeared in the journalism dataset, or whether they come from the disinformation actors dataset (or both of them). Indeed, this dataset contained information from both kinds of networks, disinformation and journalism networks, while also providing information regarding which users appear in both datasets, as they could be of potential interest to both networks (journalists and disinformation actors). However, in this research, we are far more interested in the analysis of the impact of RAs in the generation and consolidation of disinformation networks, thus we only create journalists' networks to establish a fair comparison of results considering the results obtained from disinformation networks.

In this manner, we use the dataset to feed, on a fixed-time-window basis (set to 7 days as it is explained in the subsequent sections), the selected RAs to generate recommendation networks. These recommendation networks contain both journalism and disinformation accounts, however, since we are mostly interested in the generation and consolidation of disinformation networks, for certain parts of our study (described in the subsequent sections), we filter the generated recommendation network to only keep nodes belonging to the original disinformation dataset (and their corresponding connections). This approach allowed us to give a rigorous answer to **RG3.RQ1**, in order to be more precise in its answering, we made use of the complete recommendation network generated by each RA, so as to understand how content spreads from disinformation accounts to journalist accounts, with a special focus on understanding the information dynamics promoted by each RA. In Section 9.3.3 we further describe in detail the generation of the required networks to answer the proposed research questions.

Nonetheless, this dataset does not contain information regarding a very basic and important relationship for RAs in social networks: the follower/followee relationship. Such a relationship is quite relevant because it gives an idea of the content that is *visible* to a user. Indeed, during a specific window of time, only a portion of the whole content flowing through the network is visible to each user, and this *visibility* can be easily modeled using the follower/followee relationship. To overcome the absence of this relationship in our dataset, we calculate a *visibility graph* per time window, which is used to determine the content visible to each user in each time window. Further details about the visibility graph are given in the following sections.

9.3.2 Recommendation algorithms selection and analysis

In the course of our research, we implemented several recommendation algorithms considered to be the most characteristic approaches in the field [ASIM18]. This diverse array included topology-based algorithms such as Collaborative Filtering, which leverages the power of collective user behavior to make recommendations, as well as content-based algorithms that utilize user preferences in text content, in the case of X publications, encoded with Word2Vec [Chu17]. Moreover, other approaches found in the literature were also used to provide a fair comparison between recommendation techniques. Among these other approaches, we implemented a Thompson Sampling Multi-Armed Bandit Recommender [AG12, SCCL19], under the paradigm of Reinforcement Learning, to carry out recommendations using a completely different approach. Similarly, we implemented DeepRank, a hybrid Deep Learning-based approach to recommend content based on both topology (i.e., user interconnections) and content (i.e., the meaning of the posts) [CZ20]. More specifically, the algorithm uses a sparse representation of both the user that ranked an item (e.g., a tweet and its retweet) as well as the item itself as an input. Such an input is then processed through an embedding layer that condenses the information into dense vectors, which are then used to carry out a prediction after the action of a series of hidden layers[1] that result in a prediction of the ranking of the item (tweet retweeted or tweet not retweeted), leveraging the use of both the content-based component of the tweet (its embedding) and the social-based component of the network (user embedding) to solve the ranking tasks of RAs through a more complex approach that can understand and capture the complex patterns and nuances of user interactions within large datasets, doing so in a very efficient manner [CZ20].

Indeed, in this research we implemented RAs from the different well-known families of RAs in the literature (content-based, structure-based, Reinforcement Learning and Deep Learning), considering in that selection process relevant aspects just as the efficiency of the algorithm (feasibility of this research) and its accuracy, as the correctness of link prediction avoids the network attrition [Sch15]. Thus, through the implementation of various state-of-the-art algorithms from each RA family, we seek to study which approaches tend to facilitate the generation and consolidation of disinformation networks, through a rigorous study that is described in the following sections.

By leveraging the implementation of the aforementioned approaches, our work conducts a comprehensive investigation into the multifaceted landscape of recommendation systems to give a rigorous answer to the proposed research questions. Furthermore, we also implemented a Random recommendation algorithm (Rnd) to provide a reference baseline with which to analyze the obtained results. This algorithm provides a random ranking of the visible tweets of each user (the definition of visible content is provided in the following sections), and conforms our baseline, as the RA does not implement an intelligent policy to perform content recommendation to the users in the social network.

Thus, for our research, we adopted a multifaceted approach to recommendation systems, each chosen for its unique strengths and ability to complement the others, so as to cover the full spectrum of RAs, with which to answer the proposed research questions. The Random approach serves as a baseline to compare against more sophisticated algorithms. Collaborative Filtering, specifically user-based with KNN (K-Nearest Neighbors) [Pet09], taps into the wisdom of the crowd, making recommendations based on user similarity metrics. Additionally, we implemented a Word2Vec-based RA that also leverages the use of additional user information, such as its popularity (followers and followees) and other pertinent

[1] The recommended depth of the hidden layer block by the authors is 3, and we used such implementation for this research.

characteristics, enriching the contextual backdrop against which recommendations are made. The Multi-Armed Bandit recommendation, employing Thompson Sampling, recommends users to each user and then weights their tweets based on neighborhood visibility, enhancing the relevance of tweet recommendations based on the network's structure. Furthermore, we implemented another hybrid RA mixing the Multi-Armed Bandit approach (Thompson Sampling) with a content-based approach (Word2Vec) to compare with the individual approaches of these algorithms and analyze possible differences. DeepRank represents our venture into deep learning, offering a hybrid approach that combines the strengths of both collaborative and content-based methods for a nuanced recommendation. LIWC (Linguistic Inquiry and Word Count) content-based recommendation delves into psychological and linguistic cues within user preferences, relying heavily on text topics encoded in LIWC categories instead of the overall wording (like w2v). Indeed, LIWC is a research tool used for text analysis to determine the percentage of words that reflect different emotions, thinking styles, social concerns, and even parts of speech [PBJB15]. It works by comparing the words in a text to a predefined dictionary of words categorized into psychologically relevant categories, and it can easily be used to encode user preferences under those predefined categories, as has already been used in the literature for other domains [AI19, Ber19, YNHF22].

In total, we implemented 8 different RAs, covering the different approaches that comprehend the core *essential* components of the recommender systems that constitute the state of the art. Indeed, this diversified approach allowed us to explore the breadth of recommendation system strategies to rigorously answer the proposed research questions.

Considering each of the implemented approaches, simulations were performed from one window to another covering the studied period (from 2019 to the middle of 2022). Each simulation consisted of a few steps. First, posts of the current window are considered (and the previous ones), together with user retweets up to the window of study (without including it because that would lead to skewed recommendations). The retweet information was used to build the aforementioned visibility graph, whose nodes are users and whose edges are directed links from the user that retweeted content to the user that posted that content. Indeed, the direct neighbors of a user (considering the direction of the edges) represented the *visible users* from the point of view of that user. It is the content from *visible users* the one recommended by the RAs, following their approach and implementation to rank the posts accordingly. The intuition behind the visibility graph is quite simple: since we lack the "follow" relationships that provide a lot of information regarding post recommendations, we replaced it with a vision of which users were visible to each user based on the past interactions, to guide the RAs.

Thus, we leveraged the use of the visibility graph to determine which posts had to be ranked by each RA in each window for each user. The RA had to then offer a number of N recommendations to each user. A percentage of these recommendations had to be related to posts of the window of study, whereas the rest of the posts could come from other windows (using a $70 - 30$ train-test split to avoid a potential data leakage in the execution of the RAs). For our research, we used $N = 10$ recommendations, with 25% of them belonging to the current window and 75% to the previous ones. We carefully allocate the distribution of recommendations to include 25% from the current window and 75% from the past content, with the rationale grounded in optimizing user engagement and content relevance. This allocation is informed by the dual objectives of ensuring users are kept abreast of the latest trends and discussions, which the 25% current window allocation achieves by surfacing recent content that might be of immediate interest [BHM$^+$19, SCC22]. Concurrently, the 75% allocation for past content leverages

the rich repository of historical data, acknowledging that valuable and engaging content is not solely confined to the present moment [BHM+19, SCC22]. This approach not only enhances the discovery of timeless content but also mitigates the recency bias often observed in recommendation systems, thereby offering a more diversified and enriching user experience. The chosen proportions are designed to strike a balanced compromise between novelty and time-tested relevance, catering to a broad spectrum of user interests and engagement patterns. Last, setting $N = 10$ for recommendations is a deliberate choice aimed at optimizing the user experience by providing a manageable number of options that is enough to offer variety without overwhelming users, ensuring engagement while maintaining decision simplicity [SCPC18, SCC22].

Once recommendations are generated for each user and window (per RA), a process of *retweet simulation* is launched. This process aims at simulating the behavior of a user that will (or not) retweet each of the provided recommendations by the RA. While this process is random, it is based on an exponential decay, because this approach models correctly the potential action of a user on the recommended content as he *scrolls* through the recommendations. The posts with a *simulated retweet* are then used to generate the *recommendation network*, which is further described in Section 9.3.3. Last, the *simulated retweets* are then compared with the actual retweeted content to evaluate the accuracy (on the test dataset) of the RAs. This information is then used to understand the relationship between aspects that can foster the generation and consolidation of disinformation networks, and accuracy metrics applied to the RAs. All the metrics used for our research are described in the following pages.

Furthermore, it is worth mentioning that, for this research, we adopted a window of 7 days for providing recommendations to users. This decision was based on the balance between recency and relevance, ensuring that the recommendations remain both timely and pertinent. A weekly cycle is particularly effective in capturing user behavior patterns, as certain news are introduced in the network, whereas old trending contents are still spread across. This temporal frame allows for the incorporation of the most recent interactions and preferences, thereby enhancing the algorithm's responsiveness to changes in user interests.

9.3.3 Network generation

As previously discussed, news in X do have a lifespan of about a day, although certain viral content and news may last for a more extended period [ORK18]. We captured this aspect by using time windows of exactly 7 days. As a consequence, the dataset we crafted for this research was divided into temporal segments that correspond to the time windows where RA simulations were performed.

For each window of the performed simulation, the aforementioned *visibility graph* was computed. Such a graph determines, for each user, the content that is visible to such a user, based on the previous experience. More specifically, the visibility graph uses data from the previous two windows to view the retweet relationships between each pair of users. Indeed, one user (user A) is connected to a second user (user B) in the visibility graph ($A \rightarrow B$) if and only if user A retweeted content from user B in the previous two windows. Thus, the visibility graph is a directed and weighted graph, in which the weight of the connections indicates the number of retweeted posts, indicating the strength of that retweet relationship from one user to another. Naturally, the visibility graph provides a correct and intuitive representation of the content that is visible to each user, and it is used to filter the content (i.e., tweets) that a user can be recommended by the RA.

Furthermore, as part of the simulation, a *recommendation network* was created during the simulation of each RA in each time window. In this manner, during the simulation of each window, RAs were used to simulate recommendations for each user. After simulating retweets from the content recommended by the RA, the recommendation network was created (for that RA in that window). Such a network is a directed and weighted graph, whose nodes represent users and whose edges ($A \rightarrow B$) represent the existence of a simulated retweet in a given direction (A retweets a post from B from the recommendations performed by the specific RA). Indeed, the recommendation network represents a simulation of the information flow in the network of journalists and disinformation actors, and it allows us to leverage Social Network Analysis techniques to answer the proposed research questions.

Thus, we created recommendation networks for each of the time windows in the studied period, and this task was also repeated considering each of the RAs that were described previously. The generated recommendation networks were used in two different fashions (as already introduced in Section 9.3.1). A reduced version of the original recommendation network defined above was used, including only users belonging to the original disinformation dataset and the edges between them. On the other hand, the full recommendation network was used, so as to understand the information dynamics from disinformation actors' accounts to journalists' accounts.

9.3.4 Network analysis

In this research, we utilize techniques that have already been used under similar circumstances and situations. Since disinformation networks are well-known for their high efficiency and density, that gives them the possibility to spread information very quickly and efficiently, as opposed to other networks not related to disinformation actors. Indeed, these characteristics of disinformation networks can be measured using well-known properties and techniques of Social Network Analysis [New10, SHW+18]. In this research, we specifically leveraged the use of the following techniques, expanding some of the network metrics and properties that were already used in the research presented in Chapter 6.

Indeed, the metrics we used include density (Δ), a measure of edge presence compared to the maximum possible; efficiency (E), which assesses how effectively information travels across the network based on the shortest paths between node pairs; and average degree ($\overline{Deg}$), indicating the mean number of connections per node. Additionally, modularity (Q) was used to evaluate the strength of community structures by comparing actual edge distribution within communities against a random distribution model. The average clustering coefficient ($\overline{CC}$) was assessed to measure the tendency of nodes to cluster, reflecting the graph's local cohesiveness. To capture node influence and relevance, we analyzed the average eigenvector centrality ($\overline{EVC}$) and the PageRank ($\overline{PR}$) of the generated networks, to consider not only the number of connections but also the significance of connected nodes. These metrics collectively offer a comprehensive view of the network's topology, efficiency, and community dynamics, which facilitate the analysis of the proposed research questions.

Furthermore, we leverage the use of other metrics aimed at measuring well-known aspects of disinformation networks, such as the capability of treating very similar topics (high resistance to topic changes), their ability to maintain a subset of users as the *core* of the disinformation network (user persistence from window to window) and to increase the relevance of these users as windows go by. These other metrics are described in the following paragraphs.

We utilize $Topics_w$ to measure the average number of topics that are treated by the users of a (disin-

formation) recommendation network, as disinformation networks are expected to treat a reduced number of topics when compared to legitimate networks. Furthermore, we calculate $\%TP$, which refers to the percentage of topics in the recommendation network of the previous window that were still treated in the current recommendation network, i.e., the percentage of topics of window $n-1$ that are still trending in window n. Similarly, we calculated the persistence of users ($\%UP$), as the ratio of users in the recommendation network generated by a RA for a window $n-1$ that also belong to the recommendation network generated for window n. In order to provide a measure of the content amplification, we calculated $\overline{CA_u}$, as the average number of retweets of the users that were recommended by the RA during each window. Last, we provide another measure of content amplification, URL_u, which is the average number of shared URLs considering all the users that conform the (disinformation) recommendation network created by a RA for a given window.

To further analyze the impact of RAs in the generation and consolidation of disinformation networks, while performing a rigorous analysis, we employed statistical tests to provide, under a given confidence threshold, that our claims are valid. In this research, we use the Mann-Whitney U test on three of the most relevant network metrics for disinformation networks: efficiency, density and modularity [New10]. With this study, we seek to understand whether there is statistical significance behind the hypothesis that the efficiency, density and modularity distributions of the (disinformation) networks generated by RAs that seem to facilitate the generation and consolidation of disinformation networks are different from the ones generated by other RAs. Indeed, in these statistical tests, our null hypothesis H_0 is that there is no significant difference between the network metric (efficiency/density/modularity) distribution generated by RAs that enhance the creation and consolidation of disinformation networks and the one generated by other RAs (pairwise). The alternative hypothesis, H_1, is exactly the opposite, i.e., there is a significant difference between both network metric distributions. For this research, we perform the Mann-Whitney U test using 95% confidence ($\alpha = 0.05$). Furthermore, we also applied the non-parametric Mann-Whitney U test between the distributions of the aforementioned network metrics (efficiency, density and modularity) on the disinformation and journalist recommendation networks created by each RA, i.e., we compared efficiency/density/modularity distributions between the disinformation recommendation networks generated by a RA, and the journalism (legitimate) recommendation network generated by the same RA. With this other statistical test, we seek to study whether RAs that seemed to facilitate the creation and consolidation of disinformation networks also had similar impacts on other network types, i.e., legitimate networks, or whether they impacted specifically disinformation networks. This is essential for our research, as we intend to gain understanding on RAs that bolster disinformation network creation and consolidation in specific, rather than detecting algorithms that enhance the network metrics that characterize disinformation networks in any network type (i.e., legitimate networks).

Additionally, to study content dynamics from the disinformation sub-part of a recommendation network to its legitimate sub-part, we employed other metrics aimed at reliably measuring these content dynamics. First, we calculated the average distance from the journalism network to the disinformation network, d_r. It is imperative to understand the direction of this relationship, because, as it was aforementioned, the recommendation network generated by a RA during a simulation relates two nodes in a directional manner. In such a relationship, a node is linked to another if and only if it receives a recommendation from it. Calculating the average distance to go from the journalism network to the disinformation network (pairwise) allows us to understand how many "steps" are required for content

to reach the legitimate network from the disinformation network. Similarly, we also computed the number of directional edges that interconnect journalists to disinformation actors. This metric, that we called *Disinformation-to-Legitimate Linkage* ($\overline{DLL}$), provides an intuition of which algorithms bolster the propagation of disinformation through the legitimate network. Furthermore, we also computed URL_{EC} and $Post_{EC}$, which refer to the URL Exposure Count and the Tweet Exposure Count from journalists to (the content of) disinformation actors. Indeed, these two measures complement the previous ones in the intuition of which RAs favor the expansion of disinformation through legitimate networks. Last, we computed the Structural Diversity Index (*SDI*)[MH24]. Such an index is used to understand the diversity in a social network from a topology-based perspective. In this case, it allowed us to understand whether each RA favors the creation of more diverse ecosystems (shared by both disinformation actors and journalists) or whether they favor the creation of less diverse ecosystems.

To further complement the answer to our research questions, we provide figures displaying the generated networks during our explanation. Such figures contain, strictly, the largest (weakly) connected component. Focusing on the largest (weakly) connected component in a network analysis is crucial for several reasons, as it highlights the core structure of the network [New10], where the majority of significant interactions occur, offering a clear view of the main patterns and behaviours without the clutter of less relevant data [New10]. This approach simplifies the analysis, reducing computational demands and enhancing the interpretability of the results. It ensures that the findings are both relevant and manageable, concentrating on the essential dynamics that define the network's overall behavior.

Last, we undertake the comparison between network and accuracy metrics to comprehensively understand the impact of certain aspects that influence the generation and consolidation of disinformation networks on these accuracy metrics. Through this comparison, our goal is to elucidate the ways in which the structural and dynamic properties of disinformation networks affect their ability to produce and maintain accurate or misleading information. This analysis is pivotal for identifying the leverage points within these networks that could be targeted to disrupt or mitigate the spread of disinformation, thereby contributing to the development of more effective strategies for safeguarding the integrity of public discourse. In order to carry out this comparison, we used accuracy metrics that have already been used in the literature for the purpose of analyzing the results of RAs. More specifically, we use four different metrics. First, we use the accuracy metric to gain an understanding of the proportion of simulated retweets that are actual retweets considering the test data. Similarly, we also used the recall metric, so as to calculate the proportion of retweeted content that was being detected by each RA. Last, we also leveraged the use of evaluation metrics that combine both of the already described accuracy and recall. More specifically, we used F1 and F2 metrics to provide a balanced view of the accuracy of each RA.

9.4 Results

In this section, we address the proposed research questions that were described in Section 9.1 based on the research performed. In order to give an answer to **RG3.RQ1** we analyzed the impact of the specified algorithms in Section 9.4.1, leveraging the use of general network properties, together with other metrics to quantify the impact of each algorithm in the generation of disinformation networks. Furthermore, we explain the kind of network that each approach tends to create in Section 9.4.2, whereas the analysis of content diffusion and information dynamics from the disinformation network to the rest of the online social network is described in Section 9.4.3. Last, we discuss the relationship between the factors that are

related to the generation of disinformation networks and the accuracy, for the studied recommendation algorithms in Section 9.4.4.

9.4.1 The impact of RAs in disinformation networks generation

To study the impact of RAs in the generation of disinformation networks, we first calculated the network metrics and properties described in Section 9.3.4, together with the standard deviation related to each metric's distribution across all time windows (for each RA). These results are described in Tables 9.1 and 9.2. In such tables, columns specify the acronyms used for each RA, whereas rows describe the average values of the calculated network metrics and properties together with the standard deviation related to them. This adjustment aligns with the discrete nature of the metrics, ensuring that their representation is accurate and intuitive. For that reason, discrete measures were rounded to the units, whereas continuous properties were rounded to the third decimal figure. The only exception to that rule is the metric $\overline{IncPR_w}$ because its values are on a significantly smaller scale. In that case, we decided to use scientific notation to facilitate its understanding. Indeed, Table 9.1 contains the topology-based (average) metrics of the (disinformation) recommendation networks generated by each RA, whereas Table 9.2 contains the content-based (average) metrics. It must be noted that in such tables, certain cells are written in red and blue colors. These cells represent the maximum or minimum values for every metric, and they are used to visually understand if they facilitate the generation of disinformation networks (red) or if they prevent their generation (blue).

Table 9.1: Recommendation network (disinformation) average network-based metrics by RA.

Metric	Rnd	LIWC	W2V	MD-W2V	CF	TS	TS-W2V	DR
Nodes	416 ± 89	408 ± 88	416 ± 89	407 ± 87	357 ± 89	361 ± 89	360 ± 89	338 ± 86
Edges	875 ± 337	648 ± 204	838 ± 306	383 ± 96	805 ± 333	541 ± 180	551 ± 185	758 ± 307
$\overline{Deg}$	8.430 ± 0.462	3.440 ± 0.175	8.412 ± 0.439	8.495 ± 0.413	7.854 ± 0.587	7.714 ± 0.540	7.670 ± 0.594	7.837 ± 0.686
D	13 ± 5	8 ± 5	11 ± 4	3 ± 2	12 ± 5	3 ± 2	4 ± 2	10 ± 5
E	0.059 ± 0.015	0.018 ± 0.004	0.047 ± 0.010	0.025 ± 0.004	0.061 ± 0.016	0.026 ± 0.005	0.026 ± 0.005	0.055 ± 0.012
Δ	0.005 ± 0.001	0.004 ± 0.001	0.005 ± 0.001	0.002 ± 0.001	0.006 ± 0.001	0.004 ± 0.001	0.004 ± 0.001	0.007 ± 0.001
Q	0.829 ± 0.039	0.846 ± 0.036	0.858 ± 0.033	0.971 ± 0.007	0.801 ± 0.044	0.903 ± 0.021	0.898 ± 0.022	0.793 ± 0.044
$\overline{CC}$	0.017 ± 0.004	0.016 ± 0.005	0.020 ± 0.005	0.018 ± 0.004	0.021 ± 0.006	0.021 ± 0.006	0.022 ± 0.006	0.024 ± 0.006
$\overline{EVC}$	0.002 ± 0.002	0.002 ± 0.002	0.002 ± 0.002	0.002 ± 0.002	0.003 ± 0.002	0.003 ± 0.003	0.003 ± 0.003	0.003 ± 0.002
$\overline{PR}$	$25e^{-4} \pm 0.001$	$26e^{-4} \pm 0.001$	$25e^{-4} \pm 0.001$	$26e^{-4} \pm 0.001$	$31e^{-4} \pm 0.001$	$30e^{-4} \pm 0.001$	$30e^{-4} \pm 0.001$	$33e^{-4} \pm 0.002$
$\overline{IncPR_w}$	$2.83e^{-5} \pm 0.001$	$1.40e^{-5} \pm 0.001$	$1.11e^{-5} \pm 0.001$	$4.24e^{-6} \pm 0.001$	$4.73e^{-5} \pm 0.002$	$3.33e^{-5} \pm 0.002$	$2.65e^{-5} \pm 0.002$	$4.22e^{-5} \pm 0.002$

Table 9.2: Recommendation network (disinformation) average content-based metrics by RA.

Metric	Rnd	LIWC	W2V	MD-W2V	CF	TS	TS-W2V	DR
$Topics_w$	1361 ± 554	1268 ± 499	1315 ± 517	1207 ± 439	1348 ± 558	1119 ± 427	1121 ± 434	1204 ± 489
%TP	0.575 ± 0.092	0.572 ± 0.091	0.578 ± 0.092	0.570 ± 0.090	0.574 ± 0.095	0.565 ± 0.092	0.567 ± 0.093	0.572 ± 0.095
URL_u	49.901 ± 16.349	48.857 ± 16.444	46.421 ± 15.512	43.495 ± 13.906	53.626 ± 17.591	40.939 ± 13.412	41.070 ± 13.698	52.178 ± 17.439
%UP	0.810 ± 0.105	0.770 ± 0.100	0.819 ± 0.104	0.758 ± 0.097	0.795 ± 0.111	0.776 ± 0.106	0.777 ± 0.105	0.770 ± 0.111
$\overline{CA_u}$	41.452 ± 15.587	36.548 ± 14.504	38.302 ± 14.662	42.349 ± 15.197	44.814 ± 17.004	30.440 ± 11.516	30.657 ± 11.814	38.345 ± 15.024

Considering the results obtained, we observe that there are significant differences between the RAs. For instance, the *Nodes* metric varies between 338 (DR) and 416 (Rnd), which makes a difference of almost 100 users. This difference can also be noticed in other metrics, such as the diameter, which takes values varying between 3 (MD-W2V, TS) and 13 (Rnd) or even the edges metric, which represents the recommendations between different pairs of users. In that case, the metric varies from 333 (MD-W2V) to 875 (Rnd). Thus, we observe that, in general, there are very different values for most of the metrics in both tables. There exist, however, metrics that remained quite similar from one approach to another, as it is the case with the percentage of topics that persist from one window to another (%TP), thus we observe how no algorithm achieves very different results when it comes to keeping the treated topics of

conversation for a long time or refreshing them very frequently, i.e, no algorithm takes very low or high values and, instead, all of them have quite similar values.

In any case, what we can extract from the results is that different approaches receive different values in the analyzed properties and metrics of disinformation networks. Indeed, some approaches tend to generate more recommendations between different users, such as the Random RA (Rnd), which explains the high number of nodes, edges, degree and the different observed metrics for the algorithm. However, some other RAs tend to focus on other aspects, precisely the Word2Vec-based recommendation system (W2V) leads to the creation of networks with fewer nodes, edges, and degree, as a result of less exploration and more exploitation in its results when compared to other algorithms, like the aforementioned Random recommendation system (that applies an opposite policy). This is further discussed in Section 9.4.2. In the following section, we further discuss these results and analyze the algorithms that contribute to facilitating the generation of disinformation networks, and also those algorithms that seem to not facilitate the generation and consolidation of these networks.

RAs that facilitate disinformation networks generation

Based on the characteristics of disinformation networks that were discussed in Section 9.2 and considering the results obtained in Section 9.4.1, we can observe that certain algorithms facilitate the generation of disinformation networks.

More specifically, DeepRank (DR) stands as the algorithm that most contribute to the generation of disinformation networks. This observation can be justified based on the computed network-based and content-based metrics shown in Tables 9.1 and 9.2. First, we observe that DR tends to create recommendation networks where fewer users are recommended by the algorithm, because it tends to focus always on the same kind of users, more specifically, users that give algorithm more accuracy to lower the loss function, which are also users with really high relevance in the generated recommendation network (highest average network PageRank) and users that tend to increase their relevance as time windows go by (second-highest PageRank increment per window). We also observe how this algorithm creates a valid number of edges between the nodes, i.e., the algorithm provides a decent number of connections between a relatively small number of users, which increases the density (Δ) of the recommendation networks (known to be a key factor in disinformation networks generation. Furthermore, DR generates very efficient networks (third algorithm with more average network efficiency) and clearly creates networks with less modularity and greater clustering coefficient. Indeed, the last observation allows us to understand that DR tends to create networks where information spreads rapidly, as there are less apparent communities with higher cohesion, when compared to the (disinformation) recommendation networks generated by other approaches. Moreover, we also observe that DR is one of the algorithms that creates networks with few topics (third algorithm in this aspect), while also being one of the algorithms that tends to keep more topics from one window to another ($\%TP = 0.572$). Additionally, the (disinformation) recommendation network that DR tends to create leads to the second-highest diffusion of URLs per user and the second-highest increase in user relevance (PageRank) from one window to another, and the third-highest score of content amplification per user ($\overline{CA_u} = 38.345$) of any approach. All in all, DR stands as the algorithm that most contributes to the generation of disinformation networks, as it creates very efficient and dense recommendation networks, where information spreads rapidly, due to a smaller fragmentation of the network and more cohesion between the users that belong to it, whose influence

tends to evolve both positively and more quickly over time when compared to other approaches.

To justify this observation, we applied the Mann-Whitney U test[2], as explained in Section 9.3.4. In this case, with 95% confidence we can say that the null hypothesis was rejected, pairwise, for all the three most relevant metrics that characterize disinformation networks (efficiency, density and modularity). The only cases where the rejection was closer were the efficiency distribution between DR and Rnd (U $= 20252.0$, $\rho = 0.0023$), the density distribution between DR and W2V (U $= 19549.0$, $\rho = 0.0176$), and the modularity distribution between DR and CF (U $= 19381.0$, $\rho = 0.275$), all of them smaller than $\alpha = 0.05$ (95% confidence), as both approaches tend to generate non-fragmented networks with very cohesive communities.

Similarly, we observe that, to a smaller extent, CF does facilitate the generation of disinformation networks for the same reasons. This approach tends to recommend a small number of users, with many connections, very high network efficiency (the highest efficiency of the implemented RAs) density and clustering coefficient, with a low modularity, which suggests the generation of (disinformation) recommendation networks where information spreads swiftly and efficiently. Moreover, the users recommended tend to experience a high relevance increase from one window to another, which is an important characteristic of the disinformation networks. Moreover, in the generation of disinformation networks, the users tend to disseminate a very high number of URLs (the highest number of average URLs disseminated per user), with a relatively high user persistence from one network to another, which is (again), a characteristic of disinformation networks.

In the same line, W2V is the last algorithm that seems to facilitate the generation and consolidation of disinformation networks, as it creates (disinformation) recommendation networks where there are very low values for user and topic persistence, i.e., the networks always lead to similar narratives treated by the same users, window after window. Furthermore, these networks are both efficient and dense ($E = 0.047, \Delta = 0.005$), while also keeping fragmentation to a moderate value ($Q = 0.858$), that devises the creation of echo chambers in the generated recommendation network, which are also a well-known characteristic of disinformation networks. Additionally, the resulting recommendation networks are well-connected ($\overline{Deg = 8.412}$) and the networks are very cohesive (high value of clustering coefficient), which are again known qualities of disinformation networks.

Concerning the rest of the approaches, although some of them exhibit characteristics that bolster the generation of disinformation networks, they also have certain aspects that are not in line with such an observation. For example, MD-W2V creates moderately efficient and dense networks, however, this approach also creates networks with higher modularity ($Q = 0.971$), less clustering coefficient (more fragmented and less cohesive communities), and, additionally, users do not experience an important network influence (PageRank) increment as the time windows pass by. This example analysis remains valid for the rest of the approaches, thus DR and CF stand as algorithms that facilitate the generation of disinformation networks.

As mentioned in Section 9.3.4, we applied the Mann-Whitney U test to understand whether the RAs that seemed to facilitate the creation and consolidation of disinformation networks had a similar impact on other network types, such as the legitimate network, or whether those RAs only had this kind of impact in the generated recommendation networks. The results obtained confirm and extend the previous discussion. First, the network metrics of the disinformation recommendation net-

[2]Between different algorithms, pairwise.

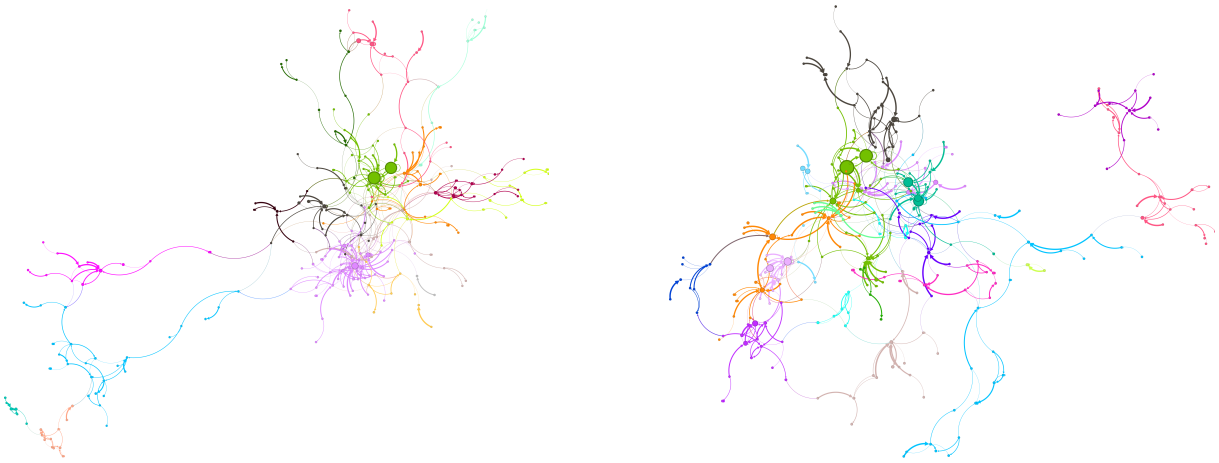
works generated by DR across all windows are significantly different to the ones that DR creates for journalists, thus demonstrating its capability to enhance the creation and consolidation of disinformation in specific, instead of having such an impact in any network type. More specifically, the Mann-Whitney U tests derived $U = 14953.0, \rho = 0.0358$ (modularity), $U = 20487.0, \rho = 0.0010$ (efficiency), and $U = 20939.5, \rho = 0.0002$ (density), allowing us to reject the null hypothesis, H_0, for all the three network metrics, accepting the alternative hypothesis, H_1, i.e., there is a significant difference between the disinformation recommendation networks generated by DR and the journalism recommendation networks it generated, in terms of how information can flow through such generated networks. Similar results were obtained for the W2V algorithm, where the generated disinformation network is significantly denser and more efficient ($U = 20281.0, \rho == 0.0021$, for the efficiency test; $U = 21513.0, \rho = 1.89e^{-5}$, for the density test), although the results of the Mann-Whitney U test involving modularity did not allow to reject the null hypothesis (with 95% confidence), as the obtained p-value was higher than 0.05 ($U = 15489.0, \rho = 0.1891$). Indeed, W2V facilitates the generation and consolidation of disinformation recommendation networks in particular, as its effect is not so pronounced in other network types, such as legitimate recommendation networks. However, its ability to create echo chambers linked to the different topics that conform the social network's conversation is similar for any network in general, as it tends to suggest content recommendations that lead to the appearance of echo chambers in those networks.

The application of statistical tests also allowed us to reject the idea of CF being a RA that facilitates the generation and consolidation of disinformation networks in specific. The obtained results ($U = 15489.0, \rho = 0.1146$ for modularity, $U = 18924.0, \rho = 0.0783$ for efficiency, and $U = 16076.0, \rho = 0.3139$ for density) demonstrate (with confidence 95%) that CF generates the same kind of recommendation networks regardless of whether the network is formed by disinformation actors or legitimate users. The approach clearly creates very efficient, dense, and barely fragmented networks, however, those characteristics are not exclusive to disinformation networks, but they are shared across any other network provided to this algorithm. As a result, although CF can generate networks with characteristics that resemble those of a disinformation network, it does not really enhance disinformation creation and consolidation in specific.

Figures 9.1a and 9.1b show recommendation networks generated by DR and W2V in the same window (100th window of the simulation starting from 1st Jan 2019). As it can be seen[3], both approaches generate networks which are very dense ($\Delta_{DR} = 0.007$ and $\Delta_{W2V} = 0.006$), quite efficient ($E_{DR} = 0.0467$, $E_{W2V} = 0.0525$), and barely fragmented, as it can be seen from their small modularity ($Q_{DR} = 0.811$, $Q_{W2V} = 0.785$). As has already been discussed, these network properties facilitate the swift spread of (dis)information through the recommendation network, helping the generation and consolidation of disinformation networks.

Last, although we observe that certain algorithms have proven that they facilitate the generation and consolidation of disinformation networks, other RAs seem to enhance only certain network properties that characterize disinformation networks, at the cost of sacrificing other properties of them. For instance, Reinforcement-Learning based algorithms, like the implemented Thompson Sampling-based approaches, seem to foster the generation of moderately dense networks, with relatively small efficiency and high fragmentation. In a sense, these networks facilitate the generation of disinformation networks

[3]The representation only considered the largest weakly connected component.



(a) DR recommendation network.

(b) W2V recommendation network.

Figure 9.1: Recommendation networks created by RAs that facilitate the consolidation of disinformation networks.

in certain aspects. For example, these networks tend to treat a very reduced set of topics, they are quite dense, and the diameter is small, which facilitates the spread of information. However, they also have other characteristics that do not follow the known characteristics of a disinformation network, like the relatively small efficiency they achieve. Similarly, other content-based approaches, like LIWC, are unable to capture the nuances of the style of speech of disinformation networks in the same way as more complex approaches do (such as DR and W2V). As a result, not only the LIWC RA achieves smaller accuracy values, but it also generates networks with characteristics of disinformation networks (such as the moderate efficiency and relatively low fragmentation), but also characteristics that are opposed to disinformation networks (such as the extremely low efficiency of the generated networks). In the end, disinformation networks are characterized by the quick spread and massive amplification of content, through slogans, hashtags, URLs, and retweets, just as we explained in Chapter 6. Algorithms like W2V and DR can capture these nuances and achieve higher accuracy values while enhancing the generation and consolidation of disinformation networks, however, the inability of LIWC to capture these nuances does not allow the algorithm to facilitate the creation and consolidation of disinformation networks.

These conclusions are maintained even when changing the main parameters in the set of RAs that can be parametrized. For instance, the implemented collaborative filtering RA is based on KNN, which is also used (to some extent) in the implemented RL-based RAs (TS and W2V-TS). All the results described until now are related to simulations that involved three neighbors ($k = 3$), however, the obtained results are consistent even when using a different parametrization (more extreme values were tested, including $k = 1, 2, 5, 7$and9). For example, using just one neighbor ($k = 1$), CF is capable of creating very efficient networks ($E = 0.062$), relatively dense networks ($\Delta = 0.039$) and moderately fragmented networks ($Q = 0.801$). Furthermore, the conclusions related to the statistical tests are also maintained. For this same example ($k = 1$), we observe that the distribution of network metrics calculated for the disinformation networks is not significantly different (considering the main network metrics: efficiency, density and modularity) to the distribution obtained for the journalists' networks. For example, considering the efficiency metric, we observe how the Mann-Whitney U test fails to reject the null hypothesis, as $U = 18351.0, \rho = 0.2288$, meaning that we do not have statistical evidence to prove that the effect of CF is significantly different in disinformation networks with respect to the journalists' networks (i.e.,

the algorithm does create more efficient, dense and less fragmented networks, but the effect is the same for both disinformation and journalism networks).

Topology and content-based RAs and disinformation networks generation

Last, we provide an answer to RG3.RQ1 based on the results shown in Section 9.4.1. As it was already discussed, certain algorithms seem to facilitate the generation of disinformation networks, such as DR and CF, however, some other algorithms seem to partially prevent and partially facilitate the consolidation of disinformation networks.

Considering the obtained results, we observe that different kinds of algorithms can facilitate the generation and consolidation of disinformation networks, but through very different approaches. First, content-based algorithms like W2V and DR (although DR is also structure-based to some extent, as the approach is hybrid) prove to foster the generation of disinformation networks in specific, as opposed to other kinds of networks, like the legitimate networks, where the effects are completely different (the generated disinformation networks are proven to be significantly more efficient, denser, and less fragmented when compared to the generated legitimate counterparts). Considering the results shown in Section 9.4.4, we observe how these algorithms can significantly boost the generation of very efficient, dense, and barely fragmented networks while keeping very high accuracy values. Indeed, disinformation networks hinge on the creation, manipulation, and distribution of content that is specifically designed to mislead, influence, or exploit the beliefs and perceptions of target audiences (maximizing content amplification), as we already know from the research related to disinformation networks presented in Chapter 6. Such an observation is represented in these results, where content-based RAs reach the highest RAs accuracy values while facilitating the generation and consolidation of disinformation networks. However, this seems to only occur in content-based RAs capable of effectively capturing the nuances of the style of speech employed in these networks, which is the reason why other approaches, like LIWC, cannot facilitate the generation and consolidation of disinformation networks.

On the other hand, we observe how disinformation also involves a community-based (social or structure-based) component, besides the well-known content-based component. In this sense, RAs like CF clearly generate extremely dense and efficient networks (with very little fragmentation), which are network properties that characterize disinformation networks. However, CF's effect on the generated networks is not specific to disinformation networks, as it also happens in legitimate networks. As a result, CF is not as dangerous as DR or W2V can potentially be, because its effect will boost the quick spread of information in both disinformation and legitimate networks, as opposed to content-based algorithms, such as DR or W2V, which only enhance the generation of disinformation networks.

Considering both the content-based and the community-based components of disinformation, it can be seen that the content-based component is slightly more relevant to disinformation networks, as content-based algorithms allow the generation of efficient, dense, and barely fragmented disinformation networks while keeping extremely high accuracy values in the RA (e.g., W2V).

Last, although we observe that certain algorithms facilitate the generation and consolidation of disinformation networks, the same does not occur with the prevention (block) of the generation of such networks. Based on the results, none of the selected RAs can be used to effectively block the generation and consolidation of disinformation networks, as they tend to offer a trade-off between network properties that characterize disinformation networks, and some others that characterize legitimate networks.

9.4.2 Recommendation networks per RA

To provide a more complete answer to the addressed research question, we aim to gain an understanding of the kind of disinformation networks that are generated by each RA. Considering the results presented in Tables 9.1 and 9.2, we can give a justified answer to this question. First, considering the baseline recommendation algorithm (**Rnd**), we observe that it **tends to generate centralized, dense, efficient and topic-diverse networks**. As it can be seen, this algorithm generates networks with many users of small relevance (decentralized network where users have very small PageRank values and Eigenvector centrality values - see Table 9.1). We observe that users tend to have many connections, and the resulting networks are both efficient and dense, although they are also less cohesive and slightly fragmented (moderate modularity). Furthermore, we observe how this algorithm bolsters topic diversity (1361 topics treated per window, on average) and favors content amplification through both URLs and RTs. This baseline algorithm is far from facilitating the generation of disinformation networks, as the approach creates a very diverse network, with many users and interconnections between them, however, the network is centralized which does not help to improve the flow of information from border users to the rest of the network, i.e., the high network diameter hinders the swift exchange of information between nodes that are not found at its center.

Continuing our analysis, we observe that the LIWC recommendation algorithm generates fragmented, monothematic, inefficient and sparse networks, characterized by having many users with a small number of connections among them. Furthermore, we observe that the network treats a rather low number of topics when compared to the networks generated by other RAs (such as Rnd or W2V). We also observe that users tend to have a higher percentage of rotation from one window to another (less persistence), which explains well the low relevance of the users within the network (fragmented networks, with users that tend to have a reduced number of connections and that rotate from one window to another) and the very small average relevance increase ($\overline{IncPR_w}$). Not only the algorithm creates networks where information spreads slowly due to the low efficiency and density, and high modularity (around 0.85), but it also prevents content amplification when compared to similar RAs, as the average URLs shared per user and window, and the average number of RTs are very small in comparison with others. As a result this RA generates networks that do not follow the well-known characteristics of a disinformation network, as the approach creates a network where users are linked to a very limited number of neighbors, restricting information dynamics and preventing the swift flow of information across the whole network.

On the other hand, the Word2Vec algorithm creates moderately fragmented networks, with high density and moderate efficiency. The most relevant aspect of the networks created by this approach is, indeed, the very high user persistence from window to window, as only 18% of the users rotate from one window to another. In other words, W2V tends to create networks where the same users are maintained over time, although due to the fragmentation of the network and its low cohesion, these users tend to not have very high relevance on average. In general, the approach creates stale networks where users and topics have very small rotation rates, i.e., they tend not to change from window to window, which can lead to the creation of echo chambers and a slightly higher fragmentation of the network when compared to other approaches [VPV21]. As it can be seen, the W2V RA generates networks where information is exchanged between the same users window after window, as it helps the same topics to be treated from one window to another. This information can be spread through the network with moderate swiftness, as the approach helps to create relatively efficient and dense networks. Indeed, the shape and information

dynamics of the generated (disinformation) recommendation networks make W2V a good algorithm for creating and consolidating disinformation networks, as it was already described in Section 9.4.1.

Considering MD-W2V, we observe how the approach generates monothematic, inefficient, sparse and fully fragmented networks($Q = 0.971$). Indeed, this algorithm brings to the extreme the ability of creating echo chambers where topics are discussed. As a result, users have lots of connections (high degree) within their echo chamber, but the network has a very high fragmentation. The most important feature is, indeed, the high content amplification, which is a direct consequence of the appearance of echo chambers that are efficiently interconnected internally, but not externally (low total network efficiency). While it is true that this approach generates more centralized networks with a very short diameter that facilitates, to some extent, the spread of information through the network, the extremely high fragmentation leads to the creation of many different user communities and significantly prevents the efficient exchange of information from one community (echo chamber) to another. Indeed, the resulting networks are not only highly fragmented, but they are also sparse and inefficient, even if the centralized shape seemed ideal for information diffusion. As a result, this RA does not help to create and consolidate disinformation networks.

Considering both Multi-Armed Bandit recommendation algorithms (TS and TS-M2V), we observe the trend of creating inefficient, sparse and centralized networks with very low user rotation. These networks are characterized by a small diameter, thus information spreads across the network without requiring much amplification, which does not help to improve the efficiency as only certain users have many connections whereas the rest are sparsely connected to the rest of the network, which is the reason why these networks have high average degree (due to the high degree of the main nodes in the network) and very low efficiency and density. Due to their characteristics, these networks make a harder task of content amplification, as both URL sharing and RTs publication reach minimum values (for both TS and TS-W2V). Last, these networks tend to treat a small number of topics that easily vary from one window to another (low persistence). The shape and information dynamics of the (disinformation) recommendation networks created by both RAs are alike the recommendation networks created by MD-W2V, as they are very centralized networks, with extremely short network diameters that help diffuse information from one side of the network to another. However, the high levels of fragmentation tend to prevent the dissemination of information between different user communities (echo chambers), and, as a result, the generated networks are both very sparse and inefficient. Thus, both Reinforcement Learning-based algorithms do a poor job of facilitating the creation and consolidation of disinformation networks.

As opposed to other structure-based algorithms, CF creates very efficient, highly dense networks that experience little fragmentation and are very cohesive. These networks tend not to have many users, but the users tend to persist over time, gaining more and more relevance as windows go by. In fact, these networks are so efficient and dense, and there is that little rotation of users, that they have the highest average user relevance increase per window of any RA (PageRank increases by $4.73e^{-5}$ on average across all windows). Furthermore, these networks span over a wide range of topics, which tend to change from window to window (low persistence). As these networks are so well-connected, content amplification is maximized, as both URL sharing and RTs publication tend to have the highest value, on average, across all simulated windows. Indeed, CF creates centralized networks with very high diameters. While it is true that information requires more "steps" to flow from one side to another side in the network, the very low fragmentation significantly improves these information dynamics, leading to the creation of extremely

efficient and dense (disinformation) recommendation networks. As a result, the generated (disinformation) recommendation networks bolster the quick spread of potentially misinformative content, thanks to the shape of the generated recommendation networks.

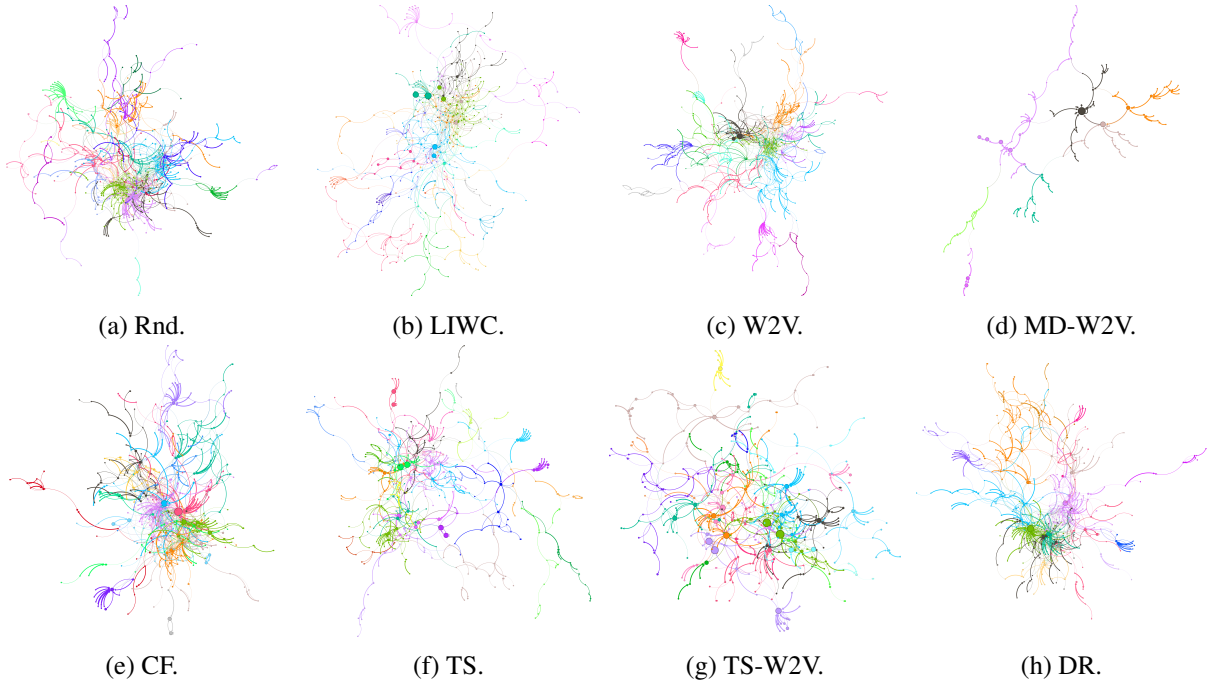
Last, we observe that DeepRank tends to create networks with similar characteristics to the ones observed in CF. These networks are very efficient and dense, while being very cohesive and less fragmented than the networks created by the other RAs. It is for those reasons that information flows swiftly and efficiently across these networks. Furthermore, users tend to be very relevant, and this relevance is improved from window to window across the users that are kept from one window to another (around 77% of them). DR recommendation networks tend to be monothematic, although these topics tend to easily change from one window to another, and content amplification is enhanced by the intrinsic characteristics of the generated networks. Similar to CF, DR stands as an algorithm capable of creating centralized networks. DR also reduces the diameter of the generated networks, fostering the swift exchange of information in the network. Moreover, the network is not only centralized but also very well-connected (very dense networks), which further enhances the dissemination of potentially misinformative and harmful content through the network. As it was explained in Section 9.4.1, the shape and information dynamics of the generated recommendation networks make this algorithm a good candidate for facilitating the creation and consolidation of disinformation networks.

An example recommendation network generated by each RA for the same window (window 158, from 11 Jan. 2022 to 17 Jan. 2022[4]) is given in Fig. 9.2, where the previous characteristics are shown. In such figures, we observe how the random algorithm creates an efficient and dense network that is not focused just on one user (decentralized), instead, the connections are well-spread between different users in the network. For the case of LIWC, we observe an inefficient and sparse network that lacks edges to successfully allow information to quickly spread through it. Additionally, such a network is also quite fragmented. Considering W2V, it is more efficient and denser than the previous network, yet it leads to the creation of echo chambers surrounding the center of the network, which proves the facility that this algorithm has for creating echo chambers. Even more extreme is the case of MD-W2V, whose network is very small as it is a compound of many small-sized weakly connected components. This network is extremely fragmented and does not provide an efficient and dense channel for information to flow through it. As opposed to these inefficient networks, we find the recommendation network created by CF. Such an algorithm creates a very efficient and dense network, which is quite cohesive and happens to not be as fragmented as the other ones. Similar features are observed for the DR network. Last, TS and TS-W2V are quite sparse networks, centralized in certain users which have a significant relevance (size of the nodes). These networks are both sparse and inefficient, and they also suffer from high fragmentation, as described previously. All in all, we observe how different RAs create different recommendation networks, whose characteristics can help facilitate (or prevent) the generation and consolidation of disinformation networks.

9.4.3 Content diffusion

In our research into the dynamics of content diffusion between disinformation and legitimate information networks, we employ the described recommendation systems and their content recommendations to trace and quantify the flow of information from users belonging to the disinformation network to users

[4]Both days included.



(a) Rnd. (b) LIWC. (c) W2V. (d) MD-W2V.

(e) CF. (f) TS. (g) TS-W2V. (h) DR.

Figure 9.2: Recommendation networks created by each RA.

Table 9.3: Average content diffusion metrics per full recommendation network of each RA.

Metric	Rnd	LIWC	W2V	MD-W2V	CF	TS	TS-W2V	DR
$\overline{d_r}$	6 ± 2	4 ± 1	5 ± 1	2 ± 1	5 ± 1	2 ± 1	2 ± 1	$5 + 2$
$\overline{DLL}$	3225 ± 677	1290 ± 270	3211 ± 675	3159 ± 657	2515 ± 637	2488 ± 632	2468 ± 633	$2376 + 628$
URL_{EC}	15582 ± 7037	12779 ± 5484	14146 ± 6154	10541 ± 4115	15369 ± 6969	10243 ± 4157	10302 ± 4258	12563 ± 5483
$Post_{EC}$	27215 ± 13653	21725 ± 10532	24922 ± 11916	20123 ± 8586	26882 ± 13548	16086 ± 7126	16314 ± 7394	20896 ± 10285
SDI	0.683 ± 0.203	0.819 ± 0.218	0.749 ± 0.183	1.813 ± 0.216	0.710 ± 0.209	1.156 ± 0.222	1.118 ± 0.208	0.704 ± 0.190

belonging to the legitimate network (journalists). Our analysis focused on identifying patterns and pathways through which disinformation infiltrates and propagates through these legitimate networks. This exploration is crucial to understand how disinformation gains credibility and spreads across different segments of the public discourse, thereby influencing societal perceptions and behavior. Considering the content diffusion metrics that were discussed in Section 9.3.4, we obtained the results in Table 9.3.

Considering these results, we observe how CF stands as one of the best algorithms for enhancing content dynamics from the disinformation network to the journalism network, as it creates a significant number of links (2515) from journalists to disinformation actors (journalists receive content recommendations from disinformation actors), while also amplifying significantly the content of the disinformation network that reaches the legitimate network. When compared to our baseline RA (Rnd), we observe how CF creates networks with a similar influence of the disinformation network on the legitimate network. It is worth noting how due to the behavior of the Rnd RA, the resulting networks maximize the impact of disinformation actors on legitimate users, as random links are created between users that would never be connected in a real scenario. This is coherent with the conclusions derived in Section 9.4.1, as CF does create networks with properties that resemble the network properties of a disinformation network (enhancing efficiency and density, while moderating fragmentation), it has the same effect over any kind of network (e.g., legitimate networks), thus the algorithm is not as dangerous as other studied RAs.

On the other hand, algorithms like DR and W2V, also create networks that maximize the impact of the disinformation network on the legitimate network. All these algorithms that seem to enhance the

dissemination of disinformation content into the legitimate network are characterized by the creation of networks with a small structural diversity index (SDI), which also characterizes disinformation networks, as it implies that information can easily travel through the network without requiring travelling long distances to reach a specific part of it. On the other hand, we also observe that certain algorithms are useful are not as good enhancers of content dissemination from the disinformation network to the legitimate network. For instance, LIWC tends to create recommendation networks with a very small amount of connections from journalists to disinformation actors, thus reducing the capability of sharing content from the disinformation network to the legitimate network. This observation looks even more extreme when compared to the baseline RA, as the difference in created network edges from legitimate users to disinformation actors is almost 2000, which is a very significant amount of interconnections between both networks. Consequently, this algorithm also leads to the creation of slightly more diverse networks when compared to the previous algorithms, as the length (normalized by the number of users in the networks) to travel from one part of the network to another is higher. Similarly, TS, TS-W2V and MD-W2V stand as approaches with reduced content dynamics from the disinformation network to the legitimate network. This observation can be derived from the reduced content amplification that is performed in this network (low values of URL_{EC} and $Post_{EC}$ metrics for all the mentioned algorithms), and also from the high structural diversity index metric values that suggest the creation of networks that foster structural diversity, as opposed to the algorithms that seemed to facilitate content diffusion into the legitimate network (CF, W2V, DR and the baseline RA - Rnd).

All in all, we observe that different RAs lead to the creation of different recommendation networks, which have a significant impact on how disinformation can potentially spread from the disinformation network to the legitimate network. Indeed, these observations must be taken into consideration together with the ones found in Section 9.4.1, as certain algorithms (like DR and CF) are algorithms that foster both the generation and consolidation of disinformation networks, but also networks where content can easily be disseminated into the legitimate network, presenting a potential threat due to the creation of network structures that bolster the quick spread of disinformation through the network.

9.4.4 Disinformation network generation vs RA's accuracy

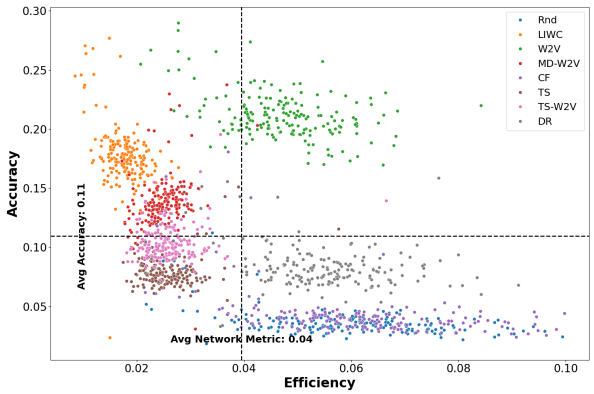
Last, we analyzed the impact of certain network properties and metrics that are known to affect disinformation (as already described during this research) on the accuracy of the RAs. In a sense, we seek to study whether certain kinds of networks, generated by specific RAs, while fostering the generation of disinformation networks do still favor the accuracy of the RA. To do so, we employed some of the network metrics described in Section 9.3.4 which are already known to facilitate the generation and consolidation of disinformation networks.

Before carrying out the aforementioned analysis, we show the average accuracy results obtained for each RA across all windows (on average). These results can be seen in Table 9.4. The bold values indicate the best-performing metric in each row for clarity and comparison.

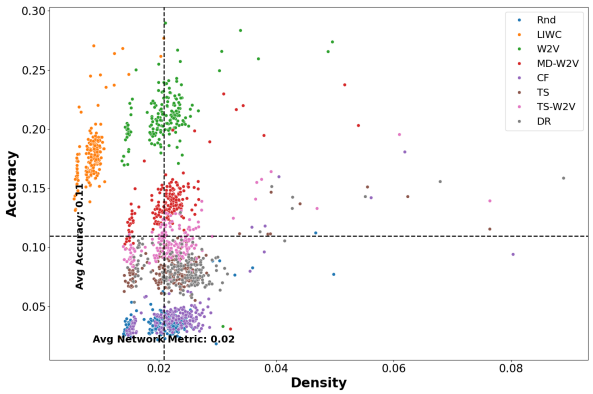
As it can be seen, W2V stands as the algorithm with the highest values for the proposed accuracy metrics. In other words, W2V is the algorithm that is more capable of delivering to each user the content that such a user expects to be recommended (content that follows its preferences). Naturally, this observation is logical, as W2V uses natural processing language techniques to encode in N-dimensional vectors the content of each tweet and the preferences of the user (to then provide content that follows as

Table 9.4: Average accuracy metrics per recommendation network generated by each RA. Bold values indicate the best-performing metric in each row.

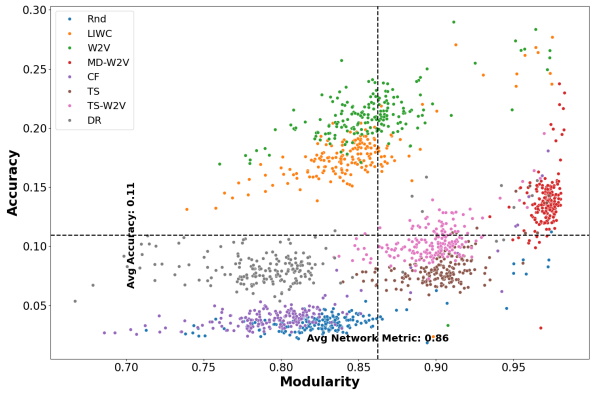
Metric	Rnd	LIWC	W2V	MD-W2V	CF	TS	TS-W2V	DR
Accuracy	0.037 ± 0.012	0.178 ± 0.0260	$\mathbf{0.211 \pm 0.024}$	0.138 ± 0.022	0.043 ± 0.020	0.081 ± 0.014	0.105 ± 0.016	0.083 ± 0.017
Recall	0.001 ± 0.001	0.002 ± 0.003	$\mathbf{0.003 \pm 0.005}$	0.002 ± 0.003	0.001 ± 0.001	0.001 ± 0.002	0.001 ± 0.002	0.001 ± 0.002
F1	0.001 ± 0.001	0.004 ± 0.006	$\mathbf{0.005 \pm 0.008}$	0.004 ± 0.005	0.001 ± 0.002	0.002 ± 0.003	0.002 ± 0.004	0.002 ± 0.003
F2	0.001 ± 0.001	0.003 ± 0.004	$\mathbf{0.004 \pm 0.006}$	0.002 ± 0.004	0.001 ± 0.001	0.001 ± 0.002	0.001 ± 0.003	0.001 ± 0.002



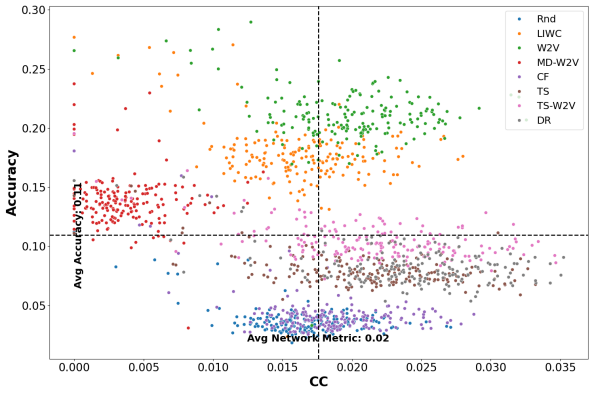
(a) Accuracy vs Efficiency.



(b) Accuracy vs Density.



(c) Accuracy vs Modularity.



(d) Accuracy vs CC.

Figure 9.3: Recommendation networks created by each RA.

much as possible those user preferences). Similarly, algorithms following a similar approach, like LIWC or MD-W2V, are capable of achieving high accuracy values, although both of them are still below W2V, which seems to solve the content recommendation problem to users in the best possible way considering all the implemented RAs.

However, we also observe that other algorithms do not seem to significantly improve the accuracy results achieved by the random algorithm. For instance, although CF is slightly better than Rnd, both have very similar accuracy metrics ($Accuracy_{Rnd} = 0.037, Accuracy_{CF} = 0.043$). On the other hand, some other structure-based and hybrid approaches do seem to provide results with decent accuracy, as it is the case with TS, TS-W2V and DR.

Following our analysis, in Fig. 9.3 we show the comparison of efficiency, density, modularity and clustering coefficient with respect to accuracy, as calculated for each window in the disinformation network generated by the RA, so as to characterize how the accuracy achieved by the employed RAs impacts the generation and consolidation of disinformation networks, based on network properties that are proven to be related to the generation and consolidation of disinformation networks.

In the observed results, we see how the algorithms that seemed to facilitate the generation and consolidation of disinformation networks (DR and W2V) stand as algorithms with high values of network efficiency and accuracy. More specifically, W2V is the algorithm that achieves very high RA accuracy values, while keeping network efficiency in very high values across many time windows (see Fig. 9.3a). While W2V duplicates the accuracy value of DR (on average), DR also achieves very high network efficiencies, while keeping accuracy at a reasonable level. Thus, these algorithms are quite dangerous (overall W2V), as they improve the accuracy while enhancing the generation and consolidation of disinformation networks, through boosting the network properties that resemble those networks (only in those network types and not in all network types, i.e., the same effect does not occur on the legitimate network). Considering other algorithms, like LIWC and MD-W2V, the RA achieves high relatively high accuracy values, however, the generated networks are way fragmented and such fragmentation compromises their efficiency. Considering CF, it generates extremely efficient networks, but compromises accuracy to a large extent, and its effect of improving efficiency, density and moderating modularity is also present on the legitimate network, thus the algorithm is not as big of a threat, as W2V and DR could be.

Similarly, with respect to the network density and accuracy comparison, we observe that certain approaches, like LIWC, tend to generate very sparse disinformation networks (due to echo chambers generation) where accuracy is relative high. This is quite understandable, as the LIWC RA recommends content from users that treat similar topics based on the LIWC categories of the discourse, however, this also creates echo chambers, which lead to sparser networks. However, some other approaches are capable of creating very dense disinformation networks where information can flow very quickly, as it is the case with DR, Rnd, or CF, as already discussed in Section 9.4.1. Again, both W2V and DR seem to offer the best trade-off between having high accuracy and high density. Considering W2V, it offers really high accuracy values ($Accuracy_{W2V} = 0.211$) while also creating networks with a high density that is well above the average ($\Delta_{W2V} = 0.05$). On the other hand, DR creates very dense networks on average ($\Delta_{DR} = 0.007$), achieving a decent average accuracy ($Accuracy_{DR} = 0.083$) value.

Considering network fragmentation, we observe how there are very significant differences between approaches. On the one hand, we can see how some algorithms experience moderate fragmentation while having high accuracy values, as it is the case with LIWC and W2V (content-based approaches). On the other hand, we also find approaches that create very fragmented networks with a moderate accuracy value, as it is the case with MD-W2V, TS and TS-W2V (structure-based and hybrid approaches). More interesting are the results shown by Rnd, CF and DR. Both Rnd and CF achieve very low accuracy values, but they create networks with very little fragmentation, where information can flow very swiftly. As opposed to all the previous approaches, DR combines the best of both worlds to provide moderate accuracy values and create networks with very little fragmentation (and high density and efficiency), where information can flow quickly. All these characteristics reinforce the conclusion that was described earlier about DR being a good RA for facilitating the creation and consolidation of disinformation networks (the same applies to W2V, which provides slightly higher modularity values in exchange of a very high RA accuracy, standing as an algorithm to be considered when it comes to facilitating the creation and consolidation of disinformation networks).

Last, we calculated the clustering coefficient vs accuracy comparison to further understand fragmentation and cohesiveness in the generated disinformation networks. Considering the results, we observe several relevant aspects. First, we can see how algorithms with very low accuracy values, such as Rnd

and CF, generate disinformation networks that take moderate values of CC, meaning that the disinformation networks they create are moderately cohesive. The same occurs with the LIWC RA, which has higher accuracy values, but moderate values of CC (similar to W2V, although W2V achieves more cohesive networks when compared to the aforementioned RAs). However, we perceive important differences concerning the clustering coefficient in the three algorithms. First, we observe that MD-W2V provides moderate accuracy values, but it generates networks with very low cohesiveness, as it could also be derived from the low diameter of the generated disinformation networks (as it can be seen in Table 9.1). As a result, we observe how this algorithm favors accuracy while, at the same time, the creation of inefficient, relatively sparse, fragmented and non-cohesive networks, which are characteristics that clearly prevent the generation and consolidation of disinformation networks. On the other hand, we observe how the approaches based on the Multi-Armed Bandit Theory (TS and TS-W2V) and DR, provide moderate accuracy values, but they generate very cohesive networks, where users tend to form clusters, i.e., echo chambers, which is one of the effects that disinformation networks can have, as it has been widely discussed in the literature [VPV21, Par11, SHW+18].

All in all, we conclude there exist RAs that facilitate the generation and consolidation of disinformation networks while keeping decent content recommendation accuracy levels, as it is the case with DR and W2V. We also observe that different RAs create recommendation networks whose content dynamics are highly influenced by the different shapes of such networks. While all RAs try to maximize recommendation accuracy, the networks they generate during their functioning can be very different in terms of their shape and properties, as they can have a varying range of densities and efficiencies, they can be highly or poorly fragmented, networks can foster the persistence of users and topics (or their renewal), etc. As a result, not all RAs create optimal networks for the creation and consolidation of disinformation networks, based on the known characteristics of a disinformation network.

9.5 Discussion

Our study aimed to understand the dynamics of disinformation networks and their relationship with the algorithms underpinning content recommendation systems on social media. Leveraging an extensive database of user publications over time, encompassing both disinformation networks and legitimate journalists, we have simulated the evolution of these networks using prevalent content recommendation approaches. Our simulations have notably highlighted that content-based approaches, particularly those focusing on textual content, significantly promote the formation and consolidation of disinformation networks while maintaining a high degree of precision. This finding underscores the critical role of language in attracting and solidifying user profiles within these networks. Using specific discourse elements, including hashtags, facilitates repeated interactions among users, forming dense and efficient networks, as proven by RAs that seem to capture this intuition correctly, such as DR and W2V.

Despite these insights, our study has limitations. While the scope of tested recommendation algorithms covers significant approaches, it could be expanded to encompass additional methods. For instance, in this research Reinforcement Learning-based methods, such as the Multi-Armed Bandit RAs, do not seem to achieve a high recommendation accuracy, nor they seem to particularly enhance the network metrics that facilitate the generation and consolidation of disinformation networks. However, only simple Multi-Armed Bandit algorithms, like the Thompson Sampling based approach, have been implemented. Indeed, context-based approaches could have been used for the research, making use

of important pieces of data, such as the follow relationships between users, as well as internal potentially available data like profile visualizations, time consumed per publication or geolocation data among others. Such information was not available for us in this research, as the datasets did not contain this information, which made us shift towards more generic algorithms such as context-free Multi-Armed Bandit approaches, as the implemented ones. Acknowledging the limitations of this research, it's pertinent to note that while the study incorporated state-of-the-art recommendation systems from the four key families—content-based, structure-based, deep learning, and reinforcement learning—the vast and evolving landscape of recommender systems includes more complex algorithms and numerous variants aimed at enhancing accuracy, user engagement in the platform and overall performance. Due to the practical constraints of this study, it was not feasible to explore every such variation, especially those focused more on refining existing models rather than offering new perspectives on recommendation network generation. This particular choice, while enabling a focused examination of innovative methodologies, would have inherently limited the breadth of the algorithmic nuances considered, as our main goal in this research has been to evaluate the role of the main core approaches in recommender systems on the formation and impact of disinformation networks.

Our research also has certain limitations related to the nature of the study. For instance, there is no access to the full dataset of the network, but just to a portion of it. As a result, we compare the ability to generate and consolidate disinformation networks of each approach with the generated legitimate recommendation networks to statistically prove such ability to generate disinformation networks. However, the availability of the full dataset could allow different and complementary analysis to further justify our claim. Furthermore, we do not have access to the actual algorithm used in the social network (i.e., X/Twitter), which can condition the retweets of the users, and, as a result, the visibility graph that was used to help rank the tweets with the different approaches.

Furthermore, our analysis was confined to a specific segment of the X network and focused solely on Spanish-language content. The recommendations for improvement offered by this study are also broad and suggest a wide array of potential enhancements. Looking forward, there are several avenues for future research. Expanding this investigation to include other social networks with different content mobilization dynamics represents a significant next step. Similarly, extending our analyses to additional datasets, particularly those from different countries or cultural contexts, could provide more comprehensive insights into the global phenomenon of disinformation. While our study encompassed a broad range of algorithms, including state-of-the-art approaches in deep learning and reinforcement learning, exploring alternative methods or novel techniques in these areas could further our understanding of how recommendation systems influence disinformation. Notably, the natural progression of this research involves developing and proposing methods to mitigate the formation of disinformation networks while ensuring that the accuracy of recommendations remains acceptable. Our proposal in this line is presented in Chapter 11, where we analyze the balance between recommendation accuracy and mitigation strategies in terms of disinformation exposure.

9.6 Conclusions

This study evaluated the influence of recommendation algorithms on the proliferation of disinformation networks within social media. Through extensive simulations over a robust dataset, we have discerned distinct patterns in how different algorithmic approaches contribute to the formation and consolidation of

the harmful disinformation networks. Our findings highlight the particularly concerning role for content-based recommendation algorithms, especially those adept at processing and amplifying certain features like hashtags and URLs—exemplified by DeepRank and Word2Vec. While achieving high levels of recommendation accuracy, these algorithms significantly bolster the creation and perpetuation of disinformation networks, posing a substantial threat to the integrity of the online conversation. The propensity of content-based algorithms to foster disinformation networks without similarly impacting other types of networks suggests a nuanced complexity in their operational mechanisms. This characteristic underscores the urgent need to reevaluate recommendation systems that inadvertently facilitate harmful information dynamics despite their effectiveness in user engagement and content relevancy. Notably, the investigation revealed that the structural aspects of networks play a pivotal role as well in this phenomenon. Hybrid algorithms like DeepRank merge content and network structure considerations to further potentiate disinformation networks, highlighting the multifaceted nature of the challenge.

Our research substantiates the assertion that specific algorithms, particularly Word2Vec, significantly promote the development of dense, efficient, and moderately modular disinformation networks due to their high precision in content recommendation. This realization points to an alarming paradox within the design of social media platforms: algorithms that enhance user experience and platform utility may also inadvertently undermine the platforms' informational ecosystem. Addressing this challenge necessitates a strategic pivot towards enhancing diversity within recommendation systems in terms of network structure and content. Integrating measures that promote exposure to a broader array of perspectives may make it possible to mitigate the formation and impact of disinformation networks. Such an approach requires a concerted effort from researchers, technologists, and policymakers alike to recalibrate the underlying algorithms that shape our digital interactions.

In conclusion, our investigation provides empirical evidence that specific text content based recommendation algorithms facilitate the emergence and consolidating of disinformation networks on online micro-blogging social networks. These insights contribute to the academic discourse on information dissemination in the digital age by underscoring the imperative need to develop more responsible and nuanced recommendation technologies as a key counter-disinformation strategy. As we move forward, it is paramount that the design and implementation of these systems are guided by a commitment to fostering a healthy and diverse informational landscape, thereby safeguarding the integrity of the public discourse in the digital realm.

Chapter 10

The Role of Recommender Systems in the Formation of Polarized Echo Chambers

10.1 Introduction

In economically developed countries, social networks have become the primary information source, replacing traditional media and transforming public discourse. This shift allows users to be both producers and consumers of real-time information, fostering diverse perspectives but also posing challenges in content quality and veracity. Algorithms now filter information, reinforcing biases and potentially limiting exposure to varied viewpoints, which can exacerbate affective polarization—segregating users into ideologically opposed groups. The research presented in this chapter examines political polarization in online networks, using the data set from Spanish elections and analyzing algorithmic impacts, network structures, and user roles. We aim to understand and mitigate polarization, promoting healthier online discourse and a more inclusive public sphere.

At the outset of this chapter, we investigate the dynamics of political polarization in online social networks, with a focus on the role of recommendation algorithms in shaping these processes. The chapter examines how various recommendation systems—including topology-based, content-based, deep learning, and reinforcement learning approaches—contribute to the formation and intensification of polarized networks. Utilizing real-world data from Spanish electoral processes spanning eight years, we identify key patterns that distinguish these algorithms in their ability to foster or mitigate polarization. By analyzing algorithmic behavior, network structures, and user dynamics, this chapter directly addresses the third and fourth research goals of this thesis: understanding the factors driving polarization and proposing algorithmic strategies to mitigate its effects. These insights support the broader aim of the research by advancing knowledge of polarization phenomena in digital ecosystems.

The chapter is organized as follows: Section 10.2 introduces the concept of affective polarization and its relevance in the context of online networks, highlighting the role of recommendation systems in influencing user behavior and network structures. Section 10.3 details the methodologies employed, including data collection, the implementation of polarization metrics, and the evaluation of various recommendation algorithms. Section 10.4 presents the findings, exploring how different algorithms affect

polarization levels and network cohesion across electoral processes. Section 10.5 provides a comparative discussion, analyzing the implications of these results and the relationship between recommendation system accuracy and polarization dynamics. Finally, Section 10.6 concludes with a synthesis of findings and recommendations for future research, emphasizing the importance of designing polarization-aware recommendation systems to promote healthier digital environments.

Research questions

Our aims in this chapter are summarized in the following research questions included in the second and third research goals of the thesis:

- **RG2:** Conduct an in-depth analysis of the main risk phenomena in political information ecosystems. Develop a set of tools for their computational analysis.
  - **RG2.RQ4:** What are the main factors contributing to the emergence of polarization?
  - **RG2.RQ5:** What is the role of different categories of users in the formation of polarization?

As well as

- **RG3:** Analyze the role or contribution of recommendation systems in promoting these phenomena (disinformation diffusion and polarization). Propose mitigation strategies through these systems.
  - **RG3.RQ2:** Which approach or family of recommendation system contributes most to the rise of polarization?

10.2 Background

Affective polarization is a social phenomenon through which individuals within a collective or community ideologically segregate into strongly differentiated sub-communities with exceptionally high hostility towards each other. The presence of polarization within a group hinders collaboration and cooperation among its members, the generation of consensus, and collective decision-making, paving the way for violent conflicts among its members [Wag21, KDK+21]. Therefore, we understand it to be an especially harmful phenomenon for a specific community and society [KvS21].

While polarization can arise around any topic of social interest, it is currently largely present in politics, especially in Western societies which are characterized by the dominance of liberal democracy and, therefore, open to debate by definition. Similarly, although this phenomenon has existed throughout history [Goe19], it has reached unprecedented heights in the era of social media[KvS21] as online social networks greatly facilitate instant self-communication; politically charged information appears and disappears quickly, communities form, and debate emerges in a natural manner [Cas13, Nec12].

Thus, polarization has become one of the main issues in human societies in recent years [KDK+21, BG20]. It hinders healthy online conversation and ultimately severely affects decision-making processes and social peace in our nations [KDK+21]. Most of the western countries under study have shown increasing levels of polarization in recent years [BG20, YRW+16].

Regarding the causes and overall dynamics of polarization in online social networks, the current consensus indicates that this phenomenon emerges due to users' inherent interest in consuming content

that aligns with their initial ideological position [FS65] (a phenomenon known as selective exposure) along with the effect of information recommendation systems [KvS21, CMMB22, RARFN22]. These systems reinforce this phenomenon by placing users in feedback loops, which can create genuinely polarized echo chambers [CMMB22, RARFN22].

Therefore, social media users, by their human condition, have inherent biases that reinforce internal congruence within ideological communities and demarcate them from outsiders. In this sense, individuals tend to align with particular social groups and ideologies (birds of a feather flock together) [MSLC01]. This process of demarcating other groups, known as homophily, is influenced by affective ideological identification, meaning they often distrust or discredit external sources of information that do not align with their group's beliefs, leading to polarization[DZ23]. Recommendation systems also influence the generation of these dynamics. Social media platforms use algorithms in their content recommendation systems to personalize the content users see. This content personalizing strategy is based on users' past behaviors, language, and preferences, which tends to create redundant exposure to similar viewpoints[CMMB22].

It is well known that these recommendation algorithms tend to reinforce existing beliefs by frequently exposing users to content that aligns with their ideological inclinations [RARFN22]. This reinforced selective exposure limits the diversity of information and perspectives users encounter, fostering echo chambers. There is a reciprocal relationship between human communication decisions and algorithmic dissemination. Users' interactions with content influence future recommendations, creating a feedback loop reinforcing existing biases and group boundaries [CMMB22, DZ23, CR22]. The interaction between socio-cognitive biases and personalized content delivery leads to ideologically homogeneous networks. These networks are characterized by frequent interactions among the same individuals, reinforcing their shared beliefs and demarcating them from opposing opinions. Personalized recommendations can exacerbate this effect by densifying intra-group communication [DZ23].

However, when studying this phenomenon in online social networks, virtually no research rigorously and thoughtfully analyzes specific types of recommendation strategies about political polarization using actual data from political processes. Existing studies do focus either on measuring polarization or on understanding their causes. On one hand, the studies focusing on its measurement are mostly based on the analysis of the network structure [GMGM18, VPV21, GGLC22], on the analysis of the content of the digital conversation [KM22, CLDB18, Mat17, YWLD17] and a minority of them on hybrid approaches [ENT$^+$20, BAB$^+$21, SIK22, HDC23]

On the other hand, the studies focusing on comprehending the phenomenon and understanding their causes are primarily based on simulation using agent-based models and on the study of polarization centered around the concept of echo chambers [MBP18, DZ23, DZ21, CR22, SKG20, DM23, SJGL24].

Agent-based models are easy to deploy as they do not need a large previously gathered data-set, and can provide a general model in order to explain the phenomenon [Mac16]. Nevertheless, they lack the scientific depth that is needed in order to generate more advanced effective solutions to the problem such as algorithmic intervention strategies. Regarding echo chambers, that is, communities of users on social networks that are strongly connected internally with few external links and focused on a limited set of topics. While we understand that such structures facilitate the emergence of affective polarization, this phenomenon extends beyond these structures, including other variables such as discourse, the political charge of the conversation, and hostility between communities. Therefore, the study of polarization

solely through the analysis of echo chambers, while providing a foundation for understanding the phenomenon, can be overly simplistic. Ideological echo chambers may exist in a social network where users group based on affinity, but this does not necessarily mean that these users are hostile towards each other or that there is no healthy ideological exchange between these communities [TBB21]. The online debate between differentiated communities is healthy and contributes to political development and community richness; ideological entrenchment and hostility harm this development. Therefore, the echo chamber scenario is, to a great extent, natural and does not have to be a significant problem as long as radical and irreconcilable positions are not reached [Bru19, RH21]. Echo chambers are a necessary but insufficient condition for the emergence of polarization. The real problem arises when these echo chambers exhibit high thematic isolation and hostility towards each other or when organized political actors external to the debate artificially polarize the network through automated accounts, information dissemination, or other information warfare strategies[MBP23].

Similarly, the study of polarization through agnostic simulation based on agent-based models helps delineate the phenomenon broadly and provides an understanding of its logic [MBP18]. However, this approach can be limited in precision and scope. Therefore, an analysis based on accurate and actual data would provide much more precise information about this phenomenon, especially if conducted on a social network like X, the leading platform for discussing politics from a polarized perspective. Moreover, it is well known that recommendation systems influence the formation of echo chambers, which are necessary for the emergence of polarization. However, not all systems operate in the same way. The existing literature still needs to resolve the influence of different recommendation strategies in promoting or mitigating this phenomenon (do all recommender systems contribute to polarization similarly?).

A similar issue arises with users. By focusing on generalities or more abstract analyses, the existing literature does not root about the role of the network's characteristics or the behavior of particular users. As we have seen, polarization exists, but can intensify or diminish at specific moments within the network. Therefore, it is advisable to study what exactly occurs in these situations, whether it be the actions of particular users or groups of users, the emergence of certain phenomena, or the formation of specific structures within the network. Understanding these network and user dynamics can significantly contribute to designing early warning systems or serious algorithmic intervention strategies to promote the "health" of network conversations.

In this chapter, we aim to gain a deeper knowledge on the overall process of affective polarization in online social networks. In order to comprehend the phenomenon including its development and potential causes, we examine polarization dynamics from a triple perspective: algorithmic, network, and user-based. Regarding social media users, we aim to understand which users, based on their behavior on the platform, contribute to increasing or decreasing polarization. Similarly, we consider which phenomena within the network (i.e., what network states) contribute to increasing polarization (or what happens in the network when polarization increases) and the overall role of the algorithmic strategy of content delivery to the user. To ensure a comprehensive and rigorous analysis, we study state-of-the-art polarization metrics and indexes in order to reliably capture the polarization phenomenon in electoral processes and explain its development and causes from the aforementioned perspectives.

10.3 Specific Methodology

In order to address the aforementioned research questions, we combine some of the contributions presented in previous chapters of this thesis, by first making use of a dataset containing tweets from Spanish political process from 2011 to 2019. Furthermore, we implement a polarization metric that proved to capture polarization effectively during electoral processes, as well as a set of recommendation algorithms from the different well-known families of recommendation algorithms existing in the literature. Last, we apply different social network analysis and statistical techniques to examine the proposed research questions.

10.3.1 Data set

To carry out the research presented in this chapter, we employed the data set related to the political process in Spain. We used two primary datasets for this, as defined in the methodology chapter (Chapter 2), as described in tables "General Processes" (see Table 4.2) and "Local Processes" (Table 4.3). Note these datasets were used in Chapter 6 to provide an analysis of the election processes and derive a narrative flow and in Chapter 8 to evaluate our proposed polarization metric.

For the context of this research, we divided the data related to each electoral process into fixed time windows of exactly 24 hours, that are used to feed the implemented recommendation algorithms (with train-test split), so that these algorithms can generate *recommendation networks* for each time window. In this sense, as defined in Section 2.17, a recommendation network is simply a weighted and directed graph where nodes represent users and edges represent recommendation links between them. Namely, a user A is linked to a user B (in that direction) if user A got recommended content from user B, and the weight of that connection is the number of posts of user B that got recommended to user A.

In our experiment simulation, each recommendation algorithm trains during each time window of a train partition, so that it can recommend content for the users employing the test partition generating, as a result, a recommendation network. This simulation is performed for each day of the 31 days surrounding each electoral process in the Spanish electoral processes dataset employed for the research, so that different recommendation algorithms can be reliably compared.

As it was aforementioned, time windows are fixed to last for exactly 24 hours due to the dynamics of information in online social networks like Twitter. Existing literature has already shown that, in Twitter, news and posts have a lifespan of about a day [ORK18], although certain viral content and news may last for longer periods. Intuitively, employing a time window whose duration matches the lifespan of the posts that conform to this research's dataset would allow us to feed the recommendation algorithms with the required data to answer the proposed research questions.

10.3.2 Measuring polarization

As it was described previously, there exist three different approaches to polarization quantification, including topology-based, content-based and hybrid approaches. Considering the existing literature, the SPIN algorithm proposed in Chapter 8 is adequate for the current study, as it is a hybrid approach that employs information from both the network structure and the content that propagates through it to capture the polarization phenomenon in political processes.

As the current study is aimed at explaining the polarization phenomenon from three different perspectives (algorithmic, network structure and user-related perspectives) and considering the employment of an electoral processes' dataset, the SPIN algorithm seems more adequate. Grounded in the principles of Information Theory, SPIN provides a global measurement of polarization that is composed of intra-community (i.e., within each community of users) and inter-community components (i.e., between different user communities), effectively capturing the polarization phenomenon within user communities and between communities through the measurement of negatively-charged information flows.

10.3.3 Recommendation algorithm selection

The study of the effect of recommendation algorithms on the polarization phenomenon starts by an in-depth analysis of the existing literature on recommendation algorithms. In order to carry out this research, we implemented recommendation algorithms to cover the four main families of current recommendation algorithms: topology, content, deep learning and reinforcement learning-based recommendation algorithms [ČLM13].

Indeed, we implemented topology-based approaches like the well-known KNN-based user collaborative filtering, the friendship-based approach proposed in [AA03], based on link recommendation of similar pairs of users, or PropFlow [LLC10], an algorithm that leverages the concept of random walks to create recommendations. Regarding content-based approaches, we implemented a wide range of algorithms that employ different techniques to represent user preferences, from word embeddings obtained with Word2Vec [Chu17], to a psychology-informed representation of the preferences based on the Linguistic Inquiry and Word Count (LIWC) framework [PBJB15]. Regarding deep learning-based recommendation algorithms, we implemented Deep Rank (as authors did in [CZ20], a state-of-the-art algorithm for top-one list-wise ranking of recommendations that significantly outperforms the other existing recommendation algorithms in the literature. Last, regarding reinforcement learning-based recommendation algorithms, we employed the Thompson Sampling-based multi-armed bandit strategy, widely used within the existing recommendation system literature [AG12, SCCL19].

10.3.4 Network analysis

In order to answer to the proposed research questions, we first calculated well-known network metrics for the recommendation networks generated by each recommendation algorithm in each electoral process studied [New18]. The employed network metrics include the number of nodes and edges of the network, the average network degree, its diameter, efficiency, density, modularity and clustering coefficient values, as well as metrics related to the average influence of the users in the created networks, like the average Eigenvector Centrality or the average PageRank (considering all the users in the network). In fact, these metrics have already been used in the context of social network analysis under similar circumstances [GAS$^+$15, New18].

As discussed previously, to measure political polarization, we employed the SPIN algorithm as presented in Chapter 8, which has proven to work significantly better than other state-of-the-art polarization metrics in the particular context of (Spanish) electoral processes. Indeed, we calculated the previous network metrics in combination to the SPIN polarization metric for each simulated recommendation network created by each recommendation algorithm in each time window (in each electoral process). The

results of those calculations were then used to answer the proposed research questions through several statistical and social network analysis techniques described in the following paragraphs.

Indeed, we utilized the Mann-Whitney U test to rigorously study which recommendation algorithms facilitate the evolution of polarization over time, as we had independent samples that did not follow normal distributions, as it was verified with the Shapiro-Wilk statistical test for normality. Precisely, we applied the mentioned statistical test to determine if there is a significant difference between polarization distributions (per recommendation algorithm) over time. Indeed, a recommendation algorithm that facilitates the increase of polarization over time should have significant differences across the different (Spanish) electoral processes (i.e., an algorithm that involves a significant monotone increasing trend over time).

Furthermore, we employed correlation analysis and linear regression fit to study whether different factors have an impact in political polarization. The resulting correlation values and the p-value of the linear regression fit allowed us to either accept or reject the hypothesis that there exists a directly (or inversely) proportional relationship between the studied factor and the resulting polarization in an electoral process. Among the factors we studied, relevant aspects like user coordination, network structure and user published content and its emotions were included.

In the same line, we applied correlation analysis and linear regression fit to study which kind of users lead to the creation of more polarized networks, so as to understand if recommendation algorithms played an important role in this other aspect. To do so, we calculated, for each electoral process, the first (Q_1) and third (Q_3) quartiles of certain user-related characteristics: the number of hashtags, URLs, and retweets employed, the number of total created posts, and the number of followers and followees (people following). Then, we calculated, for each network and characteristic study, the number of users with that characteristic over Q_3 and under Q_1, so that we could study the correlation between the number of users of a given type (e.g., users with very high usage of hashtags) and the resulting polarization (i.e., SPIN). The results of these analyses give better insights on which kinds of users increase network polarization.

Last, we analyzed the relationship between polarization and recommendation accuracy through the creation of scatterplots (one per electoral process) where the results obtained by each recommendation algorithm in each time window are plotted to visually understand the relationship between recommendation accuracy and polarization and how certain approaches achieve interesting relationships between both variables. Although we employed recommendation accuracy for this study, other metrics could have been used in its place, including Recall and F2 metrics, which have been widely used in the literature so far [GSY12, ZB22].

10.4 Results

10.4.1 Network metric summary per RA and electoral process

As mentioned in Section 10.3, network metrics were computed on the recommended networks to characterize the kind of networks that each recommendation algorithm creates, which is useful information to analyze why certain algorithms are related to an increase of polarization (**RG3.RQ2** and **RG2.RQ4**). In this case, Tables 10.1 and 10.2 (together with those included in Section 10.7, as additional results, as Tables 10.4-10.7) show average network metric values per recommendation algorithm and electoral process.

Table 10.1: Average network metrics in the recommendation networks generated by each RA (general nov 2019). Best values per row in bold.

Metric	Rnd	LIWC	W2V	MD-W2V	CF	TS	TS-W2V	DR	Friendship	PropFlow
Nodes	538.1667	533.0000	535.7000	494.0333	479.1333	479.3333	475.7333	500.8667	514.6667	**528.9667**
Edges	**1136.7667**	1113.1667	697.5000	1037.5667	1054.3333	869.6333	859.7333	1080.5667	675.8667	767.9333
$\overline{Deg}$	8.3643	8.4258	8.3635	7.3442	7.7393	7.7041	7.5470	7.6547	**8.4286**	8.4949
D	8.1333	7.5000	5.9667	6.9000	7.9667	**5.2000**	4.9333	6.2667	5.2000	5.3333
E	0.0351	0.0319	0.0266	0.0290	**0.0354**	0.0267	0.0263	0.0292	0.0283	0.0237
Δ	0.0039	0.0039	0.0025	0.0040	**0.0046**	0.0038	0.0039	0.0043	0.0026	0.0028
Q	0.8422	0.8551	**0.9313**	0.8053	0.8265	0.8544	0.8520	0.8209	0.8981	0.8982
$\overline{CC}$	0.0512	0.0669	0.0320	0.0568	0.0543	**0.0733**	0.0757	0.0701	0.0413	0.0353
$\overline{EVC}$	0.0014	0.0020	0.0017	0.0020	**0.0023**	0.0022	0.0018	0.0023	0.0025	0.0019
$\overline{PR}$	0.0019	0.0019	0.0019	0.0019	**0.0021**	0.0021	0.0021	0.0020	0.0020	0.0019

Table 10.2: Average network metrics in the recommendation networks generated by each RA (local 2019).

Metric	Rnd	LIWC	W2V	MD-W2V	CF	TS	TS-W2V	DR	Friendship	PropFlow
Nodes	**451.7667**	448.1000	443.4667	425.4667	402.9000	397.8333	395.6667	415.9000	430.0333	439.5667
Edges	**866.4667**	861.3667	536.2667	800.6000	796.0667	693.4333	690.8333	823.6333	557.3333	599.7667
$\overline{Deg}$	7.9105	7.9419	8.0071	6.9166	7.0427	7.0803	6.7909	6.8386	7.8893	**8.1028**
D	7.0667	7.1333	**4.1000**	5.7000	6.6333	5.3667	5.1333	5.6667	4.0667	4.6333
E	0.0355	0.0339	**0.0258**	0.0301	0.0349	0.0288	0.0279	0.0300	0.0278	0.0257
Δ	0.0042	0.0043	**0.0028**	0.0044	0.0049	0.0045	0.0045	0.0048	0.0031	0.0031
Q	0.8704	0.8828	**0.9487**	0.8587	0.8543	0.8789	0.8755	0.8454	0.9204	0.9179
$\overline{CC}$	0.0510	0.0584	**0.0313**	0.0491	0.0496	**0.0641**	0.0630	0.0551	0.0444	0.0348
$\overline{EVC}$	0.0024	0.0022	0.0024	**0.0027**	0.0022	**0.0029**	0.0025	0.0021	0.0030	0.0026
$\overline{PR}$	0.0023	0.0023	0.0023	0.0024	**0.0026**	0.0026	0.0026	0.0025	0.0024	0.0023

Considering the results, we first observe that the number of connections greatly varies between the oldest electoral process (close to the creation of Twitter) and the rest of electoral processes. This suggests that during the first electoral process, finding recommendations for the user was harder than it is in the rest of electoral processes (either because of the creation of more posts, or because of the user interaction with a larger subset of users, i.e., neighbors).

Moreover, we also observe that some algorithms reduce the number of edges with respect to other algorithms, which remains true for all the different electoral processes. While certain recommendation algorithms, such as the recommendation algorithm based on Collaborative Filtering or DeepRank, tend to create recommendation networks with many edges (more user-to-user recommendations), other algorithms, like Word2Vec, PropFlow, or Friendship tend to have a smaller variety of user recommendations (which makes sense considering the nature of those algorithms).

Furthermore, we also see how certain algorithms create more efficient, dense and less modular networks, as it is the case with DeepRank and Collaborative Filtering, whereas some other recommendation algorithms tend to create more sparse networks, with higher fragmentation (modularity) and networks where information flows less efficiently (overall), as it is the case with the Reinforcement Learning-based recommendation algorithms (TS and TS-W2V), and the Friendship and PropFlow algorithms (this makes sense considering their nature, as these are greedy algorithms that tend to recommend content to one user based on the users that gave better results in the past.

As a result, we can also derive that the recommendation networks that are created through an electoral process are dependant on many aspects, including the recommendation algorithm itself, as we can see from the results obtained in these results' tables.

10.4.2 Polarization evolution per recommendation algorithm

Furthermore, we performed a subsequent social network analysis in order to understand polarization dynamics and the impact of recommendation algorithms in these polarization dynamics, to further address the last research question considered in this chapter (**RG3.RQ2**). The results of this analysis are summarized in Fig. 10.1 (for general electoral processes) and 10.2 (for local electoral processes), although daily polarization results are shown in Fig. 10.3.

The results denote a clear increasing trend of political polarization during general electoral processes (obtained exclusively from the generated recommendation networks per process). As it can be seen, all the recommendation algorithms experienced a significant increase in polarization from 2011 to November 2019, reinforcing the idea of an increasing political polarization in the current society [RSM22].

Focusing solely on general electoral processes and understanding the Spanish general electoral process of 2016 as a "past reference", we observe how most of the recommendation algorithms do not generate very significant differences (e.g., PropFlow, Word2Vec, LIWC or the Collaborative Filtering algorithm are characterized by very similar polarization values between 2016 and November 2019). On the other hand, the deep learning-based approach, DeepRank, stands as the only algorithm that keeps the increasing trend in polarization across all the different Spanish general electoral processes. While it is true that the polarization change from April 2019 to November 2019 is small, so is the time-lapse between both processes, and, if both polarization values are compared with past polarization records, we do observe a very significant increasing trend in polarization. The observation that DeepRank leads to a monotone increasing trend of polarization is aligned with the existing literature that describes deep learning-based approaches as algorithms that foster the polarization phenomenon due to their capability of effectively representing the preferences of the users in the network [GK21].

On the other hand, we observe a completely different trend in local electoral processes, where most of the recommendation algorithms create similar (or slightly less) polarized recommendation networks from 2015 (local) to 2019 (local) electoral processes. The only two algorithms that are outside that trend are the two Reinforcement Learning-based recommendation algorithms (Thompson Sampling Multi-Armed Bandit recommendation algorithm, and the extension of the previous one that uses Word2Vec to weight the posts of the "best arms" (best user-to-user links) for each user. This observation also reinforces the intuition that the polarization phenomenon is a social phenomenon. In general electoral processes, a large mass of voters create an opinion on a very reduced set of options, whereas in local electoral processes, there is a higher number of (smaller) masses of voters that create their own opinion on a wider set of options [Qua17].

Conversely, the analysis of the results shown in Fig. 10.3 reflects how each recommendation algorithm leads to networks with higher or lower polarization values. Concerning the results, we observe how different strategies lead to different polarization results. First, considering the topology-based algorithms, KNN Collaborative Filtering, Friendship and Propflow), we observe similar polarization trends on different scales. While the Friendship and Propflow algorithms result in networks with moderate polarization values, whereas collaborative filtering reaches really high polarization values in each electoral process. These results are understandable, as collaborative filtering tends to recommend content to each user based on what users with similar preferences have consumed previously, reinforcing the creation of echo chambers in the network from the beginning of the process, and difficulting the creation of efficient, dense and cohesive networks, which are key characteristics of non-polarized networks, as it is described

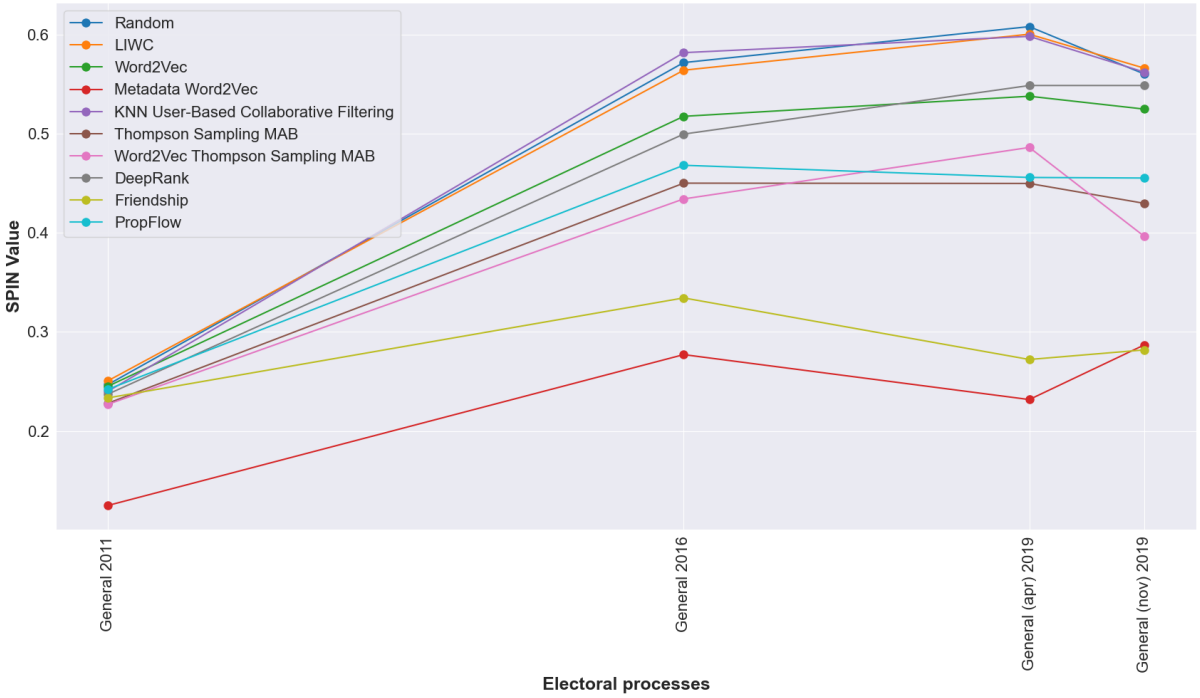


Figure 10.1: Average polarization evolution across Spanish general electoral processes from 2011 to 2019.

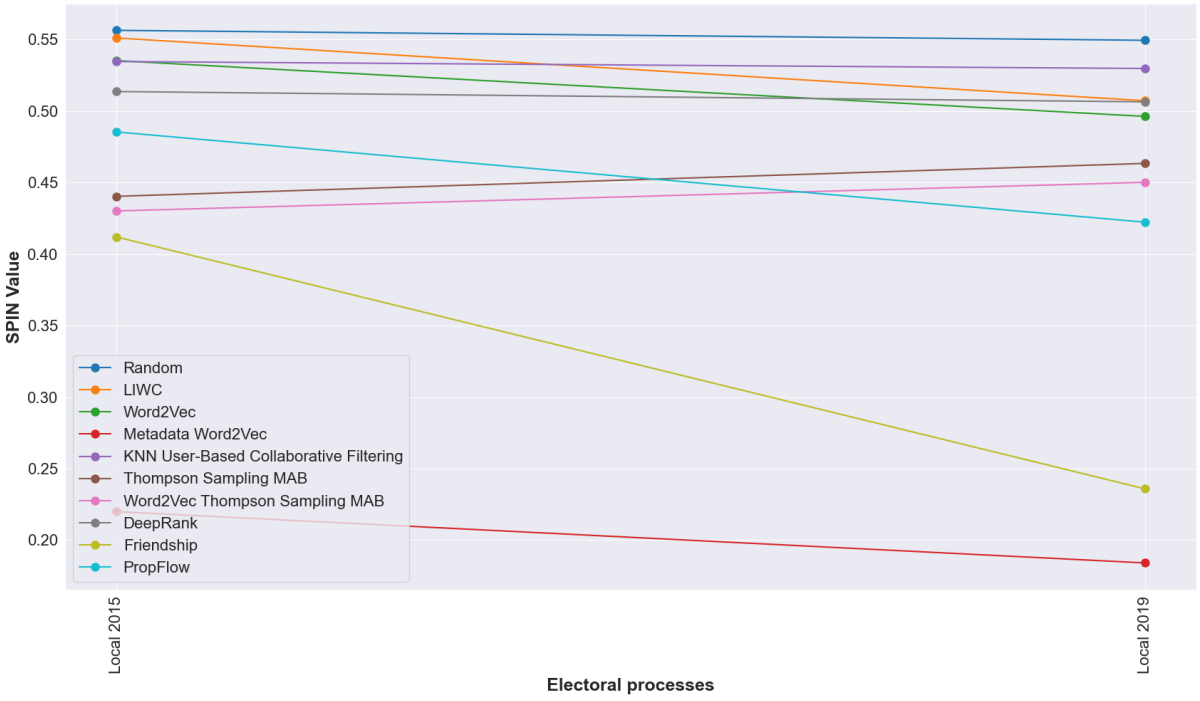


Figure 10.2: Average polarization evolution across Spanish local electoral processes from 2011 to 2019.

in the following subsection of this work. Regarding Propflow, its moderate-to-high polarization values are the logical result of an algorithm that uses the probability of information flows between users as a heuristic to carry out recommendations. Indeed, Propflow reinforces the creation and consolidation of echo chambers by recommending content that interrelates users with shared information flows. However, less common information flows are still given a probability that fosters recommendation diversity, effec-

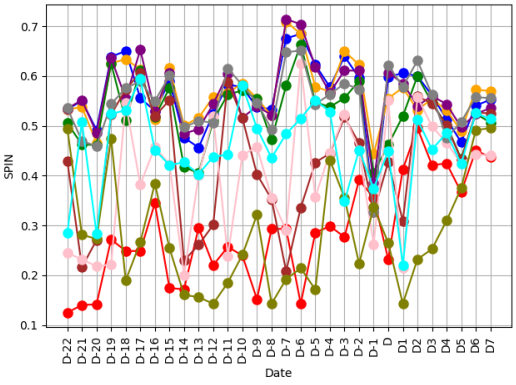
tively reducing the polarization of the resulting networks, as discussed in existing literature [LCKK14].

Furthermore, content-based algorithms also experience a variety of polarization trends. The metadata-based Word2Vec recommendation algorithm provides minimum network polarization values across all given electoral processes, because it tends to recommend content close to the preferences of a given user, but introducing variables like user popularity in terms of followers and followees, and whether the tweet references any of the existing Spanish political parties or not. Such metadata promotes wider diversity than just the preferences of the user, as a user can get recommendations from like-minded users, even if the content of the tweet does not fully match user preferences. As can be observed, the introduction of metadata results essential for minimal polarization result. Indeed, the Word2Vec recommendation algorithm, which uses the same word embedding algorithm without including meta-data of any kind, creates recommendation networks with high polarization values as a consequence of less diverse recommendations, which create less efficient, and more sparse and fragmented networks that correlate well with high polarization values, as it can be seen in the next section of this work. The last content-based recommendation algorithm, based on the LIWC framework, creates highly polarized networks, as a result of recommending content with the same psychological background as the content created by each user. Indeed, the LIWC recommendation algorithm represents user preferences as a vector of psychological categories, effectively recommending links between like-minded users, promoting the feedback loop phenomenon in the network [TALB21].

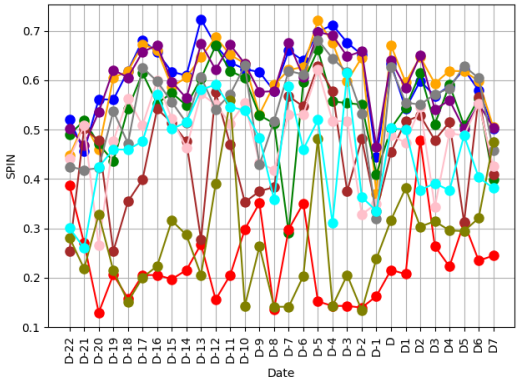
Regarding reinforcement learning-based recommendation algorithms, we observe moderate polarization values for both general and electoral processes. These results reflect the relevance of the age-old exploration-exploitation dilemma [ČLM13], and its impact on polarization. While certain strategies can maximize accuracy through the exploitation of already extracted patterns, like Word2Vec, achieving certain balance between exploration and exploitation introduces more diverse user recommendations, preventing the creation of echo chambers and creating more dense, efficient and cohesive networks, which tend to be less polarized as shown in the following section.

Last, the deep learning-based recommendation algorithm, DeepRank, reflects a high and increasing polarization trend across electoral processes, as it was discussed previously. Indeed, deep learning allows an optimal representation of user preferences, achieving decent accuracy values while fostering recommendation networks in which echo chambers predominate, facilitating and potentially aggravating network polarization through the reinforcement of user preferences, i.e., through fostering the well-known feedback loop [TALB21].

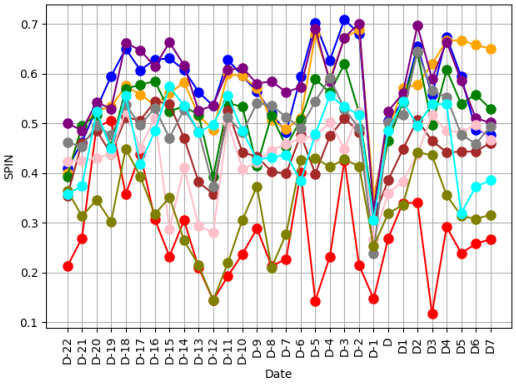
We conclude this algorithmic-level analysis by understanding the daily polarization evolution comparison shown in Fig. 10.3. We observe how all recommendation algorithms agree on a basic set of terms, however, each recommendation algorithm creates recommendation networks that lead to a unique polarization evolution over the electoral process. Indeed, all recommendation algorithms reflect that polarization increases during pre-campaign and campaign phases, reaches a minimum peak during the blackout period, and a maximum peak at the day of election, before vanishing as post-campaign progresses. However, certain strategies promote networks with higher or lower polarization due to their underlying behavior, effectively reflecting that recommendation algorithms impact the polarization phenomenon. While they are, to a large extent, influenced by external factors [DL22], social networks dynamics can clearly impact and, potentially, amplify such phenomenon as it can be seen from the presented results.



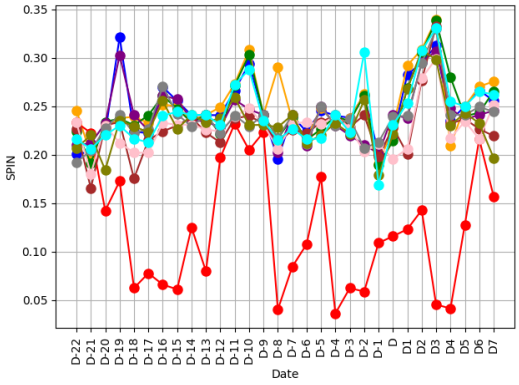
(a) General process November 2019.



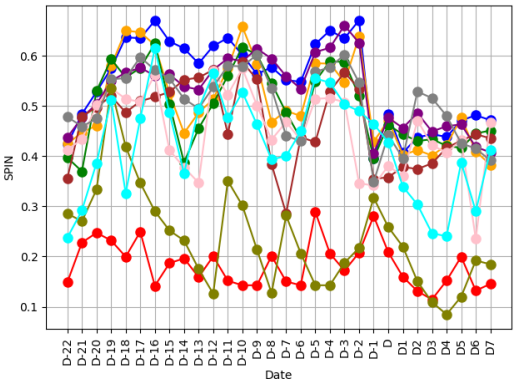
(b) General process April 2019.



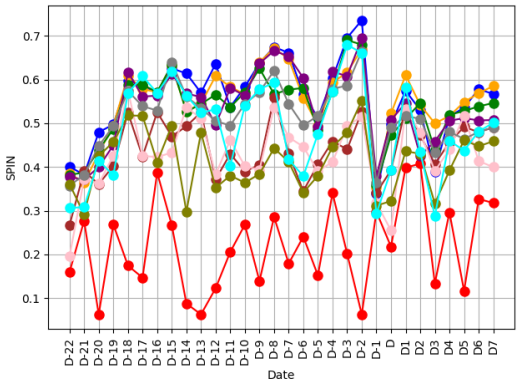
(c) General process 2016.



(d) General process 2011.



(e) Local process 2019.



(f) Local process 2015.

(g) Recommendation Algorithm legend.

Figure 10.3: Daily evolution of network polarization during Spanish electoral processes.

To further strengthen our conclusions, we applied the Mann-Whitney U test as explained in Section 10.3. This statistical test seemed to be optimal considering that polarization distributions did not follow a normal distribution (checked with the Shapiro-Wilk normality test), samples are independent (as they belong to groups which may or not be the same, and measured in different time moments) and the nature of data is continuous.

Conversely, we first applied the Mann-Whitney U test on the Deep Learning-based recommendation algorithm (see Table 10.3), as it was the only recommendation algorithm with a monotone increasing polarization trend from 2011 to November 2019. The statistical test derives that the polarization phenomenon significantly over time (as statistical evidence was found). Although polarization did not vary significantly between the electoral processes of 2019 (reduced time difference), considering general electoral processes from different years, there is a clear increasing trend of polarization supported by statistical significance, which further confirms that the deep learning-based recommendation algorithm, Deep Rank, facilitates the increase of polarization over time (at least, for general electoral processes).

We applied the same test to the reinforcement learning-based recommendation algorithms that showed an increase of polarization within the studied local electoral processes. However, in this case, no statistical evidence was found to guarantee that those algorithms facilitate the increase of political polarization over time, as the null hypothesis could not be rejected (using $\alpha = 0.05$). Indeed, the results show how the nature of the electoral process plays a fundamental role in polarization dynamics, and in the algorithms that facilitate their increase over time, as it is already described in the existing literature [PK03].

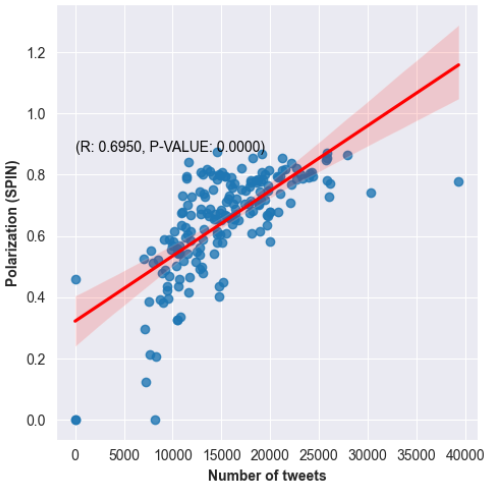
Table 10.3: Statistical significance test between Spanish general electoral processes.

Mann-Whitney U	**2011**	**2015**	**2016**	**Apr. 2019**
Nov. 2019	$U = 1$, $\rho = 3.30e^{-}11$	DATA	$U = 233$, $\rho = 0.0014$	$U = 472$, $\rho = 0.7506$

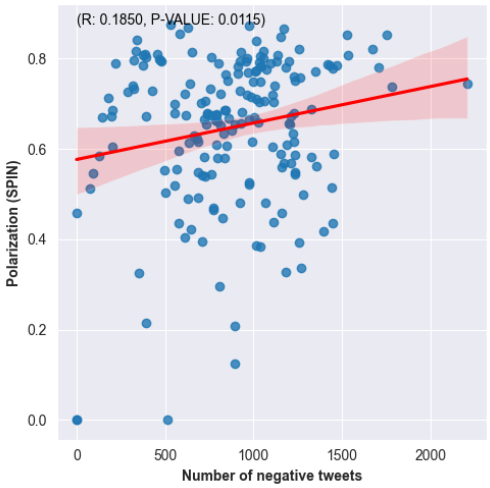
10.4.3 Network-level factors related to polarization increase

Considering the observed results, the nature of the electoral process is a factor that clearly affects polarization. As a result, an algorithm that facilitates the increase of polarization over time in (Spanish) general electoral processes (like DeepRank) can have a contrary effect in local processes (as it was derived from the results presented in this research).

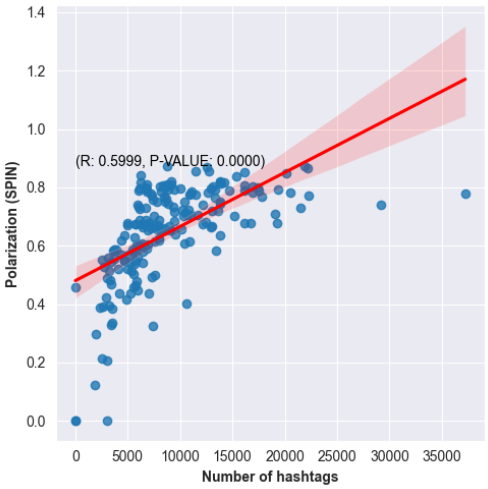
As mentioned in Section 10.3, we studied which factors (besides the nature of the electoral process) had an impact in polarization results (**RG3.RQ2** and **RG2.RQ5**). We carried out this study employing the techniques described in the methodology section, namely, correlation analysis and linear correlation fit to understand, through the p-value statistic, whether there exists or not a relationship between polarization and other factors, like network efficiency, density and modularity, or the number of hashtags and URLs created by the users recommended by the recommendation algorithm in each simulated time window. The results, shown in Figs. 10.4 and 10.5, allow us to derive several conclusions. First, we can observe the relevant role that created content (i.e., the tweets) plays in social network polarization, as there is a very strong direct linear relationship between polarization and the number of tweets in the network. However, the strength of the relationship is quite independent of the emotions that the content may have. This clearly reflects the idea that social network polarization occurs even when negative emotion are not present in the information that spreads through the network. Again, this observation remarks how



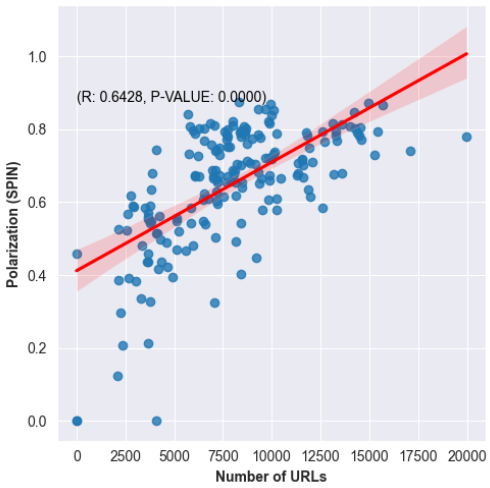
(a) Polarization vs Total number of tweets.



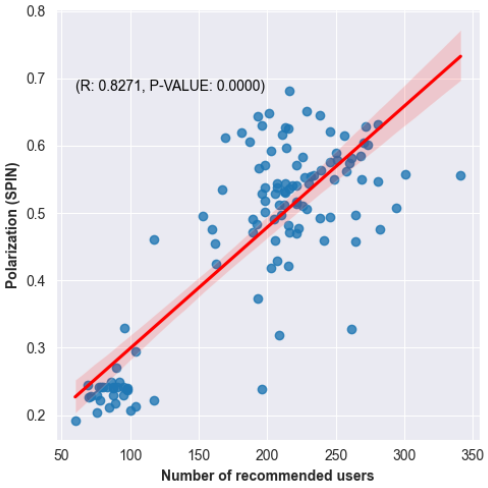
(b) Polarization vs Total number of neg. tweets.



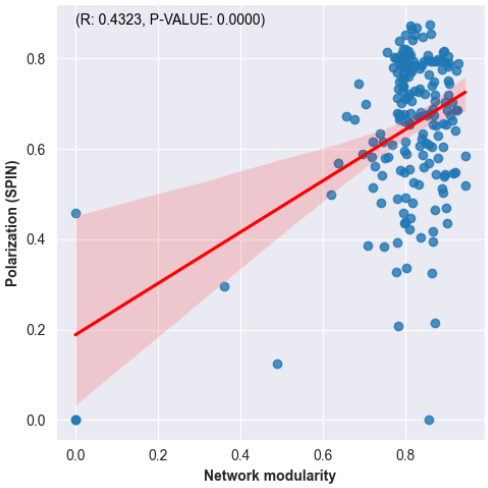
(c) Polarization vs Total number of hashtags.



(d) Polarization vs Total recommended users.



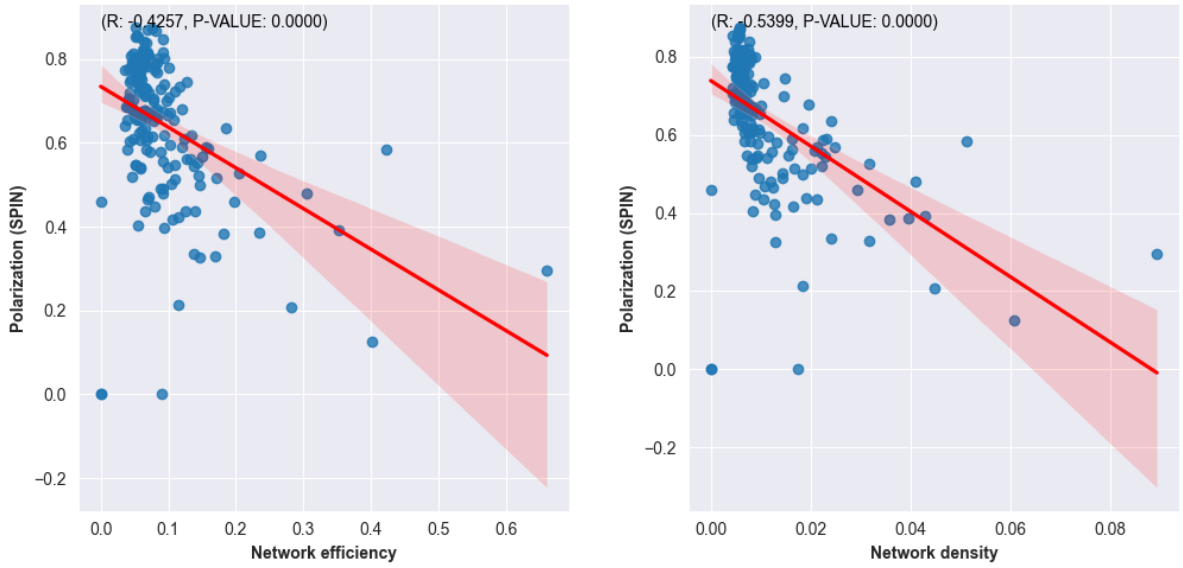
(e) Polarization vs Total recommended users.



(f) Polarization vs Modularity.

Figure 10.4: Analysis of different factors and their impact on polarization (I).

the polarization phenomenon is heavily influenced by external factors, however, our analysis proves that factors internal to the network can also influence such phenomenon.



(a) Polarization vs Efficiency.

(b) Polarization vs Density.

Figure 10.5: Analysis of different factors and their impact on polarization (II).

Conversely, we observe a very strong relationship between the degree of coordination between users and network polarization. Namely, the number of hashtags and URLs, which typically denote certain degree of coordination between users [WN20], strongly correlates with network polarization, reinforcing the idea that user coordination plays a fundamental role in network polarization, reflecting the relevance of the echo chamber effect. Users tend to cluster in very cohesive groups where they further reinforce their thoughts and opinions (e.g., through the creation and spreading of posts including hashtags and URLs that further confirm their viewpoints and reinforce the shared narrative they defend), facilitating the increase of network polarization (as it can be seen from the strong linear relationship between these variables, reflected by the correlation analysis).

Last, we observe how relevant network metrics, like efficiency, density and modularity, further reinforce the previous intuition. Concerning the results, efficiency and density negatively correlate with polarization, whereas modularity (i.e., network fragmentation) positively correlates with this phenomenon. These results reinforce the relationship between network structure and polarization. While efficient, dense and cohesive networks minimize polarization, inefficient, sparse and heavily fragmented networks (like the ones characterized by the emergence of echo chambers) lead to networks with higher polarization, as it can be seen from the extracted polarization results. Conversely, the p-value obtained from the linear regression fit in all the three cases (efficiency, density, and modularity) does not allow to reject the null hypothesis, which represents that variables are independent of each other. Indeed, the results reinforce the idea of these network properties being strongly related to the increase and decrease of polarization during the electoral process, remarking the relevance of network structure in the interplay with network polarization.

All in all, we can perceive how recommendation algorithms play a key role in the polarization perceived from social networks, as many of the aspects that seem to impact polarization are related to the recommendation networks that these algorithms generate (such as the density and efficiency of these networks, their fragmentation).

10.4.4 User-level factors related to polarization

To further complete the analysis (**RG2.RQ5**), we studied the characteristics of the recommended users of recommendation networks and their related polarization value (SPIN result). Several conclusions are derived from this analysis, shown in Fig. 10.6. First, we observe that networks in which there are more recommended users with high followers (over Q_3) result in networks with higher polarization ($R = 0.4125$). In fact, there is a moderate directly proportional linear relationship between the number of recommended users with high followers and the resulting polarization. We perceive a similar behavior with the opposite ("user-following") relationship. The result denotes that user popularity plays a relevant role in the resulting network polarization. Indeed, networks with more popular recommended users are networks with higher polarization values, whereas networks with less popular recommended users are networks with less polarization.

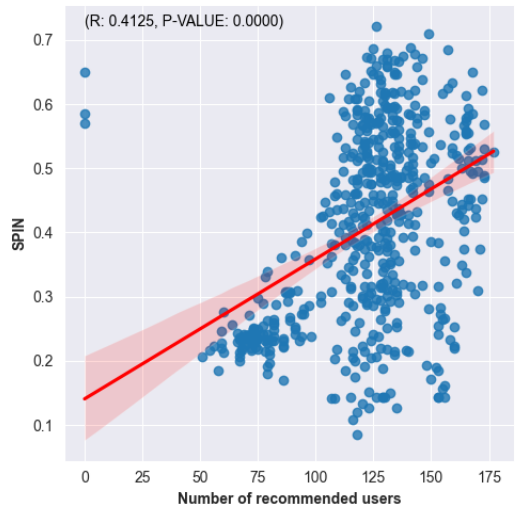
Furthermore, we observe that the recommendation of users with more activity (more posts) leads to the creation of networks with higher polarization. Indeed, users with higher activity, like political parties or activists, are related to networks with increased polarization. The case of political parties is of special interest. These users tend to have a really high activity [Kal16, BMV11], often reaching the highest quartile in terms of daily user activity. However, the only day when these users (i.e., political parties) have no activity at all, the blackout period, is the day with a local (or usually global) minimum of polarization, as it can be seen from the daily evolution of polarization per political process and recommendation algorithm show in Fig. 10.3. Indeed, political parties are users related to networks with higher polarization, as it is reflected in the presented work.

Additionally, users with a clear coordinated behavior (employing hashtags and URLs in a high proportion - over Q_3) also lead to networks with higher polarization. In this aspect, it is worth noting that users with higher URL and hashtag sharing lead to more polarized networks ($R_{hashtags} = 0.3988$, $R_{URLs} = 0.4359$) than those achieved with retweets (whose correlation does reach $R = 0.35$). In this sense, we observe that certain coordination strategies, namely hashtags and URLs, create more polarized networks than others (like retweets).

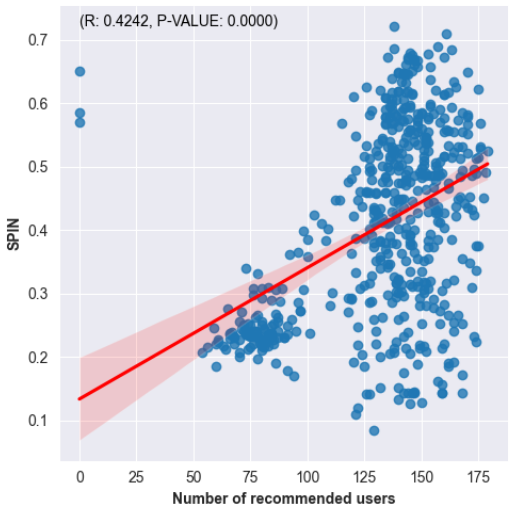
10.4.5 Recommendation accuracy and polarization

Last, we analyzed the relationship between polarization and accuracy. Considering the results (see Fig. 10.7), we can observe how the average accuracy values have increased over time for both general and electoral processes, considering the recommendation networks of all RAs. However, polarization only shows an increased average value for general electoral processes and not in local electoral process, which resembles the relevance of the nature of the political process in relation to the polarization of the process, as it was previously discussed.

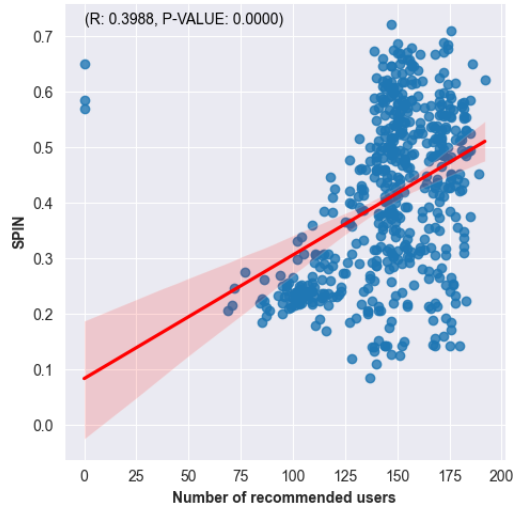
Furthermore, we do observe variability in both accuracy and polarization distributions. Regarding accuracy, variability is expected, as the accuracy of the recommendation algorithm heavily depends on the algorithm itself and the strategy that it follows. On the other hand, the polarization variability, which we can observe by considering polarization values from different recommendation algorithms, is concordant with the discussion presented in this research and in the existing literature: although the polarization phenomenon is to a great extent external to the network and the platform, social network dynamics still impact and potentially enhance and aggravate the phenomenon.



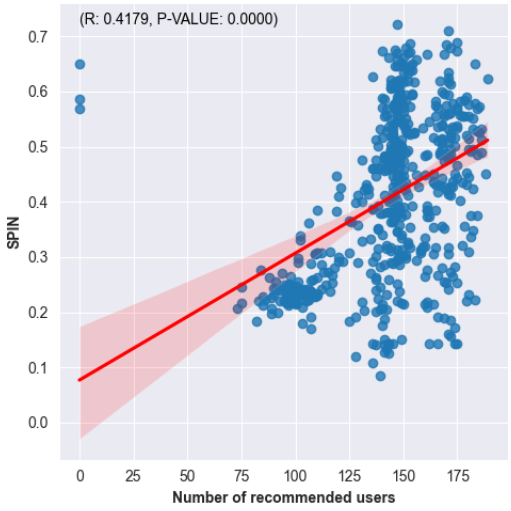
(a) Polarization vs Users with followers over Q3.



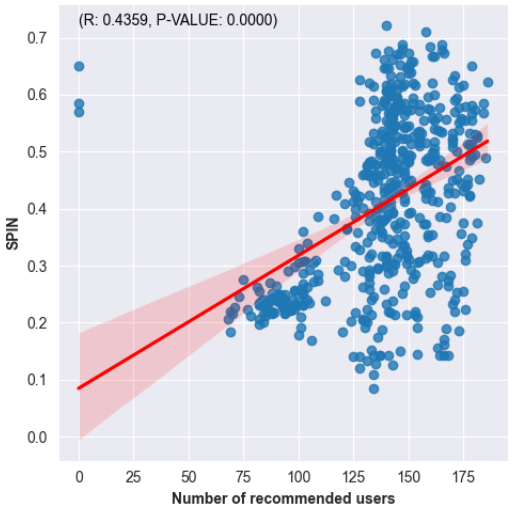
(b) Polarization vs Users with following over Q3.



(c) Polarization vs Users with hashtags over Q3.



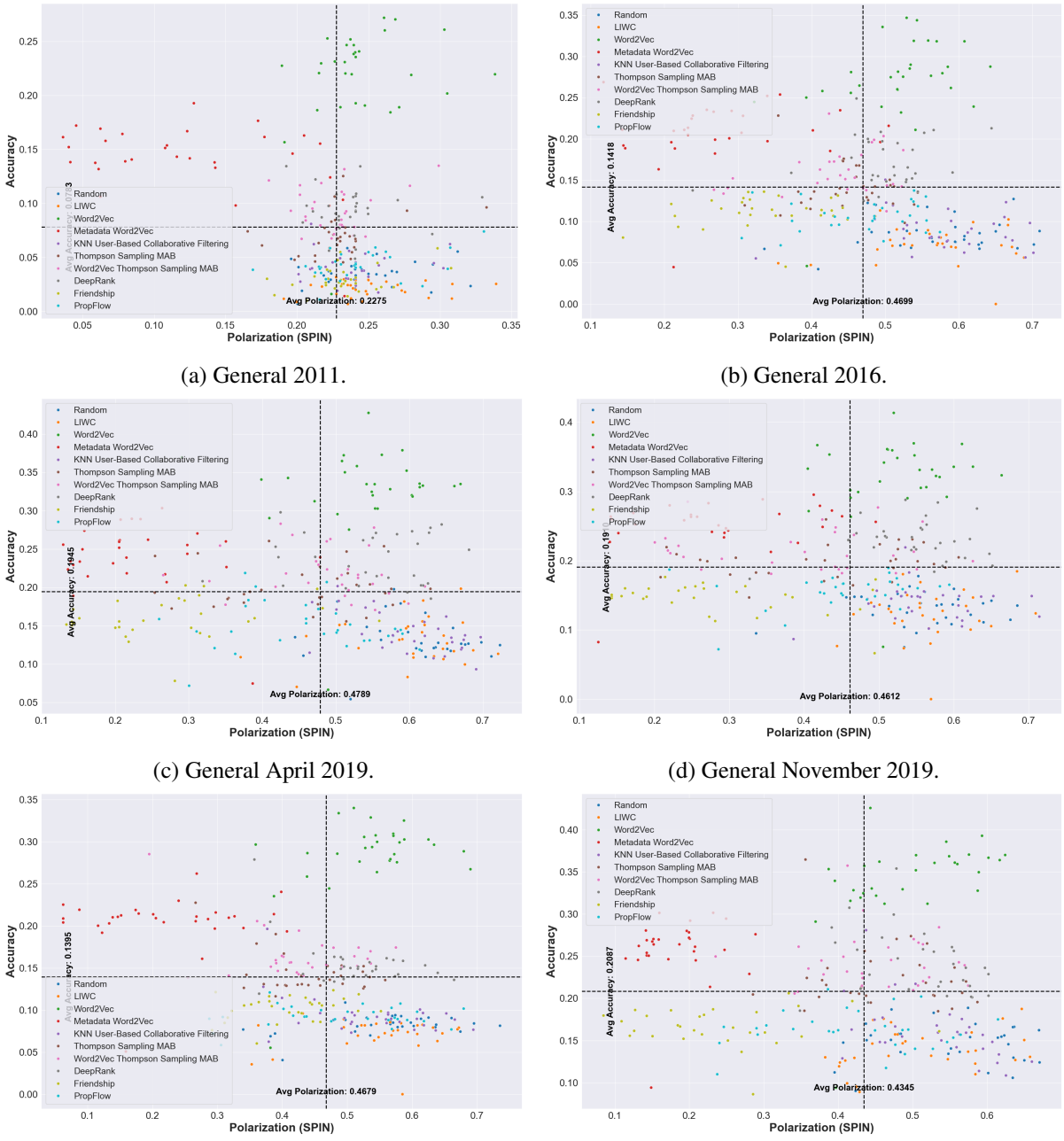
(d) Polarization vs Users with posts over Q3.



(e) Polarization vs Users with URLs over Q3.

Figure 10.6: Analysis of the impact of users in the network's polarization.

Additionally, we observe that algorithms with higher polarization values, like W2V, DeepRank or the reinforcement-learning based recommendation algorithms, reach high polarization values too. The only exception to this rule is the metadata-based Word2Vec strategy, which favors certain content diversity through the inclusion of popularity metadata, among other features described in Section 10.3, favoring the introduction of content that does not necessarily follow user preferences, but comes from users with similar popularity. Indeed, such a strategy proves to reach high accuracy values, while keeping very low polarization values. Conversely, the rest of algorithms with high accuracy are related to highly polarized networks, and vice-versa. Indeed, this observation further confirms the intuition that polarization is affected by factors that are both external and internal to the social network.



(a) General 2011.

(b) General 2016.

(c) General April 2019.

(d) General November 2019.

(e) Local 2015.

(f) Local 2019.

Figure 10.7: Accuracy and Polarization relationship per electoral process and algorithm.

10.5 Discussion

In this work we present a threefold analysis of polarization, including algorithm, network and user-level analysis. Such enriched approach allowed us to understand which recommendation algorithms seem to facilitate polarization evolution over time, understanding the reasons behind such relationship.

As opposed to many existing studies, we bring the polarization phenomenon to the social network level, allowing us to analyze the phenomenon through the observation of the social network dynamics and their behavior when different recommendation algorithms, working under different strategies, provide the recommendation between users of the network. While we implemented different recommendation algorithms covering the full spectrum of families (topology-based, content-based, deep learning-based and reinforcement learning-based), more and more complex recommendation algorithms could be introduced to further explain the algorithmic-level layer of the presented work.

Similarly, although we focused on the SPIN polarization metric, which has proven to improve state-of-the-art results in the quantification of polarization within electoral processes, other polarization measures and indexes could also be explored to further enrich the analysis, including purely topology-based measurements, purely content-based measurements, and hybrid measures (as it is the case with the employed polarization metric).

Furthermore, in this research we employed a dataset containing Spanish electoral processes from 2011 to 2019, however, other datasets could be used to further reinforce our claims, including electoral processes from other western democracies following similar electoral systems (multi-party systems) or other democracies with different electoral systems (e.g., two-party systems, as the one we can find in the United States).

Concerning future research, besides the already described elements, the presented work opens the door to the creation of polarization aware (i.e., harm aware) recommendation systems to achieve election security and prevent users from further amplifying the polarization phenomenon. Indeed, future research can be aimed at developing strategies to foster and further promote network diversity and healthiness, avoiding as much as possible the amplification of the polarization phenomenon. Such research would not only be useful for social networks and their recommendation algorithms, but, in general, many information retrieval tasks, including search engines and their capability of promoting healthier environments through their recommendations (following harm aware recommendation strategies). Chapter 12 presents our proposal to address this problem, by exploiting Information Theory concepts to reduce the polarization of the network. We shall see then how accuracy and polarization are balanced depending on different mitigation strategies.

10.6 Conclusions

In this work, we have presented a threefold analysis on polarization, examining the roles played by network structure (network level), recommendation algorithms (algorithmic level), and the users themselves (user level) in this phenomenon. Our observations demonstrate that, although the polarization phenomenon is affected by external factors, it is also influenced by all aforementioned levels. Regarding the algorithmic-level, we found that some algorithms, such as DeepRank (based on Deep Learning), can lead to an increasing trend of polarization over time. This is due to their enhanced ability to represent

user preferences and reinforce their interests with more suitable content, resulting in a feedback loop and the creation of echo chambers, which are necessary conditions for the formation of polarized networks.

Conversely, different recommendation algorithm strategies have a unique effect in polarization dynamics throughout the electoral processes. While collaborative filtering rapidly reaches high polarization values through the creation and reinforcement of communities of very like-minded users, i.e., echo chambers, reinforcement learning-based approaches utilize the exploration-exploitation dilemma to introduce more diverse recommendations, creating more efficient, dense and cohesive networks, which highly correlate with lower polarization values. Conversely, content-based algorithms also remark the relevance of recommendation diversity. While a metadata content-based recommendation algorithm favors content recommendation even if it does not strictly follow user preferences, achieving minimal polarization values, the same algorithm, Word2Vec, applied strictly on user preferences leads to highly polarized networks. Similarly, reinforcement learning-based recommendation algorithms further reinforce our claim on the relevance of recommendation diversity and its effect on polarization. In light of these observations, the polarization phenomenon proves to be affected by the underlying social network algorithm, even if it does not emerge from such network.

Our study also demonstrated that the social network level is also related to the polarization phenomenon, as more efficient, denser, and less fragmented networks tend to exhibit less polarization. This finding is supported by a correlation analysis showing the relationship between these network characteristics and the polarization phenomenon. This observation aligns with existing literature, indicating that less efficient, more dispersed, and fragmented networks facilitate the generation of echo chambers. In these network structures, users reinforce each other's viewpoints, potentially facilitating the emergence of polarized networks.

Additionally, our analysis showed that networks with higher polarization are also those where a greater number of messages are exchanged, not necessarily negative ones. Indeed, network dynamics influence the polarization phenomenon, even when those dynamics are not necessarily polarized. Furthermore, the role of political parties in the polarization phenomenon is notable, as highly active users significantly contribute to polarization. Remarkably, on election process days when their activity is minimal or nearly negligible, such as the day of reflection, polarization drops to local or even global minimum values, depending on the process. Moreover, coordination strategies also proved to be positively correlated with polarization. Users who coordinate and promote their viewpoints within small groups (echo chambers) where mutual reinforcement occurs (feedback loop) are associated with more polarized networks. This is evidenced by a relatively high correlation, indicating a significant relationship between these phenomena.

These findings are also consistent with the results obtained from the user-level analysis, which also seems to influence polarization, as users with higher activity levels and more coordinated behavior, reflected through the usage of hashtags and URLs, are related to more polarized networks. Additionally, the importance of the user within the network, i.e., their prestige or popularity, also plays a significant role, showing a positive and relatively high correlation with polarization. Therefore, the predominance of more popular users is associated with more polarized networks. This observation aligns with the finding that political parties, due to their activity, coordination, and significance within the network, are linked to networks with greater polarization, unlike networks where they do not appear, such as those on reflection days.

10.7 Additional Results

Table 10.4: Average network metrics in the recommendation networks generated by each RA (general 2011).

Metric	Rnd	LIWC	W2V	MD-W2V	CF	TS	TS-W2V	DR	Friendship	PropFlow
Nodes	241.8000	244.0000	234.3000	227.8667	181.1667	188.5000	187.7333	165.4333	227.5000	233.7333
Edges	350.4000	344.8333	198.7333	315.7667	286.3000	233.9667	233.6667	255.7667	232.7333	237.4667
$\overline{Deg}$	7.2056	7.1344	7.4234	6.8618	6.1650	5.8839	5.8238	6.2709	7.4940	7.4551
D	3.3000	3.9667	1.7667	3.0333	3.3000	1.8333	1.7000	2.2333	1.9333	2.2000
E	0.0520	0.0507	0.0445	0.0496	0.0571	0.0417	0.0421	0.0580	0.0491	0.0471
Δ	0.0060	0.0058	0.0037	0.0062	0.0090	0.0069	0.0069	0.0097	0.0046	0.0044
Q	0.8131	0.8317	0.9417	0.8093	0.7561	0.8375	0.8323	0.7419	0.8889	0.8948
$\overline{CC}$	0.0168	0.0195	0.0044	0.0170	0.0200	0.0199	0.0198	0.0224	0.0099	0.0096
$\overline{EVC}$	0.0063	0.0060	0.0052	0.0073	0.0115	0.0095	0.0095	0.0120	0.0054	0.0056
$\overline{PR}$	0.0042	0.0042	0.0044	0.0045	0.0058	0.0056	0.0056	0.0063	0.0045	0.0044

Table 10.5: Average network metrics in the recommendation networks generated by each RA (general 2016).

Metric	Rnd	LIWC	W2V	MD-W2V	CF	TS	TS-W2V	DR	Friendship	PropFlow
Nodes	487.1333	483.1000	483.4667	453.3667	439.1000	436.1333	432.2000	438.6667	468.3333	479.8667
Edges	1064.4000	1014.6667	589.0000	940.3000	1007.5667	721.2333	710.2333	894.8000	574.4667	689.3667
$\overline{Deg}$	8.6515	8.7128	8.6917	7.8725	8.1836	8.1943	8.1183	8.2869	8.8008	8.7686
D	8.3000	8.7000	6.8333	7.5612	7.9000	4.3000	4.8667	5.6333	4.0000	5.7667
E	0.0459	0.0410	0.0323	0.0376	0.0483	0.0317	0.0315	0.0357	0.0331	0.0319
Δ	0.0045	0.0044	0.0025	0.0043	0.0052	0.0038	0.0039	0.0047	0.0027	0.0030
Q	0.8167	0.8443	0.9371	0.7898	0.8019	0.8531	0.8532	0.8242	0.9033	0.8810
$\overline{CC}$	0.0574	0.0690	0.0322	0.0596	0.0609	0.0790	0.0807	0.0899	0.0328	0.0398
$\overline{EVC}$	0.0025	0.0036	0.0032	0.0030	0.0033	0.0028	0.0029	0.0034	0.0029	0.0019
$\overline{PR}$	0.0021	0.0021	0.0021	0.0021	0.0023	0.0023	0.0024	0.0023	0.0022	0.0021

Table 10.6: Average network metrics in the recommendation networks generated by each RA (general apr 2019).

Metric	Rnd	LIWC	W2V	MD-W2V	CF	TS	TS-W2V	DR	Friendship	PropFlow
Nodes	441.5000	433.3667	431.7667	425.9333	406.2667	395.3333	393.8333	411.3333	420.2333	428.8000
Edges	956.2000	927.4333	567.6000	903.4333	903.4333	717.0667	710.5333	878.9333	584.1333	634.2667
$\overline{Deg}$	8.5097	8.6530	8.6582	7.6065	7.8543	8.0390	7.8375	7.8512	8.6125	8.7374
D	7.4333	7.1333	5.5333	6.0333	6.2333	5.3000	5.4333	5.5000	4.7667	4.6000
E	0.0414	0.0382	0.0321	0.0339	0.0400	0.0329	0.0321	0.0344	0.0354	0.0284
Δ	0.0049	0.0049	0.0031	0.0050	0.0055	0.0046	0.0046	0.0052	0.0033	0.0035
Q	0.8417	0.8616	0.9326	0.8368	0.8300	0.8568	0.8560	0.8307	0.8966	0.8955
$\overline{CC}$	0.0613	0.0743	0.0362	0.0646	0.0625	0.0849	0.0842	0.0792	0.0509	0.0425
$\overline{EVC}$	0.0028	0.0025	0.0029	0.0025	0.0025	0.0031	0.0031	0.0031	0.0031	0.0021
$\overline{PR}$	0.0023	0.0023	0.0023	0.0024	0.0025	0.0025	0.0026	0.0024	0.0024	0.0023

Table 10.7: Average network metrics in the recommendation networks generated by each RA (local 2015).

Metric	Rnd	LIWC	W2V	MD-W2V	CF	TS	TS-W2V	DR	Friendship	PropFlow
Nodes	440.2333	440.9333	434.5333	409.3000	388.1333	386.5333	384.6667	385.1000	422.7000	430.8000
Edges	829.5000	795.7333	467.6667	737.9000	755.6333	601.5000	594.3333	720.9667	500.0333	553.0667
$\overline{Deg}$	8.0147	7.9816	8.0956	7.1432	7.1449	7.1315	6.9456	7.1308	8.0684	8.1787
D	7.6667	7.0333	4.2333	6.7000	6.6333	4.8333	5.3000	5.4667	3.9667	5.2333
E	0.0428	0.0383	0.0301	0.0344	0.0395	0.0310	0.0302	0.0354	0.0312	0.0341
Δ	0.0043	0.0041	0.0025	0.0041	0.0050	0.0041	0.0041	0.0049	0.0028	0.0030
Q	0.8302	0.8619	0.9496	0.8048	0.8083	0.8761	0.8761	0.8112	0.9189	0.9003
$\overline{CC}$	0.0421	0.0543	0.0300	0.0396	0.0418	0.0616	0.0612	0.0494	0.0337	0.0323
$\overline{EVC}$	0.0029	0.0021	0.0027	0.0036	0.0043	0.0042	0.0042	0.0043	0.0029	0.0030
$\overline{PR}$	0.0023	0.0023	0.0023	0.0023	0.0026	0.0026	0.0026	0.0026	0.0024	0.0023

Part IV

Algorithmic Intervention Strategies

"I'm sorry, Dave. I'm afraid I can't do that."
— HAL 9000, *2001: A Space Odyssey*

Chapter 11

Algorithmic Approaches to Break Disinformation Networks

11.1 Introduction

Social networks are increasingly replacing traditional media as the primary source of information and public debate among citizens. This shift has facilitated the migration and amplification of phenomena such as disinformation onto these platforms, posing a significant threat to our political systems. In this context, an information network is a structured system of interconnected users who share and disseminate information. When such networks are used for spreading disinformation, as already discussed before in this thesis, they become particularly dangerous as they create echo chambers that can mislead and radicalize large audiences, with persistent and long-lasting effects. Disinformation networks, which are more dangerous than isolated actors, not only distribute fake news but also create authentic disinformative echo chambers. These echo chambers can confuse and radicalize large audiences and are persistent over time. To address this issue, we employed an information theory-based approach to maximize the structural diversity of text, thus generating a recommendation system capable of curbing the growth of disinformation networks while maintaining high levels of precision. We tested our algorithm using data spanning over three years of publications from both a disinformation network and a legitimate network of journalists in Spain. Our results demonstrate that our approach effectively curbs the growth of disinformation networks while maintaining stability and general accuracy in the legitimate journalist network. Our algorithm is innovative, as this is our method of addressing the disinformation problem from a network perspective. Our approach surpasses the current state-of-the-art in combating disinformation.

This research posits that it is feasible to devise a recommendation algorithm capable of maintaining satisfactory levels of accuracy while simultaneously reducing the formation and reinforcement of disinformation networks through an approach focused on optimizing textual diversity in the network's recommendations. The underlying rationale for this hypothesis is straightforward. In social networks, such as X, "text communities" emerge where users share opinions and interact [HYYP13, LJB19]. These communities are formed based on the similarity of their posts, with some closely aligned and others distinctly separate. A recommendation algorithm prioritizing accuracy alone may inadvertently foster echo chambers, where communities become insular, exchanging information only within a closed loop

[RARFN22]. Conversely, an algorithm that emphasizes diversity to the exclusion of accuracy might yield irrelevant recommendations, potentially disengaging users and leading to attrition due to uninteresting content [IFO15].

Our study introduces a recommendation algorithm that employs principles of Information Theory in order to bridge disparate text communities. By priming weak ties between these communities, the algorithm promotes interactions that introduce diversity in the network while keeping a relevant recommendation accuracy. We consider this approach to be novel because, to our knowledge, it is the first study that tries to solve the problem of disinformation propagation in social networks by focusing on the disinformation networks phenomenon and the prevention of their formation, as opposed to strategies based on fact-checking and the study of fake news diffusion through the network, which have already been proposed in the existing literature [Sal22, WXZ+22, IAP24].

This chapter delves into the phenomenon of disinformation in online social networks, aligning with the broader objectives of this thesis to understand, analyze, and mitigate the dynamics of harmful information networks. Specifically, it addresses the third research goal of the thesis, which focuses on analyzing the role of recommender systems in promoting or mitigating the spread of disinformation. By exploring these interconnected aspects, the chapter supports the overarching aim of the thesis to provide innovative, algorithmic strategies to enhance the resilience of digital information ecosystems.

The chapter is organized as follows. Section 11.2 introduces background terminology, that, although it has already been mentioned before in the thesis, it is important to have in mind to properly understand this chapter. Section 11.3 details the methodology, including the datasets used and the design and implementation of recommendation algorithms. Section 11.4 introduces our proposal for recommendation algorithms to break disinformation networks. Section 11.5 presents the experimental setting and Section 11.6 shows the results, showcasing the effectiveness of a novel information-theory-based recommendation algorithm in reducing the formation of disinformation networks while maintaining high levels of recommendation accuracy. Section 11.7 discusses these findings in light of the thesis's research goals, particularly the balance between recommendation accuracy and mitigation strategies outlined in the third research goal. Finally, Section 11.8 concludes by summarizing the chapter's contributions to the thesis and emphasizing its role in advancing the understanding and prevention of disinformation propagation in online social networks.

Research questions

Our aims in this chapter are summarized in the following research questions included in the third research goal of the thesis:

- **RG3:** Analyze the role or contribution of recommendation systems in promoting these phenomena. Propose mitigation strategies through these systems.
  - **RG3.RQ3:** How can recommendation systems help to mitigate these phenomena?
  - **RG3.RQ4:** How do the proposed mitigation strategies affect recommendation accuracy?

11.2 Background

11.2.1 Online Social Networks as a source of information

Since their emergence and mass adoption around 2006, online social networks have progressively supplanted traditional media as primary information sources for a significant portion of the population [BD16]. The ease of use, combined with widespread smartphone adoption and the capability for rapid, short-form communication, has transferred conventional political participation dynamics to the digital realm, resulting in expanded forms of political engagement and new phenomena [RARFN22]. Online social networks facilitate narratives that bypass the editorial control of traditional media, often linked to the political establishment, thus consolidating alternative powers [BM13, BD16]. Platforms like X have emerged as key venues for political discourse, amplifying participation and enabling digital activism, such as launching campaigns, connecting remote activist communities, and coordinating large-scale political projects [Tuf17, Sán15]. Additionally, social networks have fostered endogenous phenomena like the rise of fake news, cyberbullying, and the creation of echo chambers, which facilitate political polarization by repeatedly exposing individuals to filtered content that aligns with their preexisting beliefs [FS16, CRA14, CRF$^+$11, MBP18, HSS13]. Among these platforms, X stands out as one of the primary social network for real-time political commentary and debate. X has been used to coordinate street protests [LNRT19], follow events of great social interest in real-time, facilitate the emergence of new political parties, and serve as a political communication platform for various actors and organizations [MH14, VIHP14]. This situation has significantly facilitated the expansion of preexisting phenomena, such as political propaganda and disinformation, onto X and similar platforms [Jon19]. Due to their rapidity and high volume of communication, these social networks provide particularly fertile ground for maximizing their distribution and generating substantial social impact.

11.2.2 Disinformation

The emergence of this new information production and consumption environment on the Internet potentiated a well-known social phenomenon, the spread of disinformation. Disinformation, as previously presented in Section 3.1.2, is the intentional dissemination of false, inaccurate, or misleading information designed to deceive or manipulate the public for purposes of causing harm or gaining profit [KCBP21]. It involves creating and propagating such content, often through digital and social media platforms, to influence public opinion, incite social discord, or achieve strategic objectives [AAB23]. The malicious intent behind disinformation distinguishes it from misinformation, which may occur without the aim to deceive. The pervasive spread and rapid dissemination of disinformation pose significant threats to democratic processes, public health, and societal stability, exacerbating existing biases and fostering environments ripe for conflict and radicalization [BFR18].

While organized disinformation as a social phenomenon has existed since the human race developed massive information spread systems such as the printed press, its political use and spread peaked in the internet era. Modern-day disinformation spread through the Internet can be articulated through various techniques, tactics, and procedures (TTPs), each with different aspects of operation and causes. Some include the design or instrumentation and spread of click baits, conspiracy theories, fake news, arbitrary relations, hoaxes, biased or partial information, impersonation, pseudoscience, gossip, fake reviews, or many forms of pranking [KCBP21, WTBC19].

11.2.3 Disinformation networks

Disinformation is habitually distributed by numerous actors, primarily with political objectives. These political objectives can be very local and niche, such as the propagation of conspiracy theories by small self-radicalized groups or involve political warfare between different national parties or political groups. On a larger scale, disinformation can be entirely strategic, used by state actors to advance their agendas in foreign policy. Naturally, when the dissemination of information is supported by large political groups or states with resources, the associated risk is maximized. The distribution of disinformation can impact society as a whole, undermining trust in the state, exacerbating public health issues, and generating various types of disturbances. Technically, disinformation can be spread by individual actors, automated accounts (bots) networks, organized collectives, or a hybrid of these. Disinformation reaches its maximum impact on social networks when distributed within "disinformation networks."

As it was previously introduced in Section 2.14, a disinformation network is a system of interconnected accounts on social media platforms that actively collaborate, either implicitly or explicitly, in order to disseminate false information and deliberately deceptive narratives [MDB24]. These networks are characterized by their high activity levels, strong interconnectedness, and coordinated efforts to spread disinformation. The primary objectives of these networks can vary widely, including political gain, social discord, discrediting individuals or organizations, or manipulating financial markets. Disinformation networks often employ sophisticated strategies such as presenting distorted facts, fabricating stories, or decontextualizing truths to influence public opinion and shape narratives. Their structure typically includes clusters of tightly-knit groups, high network density, and influential nodes or "conversation leaders" that facilitate rapid information propagation, and they present a set of topics of special coverage and a common narrative around them [MDB24]. They also present a very characteristic discursive style as they often employ slogans, hashtags, jargon, and specific wording such as "dog whistles" in order to enhance their own network coherence.

Due to their echo chamber structure, high levels of activity, and coherent narratives around politically charged topics targeted by disinformation, these networks pose a significant risk to healthy political and social debate on online social networks. Non-radicalized users can fall into these networks and be misinformed or even radicalized due to the high levels of activity they exhibit. False information reaches these users more quickly and from more "sources" than truthful information. Similarly, these high activity levels allow such networks to significantly impact the overall conversation, contaminating the debate. Naturally, content distribution systems on online social networks like X play a fundamental role in mediating the formation of these networks and, in general, in the dissemination of disinformation.

11.2.4 The role of recommender systems

As mentioned several times throughout this thesis, recommender systems in online social networks are software tools designed to suggest content to users based on their preferences and behaviors [RRS11]. These systems analyze user ratings, interactions, and profiles to predict and recommend products, services, or content that the user will likely find appealing. They are fundamental in personalizing user experiences by automatically curating information presented in formats like "news feeds" or lists of recommended contacts. Regarding the existing literature, as introduced in Section 2.16, there are several types of recommender systems including collaborative filtering, content-based, and hybrid recommenders [BOHG13, Bur02], although other studies in the literature propose a taxonomy based on four

different families of recommendation algorithm: Collaborative Filtering, Content, Deep Learning, and Reinforcement Learning-based recommendation algorithms [ASIM18].

Content-based recommender systems generate recommendations by analyzing items' attributes and users' past interactions with similar items. They build a profile for each user based on the characteristics of the items they have positively interacted with. This approach effectively addresses the cold-start problem for items by recommending new items similar to those the user has previously liked [BCC+17].

Collaborative filtering predicts user preferences by analyzing past behaviors and similarities between users or items. It uses historical data to recommend items liked by similar users [BOHG13]. Techniques include user-based and item-based filtering. User-based collaborative filtering recommends items that similar users have liked, while item-based collaborative filtering recommends items similar to those the user has liked. This method does not require item attribute data, making it versatile but susceptible to the cold-start problem for new users and sparsity issues in the user-item interaction matrix [SK09], although techniques like matrix factorization, including Singular Value Decomposition (SVD) and its variants, have been extensively used to address issues of scalability and sparsity in collaborative filtering [Hut09, KCN+18, RD22].

Hybrid methods combine the strengths of content-based and collaborative filtering approaches to mitigate their limitations. These methods can be integrated in various ways, such as a weighted combination, sequential combination, or even switching between methods depending on the context. Hybrid systems often show improved performance and accuracy, leveraging the benefits of both underlying methods [Bur02, KLPC22]. Recent advancements have integrated machine learning and deep learning techniques, significantly enhancing the capabilities of recommender systems by capturing non-evident user-item interactions and temporal patterns in data [KLPC22].

Recommender systems face particular challenges and opportunities in social networks and news applications. Social networks provide rich user interaction data, which can be leveraged to build precise user-profiles and enhance recommendation accuracy. Graph-based collaborative filtering and social influence models have shown promising results in this domain by effectively utilizing the network structure and user interactions. In news applications, the continuous influx of new content and the need for real-time recommendations pose significant challenges. Hybrid methods that combine content-based filtering with collaborative approaches are particularly effective here, as they can quickly adapt to new articles while refining recommendations based on user interaction data [RD22].

Naturally, as the entities responsible for content distribution on a social network, recommender systems and especially their underlying algorithms have a significant responsibility in shaping the structure of the social networks within the platforms, they also bear a particular responsibility in promoting or moderating the dissemination of harmful information on the network [WXZ+22].

11.2.5 Algorithmic intervention

Recommender systems, as the primary engine of social media's information landscape, significantly influence the spread of disinformation due to algorithmic biases such as popularity bias and the reinforcement of filter bubbles [PS24, DRDB17]. These systems, often trained on biased datasets, create a self-perpetuating loop that isolates users in intellectual silos, increasing their exposure to content that aligns with preexisting beliefs and making them more susceptible to misinformation [E+20]. Studies reveal that algorithms emphasizing popularity disproportionately recommend misinformation, acceler-

ating its propagation, while neighbor-based methods distribute it more uniformly but still contribute to its spread [FB20]. Efforts to mitigate these effects include reranking strategies that prioritize diversity, such as modified collaborative filtering and greedy optimization techniques, which balance recommendation accuracy with reducing echo chambers [BNS18, GSM+22, AK12]. However, current research, as already discussed in Section 2.18, such as the OHARS initiative, highlights the limited focus on disinformative communities and the broader challenges in addressing echo chambers, underscoring the need for more targeted interventions [TGZ21].

11.3 Specific Methodology

In this section, we present the dataset that we employed to carry out this work. We also detail the set of recommendation algorithms that we implemented to evaluate our proposal, as well as the methods and techniques that we employed in such evaluation.

11.3.1 Data set

For this study, we employed the dataset "Information disseminators" detailed in the thesis methodology (Chapter 4) and composed of two main sub datasets: one containing X accounts linked to verified journalists (Table 4.16) and another focusing on disinformation actors (Table 4.15). Note these datasets were used in Chapter 7 to analyze the behavior of disinformation networks, and in Chapter 9 to understand the impact of recommendation algorithms in this context.

Here, we were interested in the discovery of a recommendation algorithm capable of maintaining a relevant recommendation accuracy while improving text diversity, as a strategy to mitigate the generation and consolidation of the aforementioned disinformation networks [MDB24]

Indeed, we employed the merged dataset to feed the selected recommendation algorithms to generate recommendation networks. As noted earlier (see Section 2.17), these "recommendation networks" are simply graphs where users are linked to one another based on the recommendations that the recommendation algorithm created. More specifically, in these networks, a user A is linked to another user B (A $\rightarrow B$) if the former user was recommended a post authored by the latter user, resulting in a weighted and directed graph. These recommendation networks contain both disinformative and non-disinformative accounts, however, since we need to evaluate the capability of our algorithm in preventing the generation and consolidation of disinformation networks, in certain parts of our study, described in the subsequent sections, we filter the generated recommendation network to only keep nodes belonging to the original disinformation dataset and their corresponding connections, consequently obtaining networks formed by disinformation agents - disinformation networks. This approach allowed us to rigorously evaluate the original hypothesis stated previously.

11.3.2 Disinformation countering recommendation algorithm selection

In order to compare our recommendation algorithm proposal, we implemented a selection of classical and state-of-the-art recommendation algorithms, covering the main different strategies and approaches described in the existing literature [BOHG13, Bur02, ASIM18].

First, we implemented a classical k-NN Collaborative Filtering approach [Pet09], making recommendations based on user similarity, calculated through cosine distance as has already been performed

in existing research [KGSS15], and employing the knowledge of the local neighborhood to create the recommendation lists of each user. Within the Collaborative Filtering-based approaches, we also implemented the Friendship and PropFlow recommendation algorithms [AA03, LLC10]. The former employs a metric to friendship between users based on the Jaccard Similarity [DDDD09] to carry out user recommendations in a Collaborative Filtering-based manner. On the other hand, the PropFlow recommendation algorithm utilizes random walks over the network topology to calculate the top-N recommendation lists of the users. We also implement the SPGreedy algorithm proposed in [ZBZ21]. Such algorithm utilizes the network structure to diversify recommendations, breaking disinformative echo chambers through the iterative selection of the best possible recommendation for each user to build the top-N ranking recommendation list per user and achieving a trade-off between relevant recommendation accuracy and a significant reduction in disinformation propagation. In this case, we selected this algorithm due to its novelty and excellent results in related problems [ZBZ21].

Regarding content-based approaches, we implemented a content-based recommender based on the employment of Word2Vec to learn embeddings from the corpora –formed by all the posts of the training partition in each time window [MCCD13]. We averaged the Word2Vec embeddings of all the words per post, so as to obtain a representation of the posts published by each user, with which we calculated N-dimensional vectors that model user preferences through a content-based approach. We then recommended to each user the posts that were most similar to the user preferences, utilizing the cosine distance to calculate the distance between the user preferences and the posts to be ranked.

Furthermore, we also implemented the state-of-the-art hybrid Deep Learning-based recommender proposed by the authors in [CZ20]. This algorithm utilizes an embedding layer to learn the representations of both users and posts, to then learn the optimal list of ranked items that each user should be recommended for each time window, utilizing the foundations of Deep Learning as the manner to optimize such ranked recommendations. While authors propose both a top-N and a pairwise-ranking variety of their original method, we implemented the latter for efficiency reasons [CZ20].

To cover the full spectrum of approaches with which to implement recommender systems, we also implemented a Multi-Armed Bandit recommender based on the Thompson Sampling strategy [CRC19a]. This Reinforcement Learning-based recommender system is based on the creation of a multi-armed bandit for each user. This bandit is in charge of learning, in a stochastic manner, the top-N recommendation list that each user should receive for each time window. To train, each tweet of the training partition is considered as an event with a known outcome (retweeted or non-retweeted), which is used to adjust the probability distributions of the arms. Then, the bandit is played for each post in the test partition to obtain the recommendation list of each user. We opted to implement this approach due to its popularity and success in well-known use cases, as it is the case with the well-known Yahoo News Recommender [CL11, LCLS10].

Besides covering the different families and approaches of recommender systems, we also implemented state-of-the-art methods that aim to break disinformation networks by minimizing or reducing echo chambers as a manner of fighting against disinformation networks in the field of recommender systems. First, we implemented approaches that are based on the application of reranking over already existing algorithms. Regarding these approaches, we implemented the MMR algorithm described in [GSM$^+$22], which is based on Collaborative Filtering and a reranking to favor diversity, breaking disinformation networks through the creation of diverse links, to counter the formation of disinformative

echo chambers. In the same line, we implemented the Greedy Optimization algorithm described by the authors in [AK12], built upon the same foundations as the previous algorithm. In this case, authors do not recommend any particular recommendation algorithm strategy for their reranking based method, Greedy Optimization, thus we created four different versions of this algorithm, covering the main four families of recommendation algorithms devised in Section 11.2. More specifically, we built a Collaborative Filtering Greedy Optimization approach, a Word2Vec content-based Greedy Optimization approach, as well as Multi-Armed Thompson Sampling Bandit and DeepRank Greedy Optimization recommenders, so as to evaluate our algorithm with the best possible setting of the Greedy Optimization recommendation algorithm strategy. These algorithms are state-of-the-art recommendation algorithms that aim at fighting against disinformation during the generation of recommendations. Some other approaches, uncovered in this research, are related to moderation strategies that occur once recommendations have been generated and that rely on manual annotation of third-party APIs to fact-check against disinformative or misinformative content on the network [PEOOAP21, KCHM23].

Through the implementation of the described recommendation algorithms, we aim to perform a rigorous evaluation of our proposed recommendation algorithm, so as to reliably evaluate its utility to break disinformation networks that occur in reality, by feeding each of the recommendation algorithms with the dataset described previously.

11.4 An information theory-based intervention for recommendation algorithms

Our proposal is based on the intuition that favoring text diversity through the creation of weak ties helps avoid certain characteristics that lead to the formation and consolidation of disinformation networks, like the appearance of disinformative echo chambers were information flows very efficiently due to the high efficiency and density of the generated networks (and their low fragmentation), as explained in [MDB24].

This foundation of employing weak ties to foster diversity in the context of recommendation systems is not new, and many works have been developed on the idea of establishing weak ties as a manner to create healthier social network ecosystems [Gra73, SC18, Wel16]. However, our research is, to our knowledge, the first one to introduce Information Theory concepts and techniques as a cornerstone to break disinformation chambers through the power of weak ties among the text communities that emerge from the social network.

As discussed in Section 11.2, a document can be represented through a dense representation using well-known natural language processing techniques [Chu17], like Word2Vec. Our algorithm further enriches this representation through the introduction of Information Theory concepts.

More specifically, our approach calculates the Shannon entropy contribution of each word in a given document [Sha93]. Using the Shannon entropy equation, the entropy contribution of each word in a word collection, i.e., a document, can be computed as described in Equation 11.1, where $p(w_i)$ represents the probability of the word i occurring in a document with N distinct words, D represents such a document, and $H_{w_i}(D)$ represents the Shannon entropy contribution of word w_i in the entropy of document D.

$$H_{w_i}(D) = -p(w_i)\log_2 p(w_i) \tag{11.1}$$

Indeed, our algorithm extends the averaged Word2Vec representation of the document with the Word2Vec representations of the words with lowest and highest entropy contributions. The intuition behind this enrichment is simple: a document can be represented with the dense representation obtained from the averaged Word2Vec algorithm, however, extra information can be provided to model specific nuances of the document, such as the "surprise word" of the document, i.e., the word with the highest entropy contribution, and the "expected word" of the document, i.e., the word with the lowest entropy contribution. This extended representation of a document further improves the document representation capability of our recommendation algorithm proposal through the employment of Information Theory, which also introduces information dynamics in our proposed recommender system, because Information Theory-based metrics, like entropy, can be used to represent and model temporal flow of information across a social network [YWC+20]. The introduction of Information Theory makes our proposal achieve good results in terms of countering disinformation networks –see Section 11.6– while introducing uncovered techniques in existing recommender systems research, as to our knowledge there have been no proposals of counter-disinformation recommender systems based on Information Theory [BOHG13, Bur02].

In general, our algorithm consists of several steps. The first step is the calculation of the document representation. First, all the documents available at a specific time window must be represented through the aforementioned document representation, leveraging the use of Information Theory to create a precise representation of the documents, where each post is described through a dense vector composed of three parts: (I) the averaged Word2Vec representation of the post, (II) the Word2Vec representation of the "surprise" word of the post, i.e., highest entropy, (III) and the Word2Vec representation of the most common word in the post, i.e., the word with the lowest entropy.

Secondly, our algorithm applies text clustering using KMeans, employing Principal Component Analysis (PCA) to avoid the dimensionality problem [Pea01]. As it can be understood, the representation of each post has many dimensions, namely, three times the size of the word embedding employed in the Word2Vec algorithm. High dimensionality is one of the most well-known problems in text clustering [AZ12], lowering the performance of the resulting groups if it is not specifically tackled [B+10]. As a result, we apply the well-known statistical technique PCA to reduce the dimensionality while keeping a relevant value of explained variance –set to 80% as has been done in existing research to tackle the trade-off between high dimensionality and relevant results [KL20]. It must be noted that we opted to use KMeans as it is a well-known algorithm for text clustering, which has already been used in similar works and facilitates further expansion and refinement through the introduction of different distance metrics and enhancements over the base algorithm [JNXH05, Lam13, Xin12].

The third step of our algorithm is calculation of user clusters from the identified text clusters. Once all the documents have been assigned a text cluster, users are assigned the text cluster that contains the majority of the documents they created. Intuitively, this approach facilitates the creation of weak ties between communities to break disinformative echo chambers through an increase in text structural diversity, because users can receive content recommendations from other clusters different to their own. In this context, weak ties are essential to dismantling disinformation networks within our proposed recommendation algorithm. By facilitating broader information diffusion, these connections ensure that users are exposed to diverse perspectives from various text communities. This diversity not only enriches the social network but also promotes the spread of varied and credible content across all communities within

the network, thereby reducing the dominance of misleading narratives and strengthening the overall resilience of the information ecosystem.

The fourth step involves the calculation of the user preferences. Indeed, user preferences are key to recommender systems, as content recommendations should be as close as possible to user preferences, so as to guarantee relevant recommendations to the users [CSS15]. For our algorithm, we calculate user preferences as the averaged representation of the documents published by the user, and the retweets available to each user within the training partition. Thus, our enriched representation of the documents has an impact in the representation of user preferences, which helps to introduce information dynamics, extracted using Information Theory, into the user profiles that will be used for ranking the available items in the evaluation partition of the dataset at each time window.

Before user-item ranking calculation, inter-cluster distances must be computed. Intuitively, we compute the distance between two text clusters as the distance between their two centroids, and these measurements allow us to introduce text diversity in the recommendations within the user-item ranking function. Consequently, the last step of our proposed recommendation algorithm is the user-item ranking calculation. Considering all the documents in the evaluation partition of the dataset at a specific time window, we rank the items for each user as expressed in Eq. 11.2, where $R(d, u)$ indicates the ranking of the document d for user u, C_u and C_v represent the clusters of the user u and v respectively (where user v refers to the author of document d), $pref_u$ represents the user preferences of user u. Furthermore, the inter-cluster distance factor is calculated based on Eq. 11.3, whereas the document similarity factor is computed using the cosine similarity function, which is a widely used function in the domain of recommender systems [SCN16]. In the ranking that we propose, a document d_1 with higher ranking value than a document d_2 for a user u is a document with better ranking, i.e., the higher the ranking value, the more suitable the document is for a user.

$$R(d, u) = (1 + \text{intercluster\_distance}(C_u, C_v)) \cdot \text{document\_similarity}(d, pref_u) \tag{11.2}$$

$$\text{intercluster\_distance}(C_u, C_v) = 1 - \text{cluster\_similarity}(C_u, C_v) \tag{11.3}$$

The intuition behind the ranking formula is simple. The diversity algorithm we propose must favor the creation of weak ties between different text communities to prevent the creation of disinformative echo chambers and create more sane and diverse networks. To achieve this, we introduce the inter-cluster distance factor, which favors the selection of documents from a cluster that is not the user cluster C_u. However, we still want to achieve a certain balance between the generated diversity and the recommendation accuracy achieved by the algorithm, so that the recommendation algorithm provides relevant recommendations for the users of the social network. For that reason, we introduce the $+1$, so as to avoid zero-valued rankings for documents whose author belongs to the same cluster as the author. Indeed, we achieve a trade-off between relevant recommendation accuracy and the improvement of structural text diversity through the combination of the document similarity and inter-cluster distance factors.

11.5 Experimental settings

11.5.1 Procedure

In order to validate our recommendation algorithm proposal, we performed simulations with the real data described in Section 11.3.1. More specifically, we divided the data into weekly tweet collections. As social network conversations are dynamic, topics that are treated today may not be treated in the near future [PMS23]. As a result, choosing a correct granularity, i.e., size of the window, for studying the behavior of recommendation algorithms is key to their correct functioning. In our case, we selected a weekly granularity for the time windows, as it offers a trade-off between having very small time intervals, like hours and days, and very big time intervals, like months. Indeed, a weekly granularity also allows the detection of topic changes from one window to another, hence we considered this decision reasonable. It is worth noting that this temporal granularity has already been devised in existing literature and employed successfully in real use cases [LHCA10].

Furthermore, we applied the same experiment to all the implemented recommendation algorithms. For each time window, we split data into training and evaluation partitions, training the recommendation algorithms with the training data to get a top-N ranking prediction per user with the evaluation partition. Such result allows us to generate a recommendation network per time window, which we model as a directed and weighted graph, whose nodes represent users and whose edges represent recommendations between users. User A is connected to user B if the top-N recommendation list of user A contains at least one tweet from user B. The weight of the relationship represents the number of recommendations that user A received from user B in its recommendation list.

Indeed, the result of the experiment is a series of networks that show how each recommendation algorithm provides top-N recommendation lists to each user across the studied intervals of time. Such result allowed us to apply statistical methods to evaluate whether the research hypothesis is correct, that is, whether it is feasible to create a recommendation algorithm built upon the idea of optimizing text diversity as the strategy to disarm disinformative echo chambers while keeping a relevant recommendation accuracy. Furthermore, this experiment also allowed us to reliably evaluate our proposal with classical and state-of-the-art recommenders aimed at breaking disinformation networks, which we use to prove the utility of our novel algorithm.

11.5.2 Validation methods and techniques

To evaluate our proposed recommendation algorithm, we first calculated the disinformation exposure metric [AGAC22], promoted by the European Commission and already devised in existing literature [AGAC22, HAN19, SADH$^+$22, LdFP$^+$21]. With such metric, we aim to calculate the average exposure of non-disinformation agents in a network, as a result of their activity and interactions with other users. The equation employed to calculate the average disinformation exposure per user is shown in Eq. 11.4, where $\overline{DE}$ represents the average disinformation exposure in the network at a specific time window, N_{ND} represents the set of non-disinformation users in the network, and $I_D(u)$ represents the interactions (if any) between a given user u and all the disinformation agents in the network.

$$\overline{\mathrm{DE}} = \frac{\sum_{u \in N_{ND}} |I_D(u)|}{|N_{ND}|} \tag{11.4}$$

Indeed, disinformation exposure gives an idea of the impact of disinformation agents in the network. We employed this metric so as to evaluate our proposed recommendation algorithm with respect to classical and state-of-the-art recommendation algorithms, including approaches tailored to reduce disinformation dissemination in the network.

While disarming disinformation networks is important, keeping a reasonable recommendation accuracy is a key aspect in the development of recommendation algorithms, as it was explained previously. Consequently, we also evaluated our proposal using the precision metric, which measures the number of correct tweet recommendations of each recommendation algorithm with respect to the total recommendations created [GS09]. It must be noted that the notion of "correct" refers to tweets that were retweeted by the user in reality (i.e., those that match the test set).

Besides the computation of disinformation exposure and recommendation algorithm precision, we computed the ratio between both metrics to calculate an intuition of balance between recommendation accuracy and disinformation dissemination reduction. Indeed, we aim to compare our proposed recommendation algorithm with the selected recommendation algorithms, so as to infer which recommendation algorithm offers the best possible trade-off between a high recommendation accuracy and a low average disinformation exposure.

Last, we compared our proposed algorithm with respect to the real data, which follows X's recommendation algorithm. For the comparison, we applied the well-known Mann-Whitney U test [MW47], a non-parametric method that does not assume data follows a specific distribution and can be used to compare the temporal series of "precision vs disinformation exposure" ratio between our proposal and the X algorithm, so as to determine if our proposal achieves significantly better results than algorithms currently employed for trending social networks, as it is the case with Twitter.

11.6 Results

Considering the implementation of the classical and state-of-the-art recommendation algorithms devised previously, and considering our proposal of recommendation algorithm to counter disinformation by breaking disinformation networks through the increase of text diversity, we carried out the experiment described in Section 11.3. First, we calculated our evaluation metric as defined in Section 11.5.2, disinformation exposure, and its evolution over time for each of the recommendation algorithms, as it can be seen in Fig. 11.1.

The obtained disinformation exposure time series allows us to understand that our proposal of recommendation algorithm (named as *Information Theory-based Countering Disinformation*) seems to effectively disarm disinformation networks, consistently reducing the levels of average user disinformation exposure over time. While our recommendation algorithm is not the only one that seems to effectively reduce disinformation exposure over time (see the multi-armed bandit-based algorithms and their results), it stands as the algorithm that reduces the most the average disinformation exposure of the network with respect to the average disinformation exposure in the original network (Twitter's network, represented in the figure as *Real Disinformation Exposure*).

As opposed to our algorithm, some other classical and state-of-the-art algorithms, such as DeepRank, do not achieve a reduction in the average disinformation exposure levels of the network. Instead, some of the implemented approaches significantly increase users' disinformation exposure. Furthermore, we observe how the recommendation algorithms that maximize the disinformation exposure metric the most

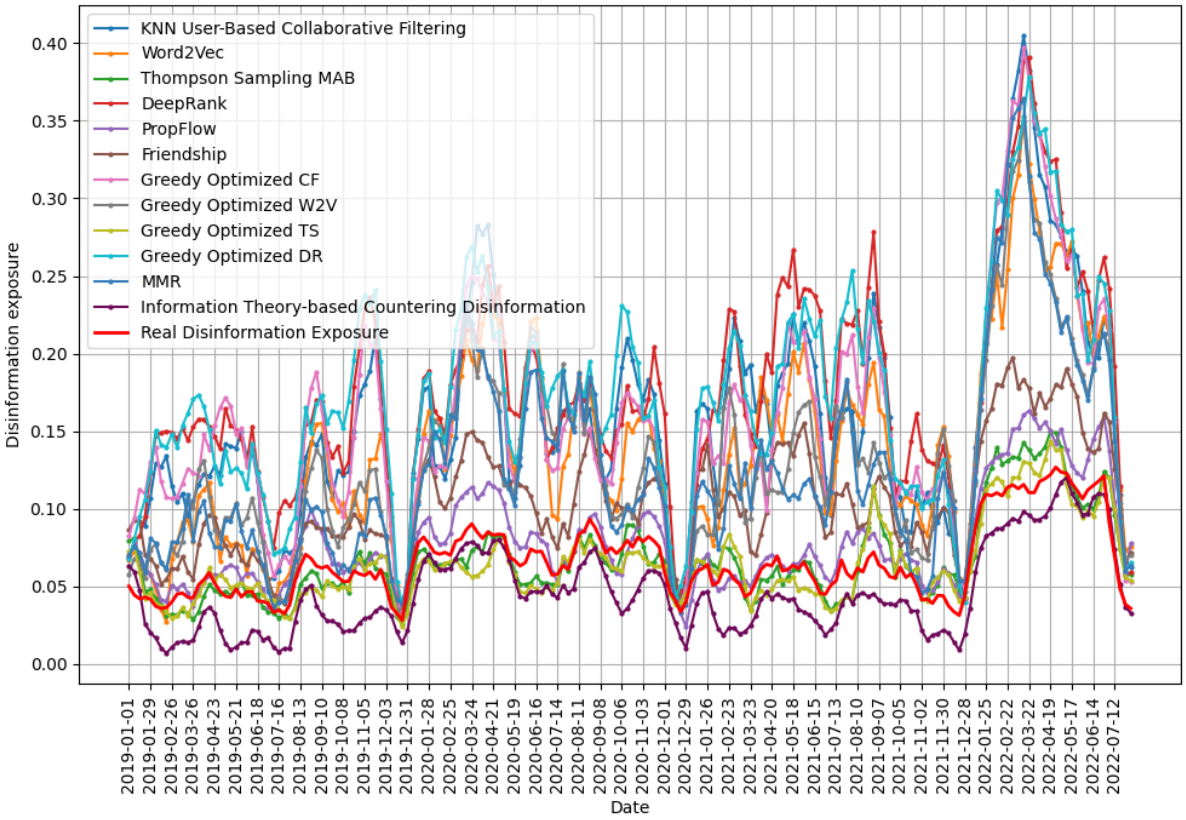


Figure 11.1: Disinformation exposure evolution from 2019 to August 2022.

belong to the family of content-based and hybrid algorithms, as it is the case with Word2Vec, DeepRank, or the greedy optimized version of the latter. As it can be seen in the accuracy results described in the following paragraphs, those algorithms seem to be capable of effectively recommending each user what it demands, at the cost of fostering the formation of disinformation networks, where disinformation can spread both easily and efficiently.

In order to guarantee that our recommendation algorithm significantly reduces disinformation exposure over time, we calculated the aforementioned one-sided Mann-Whitney U test, in order to check whether the real disinformation exposure time series median is significantly greater than the disinformation exposure median of the time series generated by our proposed recommendation algorithms. This statistical test delivered a statistic $U = 25913$ and p-value $= 2.6378e^{-15}$. Thus, using a confidence level $\alpha = 0.05$, we can assure that there is statistical significance in the fact that our proposed recommendation algorithm effectively reduces disinformation exposure with respect to its original levels in Twitter's (X's) social network.

However, our proposed recommendation algorithm must still achieve decent recommendation accuracy levels to be useful, because recommendation accuracy has a direct impact on user satisfaction, which can affect the user base of the social network if recommendation accuracy levels are too low. In the light of this need, we first calculated the ratio that relates the algorithm's recommendation accuracy with the disinformation exposure metric, so as to have a metric to measure the balance between having a good recommendation accuracy while still lowering disinformation exposure over time (see Fig. 11.2).

Considering the obtained results, we observe how our proposed recommendation algorithm achieves a good balance between recommendation accuracy and disinformation exposure, surpassing state-of-the-

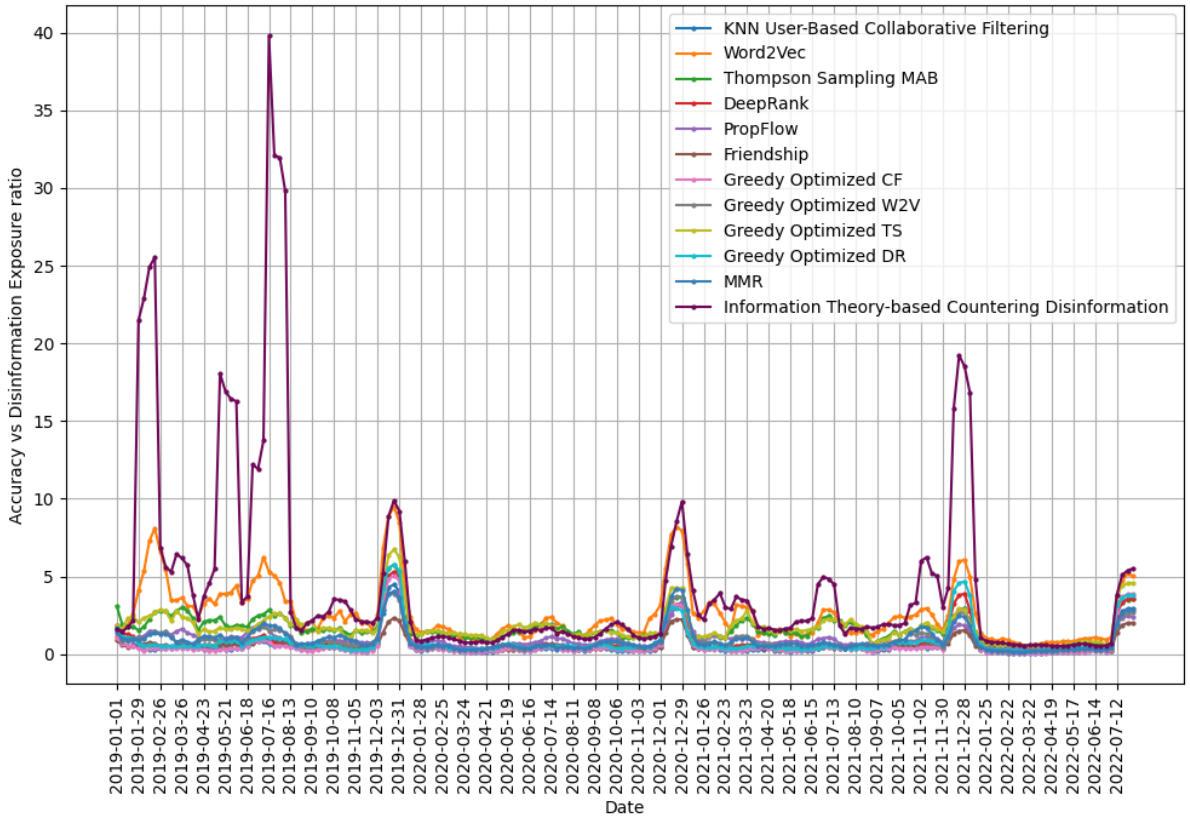


Figure 11.2: Recommendation accuracy vs disinformation exposure ratio evolution from 2019 to August 2022.

Table 11.1: Recommendation accuracy summary, sorted by accuracy values.

Algorithm	**Accuracy**
Word2Vec	0.2174
Greedy Optimized DR	0.0840
DeepRank	0.0837
Greedy Optimized TS	0.0810
Thompson Sampling MAB	0.0806
Information Theory-based Countering Disinformation	0.0684
MMR	0.0672
Greedy Optimized W2V	0.0662
PropFlow	0.0554
Greedy Optimized CF	0.0451
KNN User-Based Collaborative Filtering	0.0434
Friendship	0.0393

art recommendation algorithms as DeepRank, which seem to achieve high recommendation accuracy values (see the accuracy results shown in the following paragraphs) but poor results in the reduction of the disinformation exposure metric. Similar to our algorithm, we find that the content-based Word2Vec recommender also achieves a good balance between both metrics, due to its very high recommendation accuracy, which compensates the increase in disinformation exposure (with respect to X recommendations) that it causes as a side effect of its application.

Last, the accuracy results of each of the implemented recommendation algorithms are found in Table 11.1, which contains the average recommendation accuracy achieved by each recommendation algorithm during the whole period of time, from January 2019 to August 2022. As it can be seen from the results, the content-based Word2Vec recommendation algorithm clearly surpasses the other recommendation algorithms. Indeed, Word2Vec maximizes recommendation accuracy at the cost of a significant increase in disinformation exposure with respect to the original disinformation exposure levels. As a result, the averaged ratio between recommendation accuracy and disinformation exposure that it achieves is the second highest one with a value of $R = 2.563$, smaller than the one achieved by our proposed recommendation algorithm $R = 4.402$. Our recommendation algorithm achieves decent recommendation accuracy values, close to other state-of-the-art algorithms, as it is the case with DeepRank or the multi-armed bandit-based approaches, however, it is the recommendation algorithm that minimizes the most the disinformation exposure levels considering all the implemented approaches.

As a result, we observe how our proposed recommendation algorithm seems to effectively break disinformation networks through the application of Information Theory techniques aimed at increasing text diversity through weak ties, with which we aim to break and prevent the formation of disinformation networks. This algorithm also achieves decent recommendation accuracy values, obtaining the best balance between recommendation accuracy maximization and disinformation exposure minimization. Thus, we consider our novel approach to be an effective policy to prevent the formation and consolidation of disinformation networks in Online Social Networks like X.

11.7 Discussion

In this study, we introduced a novel algorithm for recommender systems that promises to mitigate the formation and consolidation of misinformation networks. While the algorithm demonstrates potential, it is important to recognize certain limitations and identify avenues for future work.

One primary limitation of this research is transversal to the whole thesis: the scope of the data used. The data set employed captures user interactions only within a defined temporal interval, not a complete network. This snapshot approach may overlook evolving user behaviors and temporal trends, which could influence the effectiveness of the recommender system. Hence, future research should apply our algorithm across different time frames and contexts to evaluate its consistency and robustness under varying conditions.

Furthermore, due to the vast array of recommendation algorithms available in the literature, our implementation was limited to a selected group that represents the major families of these algorithms. While this approach allowed us to cover a broad spectrum of methods, it necessarily excluded many state-of-the-art algorithms that could potentially enhance or provide new insights into the efficacy of our proposed solution.

Several critical areas for future work emerge from the current study. First, further refinement of the algorithm is needed to better balance disinformation exposure and recommendation accuracy. This could involve the introduction and adjustment of hyperparameters that control the generation of weak ties between different text-based communities. By introducing and fine-tuning these parameters, we can explore how different levels of connectivity impact the dissemination of information and misinformation.

Additionally, a comparative analysis with a wider array of state-of-the-art recommendation algorithms would be beneficial. Such an analysis could provide deeper insights into where our algorithm

stands relative to the cutting edge in recommender system research, highlighting strengths and uncovering potential weaknesses.

11.8 Conclusions

Our research has contributed significantly to the field of recommender systems and their role in preventing the formation of disinformation (**RG3.RQ3**), particularly in addressing the dual challenges of maintaining relevance and enhancing diversity in recommendations (**RG3.RQ4**). We have developed an algorithm that not only rivals the precision of state-of-the-art approaches like DeepRank and the Multi-Armed Bandit algorithm family but also excels in maximizing diversity within recommendations and effectively reducing disinformation exposure with respect to the disinformation exposure levels of the original network. This dual capability is critical in the sense that, while other algorithms that promote diversity (such as a random recommendation algorithm or collaborative filtering approaches) do manage to broaden the range of topics covered, they often sacrifice precision. Such a trade-off results in lower user engagement, which is detrimental to effectively countering the formation of misinformation networks.

Unlike algorithms that primarily focus on precision, such as Word2Vec, which indeed scores high on accuracy but contributes to the formation of disinformation networks, our algorithm provides a robust solution by creating more diverse connections (**RG3.RQ3**). It fosters linkages between different textual communities, thereby enhancing the overall diversity of the social network without compromising the accuracy of recommendations (**RG3.RQ4**). Our results are backed by statistical testing, affirming the effectiveness of our algorithm in preventing the emergence and consolidation of misinformation networks. This validation underscores the algorithm's practical relevance and its potential impact on enhancing the integrity of social networks.

The unique aspect of our algorithm lies in its ability to maintain high precision while actively reducing disinformation exposure. This is achieved through strategic generation of weak ties across text communities, which naturally encourages a more diverse array of content within social networks. By doing so, our algorithm fills a previously unaddressed gap in recommendation systems, offering a novel solution that balances user engagement with informational integrity.

As social networks continue to grow and play a crucial role in information dissemination, the importance of developing robust recommender systems that can prevent disinformation while engaging users is paramount. Our algorithm stands as a pioneering approach in this regard, providing a blueprint for future research aimed at creating more resilient and inclusive digital social platforms.

Chapter 12

Algorithmic Approaches to Depolarize Social Networks

12.1 Introduction

As already argued throughout this thesis, the Internet, particularly social media, has become the new political agora since its widespread adoption worldwide, especially in the West, starting in the mid-1990s. From the earliest forums that appeared in 1996 to the first news portals in 1998, the Internet has evolved into a primary platform for increasingly broad segments of the citizenry to obtain current information. It has also become a space for generating conversations, forming friendships, and connecting around various projects. The Internet's most distinctive feature in that aspect is its capability to consolidate as an alternative source of information and interaction, distinct from traditional media—characterized by its marked narrative tendencies, editorial control, professionalism, deontology, and hierarchical, almost unidirectional relationship with the reader. This characteristic has enabled the emergence and enhancement of various social phenomena.

Thus, the main difference between media and communication forums on the Internet and traditional (analog, if you prefer) information spaces prior to their existence lies in the lack of editorial control and the capacity for self-communication or multi-directional communication among "reader" users. These users no longer only consume information from the media but can also produce it autonomously and react in real time to the rest of the system [Cas13]. This became especially relevant with the emergence of online social networks, particularly since 2006, with the appearance of micro-blogging networks such as X, which are most significant.

This phenomenon has had a massive impact on various forms of collective action, from business entrepreneurship to forming friend groups and political participation. These aspects are exciting as they have been enhanced by online communication. Among them, we find recruitment for protest participation, the articulation of ideologies or alternative political projects, signature or fundraising campaigns, and various forms of digital protest or pressure campaigns [Tuf17]. Similarly, political debate has progressively expanded onto the Internet. Social networks like X currently serve as the digital agora we mentioned earlier. Due to this capacity for self-communication, users on micro-blogging sites comment in real-time on political developments in their society, consuming content and generating debate [EVHH11].

Due to the speed at which content is produced and consumed, its format, and the algorithmic mediation of the platforms themselves, among other endogenous and exogenous factors to the network, users of these types of networks in political contexts have experienced an increasing polarization process in recent years [RARFN22]. They become entrenched in opposing political positions, generating hostility among themselves. Polarization, therefore, has extended within and beyond social networks, becoming a significant problem, especially in Western democracies, as it hinders decision-making and undermines social peace.

In this chapter, as an application of Chapter 8, where we evaluated the various approaches to measuring polarization on social networks, we now explore the possibility of mitigating this phenomenon through algorithmic intervention strategies based on content recommendation systems in micro-blogging networks. We employ an information theory-based approach to reduce the degree of network polarization while maintaining acceptable levels of accuracy. Finally, we compare our approach with the state-of-the-art, demonstrating that our method can achieve the best ratio between high accuracy and low polarization.

At the outset of this chapter, we explore the intricate relationship between online social networks, recommendation systems, and the phenomenon of political polarization. With the Internet evolving into a dominant medium for political discourse and collective action, the rise of social media has profoundly reshaped how individuals access information and engage with each other. This chapter examines the mechanisms behind these platforms, emphasizing how recommendation systems contribute to the formation of echo chambers and exacerbate political polarization, as a natural continuation of Chapter 10. By addressing these dynamics, this chapter directly supports the third research goal of the thesis: analyzing how recommendation systems influence polarization and proposing strategies to mitigate it. Additionally, the chapter aligns with one of its objetives by evaluating the trade-off between algorithmic interventions and recommendation accuracy, advancing the overall aim of understanding and countering polarization in online environments.

The chapter is organized as follows: Section 12.2 reviews polarization quantification methods, focusing on network topology and content-based approaches, while exploring the role of recommendation systems, detailing how they influence user engagement and contribute to polarization. Section 12.4 introduces a novel information-theory-based algorithm designed to counter polarization by fostering diversity while maintaining recommendation accuracy. Section 12.3 presents the methodology followed to obtain the results of benchmarking this algorithm against state-of-the-art methods, which are included in Section 12.5, using datasets from Spanish electoral processes. Finally, Section 12.6 discusses the implications of these findings in relation to the thesis's third research goal, while Section 12.7 concludes by highlighting the chapter's contributions to understanding and mitigating polarization in online environments.

Research questions

Our aims in this chapter are summarized in the following research questions included in the third research goal of the thesis:

- **RG3:** Analyze the role or contribution of recommendation systems in promoting these phenomena (disinformation diffusion and polarization). Propose mitigation strategies through these systems.
  - **RG3.RQ3:** How can recommendation systems help to mitigate these phenomena?
  - **RG3.RQ4:** How do the proposed mitigation strategies affect recommendation accuracy?

12.2 Background

12.2.1 Online Social Networks as political platforms

As previously described in Section 2.6, since their emergence and mass adoption by large groups of users around 2006, online social networks have become genuine platforms facilitating political participation [BD16]. Their ease of use, combined with the widespread adoption of smartphones capable of accessing these platforms and the format of short and rapid communication that any user can perform, quickly transferred conventional political participation dynamics to the digital space, leading to expanded forms of these dynamics and new political phenomena [RARFN22]. Online social networks have generated narratives capable of consolidating alternative powers or counter-powers by escaping the editorial control of media outlets that are sometimes (depending on the country and historical moment) linked to the traditional political establishment. In that aspect, X has consolidated as the main online social network when it comes to political discussion [BM13, BD16]. Similarly, they have amplified political participation, serving to launch digital activism campaigns such as positioning social demands or connecting remote activist communities to articulate large political projects like coordinating street protests [Tuf17, Sán15].

Regarding phenomena endogenous to social networks, we can highlight a series of occurrences directly linked to the transfer of political debate to online social networks from the perspective that users can both emit and consume information in real-time [Cas13] (see also Section 2.7 regarding digital activism). The primary documented phenomena in this regard include the emergence of fake news or disinformation accounts, cyberbullying or "trolling" campaigns [FS16], and the creation of communicative echo chambers composed of individuals repeatedly exposed to highly filtered content [CRA14]. These closed echo chambers significantly facilitate the development of polarization on online social networks.

Political polarization, as already discussed in previous chapters and, in particular, in Section 2.13, is a phenomenon that exists beyond online social networks and has been present since before the popularization of the Internet [CRF$^+$11]. However, it has been maximized following the emergence of social networks [CRF$^+$11]. In this sense, polarization is transferred to and generated within social networks, as in these online platforms the creation of tight and segregated user communities known as echo chambers emerges naturally [MBP18]. The natural human interest in aligning with groups or "communities" composed of similar individuals, as well as voluntarily accessing (selective exposure) content favorable to their tastes, naturally facilitates the emergence of these echo chambers and thus polarization in digital spaces, even without considering the effect of recommendation systems [HSS13].

12.2.2 Polarization

One of the main phenomena of recent interest and significant impact on online social networks is political polarization, structured around echo chambers within these networks. In the broad conventional political context, political polarization is understood as the process by which the ideological and programmatic distance between political parties and their electoral bases expands markedly, leading to irreconcilable differences that can be ideological, personalistic, or both [CBR21, Tal21]. This phenomenon is evident not only in the widening chasm between the political positions of parties but also in the intensified hostility and division among voters. The emergence of anti-establishment, populist, or extreme parties often exacerbates this polarization [CBR21, McC19].

When translated to online social networks, political polarization can be defined as the process by which the ideological and programmatic distance between individuals and groups significantly increases due to interactions and content consumption on digital platforms [McC19]. This phenomenon is manifested in the expanding gap between political views, reinforced by algorithmically driven echo chambers that expose users to homogeneous and often extreme perspectives [LRT22, HT12]. The escalation of hostility and division among users is exacerbated by the proliferation of populist and extreme content, leading to increased engagement and deepening societal fractures. Online social networks amplify these effects by promoting sensationalist and polarizing content, thus accelerating the process of political polarization [XWQ+22, LRT22].

Polarization has notably increased worldwide, especially in the West, generating significant concern across various social sectors, including academia, media, and government. Alongside the rise of disinformation, numerous national and international institutions in the West have recognized polarization as one of the primary challenges currently associated with online social networks. This worldwide increase of polarization has already been subject of study in existing research [GW17, GLM20, CV18], as it raises concern in western societies and unfolds the need of creating and developing strategies that can help reduce polarization in the Online Social Networks domain, informing and enhancing existing regulations for such purpose, as it is the case with the European Digital Services Act (DSA) [C+24]. To carry out such reduction, it is first required to quantify network polarization. Existing research details different approaches to measure polarization in Online Social Networks, grouping the existing methods into three different categories: topology-based, content-based, and hybrid algorithms to quantify polarization [YWLD17, Gar18]. For an in-depth review of polarization approaches and, in particular, for techniques categorized among these groups, we refer the reader to Chapter 8.

12.2.3 The role of recommendation systems

The vast majority of content we consume on current social networks, including micro-blogging networks, is mediated by recommendation systems. These systems are responsible for suggesting both contacts (users) and content (posts). Recommender systems significantly impact polarization in social networks by amplifying existing biases and creating feedback loops reinforcing extreme views, as already analyzed in Chapter 10. Studies by [MAP+20] and [CMMB22] have demonstrated that recommendation algorithms like User-based Collaborative Filtering, Item-based Collaborative Filtering, and Singular Value Decomposition amplify popularity bias, leading to decreased aggregate diversity and the homogenization of user preferences. This process reduces the diversity of recommendations, skewing user profiles towards more homogeneous, polarized groups, mainly affecting minority perspectives and increasing societal biases. Additionally, empirical research has shown that platforms prioritizing engagement, such as X and Facebook, exacerbate polarization by promoting negative and divisive content due to the negativity bias inherent in human behavior, leading to more confrontational exchanges and misinformation.

Further studies highlight how recommender systems' design choices directly influence the polarization of online environments. [C+24] provide evidence that self-learning recommenders aiming to maximize user engagement through popularity and sentiment analysis foster algorithmic negativity bias and ideological fragmentation. These systems lead to a concentration of social power among the most toxic users, increasing their influence on public discourse and further polarizing the opinion landscape.

This phenomenon was observed in empirical calibrations on Twitter, revealing that even neutral recommendation modes inadvertently amplify harmful content due to user tendencies to engage more with such content. [KBKW21] discusses how preference amplification dynamics in matrix factorization-based recommender systems can lead to echo chambers and reduced content diversity, further contributing to polarization. [FENKW22] highlight that specific recommendation algorithms can disproportionately disadvantage minority groups by reinforcing majority preferences, reducing visibility and engagement for minority content and perspectives. [MTF20, WLRV21] analyzes the role of recommender systems in amplifying extremist content on platforms like YouTube, demonstrating the complexity of addressing algorithmic amplification of extremism and underscoring the need for nuanced regulatory approaches.

12.2.4 Algorithmic intervention

Addressing the issue of polarization on social networks requires interventions both in platform design and in the algorithms responsible for content dissemination. While platform-level changes, such as limiting post immediacy and volume, could theoretically mitigate polarization, these adjustments often come with significant challenges, including altering the user experience and risking a decline in platform popularity [LVhLH20]. Consequently, the focus has shifted towards algorithmic strategies, particularly in recommendation systems, to counter polarization effectively, as previously discussed in Section 2.18.

One approach is the Pre-Recommendation Counter-Polarization (*PrCP*) strategy, which modifies user-item ratings through stochastic mapping prior to generating recommendations. This method has demonstrated a notable reduction in polarization compared to traditional techniques like Non-Negative Matrix Factorization [BNS18]. Another prominent strategy is the Maximal Marginal Relevance (*MMR*) algorithm, which balances relevance and diversity in recommendations. By controlling the trade-off through a parameter λ, the algorithm mitigates echo chambers while maintaining user satisfaction [GSM$^+$22].

Content-based moderation strategies, such as filtering out hostile content exceeding a predefined threshold, have also been proposed. While these strategies can depolarize networks, they raise concerns about limiting user freedoms and content accessibility [Str22, LVhLH20]. Optimization approaches, like greedy algorithms and integer programming, rerank existing recommendations to introduce diversity, reducing polarization while preserving recommendation accuracy [AK12, AK14]. These techniques, though classical, remain relevant in contemporary research due to their effectiveness in mitigating network polarization [YB22].

12.3 Specific Methodology

12.3.1 Data set

We used two primary datasets for this analysis, as defined in the methodology chapter of the thesis (Chapter 2), for tables "General Processes" (see Table 4.2) and "Local Processes" (Table 4.3). Note these datasets were used in Chapter 6 to provide an analysis of the election processes and derive a narrative flow, in Chapter 8 to evaluate our proposed polarization metric, and in Chapter 10 to understand the behavior of recommendation networks in this context.

We decided to employ this dataset because existing research, as the one discussed earlier in the thesis, already devised an existing polarization trend in the social networks generated during each Spanish electoral process from 2011 to 2019. Indeed, any recommendation algorithm proposal aimed at reducing

polarization should present a decrease of network polarization in these networks, which are already known to be polarized.

12.3.2 Experiment description

To benchmark the recommendation algorithm proposal and its effectiveness and utility for countering polarization, we designed an experiment based on executing each of the implemented recommendation algorithms, including the proposal, with real data from the described dataset. More specifically, each electoral process dataset is split daily. For each day, recommendation algorithms are executed on a training partition of the dataset and evaluated on an evaluation partition.

By feeding each recommendation algorithm with data daily, each recommendation algorithm generates a ranked list of tweets that conform to the recommendations of each user per day. This allows us to generate a *recommendation network*, which we define as a directed and weighted graph whose nodes are the users of the social network and whose edges represent recommendation relationships between users. Precisely, user A is connected to user B (directional relationship) if and only if user A receives a post recommendation authored by user B. The weight of the relationship represents the number of recommendations that user A received from user B, ranging from 1 to N, with N the maximum length of the ranked tweet list each user receives daily.

To benchmark our recommendation algorithm, we calculate the network polarization by applying the SPIN metric on the recommendation network generated by each recommendation algorithm on each day of each electoral process. The only exception is the first day of each electoral process, employed as the cold start for the recommendation algorithms [SM21].

Considering the described experiment, we leveraged statistical techniques to further understand the utility of our algorithm, including an analysis of the calculated network polarization and recommendation accuracy temporal series per electoral process, as well as a ratio measurement that we computed to analyze the trade-off between accuracy and polarization performed by each implemented recommendation algorithm. By carrying out this experiment, we expect to shed light on whether the proposed algorithm can produce minimal polarization values in the different electoral processes while keeping a relevant recommendation accuracy, effectively countering the polarization phenomenon.

To assure statistical significance in our results, we applied one-sided significance tests to check if the network polarization achieved in the real data, i.e., the X algorithm, is significantly greater than the polarization achieved by our proposed algorithm. More specifically, we first applied a normality test to determine if the polarization results per election followed a normal distribution. This normality test is particularly relevant for the selection of the statistical test with which to check if X algorithm's network polarization is significantly greater than our proposed algorithm's network polarization [RGK12]. Under the assumption of normal data, the t-test can be used to obtain statistical significance to prove that our algorithm effectively lowers network polarization [Bio08]. Student's t-test can be applied to determine if the mean of a given distribution is significantly greater than that of another different distribution, as has already been applied in related literature [MBBRD24]. Conversely, under the assumption of non-normally distributed data, the non-parametric Mann-Whitney U test, also known as the Wilcoxon rank sum test, can be used to determine whether a distribution is significantly greater than another distribution [MW47]. However, since this test is an ordinal test, it is recommended to compare the medians of both data distributions instead of the means, as opposed to the t-test statistic described previously [RRMPT13].

Consequently, we first employed the Shapiro-Wilk test to determine if the two network polarization distributions follow a normal distribution in each electoral process. We decided to use the Shapiro-Wilk test due to the size of the available data per electoral process. As each electoral process provides 30 different measurements of network polarization, i.e., one network polarization value per day, the Shapiro-Wilk test is more suitable than other similar statistical tests, as it is the case with the Kolmogorov-Smirnov test, which requires $n \geq 50$ [MPS+19].

For all the performed statistical tests, we employed a significance level $\alpha = 0.05$, to obtain statistical significance in the presented conclusions. Indeed, through the application of these statistical tests we aim to prove that our algorithm can significantly lower network polarization, while keeping a relevant recommendation accuracy, outperforming current state-of-the-art approaches in this trade-off between network polarization and recommendation accuracy.

12.4 Proposal to depolarize social networks

In this section, we present our proposal to depolarize online political social networks. We first need to select a measurement to capture polarization, this is described in Section 12.4.1. Then, Section 12.4.2 lists the recommendation algorithms that will be considered by our approach. Finally, in Section 12.4.3, we introduce our proposal to counter polarization based on a greedy optimization approach.

12.4.1 Polarization index selection

In order to select a polarization index, we researched the existing literature on polarization quantification and compared the existing algorithms with respect to different criteria, as it can be seen in Table 12.1. More specifically, we compared the algorithms in terms of whether they employ information related to the network topology and whether they utilize the disseminated content to quantify polarization across the network. Furthermore, we added a specific criterion related to the dataset employed for this research, based on Spanish political electoral processes between 2011 and 2019. Indeed, the last criterion refers to the adaptation of the polarization index to the political context.

We leverage the previous criteria to select the most adequate and reliable polarization metric with which to carry out our research. As it can be seen from the table, only the novel SPIN polarization index matches all given criteria. This algorithm utilizes Information Theory basics and both network topology and content to quantify network polarization, through an index that is specifically tailored to electoral processes, as it is the case with the datasets that we employ for this research. Thus, we opted to use SPIN to benchmark the proposed recommendation algorithm with respect to the state-of-the-art counter polarization recommendation systems, as it seemed to be the most adequate polarization metric.

12.4.2 Polarization countering recommendation algorithm selection

Following the selection of the polarization index, we implemented the state-of-the-art recommendation algorithms with which to compare our proposed algorithm. Such implementation included the approaches described in Section 12.2.4, which were used to benchmark our proposal of counter-polarization recommendation algorithm.

Table 12.1: Comparison of state-of-the-art polarization quantification indexes.

Polarization metric	Topology	Content	Political polarization-specific
RWC [GW17]	✓	✗	✗
EC [GW17]	✓	✗	✗
GMCK [GJCK13, GW17]	✓	✗	✗
MBLB [GW17, MBLB15]	✓	✗	✗
BCC [GW17]	✓	✗	✗
P-score [GJCK13]	✓	✗	✗
ARWC [VPV21]	✓	✗	✗
DRWC [VPV21]	✓	✗	✗
ERIS [GGLC22]	✓	✗	✗
Multi-Opinion score [SIK22, MBBRD24]	✓	✓	✗
GE [HDC23]	✓	✓	✗
Normalized Cut Score [YWLD17]	✗	✓	✗
NLP-based [DEB96, CLDB18, Mat17, WMZK18]	✗	✓	✗
DL sentiment analysis [KM22]	✗	✓	✗
BRW [ENT$^+$20]	✓	✓	✗
DiffPool [BAB$^+$21]	✓	✓	✗
SPIN [MBBRD24]	✓	✓	✓

Considering the implementation of the optimization approaches described by the authors in [AK12], we opted to implement the Greedy Optimization strategy, as it had similar foundations to already selected algorithms (e.g., MMR), while effectively reducing network polarization through a reranking approach.

This last algorithm, Greedy Optimization, is solely based on a reranking of the original ranking created by an already existing recommendation algorithm. For this reason, we implemented four variants of the algorithm, one per main family of recommendation algorithms: collaborative filtering, content, Deep Learning, and Reinforcement Learning-based [ASIM18]. Regarding the collaborative filtering variant, we implemented a simple kNN-based collaborative filtering recommendation algorithm [Pet09] using cosine distance as the distance function [DDDD09], reranked according to the greedy optimization strategy. Concerning the content-based strategy, we implemented an approach where user preferences are encoded using the average Doc2Vec representation of all the tweets posted by each user [Chu17], employing the averaged Word2Vec representation of each word in each post as the Doc2Vec representation of each user publication. Similarly, the cosine similarity distance was used to recommend the publications that are closest to the user preferences for each user. Concerning the Deep Learning-based recommendation algorithm, we implemented the novel DeepRank algorithm, a hybrid Deep Learning-based approach to recommend content based on both network topology and content [CZ20]. More specifically, the algorithm uses a sparse representation of both users and tweets as an input. Such an input is then processed through an embedding layer that condenses the information into dense vectors, which are then used to carry out a prediction after the action of a series of hidden layers that result in a prediction of the ranking of the item, leveraging the use of both the content-based component of the tweet and the social-based component of the network to solve the ranking tasks. Last, we implemented a Thompson Sampling-based Multi-Armed Bandit as the reinforcement learning-based recommendation algorithm. In this approach, a Multi-Armed bandit recommender is created per user, with as many arms as the social network has users. The bandit plays the arms based on the real events, i.e., the real tweets, using their ranking (retweet) as the reinforcement with which the algorithm learns which tweets are most relevant for each user.

As it can be seen, our selection of recommendation algorithms includes at least one recommendation algorithm per main family of recommendation algorithms, covering many different and recent strategies to counter network polarization to provide a reliable benchmark for the proposed recommendation algorithm.

12.4.3 Depolarizing social networks

In this research, we propose a recommendation algorithm that effectively counter polarization. This recommendation algorithm is based on a hypothesis already devised in the existing literature: reducing the number of links that can introduce opposing viewpoints in a community leads to the creation of echo chambers, where each community further reinforces its own opinion, potentially causing higher network polarization [SLL21, Tin17, SSJ+22]. Most existing polarization-counter recommendation algorithms follow this same approach of diversifying viewpoints and creating healthier social networks by reinforcing diverse user interactions [ZBZ21, CMMB22, Ser24].

We consider our proposal a compelling novel approach to counter polarization with two main contributions. First, to our knowledge, it is the only information theory-based recommendation algorithm to counter polarization. While Information Theory concepts have already been the subject of research on the topic of polarization [MBBRD24], it still has not been applied to creating a recommendation algorithm aimed at minimizing network polarization based on the information that flows through the social network. Secondly, this algorithm considers the existence of echo chambers in Online Social Networks, following the hypothesis above of current research that they could facilitate the creation of more polarized network dynamics due to the creation of recommendations that further reinforce the existing echo-chamber-fragmented network dynamics [DSWS24, VGPG22, NLJS20].

Our proposed recommendation algorithm is defined through four main steps that are applied consecutively to obtain the final user-item rankings with which to create top-N recommendation lists for each user. The first step of our algorithm is the network content preprocessing. More specifically, the content published by each user on the social network must be preprocessed to reduce noise, standardize word representations and improving the performance of the algorithm, as already devised in existing Natural Language Processing research [NKLL22, CRC19b].

Secondly, user preferences are computed. User preferences are key to existing recommendation algorithms [RAC+02, JWK14, RKR08]. In our proposal, we calculate the preferences of a user based on the word probability distribution of that user, based on its preprocessed publications. Indeed, user preferences are represented as a sparse vector with as many dimensions as words are considered in the training corpora, where each dimension's value represents the probability that the user employs that word in a new post.

The third step of our algorithm is to carry out community detection, so that the recommendation algorithm can provide more diverse recommendations by creating bridges between different user communities. To carry out community detection, we employ Hierarchical and Agglomerative Clustering (*HAC*) using the *complete* linkage criterion, which computes the distance between the clusters as the maximum distance between any two points of the given clusters [DMB05]. We employed this approach because it is one of the most explainable clustering algorithms in the literature, and it has already been used for community detection for similar reasons in the literature [DPPS24, XNT14].

The intuition behind the introduction of a community detection algorithm is that users of the same community will have similar viewpoints due to the existence of echo chambers that further reinforce their opinion, as devised in existing literature [RARFN22]. Consequently, we introduced a method that can carry out community detection in a very explainable manner, as it is the case with HAC. While this approach is fully explainable [BK05], it is also suitable for the application of Information Theory-based metrics, because users are clustered based on their preferences, which are encoded as word probability distributions represented as N-dimensional vectors. These N-dimensional vectors extracted from text are a suitable application for Information Theory, as it can be used to effectively capture the underlying patterns and distributions of user preferences that change dynamically as the users operate on the network.

More specifically, we employ the Jensen-Shannon distance as an Information Theory-based metric to measure distance between user preferences, so as to carry out community detection in an accurate and explainable manner, as it has proven to be effective in several applications in broad social sciences [DHKH13, BB21], although none of these applications belong to the recommendation algorithms field, which is why we consider our approach to be novel. This metric was originally derived from another Information Theory-based concept, the Kullback-Leibler divergence, which can be employed to calculate how different a given probability distribution is with respect to another probability distribution [Kul51]. However, Kullback-Leibler divergence is a non-symmetric measurement, which makes it less adequate to its application in a problem focused on distance computation. In our case, probability distributions model the word probability distributions of the users, i.e., the probability that each user employs each word of the corpora in a new publication.

Indeed, the intuition behind our algorithm is to detect groups of users that employ the same speech and treat the same topics through the application of clustering techniques and Information Theory-based metrics, so that the user-ranking of the items can diversify recommendations interconnecting different clusters and creating healthier social networks. The equation of the Jensen-Shannon distance is shown in Eq. 12.1, where P and Q represent the probability distributions whose distance is being computed, M represents the point-wise mean of both distributions and D_{KL} represents the Kullback-Leibler divergence (see Eq. 12.2).

$$\mathrm{JS}(P \parallel Q) = \sqrt{\frac{1}{2}\left(D_{\mathrm{KL}}\left(P \parallel M\right) + D_{\mathrm{KL}}\left(Q \parallel M\right)\right)} \tag{12.1}$$

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} \tag{12.2}$$

Last, the final step of our algorithm is the calculation of user-item rankings. To carry out this step, we introduce a mathematical function that facilitates the creation of user-to-user relationships between users of different communities, establishing balance between recommendation accuracy and content diversity, which can be adjusted through the diversification parameter $\lambda_{diversity}$. This user-item ranking function is described in Eq. 12.3, where *R(u, i)* represents the ranking of item *i* by user *u*, *JS* represents the application of the Jensen-Shannon distance between the word probability distributions of *u*'s user preferences and *i*'s content, and $\lambda_{diversity}$ is the diversity factor that increases the rank of tweets whose author belongs to a cluster different to the cluster of the user *u*, which can be used to foster the creation of user-to-user relationships between users of different clusters, increasing recommendation diversity and

helping to create healthier networks. The diversification parameter of our algorithm is normalized, thus it takes values in the interval $[0, 1]$, where value 0 leads to the creation of recommendations that focus on achieving maximum accuracy with minimum diversity, whereas value 1 maximizes diversity at the cost of recommendation accuracy. Indeed, $\lambda_{diversity}$ establishes the balance between recommendation accuracy and diversity, which is used as the cornerstone upon which our algorithm minimizes network polarization.

$$R(u,i) = \begin{cases} JS(u,i) & \text{if i author's cluster equals u's cluster} \\ JS(u,i) \cdot (1 + \lambda_{diversity}) & \text{otherwise} \end{cases} \tag{12.3}$$

In order to achieve the best trade-off between recommendation accuracy and polarization reduction, we compared the results of our recommendation algorithm under different values of $\lambda_{diversity}$. More specifically, we employed values in the entirety of its range $[0, 1]$, in steps of 0.05 to identify its best possible value, at least for the employed dataset, through a computationally feasible method. We present the obtained results in Table 12.2, where we show the average daily accuracy-polarization trade-off value for each electoral process when executing our proposed algorithm with different values of $\lambda_{diversity}$. To present our results in a concise manner, we represent the results in steps of 0.2, providing 0.05 and 0.95 as the limit values of diversification parameter.

Table 12.2: Selection of $\lambda_{diversity}$ using the Accuracy-SPIN ratio as the decision metric.

Electoral process	$\lambda_{diversity} = 0.05$	$\lambda_{diversity} = 0.20$	$\lambda_{diversity} = 0.40$	$\lambda_{diversity} = 0.60$	$\lambda_{diversity} = 0.80$	$\lambda_{diversity} = 0.95$
Nov. 2019 (G)	0.5121	**0.5309**	0.4863	0.4727	0.4405	0.3972
Apr. 2019 (G)	0.5048	**0.5220**	0.4658	0.4545	0.4239	0.3833
2016 (G)	0.3349	**0.3482**	0.3343	0.3336	0.3133	0.2819
2011 (G)	0.3049	**0.3387**	0.3020	0.2728	0.2455	0.2047
2019 (L)	0.5935	**0.6255**	0.5626	0.5239	0.4867	0.4373
2015 (L)	0.3368	**0.3456**	0.3394	0.3151	0.2932	0.2630

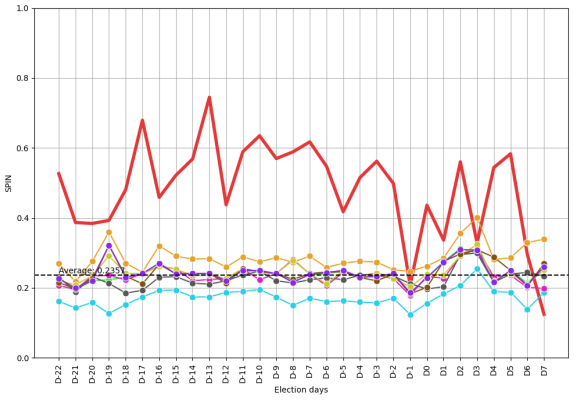
As it can be seen from the results, $\lambda_{diversity} = 0.20$ achieves the best possible trade-off between lowering network polarization and achieving a relevant recommendation accuracy. Consequently, we employ $\lambda_{diversity} = 0.20$ for the rest of our analysis, including the benchmark performed against the rest of state-of-the-art counter polarization approaches.

12.5 Results

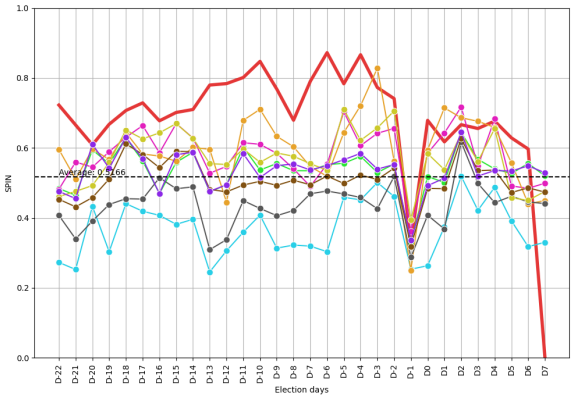
Considering the implemented state-of-the-art recommendation algorithms and our proposal of an Information Theory-based recommender to counter polarization, we carried out the experiment described in Section 12.3. Regarding the recommendation networks generated by each recommendation algorithm, we first calculated the SPIN polarization metric on each of them, obtaining the temporal series shown in Fig. 12.1.

The results allow us to derive multiple conclusions. First, we observe how the strategy that is purely based on changing the original ratings, Pre-Recommendation Counter Polarization (PrCP), clearly achieves average polarization results, showing that only focusing on the original ratings is insufficient to effectively counter network polarization.

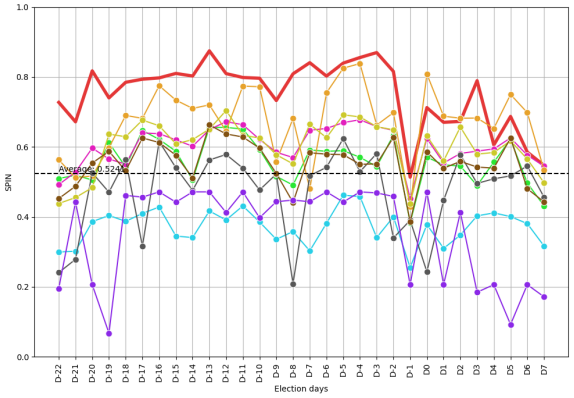
Similarly, we observe how the reranking-based strategies implemented through Greedy Optimization tend to fail for most recommendation algorithm families: content-based, Deep Learning and collaborative filtering. In all three cases, despite employing a reranking strategy to introduce completely new



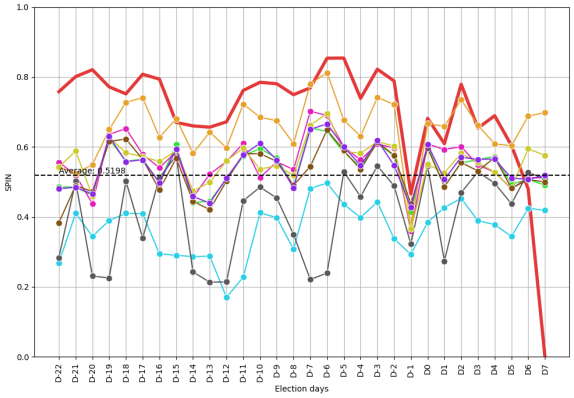
(a) Spanish elections of 2011 (general).



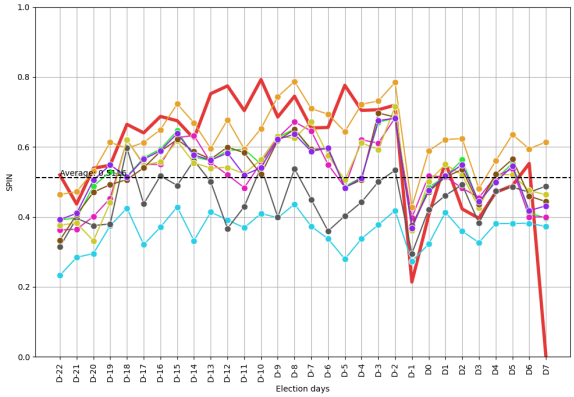
(b) Spanish elections of 2016 (general).



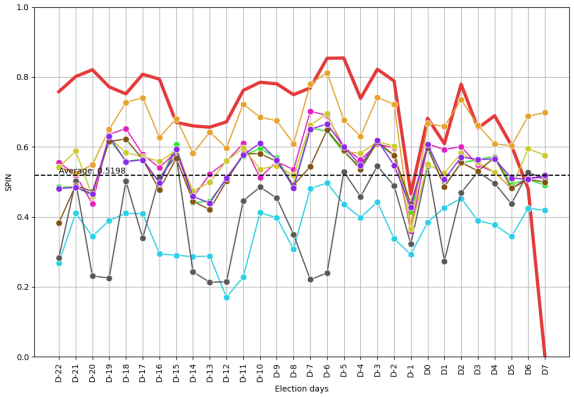
(c) Spanish elections of April 2019 (general).



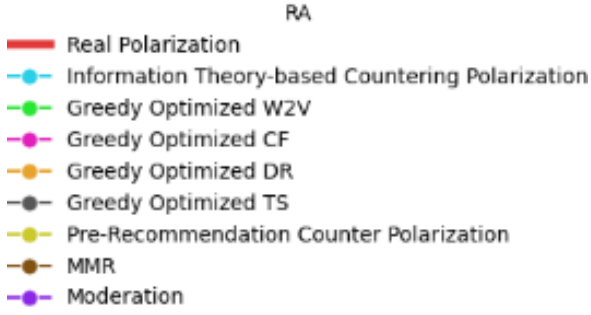
(d) Spanish elections of November 2019 (general).



(e) Spanish elections of 2015 (local).



(f) Spanish elections of 2019 (local).



(g) Legend.

Figure 12.1: Polarization evolution across each electoral process per recommendation algorithm.

and unpopular publications, polarization is still maximized. The explanation to this phenomenon is the inherent nature of all the three algorithms, as their ranking strategies are purely based on recommendation accuracy without considering their effect on network polarization. In the case of the Greedy Optimized Collaborative filtering, this behavior is clear, as each user will receive recommendations from users with similar tastes without restrictions. Similarly, the Greedy Optimized Word2Vec-based recommender promotes publications that are as much similar as possible to the user preferences. Last, the Greedy Optimized Deep Learning strategy utilizes gradient descent to optimize both network structure and publication content to promote content as much close as possible to user preferences.

Conversely, the Greedy Optimized Reinforcement Learning-based approach (Greedy Optimized TS) succeeds in the reduction of the SPIN network polarization metric. As opposed to the previous approaches, the Multi-Armed Bandit strategy follows a stochastic strategy, that favors the arm, that is, the user that is most likely to receive a retweet by a given user u, however, the strategy is based on an exploration-exploitation dilemma that gives the bandit the chance to choose what could seem like a sub-optimal arm (user). Such exploration-exploitation balancing introduces diversity in the top-N recommendations of each user, which is translated into a clear reduction of the SPIN polarization metric.

Regarding the MMR recommendation algorithm, it provides average-to-high polarization values across all electoral processes, failing to effectively reduce the SPIN polarization metric consistently. This observation matches the results obtained by other reranking-based recommendation algorithms, such as the Greedy Optimized DR, CF and W2V algorithms, deriving the idea that the original ranking puts too much focus on recommendation accuracy, which impacts the recommendation diversification performed during the reranking stage. Indeed, both optimal polarization reduction algorithms regarding the results obtained in this dataset (Greedy Optimized TS and our proposal) promote recommendation diversification besides reranking.

Concerning the content-moderation strategy, we observe that its effectiveness heavily depends on the electoral process studied. While it clearly succeeds in the reduction of network polarization during certain electoral processes, as it is the case with the Spanish general electoral process of April 2019, it clearly fails achieving reduced network polarization results in other processes, as it is the case with the Spanish general electoral process of 2016. Besides its limited capability to reduce network polarization, it directly affects user freedom, as certain publications are not available to the users, which prevents this strategy from being optimal.

Last, as it can be seen from the results, our recommendation algorithm proposal achieves minimum polarization results for each Spanish electoral process studied. The results shed light on the success of our Information Theory-based strategy to reduce polarization, built upon the idea that the introduction of community-to-community bridges helps to effectively reduce network polarization, preventing the creation of echo chambers and the further reinforcement of radical viewpoints. These results further remark that Information Theory can effectively be applied to create polarization-aware systems that significantly reduce network polarization in a variety of manners, ranging from the analysis of existing information flows in the network to the quantitative analysis of text distributions related to each user and the further analyzes of their preferences.

In this case, $\lambda_{diversity} = 0.2$ was used as it proved to be optimal in the discussion above, minimizing network polarization while achieving relevant accuracy results across all electoral processes, as it can be seen in Fig. 12.2.

A summarized version of the average network polarization values that each recommendation algorithm generates is shown in Table 12.3, which contains the percentage of SPIN polarization reduction with respect to the SPIN polarization value obtained from the application of the metric on the real data for each studied Spanish electoral process. As it can be seen from the table, our proposed recommendation algorithm surpasses existing counter-polarization recommendation algorithms, reducing the SPIN polarization metric by approximately 45-50% depending on the electoral process.

Table 12.3: Polarization reduction percentage per recommendation algorithm.

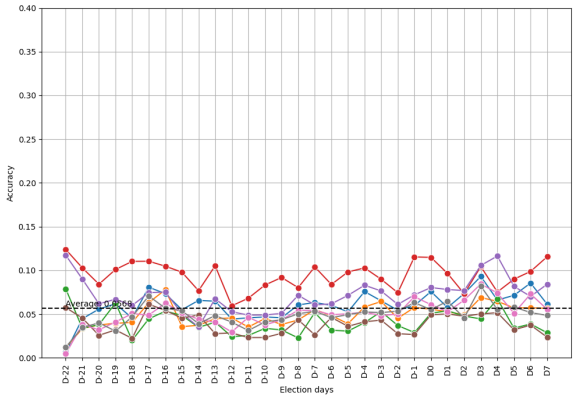
Algorithm	**2011 (G)**	**2016 (G)**	**2019-Apr (G)**	**2019-Nov (G)**	**2015 (L)**	**2019 (L)**
Information Theory-based	**45.83%**	**50.50%**	**47.63%**	**38.22%**	**48.18%**	**64.55%**
Greedy Optimized CF	15.01%	19.87%	20.30%	9.83%	18.95%	51.74%
Greedy Optimized DR	13.49%	10.90%	6.75%	-7.44%	7.72%	40.81%
Greedy Optimized TS	36.11%	36.95%	42.14%	22.88%	31.23%	53.26%
Greedy Optimized W2V	21.49%	25.70%	22.95%	8.31%	29.06%	50.65%
MMR	25.79%	26.46%	24.17%	8.79%	29.65%	49.95%
Moderation	21.66%	52.27%	22.36%	8.68%	28.59%	49.89%
PrCP	16.31%	20.00%	20.32%	9.39%	22.50%	49.89%

Considering the accuracy results (see Fig. 12.2), we can observe, again, different trends between the implemented recommendation algorithms. On the one hand, the Greedy Optimized Deep Learning-based recommendation algorithm, DeepRank, maximizes the accuracy across all the electoral processes studied, followed closely by our recommendation algorithm proposal. This clearly derives that both strategies find success in their recommendations and contribute to the engagement of the users with the social network [PLZ24]. In the case of Greedy Optimized DR, this success comes from the optimization of the underlying mathematical function that decides the ranking that user u will give to item i (i.e., retweet probability), although it comes at the cost of a very high network polarization. Conversely, our proposal of counter-polarization recommendation algorithm achieves minimal network polarization while maintaining a relevant recommendation accuracy, thus achieving a better balance between polarization and accuracy (see Table 12.4).

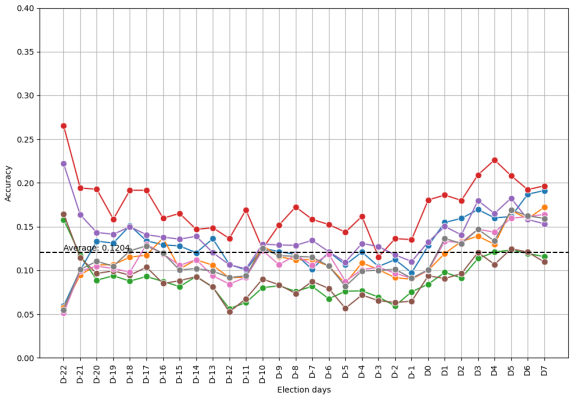
Regarding the rest of greedy optimized approaches, we observe that the Greedy Optimized Multi-Armed Bandit approach not only succeeds to minimize polarization, but it also achieves a high recommendation accuracy. Its success follows the intuition described before: each user is, *typically*, recommended the user whose content will most likely follow the preferences of the user based on the historical interactions between users. It must be noted that we say *typically*, because the exploration-exploitation dilemma, together with the stochastic nature of the multi-armed bandits, can promote content diversity, achieving both relevant recommendation accuracy and low network polarization. Conversely, Greedy Optimized W2V and CF fail to learn user preferences and yield both high network polarization and low recommendation accuracy. Thus, both approaches seem non-effective to counter polarization.

Concerning MMR, we observe this recommendation algorithm achieves high recommendation accuracy. Intuitively, these results are identical to the Greedy Optimized DR, another counter-polarization algorithm that is solely based on reranking to introduce diversity and reduce polarization. Both algorithms achieve a high recommendation accuracy, however, both algorithms fail to reduce network polarization as the reranking strategy seems insufficient to reduce network polarization significantly.

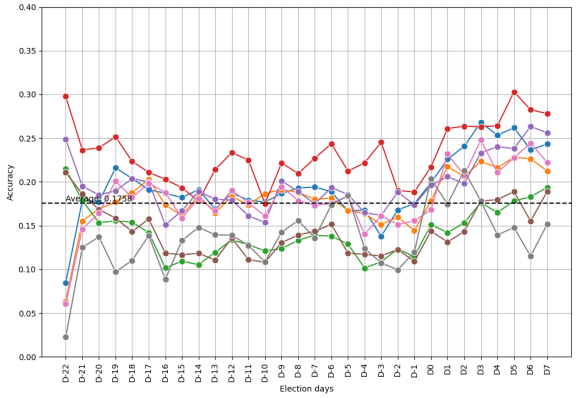
Last, regarding the content-moderation recommendation algorithm, we observe it achieves average accuracy values, while also providing average network polarization values. Hence, it balances network



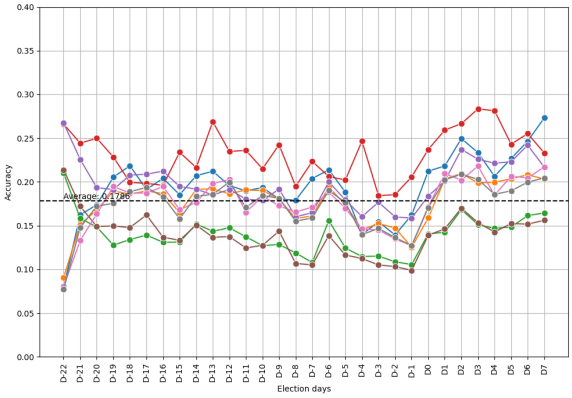
(a) Spanish elections of 2011 (general).



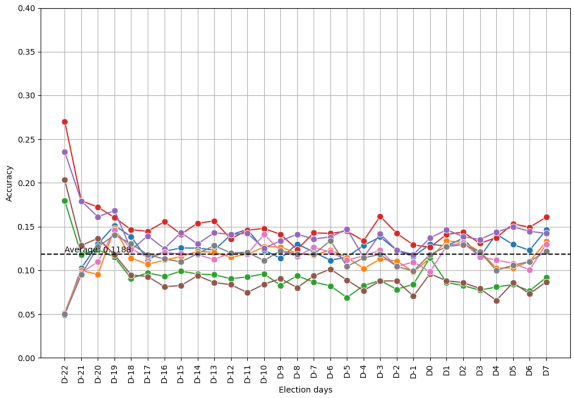
(b) Spanish elections of 2016 (general).



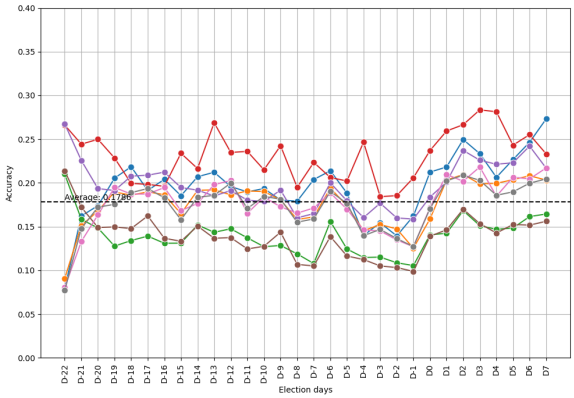
(c) Spanish elections of April 2019 (general).



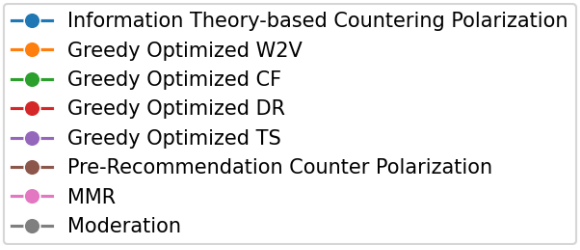
(d) Spanish elections of November 2019 (general).



(e) Spanish elections of 2015 (local).



(f) Spanish elections of 2019 (local).



(g) Legend.

Figure 12.2: Accuracy evolution across each electoral process per recommendation algorithm.

polarization and recommendation accuracy, however, it does so at the cost of user freedom, as certain publications are not recommended to users following a shadow-banning strategy [Sav22].

Table 12.4: Accuracy/Polarization ratio per recommendation algorithm and election.

Algorithm	2011 (G)	2016 (G)	2019-Apr (G)	2019-Nov (G)	2015 (L)	2019 (L)
Information Theory-based	**0.3571**	**0.3595**	**0.5294**	**0.5536**	**0.3478**	**0.6477**
Greedy Optimized CF	0.1772	0.1600	0.2404	0.2547	0.1887	0.3059
Greedy Optimized DR	0.3348	0.3019	0.3573	0.3622	0.2482	0.3987
Greedy Optimized TS	0.3193	0.3269	0.4538	0.5441	0.3304	0.5109
Greedy Optimized W2V	0.2044	0.2133	0.3277	0.3309	0.2170	0.4193
MMR	0.2103	0.2241	0.3322	0.3337	0.2203	0.4252
Moderation	0.2041	0.2144	0.4850	0.3214	0.2187	0.4149
PrCP	0.1600	0.1655	0.2456	0.2529	0.1875	0.3172

We provide a summarized version of both polarization and accuracy results in Table 12.4, which shows the average accuracy over polarization (SPIN) ratio across all days of each of the studied electoral processes. In such table, the general electoral processes are denoted by (G), whereas the local electoral processes are denoted by (L). The bold-weighted cells represent the algorithm that best performs a trade-off between recommendation accuracy and network polarization for each electoral process.

As it can be seen from the table, the ratio between recommendation accuracy and the SPIN network polarization metric shows that our proposal of counter-polarization recommendation algorithm surpasses the existing approaches to counter polarization. While it is close to existing state-of-the art counter-polarization recommendation algorithms, such as Greedy Optimized TS/DR, our algorithm successfully applies a trade-off between relevant recommendation accuracy and recommendation diversity as a strategy to counter echo chambers and the reinforcement of radical viewpoints, as opposed to other state-of-the-art recommendation algorithms (**RG3.RQ3** and **RG3.RQ4**), like MMR and Greedy Optimized DR, which achieve high recommendation accuracy but fail to effectively reduce network polarization consistently across the different electoral processes. In the light of these results, both Greedy Optimized TS and our proposal of counter-polarization recommendation algorithm perform the best, regarding the trade-off between recommendation accuracy and network polarization, for the Spanish electoral processes dataset.

Last, we applied the aforementioned statistical tests to demonstrate that our algorithm significantly reduces network polarization with respect to the real data obtained from the Twitter/X algorithm. In Table 12.5, we observe that the only case where it is not possible to reject the null hypothesis of the Shapiro-Wilk normality test, i.e., both distributions are assumed to be normal, is the Spanish general electoral process of 2011, as the P-Value obtained from the Shapiro-Wilk test applied on Twitter's network polarization results P-Value$_{Twitter} = 0.3705$ is greater than the selected confidence level $\alpha = 0.05$ and the proposed algorithm's network polarization results P-Value$_{IT-based} = 0.0993$, which is also higher than the aforementioned confidence level. Indeed, for most electoral processes we cannot assume that network polarization follows a normal distribution and, as a result, we decided to apply the non-parametric Mann-Whitney U test as described in Section 12.3.

The results of the Mann-Whitney U test are described in Table 12.6, where we observe how the null hypothesis is rejected for the selected confidence level across all the electoral processes. Consequently, we observe how there is statistical significance that our proposed recommendation algorithm effectively reduces network polarization with respect to the network polarization generated by Twitter's (**RG3.RQ3**). Such reduction is also represented in Fig. 12.1, where we observe how both distributions are most separated from each other as possible when compared to the rest of recommendation algorithms, regardless of the electoral process being studied.

Table 12.5: Shapiro-Wilk normality test results per electoral process between our approach and the original data from Twitter/X.

	Twitter		Information Theory-based	
Electoral process	**W**	**P-Value**	**W**	**P-Value**
November 2019 (G)	0.6973	0.0000	0.9513	0.1836
April 2019 (G)	0.8885	0.0045	0.9704	0.5495
2016 (G)	0.7112	0.0000	0.9577	0.2698
2011 (G)	0.9631	0.3705	0.9414	0.0993
2019 (L)	0.9245	0.0352	0.9853	0.9426
2015 (L)	0.8725	0.0019	0.9484	0.1535

Table 12.6: Mann-Whitney U test results per electoral process.

Electoral process	**U**	**P-Value**
November 2019 (G)	866.0	4.0507e-10
April 2019 (G)	900.0	1.5099e-11
2016 (G)	852.0	1.4608e-09
2011 (G)	870.0	2.7848e-10
2019 (L)	895.0	2.4876e-11
2015 (L)	822.0	1.9824e-08

12.6 Discussion

In this chapter, we propose a novel algorithm designed to effectively mitigate network polarization while keeping users engaged in the platform. The uniqueness of our approach lies in the introduction of Information Theory metrics and the foundational understanding that echo chambers reinforce radical viewpoints, potentially exacerbating polarization. Building on this idea, we create a novel counter-polarization recommendation algorithm utilizing Information Theory to quantify the distance between the probability distributions modeling the likelihood of a user using any word from the corpus in a new publication, i.e., the distance between the word-usage probability distribution of all network users.

Specifically, our algorithm employs the Jensen-Shannon distance to quantify the dissimilarity between users. This metric is then utilized by the HAC algorithm, with which we perform a precise and explainable user community detection based on the divergence in each user's word-usage probability distributions, computed employing Information Theory.

Unlike most existing approaches, which overlook the existence of echo chambers and introduce diversity through reranking or similar strategies [BNS18, GSM$^+$22, Str22, AK12, AK14], our algorithm addresses the core issue by bridging echo chambers, through the creation of user-to-user diversification links between different echo chambers, powered by the application of Information Theory. This approach enhances the diversity of discourse and introduces new topics within user communities, fostering healthier and more diverse networks. By leveraging Information Theory, our method identifies and quantifies the polarization and actively works to reduce it by dynamically adjusting to the evolving patterns of user interactions and content generation.

However, our study has several limitations. While we analyzed network polarization using a dataset from Spanish electoral processes spanning 2011 to 2019, future studies could extend this work by applying the same experimental framework to more recent Spanish electoral processes, such as the general and local elections 2023. Additionally, it would be valuable to explore network polarization in countries with different electoral systems, such as the United States. While our recommendation algorithm

demonstrated effectiveness within the employed dataset, future research is necessary to validate its utility across diverse datasets and contexts.

Furthermore, future research should incorporate more state-of-the-art counter-polarization recommendation algorithms to compare against our proposed method. Although we implemented a diverse range of strategies in our study, exploring additional algorithms could enhance the performed benchmark and provide a more robust evaluation of our algorithm's effectiveness.

12.7 Conclusions

Network polarization is a topic of significant interest in Western societies and has recently become a subject of a wide societal debate. Therefore, research aimed at better understanding network polarization, as well as developing strategies to counteract it, such as the one presented in this work, is particularly valuable. Consequently, we consider our research to be of interest not only for advancing academic knowledge in strategies that can effectively work to minimize network polarization, but also for informing and potentially enhancing existing regulations related to social network sites, such as the European Digital Services Act (DSA).

Considering our work, we first performed a comprehensive analysis of existing counter-polarization strategies, which were used to benchmark our proposal. More specifically, the results reflected that both the Multi-Armed Bandit with Greedy Optimized reranking and our proposed algorithm achieve significant performance, offering an excellent balance between recommendation accuracy and network polarization across all examined electoral processes. Conversely, other recommendation algorithms fail to achieve similar outcomes due to their inherent strategic limitations.

Our findings highlight several insights of interest. Firstly, strategies that solely rely on reranking, such as MMR, Greedy Optimized CF, DR, and W2V, do not introduce sufficient diversity to mitigate network polarization within the studied electoral contexts significantly. Moreover, strategies that primarily address network polarization from the outset, including content moderation and PrCP, do not succeed in achieving high recommendation accuracy alongside low network polarization. These methods result in average trade-offs, failing to maximize accuracy or minimize polarization, and proving less effective than the Greedy Optimized TS and our proposed recommendation algorithm.

Additionally, ethical considerations must be noted for specific strategies that employ *shadowbanning* to prevent the propagation of specific content. While such moderation techniques may reduce network polarization, they affect users' fundamental right to access information [MTF20, Ess13], as content exceeding a certain hostility threshold may not be visible or recommended to users.

Overall, our results substantiate the efficacy of our recommendation algorithm answering **RG3.RQ3**, which is grounded in information theory. By fostering connections between different user communities, our approach enhances discourse diversity and introduces new topics into conversations. This method prevents the reinforcement of radical viewpoints within echo chambers, fostering a healthier and more diverse ecosystem that effectively minimizes polarization. Information theory plays a critical role by providing computational capabilities to accurately measure dynamic user preferences and their differences in an explainable way. It helps calculate the distance between word likelihood distributions per user, which the employed HAC community detection algorithm uses to cluster users into communities. These communities experience reinforced interactions through diversified user-to-user recommendations obtained by balancing recommendation accuracy and diversity in user-item rankings answering **RG3.RQ4**.

Part V

Conclusions

"Everywhere we remain unfree and chained to technology, whether we passionately affirm or deny it. But we are delivered over to it in the worst possible way when we regard it as something neutral; for this conception of it, to which today we particularly pay homage, makes us utterly blind to the essence of technology."
— Martin Heidegger, *On Technology*

Chapter 13

Conclusions and Future Work

13.1 Research Conclusions

This thesis investigates the role of content recommendation algorithms in social networks in facilitating online political participation, particularly in relation to significant social phenomena, such as the spread of misinformation and political polarization. Additionally, a clear and concise taxonomy has been developed to classify the various actors involved in social networks during relevant political events, such as online electoral campaigns, allowing for the identification and analysis of the specific roles and behaviors of each group in the dissemination and reception of information.

Similarly, the phenomenon of misinformation has been explored through the definition of the concept of an "information network," which aids in understanding how misleading or false messages are structured and propagated in these digital environments. Furthermore, a proprietary metric has been developed to measure polarization in social networks, a fundamental tool for quantifying and analyzing the fragmentation of opinions online and assessing the impact of algorithm-mediated interactions.

13.1.1 Online Social Networks and political participation

Social networks have profoundly transformed how we relate to each other and engage with information. By becoming alternative sources of information to conventional media, they have altered our information ecosystems and even given rise to new social phenomena. These information-sharing applications have been used since their inception in the late 1990s to disseminate political and current affairs information. Unlike traditional media, they lack editorial filters, and information disseminators are not bound by ethical codes. Similarly, communication on these networks is high-speed, multidirectional, and real-time.

This environment facilitates the emergence of specific user profiles and dynamics regarding information exchange. Traditional actors, such as political parties or opinion leaders, move into social media, generating and amplifying messages. Ideological spaces materialize in dense networks where we find ideology creators, information emitters (who introduce information into the network), and amplifiers, all vying for the attention of large audiences. Traditional profiles, such as political activists, migrate online, increasingly incorporating techniques and procedures associated with digital marketing for disseminating their political content.

In particular, within this research, we have identified a new type of online figure, the **influencer-**

activist that evolves from the combination of cyber-activism and influencer figures. As discussed in Chapter 5, they aim at embedding themselves at the center of a networked structure on social media, to influence a large base of followers, and to maximize their possibilities of spreading a message. We also detected some general patterns in their behavior such as a strong ideological independence, platform diversification, and inter-related, less-structured relationships with other activists, with a will to professionalize their activism to some degree.

To further understand the different user types and their roles in the political conversation in online political ecosystems, in Chapter 6 we found that the quantitative dominance of the conversation, held by activists, does not necessarily translate into actual dominance, as media outlets and political parties play central roles by introducing news and narratives that rapidly spread through the network, reaching user categories with a more amplifying role, such as activists and politicians (particularly during the week of elections). We also identified a clear **information flow** that explains the life cycle of narratives, beginning with creating political parties' posts containing website links to news previously shared by media outlets. This flow is then continued by narrative amplification and a shift in speech dimensions, primarily driven by politicians and activists to enhance the social component of the narrative, making it more relatable to the end-users (the spectators). We discovered that spectators tend to have a more commentator role in this conversation, as they create posts around the main topics of discussion in the network, as defined by the narratives that spread across it.

This new mode of communication, among other effects, maximizes phenomena like the spread of disinformation and polarization around political issues when transferred to the online social media environment, which covers the rest of the issues considered in this thesis. An overview of these problems, from a Spanish perspective, was published in [MDB23].

13.1.2 Disinformation networks

In transitioning to a high-intensity information exchange architecture, the network component of social media applications, combined with algorithmic ease in creating connections, has transformed disinformation and misinformation on social networks from a relatively stable set of individual actors or small groups into large disinformation networks. These networks emerge and capture vast numbers of users, often using echo chambers.

Consequently, disinformation networks are highly interconnected and cohesive groups of users who typically disseminate disinformation in a coordinated manner. This coordination can be formal, mediated by disinforming agents associated with states or political organizations actively directing it, or informal, driven by shared sympathies toward specific disinformative topics. This disinformation centers on major current political issues, aiming to promote hostile narratives, mainly against Western liberal democracies, while supporting autocratic states such as Russia, Venezuela, Cuba, Iran, or China, among others. As observed in Chapter 7, the **high level of cohesion, coordination, and density of connections** within these networks makes it easy for neutral users—content consumers—to become ensnared, receiving (dis)informative inputs from multiple sources at a rate that often outpaces the dissemination of legitimate content by professional journalists or other credible communicators.

Regarding the role of content recommendation algorithms in the emergence and consolidation of these networks, our investigation presented in Chapter 9 has allowed us to **identify how algorithms** based on text similarity recommendations, especially those powered by deep learning, **play a promi-**

nent role in forming and activating these networks. Disinformation networks are thus characterized by highly homogeneous and specific language, along with coordination through hashtags and slogans centered on a limited set of highly politicized topics. Text-based recommendations accelerate the formation of these networks and sustain them over time, as disinformative users can better find each other under this dynamic.

Some results of this specific research have been published in the Applied Network Science journal [MDB24].

13.1.3 Political polarization

As noted in the first part of this thesis, polarization has existed since the advent of mass media systems such as newspapers and radio and is strongly interconnected with the contemporary political system based on representative democracies. This phenomenon is especially evident in Western liberal democracies, where public opinion holds significant weight in the political process.

However, when transferred to the social media environment, this phenomenon is intensified due to the ease with which social networks form echo chambers and the continuous exposure to politically charged content that manifests particularly strongly on these platforms. While existing literature confirms that social media is not the sole cause of polarization, it is a highly relevant factor in its development, as this thesis has uncovered.

In this context, in line with the dynamics observed in Chapter 6, political parties play a polarizing role, especially during electoral processes, by introducing narratives into the network based on current events and creating division among citizens. When these narratives are released, they are often picked up by more radical profiles—activists who, while not formally political, strongly support their preferred political option. These profiles amplify the message, conveying it in a more radicalized form to a broader audience, using a more confrontational style.

More specifically, thanks to the polarization metric proposed and benchmarked in Chapter 8, we are able to encapsulate this concept using a perspective from Information Theory while accounting for its temporal dimension, **tracking the evolution and flow of information over time**. This **polarization metric**, named SPIN (*Social-political Polarization analysis by INformation theory*), allows us to model and characterize polarization along several Spanish electoral processes, while aligning with the assumptions from political literature.

On top of this contribution, in Chapter 10 we observe and quantify the **behavior of different recommendation algorithms with respect to their accuracy and the polarization** spread in the network. Our work found that different recommendation algorithm strategies have a unique effect in polarization dynamics throughout the electoral processes; for instance, collaborative filtering rapidly reaches high polarization values through the creation and reinforcement of communities of very like-minded users. One of our main results is that the polarization phenomenon proves to be affected by the underlying social network algorithm, even if it does not emerge from such network.

Part of the results of this research have been published as an article in the EPJ Data Science journal [MBBRD24].

13.1.4 The role of Recommender Systems

Recommendation systems, in turn, can potentially accelerate the polarization formation and disinformation spread within networks, as mentioned in previous sections. Although polarization levels will always increase with the actions of political parties and the amplification of messages by activists, this amplification is significantly greater with deep learning-based approaches, as these are optimized algorithmically to maximize user engagement with content. A similar situation happens with respect to disinformation.

As a last contribution of this thesis, we demonstrate that recommendation algorithms also contribute to the mitigation of these effects, when intervention strategies are considered. Whereas these systems can accelerate or catalyze polarization and disinformation processes, they can also contribute to their mitigation if they are properly modeled and additional conditions are considered when put in production.

Therefore, the final part of this thesis focuses on **designing and evaluating "algorithmic intervention" strategies** to modulate these phenomena, aiming to create healthier and more constructive social networks for both their users and overall society. In this regard, the chosen approach for this contribution combines Information Theory with Social Network Analysis.

Regarding disinformation, in Chapter 11 we propose an algorithmic intervention strategy to **promote structural diversity in text**, applying information theory principles to text results and user clustering. This approach generates significantly weaker disinformation networks, reducing their impact while **maintaining high stability** across the network, explicitly preserving legitimate journalist networks.

Regarding polarization, Chapter 12 introduces an information theory-based recommendation algorithm designed to counteract online social network polarization. This algorithm uses hierarchical clustering techniques to detect communities and employs Jensen-Shannon distance to calculate user preferences, generating diversified recommendations. We significantly reduce polarization by **bridging different user communities without compromising recommendation accuracy**. Our results show a 45-50 percent reduction in network polarization across various electoral processes, demonstrating the effectiveness of this approach in mitigating polarizing dynamics within fragmented networks.

13.2 Limitations and Future Research

While this work provides a deep analysis of the dynamics of information on online social networks, particularly on micro-blogging platforms, offering both analytical tools and meaningful conclusions, it is important to recognize the study's limitations and the potential for future research. One significant limitation is that this study is confined to the social network X. Although X is the principal platform for real-time commentary on current events, particularly in political discourse, it is not the only social network of importance. Other micro-blogging platforms, such as Tumblr, Gab, Parler, and Truth, are also utilized for political commentary but were not examined in this work. Even though X is the most studied platform, it would be interesting to expand future research to include these alternative networks, which could yield significant insights.

Moreover, other major social networks, such as Facebook, with its closed groups, or platforms like YouTube, Instagram, and TikTok, play a critical role in disseminating current information and political communication. These platforms differ from X in their interface and the dynamics of information exchange, offering new avenues for studying distinct information flows. However, these networks were excluded from the present analysis, allowing future research to explore these differences.

In addition, instant messaging applications or hybrids between social networks and messaging services, such as Telegram, play an especially prominent role in disseminating disinformation and shaping political opinion. Due to this thesis's spatial and temporal constraints, these platforms were also excluded from the analysis. Nevertheless, studying them could further illuminate the phenomenon. Similarly, the cross-platform dynamics between such messaging platforms and social networks like X, as well as the interaction between online platforms and offline information sources, could provide a more comprehensive understanding of the dynamics of political information dissemination.

Regarding the phenomena studied, disinformation and polarization, are the political phenomena related to the internet that generate the most social concern. The rise of social networks strongly influences them and is primarily developed within them. The same applies to social influence and the processes of political opinion formation, which are highly relevant in this context, as everything begins and ends with influence. The generation of influence on social networks enables polarization and the spread of propaganda and disinformation. However, other phenomena could have equally been selected for study, such as radicalization or the dissemination of hate speech. Although these are closely linked to the dynamics of disinformation and polarization, they could also be explored in greater depth in future research.

Regarding the data used, the dataset corresponds to the most significant Spanish electoral processes between 2011 and 2019 and the tracking of a disinformation network compared with a network of legitimate journalists (and the audiences influenced by both groups) over approximately 4 years. This dataset is highly robust and well-suited to study these phenomena. Nonetheless, it is limited both temporally and geographically. The data analyzed pertains primarily to Spain and spans nearly a decade. While the data was limited to Spain for reasons of availability and ease of understanding and because the thesis was completed at a Spanish university, the research would greatly benefit from a more cross-sectional dataset. This would facilitate stronger comparisons and add robustness to the study. Similarly, expanding the study period and the number of accounts analyzed would provide further depth to the research.

Regarding the algorithmic study, this thesis explored the role of content distribution and filtering systems in social networks and their impact on the rise, modulation, or mitigation of the phenomena under study. We implemented and evaluated the most advanced algorithmic approaches available today, employing state-of-the-art techniques in social network simulation. Additionally, key state-of-the-art algorithms proposed for addressing disinformation and polarization were implemented, explicitly aiming to reduce these phenomena through recommendation systems. However, the recommendation field is extraordinarily complex and vast. For instance, deep learning-based recommendation approaches are highly varied, and implementing them presents an insurmountable challenge. This work used a representative selection, but it remains an area ripe for expansion. Similarly, methods based on singular value decomposition, matrix factorization, and reinforcement learning could have been explored more deeply.

Moreover, while the now open-sourced algorithm of the social network X could have been studied more extensively, it should be noted that much of it relies on a neural network trained on X's internal data, which makes replication and analysis difficult from an external perspective. In conclusion, despite the limitations outlined, it is essential to recognize that social networks and recommendation systems constantly evolve. As such, research in this area remains highly dynamic and must continue to advance to keep this knowledge current. Hence, future research could further explore new recommendation methods and their impact on the phenomena studied in this thesis. This could include emphasizing deep learning-based recommendation systems or examining how large language models (LLMs) influence

users through recommendations or how their users might bias them. Additionally, more attention could be directed toward information retrieval systems, such as search engines like Google, to investigate how their algorithms influence information dynamics and user behavior.

Regarding the strategies used to study these phenomena, particularly in mitigating them for the benefit of the social network and its user base, this thesis applied network science and information theory principles to develop advanced recommendation strategies to curb harmful phenomena. These strategies primarily focused on enhancing diversity while maintaining accuracy. Although the results presented in this work are positive and contribute to the state-of-the-art modeling of these phenomena, alternative approaches could be considered. These might include content moderation strategies, the incorporation of fact-checking, or other recommendation strategies based on different theoretical frameworks. Such approaches should be explored in subsequent research to continue advancing the field.

Chapter 14

Conclusiones y Trabajo Futuro

14.1 Conclusiones de la Investigación

En esta tesis se ha investigado el papel de los algoritmos de recomendación de contenido en redes sociales para facilitar la participación política en línea, especialmente en relación con fenómenos de interés social significativo, como la propagación de desinformación y la polarización política. Además, se ha elaborado una taxonomía clara y concisa que clasifica a los distintos actores involucrados en las redes sociales durante eventos políticos relevantes, como las campañas electorales online, lo que permite identificar y analizar los roles y comportamientos específicos de cada grupo en la difusión y recepción de información.

De igual forma, se ha profundizado en el fenómeno de la desinformación mediante la definición del concepto de "red de información", que facilita la comprensión de cómo se estructuran y propagan los mensajes engañosos o falsos en estos entornos digitales. Complementariamente, se ha desarrollado una métrica propia para medir la polarización en redes sociales, herramienta que resulta fundamental para cuantificar y analizar la fragmentación de las opiniones en línea y evaluar el impacto de las interacciones mediadas por algoritmos.

14.1.1 Redes sociales en línea y participación política

Las redes sociales han transformado profundamente la manera en que nos relacionamos e interactuamos con la información. Al convertirse en fuentes alternativas a los medios tradicionales, han modificado nuestros ecosistemas informativos y generado nuevos fenómenos sociales. Estas aplicaciones de intercambio de información se han utilizado desde su creación a finales de los años 90 para difundir información política y de actualidad. A diferencia de los medios tradicionales, carecen de filtros editoriales y los difusores de información no están sujetos a códigos éticos. De igual modo, la comunicación en estas redes es de alta velocidad, multidireccional y en tiempo real.

Este entorno propicia la aparición de perfiles y dinámicas específicas en este proceso de intercambio de información. Actores tradicionales, como partidos políticos o líderes de opinión, se trasladan a las redes sociales, generando y amplificando mensajes. Se crean espacios ideológicos densos donde encontramos creadores de ideas, emisores de información (quienes introducen contenido en la red) y amplificadores, todos compitiendo por captar la atención de grandes audiencias. Perfiles tradicionales, como los activistas políticos, migran al entorno digital, integrando cada vez más técnicas propias del marketing digital para difundir su contenido político.

En esta investigación se identificó un nuevo tipo de figura en línea: el **influencer-activista**, que surge de la combinación del ciberactivismo y los influencers. Como se discutió en el Capítulo 5, estos buscan posicionarse en el centro de estructuras en red dentro de las redes sociales, con el objetivo de influir en una gran base de seguidores y maximizar la difusión de sus mensajes. Se observaron patrones generales en su comportamiento, como independencia ideológica, diversificación de plataformas y relaciones menos estructuradas con otros activistas, con un esfuerzo por profesionalizar su activismo.

Para entender mejor los distintos tipos de usuarios y sus roles en las conversaciones políticas en ecosistemas digitales, en el Capítulo 6 se concluyó que, aunque los activistas dominan cuantitativamente las conversaciones, esto no implica necesariamente una dominancia efectiva, ya que los medios de comunicación y los partidos políticos desempeñan un rol central al introducir narrativas y noticias que se expanden rápidamente por la red. Estas llegan a usuarios con roles amplificadores, como activistas y políticos (particularmente durante semanas electorales). También se identificó un claro **flujo de información** que explica el ciclo de vida de las narrativas en las redes sociales en línea, iniciándose con publicaciones de partidos políticos que enlazan a noticias de medios, seguido por la amplificación de estas narrativas y un cambio en las dimensiones discursivas liderado por políticos y activistas, quienes refuerzan su componente social para hacerlas más comprensibles para los usuarios finales. Estos últimos, los espectadores, tienden a asumir un rol comentador al crear publicaciones alrededor de los temas principales definidos por las narrativas en expansión.

Este nuevo modelo comunicativo maximiza, entre otros efectos, fenómenos como la propagación de desinformación y la polarización política, particularmente en entornos digitales, temas que se abordan en el resto de esta tesis. Una visión general de estos problemas desde una perspectiva española se publicó en [MDB23].

14.1.2 Redes de desinformación

Al transitar hacia una arquitectura de intercambio de información de alta intensidad, la componente de red de aplicaciones de medios sociales, junto con algoritmos que facilitan la creación de conexiones, ha transformado la desinformación y misinformación en redes sociales de un relativamente estable conjunto de actores individuales o pequeños grupos a grandes redes de desinformación. Estas redes emergen y captan a una gran cantidad de usuarios, a menudo utilizando cámaras de eco.

Las redes de desinformación están formadas por grupos de usuarios altamente interconectados y cohesionados que difunden desinformación de manera coordinada. Esta coordinación puede ser formal, dirigida por agentes desinformativos vinculados a estados u organizaciones políticas, o informal, motivada por afinidades hacia ciertos temas. Estas redes suelen centrarse en temas políticos actuales y buscan promover narrativas hostiles contra democracias liberales occidentales mientras apoyan a estados autocráticos como Rusia, Venezuela, Cuba, Irán o China. Como se muestra en el Capítulo 7, el **alto nivel de cohesión, coordinación y densidad de conexiones** facilita que los usuarios neutrales (consumidores de contenido) queden atrapados, recibiendo elementos (des)informativos a un ritmo que supera la difusión de contenido legítimo por periodistas profesionales u otros comunicadores fiables.

En cuanto al papel de los algoritmos de recomendación en la consolidación de estas redes, nuestra investigación presentada en el Capítulo 9 identificó cómo los **algoritmos basados en similitud textual**, especialmente los impulsados por aprendizaje profundo, **desempeñan un papel destacado en la formación y activación de estas redes**. Las redes de desinformación están, de esta manera, caracterizadas

por un lenguaje homogéneo y específico, junto con la coordinación a través de hashtags y eslóganes centrados en un conjunto limitado de temas altamente politizados. Las recomendaciones basadas en texto aceleran la formación de estas redes y las sostienen en el tiempo, ya que los usuarios desinformadores pueden encontrarse, esto es, identificarse mútuamente y establecer relación de mejor manera en esta dinámica.

Algunos resultados de esta investigación se publicaron en la revista *Applied Network Science* [MDB24].

14.1.3 Polarización política

Como se señala en la primera parte de esta tesis, la polarización ha existido desde los sistemas de medios masivos como la prensa y la radio, y está profundamente interconectada con los sistemas políticos democráticos contemporáneos. Sin embargo, este fenómeno se intensifica en las redes sociales debido a la facilidad con la que estas forman cámaras de eco y exponen continuamente a los usuarios a contenido politizado.

Sin embargo, cuando se traslada al ámbito de las redes sociales, este fenómeno se intensifica debido a la facilidad con la que las redes sociales forman cámaras de resonancia y la exposición continua a contenidos de carga política se manifiesta con especial fuerza en estas plataformas. Si bien la literatura existente confirma que las redes sociales no son la única causa de la polarización, sí son un factor muy relevante en su desarrollo, como ha descubierto esta tesis.

En este contexto, en línea con las dinámicas observadas en el Capítulo 6, los partidos políticos juegan un rol polarizador durante los procesos electorales, introduciendo narrativas en la red basadas en eventos actuales y creando división entre los ciudadanos. Cuando se divulgan estas narrativas, suelen ser retomadas por perfiles más radicales: activistas que, si bien no son formalmente políticos, apoyan firmemente su opción política preferida. Estos perfiles amplifican el mensaje y lo transmiten de una forma más radicalizada a un público más amplio, utilizando un estilo más confrontativo.

Más concretamente, gracias a la métrica de polarización propuesta y analizada en el Capítulo 8, podemos resumir este concepto utilizando una perspectiva de la Teoría de la Información, teniendo en cuenta su dimensión temporal, **rastreando la evolución y el flujo de información a lo largo del tiempo**. Esta **métrica de polarización**, denominada SPIN (del inglés, *Social-political Polarization analysis by INformation theory*), nos permite modelar y caracterizar la polarización a lo largo de varios procesos electorales españoles, alineándonos al mismo tiempo con los supuestos de la literatura política.

Además de esta contribución, en el Capítulo 10 observamos y cuantificamos el **comportamiento de diferentes algoritmos de recomendación con respecto a su precisión y la polarización** extendida en la red. Nuestro trabajo encontró que diferentes estrategias de algoritmos de recomendación tienen un efecto único en la dinámica de polarización a lo largo de los procesos electorales; por ejemplo, el filtrado colaborativo alcanza rápidamente altos valores de polarización a través de la creación y el refuerzo de comunidades de usuarios con ideas muy afines. Uno de nuestros principales resultados es que el fenómeno de polarización resulta afectado por el algoritmo de red social subyacente, incluso si no emerge de dicha red.

Parte de los resultados de esta investigación se han publicado como artículo en la revista *EPJ Data Science* [MBBRD24].

14.1.4 El rol de los sistemas de recomendación

Los sistemas de recomendación, a su vez, pueden acelerar potencialmente la formación de polarización y la difusión de desinformación dentro de las redes, como se mencionó en secciones anteriores. Si bien los niveles de polarización siempre aumentarán con las acciones de los partidos políticos y la amplificación de los mensajes de los activistas, esta amplificación es significativamente mayor con los enfoques basados en aprendizaje profundo, ya que estos están optimizados algorítmicamente para maximizar la interacción del usuario con el contenido. Una situación similar ocurre con respecto a la desinformación.

Como última contribución de esta tesis, demostramos que los algoritmos de recomendación también pueden contribuir a la mitigación de estos efectos, cuando se consideran estrategias de intervención. Si bien los sistemas de recomendación pueden acelerar o catalizar estos procesos de polarización y desinformación, también pueden contribuir a su mitigación si se modelan adecuadamente y se consideran condiciones adicionales al ponerlos en producción.

Por ello, la parte final de esta tesis se centra en el **diseño y evaluación de estrategias de "intervención algorítmica"** para modular estos fenómenos, con el objetivo de crear redes sociales más saludables y constructivas tanto para sus usuarios como para la sociedad en general. En este sentido, el enfoque elegido para esta contribución combina la teoría de la información con el análisis de redes sociales.

En relación con la desinformación, en el Capítulo 11 proponemos una estrategia de intervención algorítmica para **promover la diversidad estructural en el texto**, aplicando principios de la teoría de la información a los resultados de texto y a la agrupación de usuarios. Este enfoque genera redes de desinformación significativamente más débiles, reduciendo su impacto mientras **mantiene una alta estabilidad** en toda la red, preservando explícitamente las redes de periodistas legítimos en este caso.

En relación con la polarización, el Capítulo 12 presenta un algoritmo de recomendación basado en la teoría de la información diseñado para contrarrestar la polarización de las redes sociales en línea. Este algoritmo utiliza técnicas de agrupamiento jerárquico para detectar comunidades y emplea la distancia de Jensen-Shannon para calcular las preferencias de los usuarios, generando recomendaciones diversificadas. Reducimos significativamente la polarización al **unir diferentes comunidades de usuarios sin comprometer la precisión de las recomendaciones**. Nuestros resultados muestran una reducción del 45-50 por ciento en la polarización de la red en varios procesos electorales, lo que demuestra la eficacia de este enfoque para mitigar la dinámica polarizadora dentro de redes fragmentadas.

14.2 Limitaciones y Líneas Futuras

Si bien este trabajo proporciona un análisis profundo de la dinámica de la información en las redes sociales en línea, en particular en las plataformas de microblogging, ofreciendo tanto herramientas analíticas como conclusiones significativas, es importante reconocer las limitaciones del estudio y el potencial para futuras investigaciones. Una limitación significativa es que este estudio se limita a la red social X. Si bien X es la plataforma principal para comentarios en tiempo real sobre eventos actuales, en particular en el discurso político, no es la única red social importante. Otras plataformas de microblogging, como Tumblr, Gab, Parler y Truth, también se utilizan para comentarios políticos, pero no se examinaron en este trabajo. Si bien X es la plataforma más estudiada, sería interesante ampliar la investigación futura para incluir estas redes alternativas, que podrían brindar información significativa.

Además, otras redes sociales importantes, como Facebook, con sus grupos cerrados, o plataformas como YouTube, Instagram y TikTok, desempeñan un papel fundamental en la difusión de información actual y la comunicación política. Estas plataformas difieren de X en su interfaz y la dinámica del intercambio de información, lo que ofrece nuevas vías para estudiar los distintos flujos de información. Sin embargo, estas redes se excluyeron del presente análisis, lo que permitirá que futuras investigaciones exploren estas diferencias.

Además, las aplicaciones de mensajería instantánea o los híbridos entre redes sociales y servicios de mensajería, como Telegram, desempeñan un papel especialmente destacado en la difusión de desinformación y la formación de la opinión política. Debido a las limitaciones espaciales y temporales de esta tesis, estas plataformas también fueron excluidas del análisis. No obstante, su estudio podría arrojar más luz sobre el fenómeno. De manera similar, la dinámica multiplataforma entre dichas plataformas de mensajería y redes sociales como X, así como la interacción entre plataformas en línea y fuentes de información fuera de línea, podrían proporcionar una comprensión más completa de la dinámica de la difusión de información política.

En cuanto a los fenómenos estudiados, la desinformación y la polarización son los fenómenos políticos relacionados con internet que generan mayor preocupación social. El auge de las redes sociales influye fuertemente en ellos y se desarrolla fundamentalmente dentro de ellas. Lo mismo ocurre con la influencia social y los procesos de formación de opinión política, que son de gran relevancia en este contexto, pues todo empieza y acaba con la influencia. La generación de influencia en las redes sociales posibilita la polarización y la difusión de propaganda y desinformación. Sin embargo, se podrían haber seleccionado igualmente otros fenómenos para su estudio, como la radicalización o la difusión de discursos de odio. Aunque estos están estrechamente vinculados a la dinámica de la desinformación y la polarización, también podrían ser explorados con mayor profundidad en futuras investigaciones.

Con respecto a los datos utilizados, el conjunto de datos corresponde a los procesos electorales españoles más significativos entre 2011 y 2019 y al seguimiento de una red de desinformación en comparación con una red de periodistas legítimos (y las audiencias influenciadas por ambos grupos) durante aproximadamente cuatro años. Este conjunto de datos es muy robusto, relevante y adecuado para estudiar estos fenómenos. No obstante, está limitado tanto temporal como geográficamente. Los datos analizados pertenecen principalmente a España y abarcan casi una década. Si bien los datos se limitaron a España por razones de disponibilidad y facilidad de comprensión y porque la tesis se realizó en una universidad española, la investigación se beneficiaría enormemente de un conjunto de datos más transversal. Esto facilitaría comparaciones más sólidas y agregaría robustez al estudio. Del mismo modo, ampliar el período de estudio y el número de cuentas analizadas proporcionaría mayor profundidad a la investigación.

En cuanto al estudio algorítmico, esta tesis exploró el papel de los sistemas de distribución y filtrado de contenidos en las redes sociales y su impacto en el surgimiento, modulación o mitigación de los fenómenos en estudio. La investigación implementó y evaluó los enfoques algorítmicos más avanzados disponibles en la actualidad, empleando técnicas de última generación en simulación de redes sociales. Además, la tesis implementó algoritmos clave de última generación propuestos para abordar la desinformación y la polarización, apuntando explícitamente a reducir estos fenómenos a través de sistemas de recomendación. Sin embargo, el campo de los sistemas de recomendación es extraordinariamente complejo y vasto. Por ejemplo, los enfoques de recomendación basados en el aprendizaje profundo son muy variados y su implementación presenta un desafío insuperable. Este trabajo utilizó una selección repre-

sentativa, pero sigue siendo un área madura para la expansión. De manera similar, los métodos basados en la descomposición en valores singulares (del inglés, *singular value decomposition*), la factorización de matrices y el aprendizaje por refuerzo podrían haberse explorado más profundamente.

Además, si bien el algoritmo de la red social X, que ahora es de código abierto, podría haberse estudiado más extensamente, cabe señalar que gran parte del algoritmo se basa en una red neuronal entrenada con los datos internos de X, lo que dificulta la replicación y el análisis desde una perspectiva externa. En conclusión, a pesar de las limitaciones señaladas, es esencial reconocer que las redes sociales y los sistemas de recomendación evolucionan constantemente. Por lo tanto, la investigación en esta área sigue siendo muy dinámica y debe seguir avanzando para mantener este conocimiento actualizado.

Por lo tanto, las investigaciones futuras podrían explorar más a fondo los nuevos métodos de recomendación y su impacto en los fenómenos estudiados en esta tesis. Esto podría incluir el énfasis en los sistemas de recomendación basados en el aprendizaje profundo o examinar cómo los grandes modelos de lenguaje (LLM, del inglés *Large Language Models*) influyen en los usuarios a través de recomendaciones o cómo sus usuarios pueden sesgarlos. Además, se podría dirigir más atención a los sistemas de recuperación de información, como los motores de búsqueda (por ejemplo, Google), para investigar cómo sus algoritmos influyen en la dinámica de la información y el comportamiento del usuario.

En cuanto a las estrategias utilizadas para estudiar estos fenómenos, en particular para mitigarlos en beneficio de la red social y su base de usuarios, esta tesis aplicó principios de la ciencia de redes y la teoría de la información para desarrollar estrategias avanzadas de recomendación para frenar los fenómenos nocivos. Estas estrategias se centraron principalmente en mejorar la diversidad manteniendo la precisión. Aunque los resultados presentados en este trabajo son positivos y contribuyen al modelado de vanguardia de estos fenómenos, se podrían considerar enfoques alternativos. Estos podrían incluir estrategias de moderación de contenido, la incorporación de verificación de hechos u otras estrategias de recomendación basadas en diferentes marcos teóricos. Dichos enfoques deberían explorarse en investigaciones posteriores para continuar avanzando en el área.

Bibliography

[A+13] Pablo Aragón et al. Communication dynamics in twitter during political campaigns: The case of the 2011 spanish national election. *Policy & internet*, 5(2):183–206, 2013.

[AA03] Lada A Adamic and Eytan Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.

[AAB23] Esma Aïmeur, Sabrine Amri, and Gilles Brassard. Fake news, disinformation and misinformation in social media: A review. *Social Network Analysis and Mining*, 13(30):1–36, 2023.

[AABM21] T. Alsinet, J. Argelich, R. Béjar, and S. Martínez. Measuring polarization in online debates. *Applied Sciences*, 11(24):11879, 2021.

[ACF22] M. Mehdi Afsar, Trafford Crump, and Behrouz Far. Reinforcement learning based recommender systems: A survey. *ACM Computing Surveys*, 55(7):1–38, 2022.

[ADM+21] S. Ajovalasit, V. Dorgali, A. Mazza, A. D'Onofrio, and P. Manfredi. Evidence of disorientation towards immunization on online social media after contrasting political communication on vaccines. results from an analysis of twitter data in italy. *Plos One*, 16(7):e0253569, 2021.

[AF18] Marina Azzimonti and Marcos Fernandes. Social media networks, fake news, and polarization. Technical report, National Bureau of Economic Research, 2018.

[AG12] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.

[AG17] H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236, 2017.

[AGAC22] Rubén Arcos, Manuel Gertrudix, Cristina Arribas, and Monica Cardarilli. Responses to digital disinformation as part of hybrid threats: a systematic review on the effects of disinformation and the effectiveness of fact-checking/debunking. *Open Research Europe*, 2, 2022.

[AI19] Haifa Alharthi and Diana Inkpen. Study of linguistic features incorporated in a literary book recommender system. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, SAC '19, page 1027–1034, New York, NY, USA, 2019. Association for Computing Machinery.

[AK12] Gediminas Adomavicius and YoungOk Kwon. Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Transactions on Knowledge and Data Engineering*, 24(5):896–911, May 2012.

[AK14] Gediminas Adomavicius and YoungOk Kwon. Optimization-based approaches for maximizing aggregate recommendation diversity. *INFORMS Journal on Computing*, 26(2):351–369, 2014.

[All22] Alliance4Democracy. Hamilton 2.0 dashboard. `https://securingdemocracy.gmfus.org/hamilton-dashboard/`, accessed July 2023, 2022.

[AR11] Manish Agarwal and David K. Round. *The Emergence of Global Search Engines: Trends in History and Competition*. PhD thesis, Competition Policy International Incorporated, 2011.

[AS22] Yousef Aldaihani and Jae-Hwa Shin. News agenda setting in social media era: Twitter as alternative news source for citizen journalism. In *The Emerald Handbook of Computer-Mediated Communication and Social Media*, pages 233–249. Emerald Publishing Limited, 2022.

[ASI20] Mohiuddin Ahmed, Raihan Seraj, and Syed Mohammed Shamsul Islam. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8):1295, 2020.

[ASIM18] Anitha Anandhan, Nor Liyana Mohd Shuib, Maizatul Akmar Ismail, and Ghulam Mujtaba. Social media recommender systems: Review and open research issues. *IEEE Access*, 6:15608–15628, 2018.

[Ast21] Elena Astakhova. The polarization of spanish society: historical reality and modern dimension. 2021. Available at SSRN 3959700.

[AZ12] Charu C Aggarwal and ChengXiang Zhai. A survey of text clustering algorithms. *Mining text data*, pages 77–128, 2012.

[B+10] Simona Balbi et al. Beyond the curse of multidimensionality: high dimensional clustering in text mining. *Statistica Applicata-Italian Journal of Applied Statistics*, 22(1):53–63, 2010.

[B+21] Javier Bernacer et al. Polarization of beliefs as a consequence of the covid-19 pandemic: The case of spain. *PloS one*, 16(7):e0254511, 2021.

[BAB+21] Samy Benslimane, Jérôme Azé, Sandra Bringay, Maximilien Servajean, and Caroline Mollevi. Controversy detection: A text and graph neural network based approach. In

Wenjie Zhang, Lei Zou, Zakaria Maamar, and Lu Chen, editors, *Web Information Systems Engineering - WISE 2021 - 22nd International Conference on Web Information Systems Engineering, WISE 2021, Melbourne, VIC, Australia, October 26-29, 2021, Proceedings, Part I*, volume 13080 of *Lecture Notes in Computer Science*, pages 339–354. Springer, 2021.

[BAS23] Parisa Bazmi, Masoud Asadpour, and Azadeh Shakery. Multi-view co-attention network for fake news detection by modeling topic-specific user and news source credibility. *Inf. Process. Manag.*, 60(1):103146, 2023.

[BB21] Dhiraj Vaibhav Bagul and Sunita Barve. A novel content-based recommendation approach based on lda topic modeling for literature recommendation. In *2021 6th International conference on inventive computation technologies (ICICT)*, pages 954–961. IEEE, 2021.

[BBF21] A. Bliuc, A. Bouguettaya, and K. D. Felise. Online intergroup polarization across political fault lines: an integrative review. *Frontiers in Psychology*, 12, 2021.

[BCC+17] Mattia Bianchi, Federico Cesaro, Filippo Ciceri, Mattia Dagrada, Alberto Gasparin, Daniele Grattarola, Ilyas Inajjar, Alberto Maria Metelli, and Leonardo Cella. Content-based approaches for cold-start job recommendations. In *Proceedings of the Recommender Systems Challenge 2017*, pages 1–5. 2017.

[BD16] Leticia Bode and Kajsa E. Dalrymple. Politics in 140 characters or less: Campaign communication, network interaction, and political participation on twitter. *Journal of Political Marketing*, 15(4):311–332, 2016.

[Ber19] Melania Berbatova. Overview on nlp techniques for content-based recommender systems for books. In *Proceedings of the Student Research Workshop Associated with RANLP 2019*, pages 55–61, 2019.

[BFR18] Yochai Benkler, Robert Faris, and Hal Roberts. *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press, 2018.

[BG20] Lucas Böttcher and Hans Gersbach. The great divide: drivers of polarization in the us public. *EPJ Data Science*, 9(1):32, 2020.

[BGLL08] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, oct 2008.

[BGS05] Monica Bianchini, Marco Gori, and Franco Scarselli. Inside pagerank. *ACM Transactions on Internet Technology (TOIT)*, 5(1):92–128, 2005.

[BHH18] A. Bruns, S. Harrington, and E. Hurcombe. Click, share, send, forget: The dynamics of news diffusion via twitter. *Journalism Studies*, 19(11):1559–1579, 2018.

[BHM+19] Dimitrios Bountouridis, Jaron Harambam, Mykola Makhortykh, Mónica Marrero, Nava Tintarev, and Claudia Hauff. SIREN: A simulation framework for understanding the effects of recommender systems in online news environments. In danah boyd and Jamie H. Morgenstern, editors, *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 150–159. ACM, 2019.

[BHMW11] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an influencer: quantifying influence on twitter. In *WSDM*, pages 65–74. ACM, 2011.

[Bio08] Student Biometrika. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908.

[BK05] Jayanta Basak and Raghu Krishnapuram. Interpretable hierarchical clustering by constructing an unsupervised decision tree. *IEEE transactions on knowledge and data engineering*, 17(1):121–132, 2005.

[BK22] Laia Balcells and Alexander Kuo. Secessionist conflict and affective polarization: Evidence from catalonia. *Journal of Peace Research*, page 00223433221088112, 2022.

[BM13] Marija Anna Bekafigo and Allan McBride. Who tweets about politics? political participation of twitter users during the 2011 gubernatorial elections. *Social Science Computer Review*, 31(5):625–643, 2013.

[BM19] M. T. Bastos and D. Mercea. The brexit botnet and user-generated hyperpartisan news. *Social Science Computer Review*, 37(1):38–54, 2019.

[BMA15] E. Bakshy, S. Messing, and L. A. Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.

[BMB20] M. T. Bastos, D. Mercea, and A. Baronchelli. The brexit botnet and user-generated hyperpartisan news. *Social Science Computer Review*, 38(1):38–54, 2020.

[BMV11] Graeme Baxter, Rita Marcella, and Evaggelos Varfis. The use of the internet by political parties and candidates in scotland during the 2010 uk general election campaign. In *Aslib Proceedings*, volume 63, pages 464–483. Emerald Group Publishing Limited, 2011.

[BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[BNS18] Mahsa Badami, Olfa Nasraoui, and Patrick Shafto. Prcp: Pre-recommendation counterpolarization. In *KDIR*, pages 280–287, 2018.

[BOHG13] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. Recommender systems survey. *Knowledge-based systems*, 46:109–132, 2013.

[Bou20] Shelley Boulianne. Twenty years of digital media effects on civic and political participation. *Communication research*, 47(7):947–966, 2020.

[BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Comput. Networks*, 30(1-7):107–117, 1998.

[Bru19] Axel Bruns. *Are filter bubbles real?* John Wiley & Sons, 2019.

[BS12] W. Lance Bennett and Alexandra Segerberg. The logic of connective action. *Information, Communication & Society*, 15(5):739–768, 2012.

[BSKNS22] A. Bell, I. Solano-Kamaiko, O. Nov, and J. Stoyanovich. It's just not that simple: an empirical study of the accuracy-explainability trade-off in machine learning for public policy. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 248–266, June 2022.

[Bur02] Robin Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12:331–370, 2002.

[BW19] D. Bischof and M. Wagner. Do voters polarize when radical parties enter parliament? *American Journal of Political Science*, 63(4):888–904, 2019.

[C+11] Michael Conover et al. Political polarization on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, 2011.

[C+24] D. Chavalarias et al. Can a single line of code change society? optimizing engagement in recommender systems necessarily entails systemic risks for global information flows, opinion dynamics and social structures. *Journal of Artificial Societies and Social Simulation*, 24(1):9, 2024.

[Cas10] Manuel Castells. *The Rise of the Network Society*, volume 1. Wiley-Blackwell, Chichester, 2 edition, 2010.

[Cas12] Manuel Castells. *Networks of Outrage and Hope: Social Movements in the Internet Age*. Polity, London, UK, 2012.

[Cas13] Manuel Castells. *Communication Power*. OUP Oxford, Oxford, UK, 1 edition, 2013.

[CBR21] Fernando Casal Bértoa and José Rama. Polarization: what do we know and what can we do about it? *Frontiers in Political Science*, 3:687695, 2021.

[CG98] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336, 1998.

[Cha13] Andrew Chadwick. *The Hybrid Media System: Politics and Power*. Oxford studies in Digital Politics, Oxford, UK, 2013.

[Cha19] Fake News Challenge. Fake news challenge: A dataset and competition for fake news detection. `http://www.fakenewschallenge.org/`, accessed July 2023, 2019.

[Chu17] Kenneth Ward Church. Word2vec. *Natural Language Engineering*, 23(1):155–162, 2017.

[CKS14] Mingming Chen, Konstantin Kuzmin, and Boleslaw K. Szymanski. Community detection via maximization of modularity and its variants. *IEEE Transactions on Computational Social Systems*, 1(1):46–65, 2014.

[CL11] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.

[CLDB18] Xi Chen, Jefrey Lijffijt, and Tijl De Bie. Quantifying and minimizing risk of conflict in social networks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 1197–1205, New York, NY, USA, 2018. Association for Computing Machinery.

[CLH+21] K. Chen, Y. Luo, A. Hu, J. Zhao, and L. Zhang. Characteristics of misinformation spreading on social media during the covid-19 outbreak in china: a descriptive analysis. *Risk Management and Healthcare Policy*, 14:1869–1879, 2021.

[ČLM13] Matej Črepinšek, Shih-Hsi Liu, and Marjan Mernik. Exploration and exploitation in evolutionary algorithms: A survey. *ACM computing surveys (CSUR)*, 45(3):1–33, 2013.

[CM01] S. H. Chaffee and M. J. Metzger. The end of mass communication? *Mass Communication and Society*, 4(4):365–379, 2001.

[CMMB22] Fabio Cinus, Marco Minici, Claudio Monti, and Francesco Bonchi. The effect of people recommenders on echo chambers and polarization. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 90–101, 2022.

[CR20] Andreu Casero-Ripolles. Political influencers in the digital public sphere Introduction. *Communication & Society-Spain*, 33(2):171–173, 2020.

[CR22] Michele Coscia and Luca Rossi. How minimizing conflicts could lead to polarization on social media: An agent-based model investigation. *PloS One*, 17(1):e0263184, 2022.

[CRA14] Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication*, 64(2):317–332, 2014.

[CRB19] Victoria Carty and Francisco G. Reynoso Barron. *Social movements and new technology: The dynamics of cyber activism in the digital age*, pages 373–397. 2019.

[CRC19a] Rocío Cañamares, Marcos Redondo, and Pablo Castells. Multi-armed recommender system bandit ensembles. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 432–436, 2019.

[CRC19b] Namrata Chaudhary and Drimik Roy Chowdhury. Data preprocessing for evaluation of recommendation models in e-commerce. *Data*, 4(1):23, 2019.

[CRF+11] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 89–96. AAAI, 2011.

[CSS15] Zohreh Dehghani Champiri, Seyed Reza Shahamiri, and Siti Salwah Binti Salim. A systematic review of scholar context-aware recommender systems. *Expert Systems with Applications*, 42(3):1743–1758, 2015.

[CV18] Gian Vittorio Caprara and Michele Vecchione. On the left and right ideological divide: Historical accounts and contemporary perspectives. *Political Psychology*, 39:49–83, 2018.

[CZ20] Ming Chen and Xiuze Zhou. Deeprank: Learning to rank with neural networks for recommendation. *Knowledge-Based Systems*, 209:106478, 2020.

[Dah21] P. Dahlgren. A critical review of filter bubbles and a comparison with selective exposure. *Journal of Media Studies*, 20(2):120–135, 2021.

[DCFG21] A. D'Ulizia, M. C. Caschera, F. Ferri, and P. Grifoni. Repository of fake news detection datasets. version 1. 4tu.researchdata. dataset. `https://doi.org/10.4121/14151755.v1`, accessed July 2023, 2021.

[DD19a] Alison Dagnes and Alison Dagnes. *Us vs. them: Political polarization and the politicization of everything*, pages 119–165. 2019.

[DD19b] Sasha Dookhoo and Melissa D. Dodd. Slacktivists or activists? millennial motivations and behaviors for engagement in activism. *Public relations journal*, 13(1), 2019.

[DDDD09] Elena Deza, Michel Marie Deza, Michel Marie Deza, and Elena Deza. *Encyclopedia of distances*. Springer, 2009.

[DEB96] Paul DiMaggio, John Evans, and Bethany Bryson. Have american's social attitudes become more polarized? *American Journal of Sociology*, 102(3):690–755, 1996.

[dGLM+15] Marco de Gemmis, Pasquale Lops, Cataldo Musto, Fedelucio Narducci, and Giovanni Semeraro. Semantics-aware content-based recommender systems. *Recommender Systems Handbook*, pages 119–159, 2015.

[DHKH13] Simon DeDeo, Robert XD Hawkins, Sara Klingenstein, and Tim Hitchcock. Bootstrap methods for the empirical study of decision-making and information flows in social systems. *Entropy*, 15(6):2246–2276, 2013.

[DL22] James N Druckman and Jeremy Levy. Affective polarization in the american public. In *Handbook on politics and public opinion*, pages 257–270. Edward Elgar Publishing, 2022.

[DM23] Stephen Davies and Justin Mittereder. An agent-based model of political polarization without party influence or centralized messaging. In *9th International Conference on Computational Social Science (IC2S2)*, Copenhagen, Denmark, 2023.

[DMB05] Peter Dawyndt, H De Meyer, and B De Baets. The complete linkage clustering algorithm revisited. *Soft Computing*, 9:385–392, 2005.

[dPD16] Donatella della Porta and Mario Diani. *The Oxford Handbook of Social Movements*. Oxford University Press, Oxford, UK, 2016.

[DPPS24] Pavla Dráždilová, Petr Prokop, Jan Platoš, and Václav Snášel. A hierarchical overlapping community detection method based on closed trail distance and maximal cliques. *Information Sciences*, 662:120271, 2024.

[DRDB17] William H Dutton, Bianca Reisdorf, Elizabeth Dubois, and Grant Blank. Social shaping of the politics of internet search and networking: Moving beyond filter bubbles, echo chambers, and fake news. 2017.

[DSWS24] Kayla Duskin, Joseph S Schafer, Jevin D West, and Emma S Spiro. Echo chambers in the age of algorithms: An audit of twitter's friend recommender system. In *Proceedings of the 16th ACM Web Science Conference*, pages 11–21, 2024.

[DWX+22] Zhenhua Dong, Zhe Wang, Jun Xu, Ruiming Tang, and Jirong Wen. A brief history of recommender systems. *arXiv*, 2209.01860v1, 2022. Version 1, September 2022.

[DZ21] Tim Donkers and Jürgen Ziegler. The dual echo chamber: Modeling social media polarization for interventional recommending. In *Proceedings of the 15th ACM Conference on Recommender Systems*, 2021.

[DZ23] Tim Donkers and Jürgen Ziegler. De-sounding echo chambers: Simulation-based analysis of polarization dynamics in social networks. *Online Social Networks and Media*, 37:100275, 2023.

[E+20] Elif Edizel et al. Understanding filter bubbles and bias in recommender systems. *Journal of Information Science*, 2020.

[Eks21] Michael D. Ekstrand. Multiversal simulacra: Understanding hypotheticals and possible worlds through simulation. *CoRR*, abs/2110.00811, 2021.

[Ell21] Jacques Ellul. *Propaganda: The formation of men's attitudes*. Vintage, New York, NY, 2021.

[ENT+20] Hanif Emamgholizadeh, Milad Nourizade, Mir Saman Tajbakhsh, Mahdieh Hashminezhad, and Farzaneh Nasr Esfahani. A framework for quantifying controversy of social network debates using attributed networks: biased random walk (BRW). *Social Network Analysis and Mining*, 10(1), nov 2020.

[ER15] Robert Epstein and Ronald E Robertson. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, 112(33):E4512–E4521, 2015.

[Ess13] Charles Ess. *Digital Media Ethics*. Polity, 2013.

[EVHH11] Robin Effing, Jos Van Hillegersberg, and Theo Huibers. Social media and political participation: are facebook, twitter and youtube democratizing our political systems? In *Electronic Participation: Third IFIP WG 8.5 International Conference, ePart 2011, Delft,*

The Netherlands, August 29–September 1, 2011. Proceedings, volume 3, pages 25–35. Springer Berlin Heidelberg, 2011.

[Fal15] Don Fallis. What is disinformation? *Library trends*, 63(3):401–426, 2015.

[FB20] Miriam Fernandez and Alejandro Bellogín. Recommender systems and misinformation: The problem or the solution? In *OHARS Workshop. 14th ACM Conference on Recommender Systems*, 2020.

[FENKW22] Antonio Ferrara, Lisette Espin-Noboa, Fariba Karimi, and Claudia Wagner. Link recommendations: Their impact on network structure and minorities. In *Proceedings of the 14th ACM Web Science Conference 2022 (WebSci '22)*, 2022.

[FGR16] S. Flaxman, S. Goel, and J. M. Rao. Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, 80(S1):298–320, 2016.

[FI20] Terry Flew and Petros Iosifidis. Populism, globalisation and social media. *International Communication Gazette*, 82(1):7–25, 2020.

[FS65] Jonathan L. Freedman and David O. Sears. Selective exposure. In *Advances in Experimental Social Psychology*, volume 2, pages 57–97. Academic Press, 1965.

[FS16] Pnina Fichman and Madelyn R. Sanfilippo. *Online trolling and its perpetrators: Under the cyberbridge*. Rowman & Littlefield, 2016.

[Gar17] R.K. Garrett. Echo chamber distraction: Disinformation campaigns and their impact. *Journal of Communication*, 67(3):451–471, 2017.

[Gar18] Kiran Garimella. *Polarization on social media*. Phd thesis, Aalto University, July 2018.

[Gar23] Juan Rodríguez Garat. Ucrania: La desinformación como arma de guerra. *Cuadernos de pensamiento naval: Suplemento de la revista general de marina*, 35:33–52, 2023.

[GAS+15] D. Garcia, A. Abisheva, S. Schweighofer, U. Serdült, and F. Schweitzer. Ideological and temporal components of network polarization in online political participatory media. *Policy & Internet*, 7(1):46–79, 2015.

[GB17] Nathan Gilkerson and Kati Berg. *Social Media, Hashtag Hijacking, and the Evolution of an Activist Group Strategy*, pages 141–155. 06 2017.

[GGL+13] Pankaj Gupta, Ashish Goel, Jimmy Lin, Aneesh Sharma, Dong Wang, and Reza Zadeh. WTF: the who to follow service at twitter. In Daniel Schwabe, Virgílio A. F. Almeida, Hartmut Glaser, Ricardo Baeza-Yates, and Sue B. Moon, editors, *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*, pages 505–514, New York, NY, 2013. International World Wide Web Conferences Steering Committee / ACM.

[GGLC22] Alexis Guyot, Annabelle Gillet, Eric Leclercq, and Nadine Cullot. *ERIS: An Approach Based on Community Boundaries to Assess Polarization in Online Social Networks*, pages 88–104. 05 2022.

[GJCK13] Pedro Henrique Calais Guerra, Wagner Meira Jr, Claire Cardie, and Robert D. Kleinberg. A measure of polarization on social media networks based on community boundaries. *Proceedings of the International AAAI Conference on Web and Social Media*, 2013.

[GJF+19] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer. Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425):374–378, 2019.

[GK20] A. Goreis and O. D. Kothgassner. Social media as vehicle for conspiracy beliefs on covid-19. *Digital Psychology*, 1(2):36–39, 2020.

[GK21] Garima Gupta and Rahul Katarya. Research on understanding the effect of deep learning on user preferences. *Arabian Journal for Science and Engineering*, 46(4):3247–3286, 2021.

[GLM20] Daniel Q Gillion, Jonathan M Ladd, and Marc Meredith. Party polarization, ideological sorting and the emergence of the us partisan gender gap. *British Journal of Political Science*, 50(4):1217–1243, 2020.

[GMGM18] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Quantifying controversy on social media. *ACM Trans. Soc. Comput.*, 1(1):3:1–3:27, 2018.

[GNL13] R Kelly Garrett, Erik C Nisbet, and Emily K Lynch. Undermining the corrective effects of media-based political fact checking? the role of contextual cues and naïve theory. *Journal of Communication*, 63(4):617–637, 2013.

[GNR19] A. Guess, B. Nyhan, and J. Reifler. Exposure to untrustworthy websites in the 2016 us election. *Nature Human Behaviour*, 3(4):1–9, 2019.

[GOBG23] Emilija Gagrčin, Jakob Ohme, Lina Buttgereit, and Felix Grünewald. Datafication markers: Curation and user network effects on mobilization and polarization during elections. *Media and Communication*, 11(3), 2023.

[Goe19] Niels D. Goet. Measuring polarization with text analysis: Evidence from the uk house of commons, 1811–2015. *Political Analysis*, 27(4):518–539, 2019.

[Goo14] Jeff Goodwin. *The Social Movements Reader*. Wiley John + Sons, Hoboken, NJ, USA, 3 edition, 2014.

[GR14] Anatoliy Gruzd and Jeffrey Roy. Investigating political polarization on twitter: A canadian perspective. *Policy & Internet*, 6(1):28–45, 2014.

[Gra73] Mark S Granovetter. The strength of weak ties. *American journal of sociology*, 78(6):1360–1380, 1973.

[Gra94] Antonio Gramsci. *Letters from Prison*. Columbia University Press, Camden, London, 1 edition, 1994.

[GRZW21] L. Guenther, G. Ruhrmann, M. C. Zaremba, and N. Weigelt. The newsworthiness of the "march for science" in germany: Comparing news factors in journalistic media and on twitter. *JCOM*, 20(02):A03, 2021.

[GS09] Asela Gunawardana and Guy Shani. A survey of accuracy evaluation metrics of recommendation tasks. *Journal of Machine Learning Research*, 10(12), 2009.

[GS11] M. Gentzkow and J. M. Shapiro. Ideological segregation online and offline. *The Quarterly Journal of Economics*, 126(4):1799–1839, 2011.

[GSM+22] Zhaolin Gao, Tianshu Shen, Zheda Mai, Mohamed Reda Bouadjenek, Isaac Waller, Ashton Anderson, Ron Bodkin, and Scott Sanner. Mitigating the filter bubble while maintaining relevance: Targeted diversification with vae-based recommender systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2524–2531, New York, NY, USA, 2022. Association for Computing Machinery.

[GSRC+16] Javier Guallar, Jaume Suau, Carlos Ruiz-Caballero, Albert Sáez, and Pere Masip. Redissemination of news and public debate on social networks. *Profesional de la información*, 25(3):358–366, jun. 2016.

[GSY12] Asela Gunawardana, Guy Shani, and Sivan Yogev. Evaluating recommender systems. In *Recommender systems handbook*, pages 547–601. Springer, 2012.

[GTC+20] S. Guarino, N. Trino, A. Celestini, et al. Characterizing networks of propaganda on twitter: a case study. *Applied Network Science*, 5(59), 2020.

[GW17] Venkata Rama Kiran Garimella and Ingmar Weber. A long-term analysis of polarization on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 2017.

[HAN19] Frederik Hjorth and Rebecca Adler-Nissen. Ideological asymmetry in the reach of pro-russian digital disinformation to united states audiences. *Journal of Communication*, 69(2):168–192, April 2019.

[HAR21] Enrique Hernández, Eva Anduiza, and Guillem Rico. Affective polarization and the salience of elections. *Electoral Studies*, 69:102203, 2021.

[HDC23] Marilena Hohmann, Karel Devriendt, and Michele Coscia. Quantifying ideological polarization on a network using generalized euclidean distance. *Science Advances*, March 2023.

[Hen43] Edgar H. Henderson. Toward a definition of propaganda. *The Journal of Social Psychology*, 18(1):71–87, 1943.

[HKP17] Kasper M. Hansen and Karina Kosiara-Pedersen. How campaigns polarize the electorate: Political polarization as an effect of the minimal effect theory within a multi-party system. *Party Politics*, 23(3):181–192, 2017.

[HPR21] Louisa Ha, Loarre Andreu Perez, and Rik Ray. Mapping recent development in scholarship on fake news and misinformation, 2008 to 2017: Disciplinary contribution, topics, and impact. *American behavioral scientist*, 65(2):290–315, 2021.

[HRC11] Richard Hanna, Andrew Rohm, and Victoria L. Crittenden. We're all connected: The power of the social media ecosystem. *Business Horizons*, 54(3):265–273, 2011.

[HSS13] Itai Himelboim, Marc Smith, and Ben Shneiderman. Tweeting apart: Applying network analysis to detect selective exposure clusters in twitter. *Communication Methods and Measures*, 7(3-4):195–223, 2013.

[HSSP13] I. Himelboim, M. A. Smith, B. Shneiderman, and S. Park. Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on twitter. *Journal of Computer-Mediated Communication*, 18(2):154–174, 2013.

[HT12] Robert Hillmann and Matthias Trier. Sentiment polarization and balance among users in online social networks. 2012.

[Huc16] Thomas Huckin. Propaganda defined. In *Propaganda and rhetoric in democracy: History, theory, analysis*, pages 118–136. 2016.

[Hut09] Joseph Huttner. *From Tapestry to SVD: A Survey of the Algorithms That Power Recommender Systems*. PhD thesis, University Name (if available), 2009.

[HYYP13] Jina Huh, Meliha Yetisgen-Yildiz, and Wanda Pratt. Text classification for assisting moderators in online health communities. *Journal of biomedical informatics*, 46(6):998–1005, 2013.

[IAP24] Andreea Iana, Mehwish Alam, and Heiko Paulheim. A survey on knowledge-aware news recommender systems. *Semantic Web*, (Preprint):1–62, 2024.

[ICMFS12] J. Ignacio Criado, Guadalupe Martínez-Fuentes, and Aitor Silván. *Social media for political campaigning. The use of Twitter by Spanish mayors in 2011 local elections*, pages 219–232. 2012.

[ID10] Nitin Indurkhya and Fred J. Damerau, editors. *Handbook of Natural Language Processing, Second Edition*. Chapman and Hall/CRC, 2010.

[IFO15] Folasade Olubusola Isinkaye, Yetunde O Folajimi, and Bolande Adefowoke Ojokoh. Recommendation systems: Principles, methods and evaluation. *Egyptian informatics journal*, 16(3):261–273, 2015.

[IK87] S. Iyengar and D. R. Kinder. *News that matters: Television and American opinion*. University of Chicago Press, Chicago, IL, 1987.

[ISL12] S. Iyengar, G. Sood, and Y. Lelkes. Affect, not ideology. *Public Opinion Quarterly*, 76(3):405–431, 2012.

[JJC22] S.M. Jones-Jang and M. Chung. Can we blame social media for polarization? counter-evidence against filter bubble claims during the covid-19 pandemic. *Journal of Information Technology & Politics*, 19(1):1–16, 2022.

[JJS12] A. Jungherr, P. Jürgens, and H. Schoen. Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, a., sprenger, t. o., sander, p. g., & welpe, i. m. (2011). *Social Science Computer Review*, 30(2):229–234, 2012.

[JNXH05] Liping Jing, Michael K Ng, Jun Xu, and Joshua Zhexue Huang. Subspace clustering of text documents with feature weighting k-means algorithm. In *Advances in Knowledge Discovery and Data Mining: 9th Pacific-Asia Conference, PAKDD 2005, Hanoi, Vietnam, May 18-20, 2005. Proceedings 9*, pages 802–812. Springer, 2005.

[Jon19] Marc Owen Jones. The gulf information war— propaganda, fake news, and fake trends: The weaponization of twitter bots in the gulf crisis. *International journal of communication*, 13:27, 2019.

[Joy10] Mary Joyce. *Digital Activism Decoded: The New Mechanism of Change*. Central European University Press, Budapest, Hungary, 1 edition, 2010.

[JSH+21] Umair Javed, Kamran Shaukat, Ibrahim A. Hameed, Farhat Iqbal, Talha Mahboob Alam, and Suhuai Luo. A review of content-based and context-based recommendation systems. *International Journal of Emerging Technologies in Learning (iJET)*, 16(3):274–306, 2021.

[Jun16] Andreas Jungherr. Twitter use in election campaigns: A systematic literature review. *Journal of Information Technology & Politics*, 13(1):72–91, 2016.

[JVHB14] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PloS one*, 9:e98679, 06 2014.

[JVS00] Nicholas W. Jankowski and Martine Van Selm. Traditional news media online: an examination of added-value. *COMMUNICATIONS-SANKT AUGUSTIN THEN BERLIN-*, 25(1):85–102, 2000.

[JWK14] Gawesh Jawaheer, Peter Weller, and Patty Kostkova. Modeling user preferences in recommender systems: A classification framework for explicit and implicit user feedback. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 4(2):1–26, 2014.

[JWS+23] Jing Jing, Hongchen Wu, Jie Sun, Xiaochang Fang, and Huaxiang Zhang. Multimodal fake news detection via progressive fusion networks. *Inf. Process. Manag.*, 60(1):103120, 2023.

[Kal16] Bente Kalsnes. The social media paradox explained: Comparing political parties' facebook strategy versus practice. *Social Media+ Society*, 2(2):2056305116644616, 2016.

[Kar15] Athina Karatzogianni. *Firebrand Waves of Digital Activism 1994-2014*. Palgrave Macmillan, Camden, London, 2015.

[Kar21] Vasileios Karagiannopoulos. *A Short History of Hacktivism: Its Past and Present and What Can We Learn from It*, pages 63–86. Springer International Publishing, Cham, 2021.

[KASW98] Ioannis Kontoyiannis, Paul Algoet, Yuri Suhov, and Abraham Wyner. Nonparametric entropy estimation for stationary processesand random fields, with applications to english text. *Information Theory, IEEE Transactions on*, 44:1319 – 1327, 06 1998.

[KB18] S. Kalathil and T. C. Boas. The rise of digital repression: How technology is reshaping power, politics, and resistance. *Journal of Democracy*, 29(3):41–55, 2018.

[KBKW21] Dimitris Kalimeris, Smriti Bhagat, Shankar Kalyanaraman, and Udi Weinsberg. Preference amplification in recommender systems. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 805–815, New York, NY, USA, 2021. Association for Computing Machinery.

[KCBP21] Eleni Kapantai, Alexandra Christopoulou, Christos Berberidis, and Vassilios Peristeras. A systematic literature review on disinformation: Toward a unified taxonomical framework. *New Media & Society*, 23(5):1301–1326, 2021.

[KCHM23] Maria Kyriakidou, Stephen Cushion, Ceri Hughes, and Marina Morani. Questioning fact-checking in the fight against disinformation: An audience perspective. *Journalism Practice*, 17(10):2123–2139, 2023.

[KCN+18] Mohammad Karimi, Fabio Crestani, Mohsen J. Nouri, Iadh Ounis, and Hideo Joho. A review of recommender systems in the recsys community. *Information Processing & Management*, 54(3):48–68, 2018.

[KDK+21] Jon Kingzette, James N Druckman, Samara Klar, Yanna Krupnikov, Matthew Levendusky, and John Barry Ryan. How affective polarization undermines support for democratic norms. *Public Opinion Quarterly*, 85(2):663–677, 2021.

[KGSS15] Anuranjan Kumar, Sahil Gupta, Sanjay Kumar Singh, and Kaushal K Shukla. Comparison of various metrics used in collaborative filtering for recommendation system. In *2015 Eighth International Conference on Contemporary Computing (IC3)*, pages 150–154. IEEE, 2015.

[Kid03] Dorothy Kidd. *Indymedia.org: A New Communications Commons*, pages 47–69. 01 2003.

[KJK+20] Ramez Kouzy, Joseph Abi Jaoude, Afif Kraitem, Molly B El Alam, Basil Karam, Elio Adib, Jabra Zarka, Cindy Traboulsi, Elie W Akl, and Khalil Baddour. Coronavirus goes viral: quantifying the covid-19 misinformation epidemic on twitter. *Cureus*, 12(3), 2020.

[KK97] George Karypis and Vipin Kumar. Metis—a software package for partitioning unstructured graphs, partitioning meshes and computing fill-reducing ordering of sparse matrices. 01 1997.

[KL20] Ferath Kherif and Adeliya Latypova. Principal component analysis. In *Machine learning*, pages 209–225. Elsevier, 2020.

[KLPC22] Hyeyoung Ko, Suyeon Lee, Yoonseo Park, and Anna Choi. A survey of recommendation systems: Recommendation models, techniques, and application fields. *Electronics*, 11(141), 2022. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

[KM22] Md Yasin Kabir and Sanjay Madria. A deep learning approach for ideology detection and polarization analysis using covid-19 tweets. In Jolita Ralyté, Sharma Chakravarthy, Mukesh Mohania, Manfred A. Jeusfeld, and Kamalakar Karlapalem, editors, *Conceptual Modeling*, pages 209–223, Cham, 2022. Springer International Publishing.

[KPM19] Gavin L. Kirkwood, Holly J. Payne, and Joseph P. Mazer. Collective trolling as a form of organizational resistance: Analysis of the #justiceforbradswife twitter campaign. *Communication Studies*, 70(3):332–351, 2019.

[KU18] Anne Kaun and Julie Uldam. Digital activism: After the hype. *New Media & Society*, 20(6):2099–2106, 2018.

[Kul51] Solomon Kullback. Kullback-leibler divergence, 1951.

[KvS21] Emily Kubin and Christian von Sikorski. The role of (social) media in political polarization: a systematic review. *Annals of the International Communication Association*, 45(3):188–206, 2021.

[Lam13] Prabin Lama. Clustering system based on text mining using the k-means algorithm. *Turku University of Applied Sciences Thesis, Information Technology*, 2013.

[LBB+18] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, and M. Schudson. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[LCKK14] Jae Kook Lee, Jihyang Choi, Cheonsoo Kim, and Yonghwan Kim. Social media, network heterogeneity, and opinion polarization. *Journal of communication*, 64(4):702–722, 2014.

[LCLS10] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[LdFP+21] Sahil Loomba, Alexandre de Figueiredo, Sarah J. Piatek, Kristen de Graaf, and Heidi J. Larson. Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa. *Nature Human Behaviour*, 5(3):337–348, Mar 2021. Epub 2021 Feb 5. Erratum in: Nat Hum Behav. 2021 Mar;5(3):407. doi: 10.1038/s41562-021-01088-7. Erratum in: Nat Hum Behav. 2021 Jul;5(7):960. doi: 10.1038/s41562-021-01172-y.

[LEC12] S. Lewandowsky, U. K. Ecker, and J. Cook. Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3):106–131, 2012.

[LEC17] S. Lewandowsky, U. K. Ecker, and J. Cook. Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, 6(4):353–369, 2017.

[LHCA10] Neal Lathia, Stephen Hailes, Licia Capra, and Xavier Amatriain. Temporal diversity in recommender systems. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 210–217, 2010.

[LHW+22] Ziyue Li, Hang Hu, He Wang, Luwei Cai, Haipeng Zhang, and Kunpeng Zhang. Why does the president tweet this? discovering reasons and contexts for politicians' tweets from news articles. *Inf. Process. Manag.*, 59(3):102892, 2022.

[Li22] Wenxiao Li. Online political participation based on a political—cultural view: evidence from a survey in china. *Journal of Innovation and Social Science Research*, 10(10), 2022.

[LJB19] Roel O Lutkenhaus, Jeroen Jansz, and Martine PA Bouman. Mapping the dutch vaccination debate on twitter: Identifying communities, narratives, and interactions. *Vaccine: X*, 1:100019, 2019.

[LKM17] Darren G. Lilleker and Karolina Koc-Michalska. What drives political participation? motivations and mobilization in a digital age. *Political Communication*, 34(1):21–43, 2017.

[LLC10] Ryan N Lichtenwalter, Jake T Lussier, and Nitesh V Chawla. New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 243–252, 2010.

[LNRT19] Jennifer M Larson, Jonathan Nagler, Jonathan Ronen, and Joshua A Tucker. Social networks and protest participation: Evidence from 130 million twitter users. *American Journal of Political Science*, 63(3):690–705, 2019.

[LRT22] Nadia Loy, Matteo Raviola, and Andrea Tosin. Opinion polarization in social networks. *Philosophical Transactions of the Royal Society A*, 380(2224):20210158, 2022.

[Luc15] Andrés Lucero. Using affinity diagrams to evaluate interactive prototypes. In *INTERACT (2)*, volume 9297 of *Lecture Notes in Computer Science*, pages 231–248. Springer, 2015.

[LVhLH20] Emma Llansó, Joris VaN hoboKeN, Paddy Leerssen, and Jaron Harambam. Content moderation, and freedom of expression. *Algorithms*, 2020.

[Lyn11] Marc Lynch. After egypt: The limits and promise of online challenges to the authoritarian arab state. *Perspectives on Politics*, 9(2):301–310, 2011.

[MA03] Marta Mccaughey and Michael Ayers. *Cyberactivism: Online activism in theory and practice*. Routledge, London, UK, 2003.

[Mac16] Charles M. Macal. Everything you need to know about agent-based modelling and simulation. *Journal of Simulation*, 10:144–156, 2016.

[MAP+20] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. Feedback loop and bias amplification in recommender systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 2145–2148, New York, NY, USA, 2020. Association for Computing Machinery.

[Mas18] Lilliana Mason. *Uncivil agreement: How politics became our identity*. University of Chicago Press, 2018.

[Mat17] Antonis Matakos. Measuring and moderating opinion polarization in online social networks. 2017.

[MBBRD24] Pau Muñoz, Alejandro Bellogín, Raúl Barba-Rojas, and Fernando Díez. Quantifying polarization in online political discourse. *EPJ Data Science*, 13(1):39, 2024.

[MBG+13] Manuel Mazzara, Luca Biselli, Pier Paolo Greco, Nicola Dragoni, Antonio Marraffa, Nafees Qamar, and Simona de Nicola. *Social networks and collective intelligence: a return to the agora*, pages 88–113. IGI Global, 2013.

[MBLB15] A. J. Morales, J. Borondo, J. C. Losada, and R. M. Benito. Measuring political polarization: Twitter shows the two sides of venezuela. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(3), mar 2015.

[MBP18] Jacob K. Madsen, Rachael M. Bailey, and Terri D. Pilditch. Large networks of rational agents form persistent echo chambers. *Scientific Reports*, 8(1):12391, 2018.

[MBP23] Danielle R. Mehlman-Brightwell and Mark J. Piwinsky. Information warfare fostering political polarization: Facebook addiction, news credibility, and concern of foreign interference. In *Social Media Politics*, pages 175–193. Routledge, 2023.

[MC20] N.S. Morais and M. Cruz. Gender perception about fake news and disinformation: Case study with portuguese higher education students. In *EDULEARN20 Proceedings*, 12th International Conference on Education and New Learning Technologies, pages 7746–7753. IATED, 6-7 July, 2020 2020.

[MC21] G. D. F. Morales and J. P. Cointet. Auditing the effect of social network recommendations on polarization in geometrical ideological spaces. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 1447–1455, 2021.

[McC19] Nolan McCarty. *Polarization: What Everyone Needs to Know?* Oxford University Press, 2019.

[MCCD13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[MDB23] Pau Muñoz, Fernando Díez, and Alejandro Bellogín. El derecho a la información en las redes sociales. In *Agenda 2030: teoría y práctica: una mirada constructiva desde la academia*, pages 213–234. Los Libros de la Catarata, 2023.

[MDB24] Pau Muñoz, Fernando Díez, and Alejandro Bellogín. Modeling disinformation networks on twitter: structure, behavior, and impact. *Applied Network Science*, 9(1):4, 2024.

[MH14] Soo Jung Moon and Patrick Hadley. Routinizing a new technology in the newsroom: Twitter as a news source in mainstream media. *Journal of Broadcasting & Electronic Media*, 58(2):289–305, 2014.

[MH24] Andrea Musso and Dirk Helbing. How networks shape diversity for better or worse, 2024.

[ML17] A. Marwick and R. Lewis. Media manipulation and disinformation online. Technical report, Data Society Research Institute, 2017.

[ML22] Oksana Moroz and Anna Loza. Youcontrol, database of russian propagandists. `https://youcontrol.com.ua/en/articles/database-of-russian-propagandists/`, accessed July 2023, 2022.

[MNC22] Thomas Magelinski, Lynnette Ng, and Kathleen Carley. A synchronized action framework for detection of coordination on social media. *Journal of Online Trust and Safety*, 1(2), Feb. 2022.

[MPS+19] Prabhaker Mishra, Chandra M Pandey, Uttam Singh, Anshul Gupta, Chinmoy Sahu, and Amit Keshri. Descriptive statistics and normality tests for statistical data. *Annals of cardiac anaesthesia*, 22(1):67–72, 2019.

[MS72] M. E. McCombs and D. L. Shaw. The agenda-setting function of mass media. *Public Opinion Quarterly*, 36(2):176–187, 1972.

[MS21] A. Maulana and H. Situngkir. Political polarization in the media landscape: the case of indonesian elections. 2021.

[MSLC01] Miller McPherson, Lynn Smith-Lovin, and James M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[MSPH20] Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. Dancing to the partisan beat: A first analysis of political communication on tiktok. In *12th ACM Conference on Web Science*, page 257–266, New York, NY, USA, 2020. Association for Computing Machinery.

[MSZS14] S. Mohd Shariff, X. Zhang, and M. Sanderson. User perception of information credibility of news on twitter. In M. et al. de Rijke, editor, *Advances in Information Retrieval. ECIR 2014. Lecture Notes in Computer Science*, volume 8416, Cham, 2014. Springer.

[MTF20] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. Recommender systems and their ethical challenges. *AI & Soc*, 35:957–967, 2020.

[MV17] Ulises A. Mejias and Nikolai E. Vokuev. Disinformation and the media: the case of russia and ukraine. *Media, Culture & Society*, 39(7):1027–1042, 2017.

[MVB20] L. Molyneux, A. C. Vasconcelos, and L. Breen. Journalists as sensemakers: Sensemaking in the context of covid-19. *Journalism Studies*, 21(13):1733–1749, 2020.

[MW47] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[NCC22] Lynnette Hui Xian Ng, Iain J. Cruickshank, and Kathleen M. Carley. Cross-platform information spread during the january 6th capitol riots. *Soc. Netw. Anal. Min.*, 12(1):133, 2022.

[Nec12] Alina Nechita. Mass self-communication. *Journal of Media Research*, 5(3):29, 2012.

[New06] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, jun 2006.

[New10] Mark E. J. Newman. *Networks: An Introduction*. Oxford University Press, Oxford, UK, 2010.

[New18] Mark Newman. *Networks*. Oxford University Press, 2018.

[NFN+20] R. K. Nielsen, R. Fletcher, N. Newman, J. S. Brennen, and P. N. Howard. Navigating the 'infodemic': How people in six countries access and rate news and information about coronavirus. *Reuters Institute for the Study of Journalism*, 2020.

[NHKE17] Vidya Narayanan, Philip N. Howard, Bence Kollanyi, and Mona Elswah. Russian involvement and junk news during brexit. Technical report, The Computational Propaganda Project, Oxford Internet Institute, 2017.

[NKLL22] Marcello Nesca, Alan Katz, Carson K Leung, and Lisa M Lix. A scoping review of preprocessing methods for unstructured text data to assess data quality. *International Journal of Population Data Science*, 7(1), 2022.

[NLJS20] Emil Noordeh, Roman Levin, Ruochen Jiang, and Harris Shadmany. Echo chambers in collaborative filtering based recommendation systems. *arXiv preprint arXiv:2011.03890*, 2020.

[NO22] Muhammad Naeem and Wilson Ozuem. Understanding misinformation and rumors that generated panic buying as a social practice during covid-19 pandemic: evidence from twitter, youtube and focus group interviews. *Information Technology and People*, 35:2140–2166, 12 2022.

[OA23] Antonia Olmos-Alcaraz. Islamophobia and twitter: The political discourse of the extreme right in spain and its impact on the public. *Religions*, 14(4):506, 2023.

[OCLB19] G. Olivares, J. P. Cárdenas, J. C. Losada, and J. Borondo. Opinion polarization during a dichotomous electoral process. *Complexity*, 2019, 2019.

[oFND18] High Level Expert Group on Fake News and Online Disinformation. A multi-dimensional approach to disinformation: Report of the independent high level group on fake news and online disinformation. *European Commission*, 2018.

[OJD22] OJD. Principales medios de comunicación en españa. https://www.ojdinteractiva.es/medios-digitales, accessed July 2023, 2022.

[ORK18] Claudia Orellana-Rodriguez and Mark T. Keane. Attention to news and its dissemination on twitter: A survey. *Computer Science Review*, 29:74–94, 2018.

[PAF98] Bruce E. Pinkleton, Erica Weintraub Austin, and Kristine K. J. Fortman. Relationships of media use and political disaffection to political efficacy and voting behavior. *Journal of Broadcasting & Electronic Media*, 42(1):34–49, 1998.

[Par11] Eli Pariser. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin, 2011.

[Pay19] Pablo Ribera Payá. Measuring populism in spain: content and discourse analysis of spanish political parties. *Journal of Contemporary European Studies*, 27(1):28–60, 2019.

[PBJB15] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of liwc2015. 2015.

[PCI+07] James Pennebaker, Cindy Chung, Molly Ireland, Amy Gonzales, and Roger Booth. The development and psychometric properties of liwc2007. 01 2007.

[PD17] Fabio Persia and Daniela D'Auria. A survey of online social networks: Challenges and opportunities. In *2017 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 614–620, 2017.

[Pea01] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

[PELTF23] Marta Pérez-Escolar, Darren Lilleker, and Alejandro Tapia-Frade. A systematic literature review of the phenomenon of disinformation and misinformation. *Media and Communication*, 11(2):76–87, 2023.

[PEOOAP21] Marta Pérez-Escolar, Eva Ordóñez-Olmedo, and Purificación Alcaide-Pulido. Fact-checking skills and project-based learning about infodemic and disinformation. *Thinking Skills and Creativity*, 41:100887, 2021.

[Pet09] Leif E. Peterson. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009.

[PGC20] Vidushi Pandey, Sumeet Gupta, and Manojit Chattopadhyay. A framework for understanding citizens' political participation in social media. *Information Technology and People*, 33:1053–1075, 6 2020.

[PGL21] Juan Manuel Pérez, Juan Carlos Giudici, and Franco Luque. pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks, 2021.

[Pic08] Victor W. Pickard. Cooptation and cooperation: institutional exemplars of democratic internet technology. *New Media & Society*, 10(4):625–645, 2008.

[PK03] Francesc Pallarés and Michael Keating. Multi-level electoral competition: Regional elections and party systems in spain. *European Urban and Regional Studies*, 10(3):239–255, 2003.

[PLCA14] Ismael Peña-López, Mariluz Congosto, and Pablo Aragón. Spanish indignados and the evolution of the 15m movement on twitter: towards networked para-institutions. *Journal of Spanish Cultural Studies*, 15(1-2):189–216, 2014.

[PLZ24] Yanni Ping, Yang Li, and Jiaxin Zhu. Beyond accuracy measures: the effect of diversity, novelty and serendipity in recommender systems on user engagement. *Electronic Commerce Research*, pages 1–28, 2024.

[PMS23] Jürgen Pfeffer, Daniel Matter, and Anahit Sargsyan. The half-life of a tweet. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1163–1167, 2023.

[PMSGA21] María Antonia Paz, Ana Mayagoitia-Soria, and Juan-Manuel González-Aguilar. From polarization to hate: Portrait of the spanish political meme. *Social media+ society*, 7(4):20563051211062920, 2021.

[PS24] Royal Pathak and Francesca Spezzano. An empirical analysis of intervention strategies' effectiveness for countering misinformation amplification by recommendation algorithms. In *European Conference on Information Retrieval*, pages 285–301. Springer, 2024.

[PSP23] Royal Pathak, Francesca Spezzano, and Maria Soledad Pera. Understanding the contribution of recommendation algorithms on misinformation recommendation and misinformation dissemination on social networks. *ACM Transactions on the Web*, 17(4):1–26, 2023.

[PvD21] Miroslava Pavlíková, Barbora Šenkýřová, and Jakub Drmola. Propaganda and disinformation go online. In *Challenging online propaganda and disinformation in the 21st century*, pages 43–74. 2021.

[Qua17] Walter Quattrociocchi. Inside the echo chamber. *Scientific American*, 316:60–63, 04 2017.

[RAC+02] Al Mamunur Rashid, Istvan Albert, Dan Cosley, Shyong K Lam, Sean M McNee, Joseph A Konstan, and John Riedl. Getting to know you: learning new user preferences in recommender systems. In *Proceedings of the 7th international conference on Intelligent user interfaces*, pages 127–134, 2002.

[RARFN22] Amy Ross Arguedas, Craig Robertson, Richard Fletcher, and Rasmus Nielsen. Echo chambers, filter bubbles, and polarisation: A literature review. 2022.

[Rau96] Richard Raucci. Gophers, web encyclopedias, and search engines. In *Netscape™ for Macintosh®: A hands-on configuration and set-up guide for popular Web browsers*, pages 89–104. 1996.

[RCM+11] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Detecting and tracking political abuse in social media. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 297–304, 2011.

[RD22] Deepjyoti Roy and Mala Dutta. A systematic review and research perspective on recommender systems. *Journal of Big Data*, 9(1):59, 2022.

[RGK12] Justine Rochon, Matthias Gondan, and Meinhard Kieser. To test or not to test: Preliminary assessment of normality when comparing two independent samples. *BMC medical research methodology*, 12:1–11, 2012.

[RH21] Thomas J. Rudolph and Marc J. Hetherington. Affective polarization in political and nonpolitical settings. *International Journal of Public Opinion Research*, 33(3):591–606, 2021.

[Rhe02] Howard Rheingold. *Smart Mobs: The Next Social Revolution*. Basic books, New York, USA, 1 edition, 2002.

[Rig97] Stephen Harold Ed. Riggins. *The language and politics of exclusion: Others in discourse*. Sage Publications, Inc, 1997.

[RKR08] Al Mamunur Rashid, George Karypis, and John Riedl. Learning preferences of new users in recommender systems: an information theoretic approach. *Acm Sigkdd Explorations Newsletter*, 10(2):90–100, 2008.

[RRMPT13] Rodolfo Rivas-Ruiz, Jorge Moreno-Palacios, and Juan O Talaveraa. Clinical research xvi. differences between medians with mann-whitney u test. *Revista medica del Instituto Mexicano del Seguro Social*, 51(4):414–419, 2013.

[RRS11] Francesco Ricci, Lior Rokach, and Bracha Shapira. Introduction to recommender systems handbook. In Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, editors, *Recommender Systems Handbook*, pages 1–35. Springer, Boston, MA, 2011.

[RRS22] Francesco Ricci, Lior Rokach, and Bracha Shapira, editors. *Recommender Systems Handbook*. Springer US, New York, NY, 2022.

[RSAP19] Ana Ruiz-Sánchez and Manuel Alcántara-Plá. Us vs. them: Polarization and populist discourses in the online electoral campaign in spain. In *Populist Discourse*, pages 103–119. Routledge, 2019.

[RSM22] Isabel Rodríguez, Diego Santamaría, and Luis Miller. Electoral competition and partisan affective polarisation in spain. *South European Society and Politics*, 27(1):27–50, 2022.

[RT14] Lavanya Rajendran and Preethi Thesinghraja. The impact of new media on traditional media. *Middle-East Journal of Scientific Research*, 22(4):609–616, 2014.

[Ruo21] Jukka Ruohonen. A few observations about state-centric online propaganda. *CoRR*, abs/2104.04389, 2021.

[SADH+22] David Smailes, Ben Alderson-Day, Catherine Hazell, Adrian Wright, and Peter Moseley. Measurement practices in hallucinations research. *Cognitive Neuropsychiatry*, 27(2-3):183–198, Mar-May 2022. Epub 2021 Nov 8.

[Sal22] Dorsaf Sallami. Personalized fake news aware recommendation system. 2022.

[Sán15] Damien Sánchez. Digital justice: An exploratory study of digital activism actions on twitter. *Journal of Educational Technology Development and Exchange (JETDE)*, 8(2):1, 2015.

[Sav22] Laura Savolainen. The shadow banning controversy: perceived governance and algorithmic folklore. *Media, Culture & Society*, 44(6):1091–1109, 2022.

[SBDMP16] Karen Sanders, Rosa Berganza, and Roberto De Miguel Pascual. *Spain: Populism From the Far Right to the Emergence of Podemos*, pages 249–260. 01 2016.

[SBLS20] Sebastian Stier, Arnim Bleier, Haiko Lietz, and Markus Strohmaier. *Election campaigning on social media: Politicians, audiences, and the mediation of political communication on Facebook and Twitter*, pages 50–74. Routledge, 2020.

[SC18] Javier Sanz-Cruzado and Pablo Castells. Enhancing structural diversity in social networks by recommending weak ties. In Sole Pera, Michael D. Ekstrand, Xavier Amatriain, and John O'Donovan, editors, *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, pages 233–241, New York, NY, 2018. ACM.

[SCC22] Javier Sanz-Cruzado and Pablo Castells. Relison: A framework for link recommendation in social networks. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022.

[SCCL19] Javier Sanz-Cruzado, Pablo Castells, and Esther López. A simple multi-armed nearest-neighbor bandit for interactive recommendation. In *Proceedings of the 13th ACM conference on recommender systems*, 2019.

[Sch15] Daniel Schall. *Social network-based recommender systems*. Springer, 2015.

[Sch20] Jen Schradie. "Give me Liberty or Give me Covid-19": Anti-lockdown protesters were never Trump puppets. *Communication and the Public*, 5(3-4):126–128, 2020.

[SCN16] Mr Sridhar Dilip Sondur, Mr Amit P Chigadani, and Shantharam Nayak. Similarity measures for recommender systems: a comparative study. *Journal for Research*, 2(3), 2016.

[SCPC18] Javier Sanz-Cruzado, Sofía M. Pepa, and Pablo Castells. Structural novelty and diversity in link prediction. In *Companion Proceedings of the The Web Conference 2018*, 2018.

[SCV+18] C. Shao, G. L. Ciampaglia, O. Varol, K. C. Yang, A. Flammini, and F. Menczer. The spread of low-credibility content by social bots. *Nature Communications*, 9(1):1–10, 2018.

[SEN18] Michael Smith and Denise Edwards-Neff. *Organizing for Advocacy*, pages 439–451. 04 2018.

[Ser24] Uroš Sergaš. Perspective diversification news recommender system. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, pages 45–49, 2024.

[SFK11] Tom Seymour, Dean Frantsvog, and Satheesh Kumar. History of search engines. *International Journal of Management & Information Systems (IJMIS)*, 15(4):47–58, 2011.

[Sha48] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[Sha93] Claude E Shannon. *Claude elwood shannon: Collected papers*. IEEE press, 1993.

[SHW+18] Chengcheng Shao, Pik-Mai Hui, Lei Wang, Xinwen Jiang, Alessandro Flammini, Filippo Menczer, and Giovanni Luca Ciampaglia. Anatomy of an online misinformation network. *Plos one*, 13(4):e0196087, 2018.

[SIK22] Maneet Singh, S. R. S. Iyengar, and Rishemjit Kaur. A multi-opinion based method for quantifying polarization on social networks, 2022.

[SJ16] Lisen Selander and Sirkka Jarvenpaa. Digital action repertories and transforming a social movement organization. *MIS quarterly*, 40(2):331–352, 2016.

[SJGL24] Beth M. Stokes, Samuel E. Jackson, Philip Garnett, and Jingxi Luo. Extremism, segregation and oscillatory states emerge through collective opinion dynamics in a novel agent-based model. *The Journal of Mathematical Sociology*, 48(1):42–80, 2024.

[SK09] Xiaoyuan Su and Taghi M Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in artificial intelligence*, 2009(1):421425, 2009.

[SKG20] Frank Schweitzer, Tamas Krivachy, and David Garcia. An agent-based model of opinion polarization driven by emotions. *Complexity*, 2020:1–11, 2020.

[SLL21] F. A. M. Santos, Y. Lelkes, and S. A. Levin. Link recommendation algorithms and dynamics of polarization in online social networks. *Proceedings of the National Academy of Sciences*, 118(50), 2021.

[SLR10] Michael A. Stefanone, Derek Lackaff, and Devan Rosen. The relationship between traditional mass media and "social media": Reality television as a model for social network site behavior. *Journal of Broadcasting & Electronic Media*, 54(3):508–525, 2010.

[SM21] Rachna Sethi and Monica Mehrotra. Cold start in recommender systems—a survey from domain perspective. In *Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI 2020*, pages 223–232. Springer, 2021.

[SMW+18] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*, 2018.

[SNB+21] Alexandra A. Siegel, Evgenii Nikitin, Pablo Barberá, Joanna Sterling, Bethany Pullen, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker. Trumping hate on twitter? online hate speech in the 2016 us election campaign and its aftermath. *Quarterly Journal of Political Science*, 16(1):71–104, 2021.

[SR18] Jorge Sola and César Rendueles. Podemos, the upheaval of spanish politics and the challenge of populism. *Journal of Contemporary European Studies*, 26(1):99–116, 2018.

[SSAW19] K. Starbird, E. Spiro, A. Arif, and T. Wilson. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–27, 2019.

[SSBW23] Naman Saxena, Adwitiya Sinha, Tanishk Bansal, and Ankita Wadhwa. A statistical approach for reducing misinformation propagation on twitter social media. *Inf. Process. Manag.*, 60(4):103360, 2023.

[SSJ+22] Jianshan Sun, Jian Song, Yuanchun Jiang, Yezheng Liu, and Jun Li. Prick the filter bubble: A novel cross domain recommendation model with adaptive diversity regularization. *Electronic Markets*, 32(1):101–121, 2022.

[SSRK22] S. Shreyashree, P. Sunagar, S. Rajarajeswari, and A. Kanavalli. A literature review on bidirectional encoder representations from transformers. In *Inventive Computation and Information Technologies: Proceedings of ICICIT 2021*, pages 305–320. Springer, 2022.

[SSRM22] Tobin South, Bridget Smart, Matthew Roughan, and Lewis Mitchell. Information flow estimation: A study of news on twitter. *Online Social Networks and Media*, 31:100231, 2022.

[Sta22] Statista. Principales medios en españa. `https://es.statista.com/estadisticas/476795/periodicos-diarios-mas-leidos-en-espana/`, accessed July 2023, 2022.

[Ste19] R. Stengel. *Information Wars: How We Lost the Global Battle Against Disinformation and What We Can Do About It*. Grove Press, New York, NY, 2019.

[Str22] Jonathan Stray. Designing recommender systems to depolarize. *First Monday*, 2022.

[Sub21] I. Subekti. Analysis twitter's as tools a political campaigns for new party during the 2020 regional head election in indonesia. 2021.

[Sun17] C. R. Sunstein. *#Republic: Divided democracy in the age of social media*. Princeton University Press, 2017.

[SV16] Jeremy Singer-Vine. Buzzfeednews: A dataset of fact-checked articles from buzzfeed news. `https://github.com/BuzzFeedNews/2016-10-facebook-fact-check`, accessed July 2023, 2016.

[SWB+22] B. Smart, J. Watt, S. Benedetti, L. Mitchell, and M. Roughan. #istandwithputin versus #istandwithukraine: The interaction of bots and humans in discussion of the russia/ukraine war. In F. Hopfgartner, K. Jaidka, P. Mayr, J. Jose, and J. Breitsohl, editors, *Social Informatics. SocInfo 2022. Lecture Notes in Computer Science*, volume 13618, Cham, 2022. Springer.

[T+18] Joshua A. Tucker et al. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*, 2018.

[Tal21] Robert B. Talisse. Problems of polarization. *Political epistemology*, 209:904, 2021.

[TALB21] Petter Törnberg, Claes Andersson, Kristian Lindgren, and Sven Banisch. Modeling the emergence of affective polarization in the social media society. *Plos one*, 16(10):e0258259, 2021.

[Tar97] Sidney Tarrow. *The Social Movement Society*. Rowman & Littlefield Publishers, Lanham, Maryland, Estados Unidos, 1 edition, 1997.

[Tas22] EastStratCom TaskForce. Euvsdisinfo, disinfo database. `https://euvsdisinfo.eu/disinformation-cases/`, accessed July 2023, 2022.

[TBB21] Ludovic Terren and Rosa Borge-Bravo. Echo chambers on social media: A systematic review of the literature. *Review of Communication Research*, 9, 2021.

[TCC20] P. Törnberg, U. Carlsson, and C. Clerwall. Disinformation and the european parliament election 2019: A case study of the european union stratcom task force. *Media and Communication*, 8(2):411–421, 2020.

[TGP20] Antonela Tommasel, Daniela Godoy, and Patricia Pesado. Do recommender systems make social media more susceptible to misinformation spreaders? In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, 2020.

[TGZ21] Antonela Tommasel, Daniela Godoy, and Arkaitz Zubiaga. Ohars: second workshop on online misinformation-and harm-aware recommender systems. In *Proceedings of the 15th ACM Conference on Recommender Systems*, pages 789–791, 2021.

[Tin17] Nava Tintarev. Presenting diversity aware recommendations. In *FATREC Workshop on Responsible Recommendation Proceedings*, 2017.

[TJLL18] E. C. Tandoc Jr., Z. W. Lim, and R. Ling. Defining ”fake news”: A typology of scholarly definitions. *Digital Journalism*, 6(2):137–153, 2018.

[TM22] Antonela Tommasel and Filippo Menczer. Do recommender systems make social media more susceptible to misinformation spreaders? *Poster presentation*, 2022.

[Tuf17] Zeynep Tufekci. *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press, 2017.

[Vam20] Davide Vampa. Competing forms of populism and territorial politics: the cases of vox and podemos in spain. *Journal of Contemporary European Studies*, 28(3):304–321, 2020.

[VBP21] M. E. D. Valle, M. Broersma, and A. Ponsioen. Political interaction beyond party lines: communication ties and party polarization in parliamentary twitter networks. *Social Science Computer Review*, 40(3):736–755, 2021.

[VdBGH22] Lawrence Van den Bogaert, David Geerts, and Jaron Harambam. Putting a human face on the algorithm: Co-designing recommender personae to democratize news recommender systems. *Digital Journalism*, pages 1–21, 2022.

[VdBH20] H. Van den Bulck and A. Hyzen. Of lizards and ideological entrepreneurs: Alex Jones and Infowars in the relationship between populist nationalism and the post-global media ecology. *International Communication Gazette*, 82(1, SI):42–59, 2020.

[vDD21] Loes van Driel and Delia Dumitrica. Selling brands while staying "authentic": The professionalization of instagram influencers. *Convergence*, 27(1):66–84, 2021.

[Ver15] Maurice Vergeer. Twitter and political campaigning. *Sociology compass*, 9(9):745–760, 2015.

[VGPG22] Antoine Vendeville, Anastasios Giovanidis, Effrosyni Papanastasiou, and Benjamin Guedj. Opening up echo chambers via optimal content recommendation. In *International Conference on Complex Networks and Their Applications*, pages 74–85. Springer, 2022.

[VHS13] Maurice Vergeer, Liesbeth Hermans, and Steven Sams. Online social networks and microblogging in political campaigning: The exploration of a new campaign tool and a new campaign style. *Party politics*, 19(3):477–501, 2013.

[VIHP14] Aaron S Veenstra, Narayanan Iyer, Mohammad Delwar Hossain, and Jiwoo Park. Time, place, technology: Twitter as an information source in the wisconsin labor protests. *Computers in Human Behavior*, 31:65–72, 2014.

[VPV21] Giacomo Villa, Gabriella Pasi, and Marco Viviani. Echo chamber detection and analysis: A topology- and content-based approach in the covid-19 scenario. *Social Network Analysis and Mining*, 11, 12 2021.

[VRA18] S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[Wag21] Markus Wagner. Affective polarization in multiparty systems. *Electoral Studies*, 69:102199, 2021.

[Wan17] W. Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. 2017.

[WD17] Claire Wardle and Hossein Derakhshan. Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe*, 2017.

[Wel16] Karen Wells. The strength of weak ties: The social networks of young separated asylum seekers and refugees in london. In *Diverse Spaces of Childhood and Youth*, pages 43–53. Routledge, 2016.

[WH18] Samuel C. Woolley and Philip N. Howard, editors. *Computational propaganda: Political parties, politicians, and political manipulation on social media*. Oxford University Press, 2018.

[WHD20] Panpan Wang, Qian Huang, and Robert M. Davison. How do digital influencers affect social commerce intention? the roles of social power and satisfaction. *Information Technology and People*, 34:1065–1086, 2020.

[WLRV21] Joe Whittaker, Sean Looney, Alastair Reed, and Francesco Votta. Recommender systems and the amplification of extremist content. *Internet Policy Review*, 10(2), 2021.

[WMZK18] Michael Wojatzki, Saif M. Mohammad, Torsten Zesch, and Svetlana Kiritchenko. Quantifying qualitative data for understanding controversial issues. In *International Conference on Language Resources and Evaluation*, 2018.

[WN20] Derek Weber and Frank Neumann. Who's in the gang? revealing coordinating communities in social media. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 89–93. IEEE, 2020.

[WP13] M. Wojcieszak and V. Price. The impact of candidate communication strategies on citizens' attitudes and behavior: A social identity framework. *Political Psychology*, 34(3):337–361, 2013.

[WTBC19] Christopher R. Walker, Sara-Jayne Terp, Pablo C. Breuer, and Courtney L. Crooks. Misinfosec. In Sihem Amer-Yahia, Mohammad Mahdian, Ashish Goel, Geert-Jan Houben, Kristina Lerman, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia, editors, *Companion of The 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 1026–1032. ACM, 2019.

[WXZ+22] Shoujin Wang, Xiaofei Xu, Xiuzhen Zhang, Yan Wang, and Wenzhuo Song. Veracity-aware and event-driven personalized news recommendation for fake news mitigation. In *Proceedings of the ACM web conference 2022*, pages 3673–3684, 2022.

[Xin12] Li Xinwu. A new text clustering algorithm based on improved k-means. *Journal of Software*, 7(1):95–101, 2012.

[XNT14] Feng Xiao, Tomoya Noro, and Takehiro Tokuda. Finding news-topic oriented influential twitter users based on topic related hashtag community detection. *Journal of Web Engineering*, pages 405–429, 2014.

[XWQ+22] Yunfei Xing, Xiwei Wang, Chengcheng Qiu, Yueqi Li, and Wu He. Research on opinion polarization by big data analytics capabilities in online social networks. *Technology in Society*, 68:101902, 2022.

[XZW23] Yi Xu, Deru Zhou, and Wei Wang. Being my own gatekeeper, how I tell the fake and the real - fake news perception between typologies and sources. *Inf. Process. Manag.*, 60(2):103228, 2023.

[Yan16] Guobin Yang. Narrative agency in hashtag activism: The case of #blacklivesmatter. *Media and Communication*, 4(4):13–17, 2016.

[YB22] Emre Yalcin and Alper Bilge. Evaluating unfairness of popularity bias in recommender systems: A comprehensive user-centric analysis. *Information Processing & Management*, 59(6):103100, 2022.

[YK01] Sung-Joon Yoon and Joo-Ho Kim. Is the internet more effective than traditional media? factors affecting the choice of media. *Journal of advertising research*, 41(6):53–60, 2001.

[YNHF22] Qi Yang, Sergey Nikolenko, Alfred Huang, and Aleksandr Farseev. Personality-driven social multimedia content recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7290–7299, 2022.

[YRW+16] JungHwan Yang, Hernando Rojas, Magdalena Wojcieszak, Toril Aalberg, Sharon Coen, James Curran, Kaori Hayashi, Shanto Iyengar, Paul K. Jones, Gianpietro Mazzoleni, Stylianos Papathanassopoulos, June Woong Rhee, David Rowe, Stuart Soroka, and Rodney Tiffen. Why are "others" so polarized? perceived political polarization and media use in 10 countries. *Journal of Computer-Mediated Communication*, 21(5):349–367, 2016.

[YWC+20] Wenwen Ye, Shuaiqiang Wang, Xu Chen, Xuepeng Wang, Zheng Qin, and Dawei Yin. Time matters: Sequential recommendation with complex temporal information. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 1459–1468, 2020.

[YWLD17] Muheng Yang, Xidao Wen, Yu-Ru Lin, and Lingjia Deng. Quantifying content polarization on twitter. In *2017 IEEE 3rd International Conference on Collaboration and Internet Computing (CIC)*, pages 299–308, 2017.

[ZAB+18] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, R. Procter, and P. Tolmie. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–36, 2018.

[ZB22] Eva Zangerle and Christine Bauer. Evaluating recommender systems: survey and framework. *ACM computing surveys*, 55(8):1–38, 2022.

[ZBZ21] Liwang Zhu, Qi Bao, and Zhongzhi Zhang. Minimizing polarization and disagreement in social networks via link recommendation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 2072–2084. Curran Associates, Inc., 2021.

[Zim17] Melissa Zimdars. Bigmclargehuge/opensources. `https://github.com/BigMcLargeHuge/opensources/blob/master/sources/sources.csv`, accessed July 2023, 2017.

[ZSBK19] Savvas Zannettou, Michael Sirivianos, Jeremy Blackburn, and Nicolas Kourtellis. The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *Journal of Data and Information Quality*, 11(3):1–27, 2019.

[ZXWY21] Cheng Zhou, Haoxin Xiu, Yuqiu Wang, and Xinyao Yu. Characterizing the dissemination of misinformation on social media in health emergencies: An empirical study based on COVID-19. *Inf. Process. Manag.*, 58(4):102554, 2021.

[ZYST19] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM Comput. Surv.*, 52(1):5:1–5:38, 2019.