POLITECNICO DI MILANO
DIPARTIMENTO DI ELETTRONICA, INFORMAZIONE E BIOINGEGNERIA
DOCTORAL PROGRAMME IN DATA ANALYTICS AND DECISION SCIENCES

CHARACTERIZATION AND DETECTION OF DISINFORMATION SPREADING IN ONLINE SOCIAL NETWORKS

Doctoral Dissertation of:
Francesco Pierri

Supervisor:
Prof. Stefano Ceri

Co-supervisor:
Prof. Fabio Pammolli

The Chair of the Doctoral Program:
Prof. Pierluca Lanzi

2021 - XXXIV Cycle

Acknowledgements

Getting to write this dissertation has been a long yet exciting journey, and there are many people who helped me along the way, going back to the beginning of my academic career.

First, I thank my mom for supporting me over the years, which includes and is not limited to playing, cooking, washing clothes, buying new clothes (especially those for graduation's days), reading, traveling, skiing, swimming, yelling, arguing, encouraging and – of course – pushing. I also thank my father for helping me cultivating both my rational and irrational sides (especially through arguing and pushing!) and my brother for sharing the burden of parental pressure and playing it down. I am grateful to my grand parents, aunts, uncles and cousins, who took care of my when I was little and still do when we meet these days, and to my great aunts and uncles – especially Zia Maria and Zio Antonio –, and all cousins in Torino who looked out for me during my university years.

I am also grateful to Betta, Mimmo, Alessandra, Giovanni, Antonella, Sandra, Nonna Ginelia and the rest of Parente and De Falco families.

I thank my advisor Stefano Ceri for guiding me over the last three years. I couldn't reach the end of this tunnel without his personal and professional advice. I am also extremely grateful to Carlo Piccardi for working with me over the last three years, and for providing useful guidance as well. I also thank my co-advisor Fabio Pammolli for giving me the opportunity to contribute, in my own little way, to the understanding of the on-going pandemic.

I am grateful to all the colleagues and students I met at Politecnico di Milano for the time spent inside (and especially outside) the office (listed in no specific ordering): Anna, Arif, Pietro, Giorgia, Marco Di Giovanni, Mattia, Marco Brambilla, Alessandro Artoni, Giada, Marco Varrone, Sara, Gaia, Eirini, Andrea Gulino, Luca, Silvio, Math-

yas, Lorenzo, Andrea Tocchetti, Andrea Flori, Giovanni, Francesco, Fabio Azzalini, Agostino, Michela, Simone Vantini, Mara Tanelli, Anna Paganoni, Piercesare Secchi.

I am very thankful to my Barzizza's friends for the ubiquitous diversions that provide me every day, and likewise to all my friends from Salerno, Torino and Milano who shared with me a lot over the years.

I am grateful to Capoeira Sul Da Bahia, in particular Mestrando Pedro, for all the days of hard training (and partying).

I am also very thankful to all the friends and colleagues I met recently in Bloomington (NaNners but not only).

Special thanks go to whoever I might be forgetting while I am writing these acknowledgements, and who helped me along the way.

Finally, I could not reach this achievement without Maddalena on my side inspiring me every day.

Abstract

In the last decade, online Social Networking Sites have become a fundamental part of our everyday life. Billions of individuals worldwide participate in such virtual communities, sharing and discussing messages, photos, videos, and other user-generated content. News consumption habits have also changed, and more and more individuals consume online news on social platforms such as Facebook and Twitter rather than traditional media such as newspapers and TV.

However, online social media also expose us and make us vulnerable to a variety of false and misleading information which erodes public trust towards institutions, with severe backlashes in the real world. One example is the ongoing COVID-19 pandemic, which has been accompanied by waves of potentially unreliable information which undermine medical intervention and governmental efforts to circumvent the spread of the disease.

In this work, we leverage a network and computer science approach to tackle the problem of disinformation – a term we use hereby as a shorthand to indicate all sorts of misleading, false and potentially harmful information – spreading in online social networks.

Focusing on Twitter and Facebook, we study the mechanisms and the actors involved in the spread of false information and other malicious content during relevant events such as political elections and the ongoing COVID-19 pandemic, when the need of reliable information for the public is higher.

We carry out a systematic comparison of reliable information, published by mainstream and traditional news websites, versus unreliable information conveyed by websites that have been repetitively flagged for sharing disinformation, misinformation, hoaxes, fake news and hyper-partisan propaganda.

We provide evidence of superspreaders of disinformation, i.e., influential users which are responsible for most of the disinformation shared online, and we unveil links with far-right communities, which oftentimes exploit fabricated information to push their agenda. At the same time, we show that reliable information accounts for the majority of news stories circulating online and that disinformation has a small yet non-negligible online prevalence which can still influence individuals' opinions and feelings.

We further investigate the interplay between vaccine-related disinformation shared on Twitter and the vaccine hesitancy and uptake rates measured across U.S. regions, following the roll-out of the COVID-19 vaccination program. Building a regression model which takes into account demographics, socio-economic and political factors, we find a significant association between online disinformation and vaccine outcomes.

Finally, drawing on the results of aforementioned analyses, we deploy a methodology to accurately classify news articles based on the interactions between users that naturally take place on Twitter. Following the intuition that users shape different diffusion patterns depending on the content they share, we train and test off-the-shelf machine learning classifiers that can classify the veracity of a news article, without the need of looking at its content.

All in all, our results contribute to a better understanding of the issue of disinformation spreading in online social media, and highlight the need for intervention by platforms and governments to address this issue in a timely fashion.

Sommario

NEGLI ultimi anni, i Social Network hanno acquisito un ruolo importante nella nostra vita quotidiana. Miliardi di persone si organizzano in comunità virtuali su scala globale, condividendo e discutendo messaggi, foto, video e altri tipi di contenuto. Anche la fruizione di notizie è cambiata, e sempre più individui si rivolgono a piattaforme come Facebook e Twitter per informarsi, abbandonando media tradizionali come giornali e televisione.

Tuttavia, i "social" ci espongono e ci rendono vulnerabili a una varietà di informazioni false e ingannevoli, e contribuiscono a erodere la fiducia nelle istituzioni con gravi conseguenze nel mondo reale. Un esempio è la pandemia di COVID-19 attualmente in corso, accompagnata da un'infodemia di notizie inattendibili che minano gli sforzi a contenere il virus.

In questo lavoro, utilizziamo un approccio a metà tra scienza delle reti e informatica per affrontare il problema della disinformazione – un termine che utilizziamo come cappello per indicare i vari tipi di informazione inattendibile, falsa e potenzialmente dannosa – che circola sui Social Network online.

Focalizzando la nostra attenzione su Twitter e Facebook, studiamo i meccanismi e gli attori coinvolti nella diffusione di disinformazione durante eventi rilevanti quali elezioni politiche e la pandemia in corso, quando la possibilità di avere accesso a informazioni attendibili è cruciale.

Effettuiamo una comparazione sistematica dell'informazione attendibile, prodotta da siti di notizie tradizionali e "mainstream", in contrapposizione all'informazione inattendibile prodotta da siti che sono stati ripetutamente richiamati per aver condiviso disinformazione, bufale, "fake news" e propaganda faziosa.

Forniamo prove della presenza di super diffusori di disinformazioni, i.e., utenti in-

fluenti che sono responsabili per la maggior parte dei contenuti di disinformazione condivisi online, e riveliamo collegamenti tra le comunità di estrema destra, che spesso fanno ricorso a notizie contraffatte per promuovere la loro ideologia. Allo stesso tempo, mostriamo che la maggioranza delle notizie che circolano online proviene da siti attendibili, e che la disinformazione ha una presenza limitata ma non trascurabile che può influenzare le opinioni e i sentimenti degli utenti online.

Investighiamo l'influenza della disinformazione relativa ai vaccini che si diffonde su Twitter sulla campagna di vaccinazione degli Stati Uniti. Utilizzando un modello di regressione lineare multipla, che tiene conto di fattori demografici, socio-economici e politici, troviamo un'associazione significativa tra la disinformazione online e le vaccinazioni.

In conclusione, basandoci sui risultati delle analisi sovra-citate, costruiamo una metodologia per classificare accuratamente le notizie sulla base delle interazioni sociali tra utenti che avvengono su Twitter. Seguendo l'intuizione che gli utenti danno vita a diversi "pattern" di diffusione a seconda del contenuto condiviso, alleniamo e testiamo dei classificatori ad apprendimento automatico che possono verificare la veracità di un articolo senza aver bisogno di guardare al contenuto.

I nostri risultati contribuiscono ad una maggiore comprensione del problema della disinformazione che circola sulle piattaforme social, e sottolineano l'urgenza di interventi da parte di piattaforme e governi per contrastare il fenomeno.

Contents

List of Figures

CHAPTER 1

Introduction

In the last decade, online Social Networking Sites (SNS) have become a pervasive presence in our everyday life. Billions of people nowadays use platforms such as Twitter, Facebook and Instagram to share text messages, photos and videos with their friends, but also to consume and disseminate news articles from news media outlets and blogs [135].

Recently, these platforms have witnessed an explosive growth of malicious and deceptive information. The research community usually refers to it with a variety of terms, such as disinformation, misinformation and most often false (or "fake") news, hardly reaching agreement on a single definition [138, 190].

Following a major shift in news consumption habits towards online content, and a drop in trust towards traditional news outlets [167], the problem of false information circulating online has become crucial, especially during periods such as political elections [79, 81, 189] or epidemics [44, 270, 276], when the prevalence of unreliable online information has severe backlashes in the real world.

One example is the on-going COVID-19 pandemic, as the world experiences an "infodemic" [276], i.e., an overabundance of information including false and misleading content, which undermines medical efforts and governmental efforts to fight the disease [270]. For example, misinformation about masks contributed to low adoption rates and likely increased disease transmission [146]. As false vaccine narratives

spread [144], vaccine hesitancy will make it difficult to reach herd immunity and prevent future outbreaks.

Several reasons explain the rise of such malicious phenomenon. First, barriers to enter the online media industry have dropped considerably and (dis)information websites are nowadays created faster than ever, generating revenues through advertisement without the need to adhere to traditional journalistic standards (as there is no third-party verification or editorial judgment for online news) [12].

Second, human factors such as confirmation biases [166], algorithmic biases [78, 138] and naive realism [203] have exacerbated the so-called echo chamber effect, i.e. the formation of homogeneous communities where people share and discuss about their opinions in a strongly polarized way, insulated from different and contrary perspectives [58, 178, 239, 240].

Third, direct intervention that could be put in place by platform government bodies for banning deceptive information is not encouraged, as it may raise ethical concerns about censorship [138, 226].

Therefore, ever since 2016 U.S. Presidential elections and UK Brexit Referendum, the research community has witnessed an explosion of interest around the issue of disinformation, misinformation and other sorts of false information spreading in social media platforms [190].

The combat against online mis/disinformation is challenged by: the massive rates at which malicious items are produced, and the impossibility to verify them all [226]; the adversarial setting in which they are created, as sources of misleading content usually attempt to mimic traditional news outlets [138]; the lack of gold-standard datasets and the limitations imposed by social media platforms on the collection of relevant data [179, 190].

Most methods for "fake news" detection are carried out by using features extracted from the news articles and their social context (notably textual features, users' profile, etc); existing techniques are built on this content-based evidence, using traditional machine learning or more elaborate deep neural networks, but they are often applied to small, ad-hoc datasets which do not generalize to the real world [190].

Important studies, featuring large-scale analyses, have produced deeper knowledge about the phenomenon, showing that: false news spread faster and more broadly than the truth on social media [226, 254]; social bots play an important role as "superspreaders" in the core of diffusion networks [225]; echo chambers are primary drivers for the diffusion of true and false content [58].

1.1 Contribution of this thesis

In this thesis, I[1] leverage a hybrid computer science and network science approach to tackle the problem of disinformation, a term I will use as a shorthand to indicate all sorts of false, misleading and potentially harmful information spreading in online social networks [190].

Gathering data from multiple social platforms, we study the mechanisms and the actors involved in the spread of disinformation and other malicious content on social media during relevant events such as political elections and the on-going COVID-19 pandemic [102, 188, 189, 270].

We carry out a systematic comparison of reliable information, published by mainstream and traditional news websites, versus unreliable information conveyed by websites which have been repetitively flagged for sharing disinformation, misinformation, hoaxes, fake news and hyper-partisan propaganda [102, 188, 189, 270].

We provide evidence of superspreaders of disinformation, i.e., influential users who are responsible for most of the disinformation shared online [270], and we unveil links with far right communities, which oftentimes exploit fabricated information to push their agenda [189]. At the same time, we show that reliable information accounts for the majority of news stories circulating online, and that disinformation has a small yet non-negligible online prevalence which can still influence individuals' opinions and feelings [188, 189, 270].

Combining data from Twitter and Facebook, we investigate the interplay between vaccine-related disinformation shared on Twitter and the vaccine hesitancy and uptake rates measured among U.S. regions, following the roll-out of the COVID-19 vaccination program [62, 192]. Building a regression model which takes into account demographics, socio-economic and political factors, we find a significant association between online disinformation and the vaccine outcomes. Further preliminary analyses show similar results for the Italian context [195] .

Finally, drawing on the results of these quantitative analyses, we deploy a methodology to accurately classify news articles based on the interactions between users that take on platforms like Twitter [193, 194]. Based on the assumption that users share differently disinformation news rather than mainstream articles, thus shaping different diffusion patterns, we train and test off-the-shelf machine learning classifiers which are able to classify the veracity of a news article, without the need of looking at its content.

All in all, our results pave the way to a better understanding of the issue of disinformation spreading in online social media, and highlight the need for intervention by platforms and governments to address this issue in a timely fashion.

[1] I will use the first person mostly in the **Structure of the thesis** section and in **Epilogue** chapter to highlight my contributions to the research presented in the thesis; I will use the first plural person throughout remaining chapters, as I was lucky to work together with other colleagues, and under the supervision of my advisors.

1.2 Structure of the thesis

In chapter 2 **Background and Related Work**, we will provide the reader with some contextual background on the terminology, the research problems and the challenges present in the literature about online disinformation, and fundamental notions of network science and machine learning, in order to better understand the results provided in later sections. We will also describe related work on the detection and the characterization of false information spreading in online social networks. Finally, we present an overview of recent literature about the recent COVID-19 infodemic. The material presented in this section is mostly based on these two published articles:

- **Francesco Pierri**, Stefano Ceri. False News On Social Media: A Data-Driven Survey. ACM SIGMOD Record Vol. 48 issue 2 (2019)
- Yang, Kai-Cheng, **Francesco Pierri**, Pik-Mai Hui, David Axelrod, Christopher Torres-Lugo, John Bryden, and Filippo Menczer. The COVID-19 Infodemic: Twitter versus Facebook. Big Data & Society (2021).

The first paper is a review article in which I analyze the literature related to the detection, characterization and mitigation of false and misleading information spreading in online social networks; I wrote it under the supervision of my advisor Stefano Ceri during the first year of my Ph.D., and it was extremely useful to me to understand which research directions to explore.

In Chapter 3 **Investigating Italian disinformation spreading on Twitter**, we present results from our investigation of Italian language disinformation spreading on Twitter in the run-up to 2019 EU Parliament elections, highlighting the role of online far-right communities and conspiracy theory advocates. The material presented in this section in based on the following publication:

- **Francesco Pierri**, Alessandro Artoni, Stefano Ceri. Investigating Italian disinformation spreading on Twitter in the context of 2019 European elections. (2020) PLoS One

In this paper I basically extended the analysis and results provided in the thesis of a M.Sc. student (Alessandro Artoni), that I co-advised together with my supervisor Stefano Ceri.

In Chapter 4 **Understanding the COVID-19 infodemic on Twitter and Facebook**, we show results from a systematic comparison of English language disinformation related to COVID-19 and shared on Twitter and Facebook throughout 2020, providing evidence of so-called "superspreaders", i.e., influential users who account for most of the misleading and harmful content shared online. The material presented in this section in based on this paper:

- Yang, Kai-Cheng, **Francesco Pierri**, Pik-Mai Hui, David Axelrod, Christopher Torres-Lugo, John Bryden, and Filippo Menczer. The COVID-19 Infodemic: Twitter versus Facebook. Big Data & Society (2021).

This paper is the first output of my international collaboration with the Observatory on Social Media at the Indiana University (IU), started as a "virtual" collaboration during my second year; at the end of my third year I went physically to Bloomington, IN. I proposed to draw upon a previous contribution of the team [271], which consisted of a small scale analysis of COVID-19 infodemic on Twitter in the early months of the pandemic, and to carry out a systematic comparison of Twitter and Facebook throughout 2020; I saw the opportunity in a call for paper of the Big Data & Society journal focused the COVID-19 infodemic [101]. My contribution was both in the design of the study and the collection, processing and analysis of Facebook data, as well as writing and revising the manuscript. The leader of the project, however, was my colleague Kai-Cheng Yang.

In Chapter 5 **The impact of vaccine-related disinformation**, we provide results from an on-going investigation of the interplay between vaccine-related disinformation shared on social media and the vaccine hesitancy/uptake rates in U.S. and Italy. An interactive visualization of results is also currently available to the public in two online dashboards, which are also described in the chapter. An important result of our analyses is a significant association between vaccine-related disinformation shared by U.S. users and the vaccine uptake rates measured in different geographical regions. The material presented in this section in based on these contributions:

- Matthew R. DeVerna, **Francesco Pierri**, Bao Truong, John Bollenbacher, David Axelrod, Torres-Lugo, Filippo Menczer and John Bryden. CoVaxxy: A Collection of English-Language Twitter Posts About COVID-19 Vaccines. In Proceedings of the International AAAI Conference on Web and Social Media (ICWSM'21)

- **Francesco Pierri**, Perry, B., DeVerna, M. R., Yang, K. C., Flammini, A., Menczer, F., & Bryden, J. The impact of online misinformation on US COVID-19 vaccinations. arXiv preprint arXiv:2104.10635 *(under review)*

- **Francesco Pierri**, Tocchetti Andrea, Corti Lorenzo, Di Giovanni Marco, Pavanetto Silvio, Brambilla Marco, Ceri Stefano VaccinItaly: monitoring Italian conversations around vaccines on Twitter and Facebook CySoc Workshop Proceedings of the International AAAI Conference on Web and Social Media (ICWSM '2021)

The first two papers result from my collaboration with the Observatory on Social Media at IU. I am responsible of the CoVaxxy[2] project, which monitors COVID-19 vaccine

[2] https://osome.iu.edu/tools/covaxxy

related conversations on Twitter, together with my IU colleagues Matthew R. DeVerna and John Bryden; the first paper is a "manifesto" of the project whereas in the second I led a joint effort to address the research questions which drive the project, i.e., can we find a correlation between online misinformation and vaccine uptake rates? Similarly, the third paper is a "manifesto" of the Italian counterpart to CoVaxxy, which I currently lead under the supervision of prof. Marco Brambilla in the context of the H2020 Periscope[3]; the aim of VaccinItaly[4] is thus to investigate the interplay between vaccine-related misinformation and anti-vax views shared on Twitter in the Italian scenario. We are currently planning to extend it to a multi-country setting where we include multiple European countries (given the broader scope of the H2020).

In Chapter 6 **A network-based approach to detect online disinformation on Twitter**, we present our methodology to automatically classify disinformation versus mainstream news articles shared on Twitter. In particular, we provide results from two different approaches which employ, respectively, a single-layer and a multi-layer representation of Twitter diffusion networks. The material presented in this section in based on these two publications:

- **Francesco Pierri**, Carlo Piccardi, Stefano Ceri Topology comparison of Twitter diffusion networks effectively reveals misleading news. (2020) Scientific Reports
- **Francesco Pierri**, Carlo Piccardi, and Stefano Ceri. A multi-layer approach to disinformation detection in US and Italian news spreading on Twitter. EPJ Data Science

These two papers are the result of my collaboration with Carlo Piccardi and Stefano Ceri, which started during my first year as a Ph.D. student. In both cases I defined the research questions and designed the study and data collection, as well performing experiments and writing the manuscript, under the supervision of my two colleagues.

In Chapter **Epilogue** we will sum up the results presented in this thesis, and we will draw conclusions and future work. Finally, in the **Appendix** we provide some additional material which serves as supporting information to the contributions presented throughout the thesis.

[3] https://www.periscopeproject.eu/start
[4] http://genomic.elet.polimi.it/vaccinitaly/

CHAPTER 2

Background and Related work

This section serves as an introduction to the topic of disinformation spreading in online social media. The material present in this section has been directly taken and adapted from our own contributions [190, 270].

In Section 2.1, we provide common terminology, describe the social media platforms where disinformation is most widespread, overview psychological and social factors that are involved, and discuss some of the effects on the real world and some open challenges. We also provide some fundamentals of network science and machine learning necessary to grasp results presented in later sections.

In sections 2.2 to 2.5, we then review state-of-the-art approaches to the automatic detection of false and misleading information, as well as contributions which provide a deeper understanding of the mechanisms behind the spread of online disinformation. We also point to existing data sources which have been used by researchers to carry out their analyses. Finally, we describe recent literature which relates to the COVID-19 infodemic, i.e., an overabundance of misleading and potentially harmful content which has followed the outbreak of the SARS-CoV-2, and the general spread of health-related misinformation on social media.

2.1 Background

2.1.1 Terminology

In recent years the term *fake news* [138] has been widely used to indicate false and problematic information in a variety of flavors: disinformation, misinformation, hoaxes, propaganda, satire, rumors, click-bait and junk news. We provide here a list of the definitions encountered in the literature which is by no means exhaustive. While there is common agreement that these terms indicate deceptive information to a certain degree, we believe that an academic formulation on the meaning of *fake news* is still missing in the literature.

Some researchers define **false news** as news articles that are potentially or intentionally misleading for the readers, as they are verifiable and deliberately false [12]. They can represent fabricated information which mimics traditional news content in form but not in the intent or the organizational process [138]. It has been highlighted how the neologism **fake news** is usually employed with a stronger political connotation with respect to the more traditional *false news* [138, 254].

Misinformation is defined as information that is inaccurate or misleading [138, 225]. It can spread either intentionally or unintentionally [78] due to honest reporting mistakes or incorrect interpretations [109, 259].

Disinformation is false information that is spread deliberately to deceive people [138] or promote biased agenda [252, 259]. According to [134, 259] it can be distinguished by misinformation which is unintentionally false. Nowadays, The term *misinformation* is commonly employed to refer collectively to all kinds of deceptive information [225].

Similarly to disinformation, **hoaxes** are intentionally conceived to deceive readers, usually described as "humorous and mischievous" (as defined in The Oxford English Dictionary) [134].

Propaganda is defined as information that tries to influence the emotions, the opinions and the actions of target audiences by means of deceptive, selectively omitting and one-sided messages. The purpose can be political, ideological or religious [251, 252].

Satirical news are written with the primary purpose of entertaining or criticize the readers, but similarly to hoaxes they can be harmful when shared out of context [50, 212]. They are characterized by humor, irony and absurdity and they can mimic genuine news [213].

Click-bait is defined as low quality journalism which is intended to attract traffic and monetize via advertising revenue [252].

The term **junk news** is more generic and it aggregates several types of information, from propaganda to hyper-partisan or conspiratorial news and information. It usually

refers to the overall content that pertains to a publisher rather than a single article [266].

Finally, we might come across several different definitions for **rumor**. Briefly, a rumor can be defined as a claim which did not originate from news events and that has not been verified while it spreads from one person to another [12, 239]. As there exists a huge literature on the subject, we refer the interested reader to [281] for an extensive review.

2.1.2 Social media platforms as news outlets

The issue of false news appearing on news outlets is by no means a new phenomenon: in 1835 a series of articles published on the New York Sun, also popular as the Great Moon Hoax, was describing the discovery of life on the moon [12]. However, nowadays the world is experiencing much more elaborated hoaxes which range from finance to politics [138].

Indeed, social media platforms exhibit unique characteristics which have favored the proliferation of fabricated news with brand new proportion and impact. It has been recently shown how most of nowadays news consumption has shifted towards online social media, where it is more comfortable to ingest, share and further discuss news with friends or other readers [97, 231, 232].

As it is easier (and faster) than ever to produce content online and monetize through advertisement, barriers for entering the online media industry have dropped [12]. This has conveyed the dissemination of low quality news which reject traditional journalistic standards and lack of third-party filtering and fact-checking [12].

A decline of general trust and confidence in traditional mass media in conjunction with the aforementioned factors has been indicated as the primary driver for the explosive growth of fake news on social media [12, 138].

Two main motivations have been proposed as to explain the rise of disinformation websites: 1) a pecuniary one, where viral news articles draw significant advertising revenue and 2) a more ideological one, as providers of fake news usually aim to influence public opinion on particular topics [12]. Besides, the presence of malicious agents such as bots, cyborgs and trolls has been highlighted as another major cause to the spreading of misinformation [133, 225].

A few social networks have attracted most of the research focus from the beginning: **Twitter**, **Facebook** and **Sina Weibo** – a popular Chinese microblogging website which is a hybrid between Facebook and Twitter. This is mainly due to the public availability of data and the existence of proprietary application programming interfaces (API) which ease the burden of collecting data. Over time, the research community has considered also other popular platforms such as **Reddit** and **Youtube** [47], as well as unmoderated ones such as **Gab** and **Parler**. In particular, the last two are known for being an "alternative" to mainstream platfroms such as Twitter and Facebook, as

they perform very little content moderation, and eventually ended up becoming echo chambers for right wing extremists and conspiracy theorists [11].

2.1.3 Human factors

Aside from the technical aspects of social network platforms, the research community has leveraged a set of well known psychological, cognitive and social aspects which are considered key contributors to the proliferation of disinformation on social media.

It appears that humans have no natural expertise at distinguishing real from false news [134]. Two major factors to explain this are notably the **naive realism** and the **confirmation bias**.

The first one refers to the tendency of users to believe that their view is the only accurate one whereas those who disagree are biased or uninformed [203].

The second factor (also present in the literature as *selective exposure*), is described as the inclination to prefer (and receive) information which confirms existing view [166]. As a consequence, presenting factual information to correct false beliefs is usually unhelpful and may increase "misperception" [172].

Some studies also mention the importance of **social identity theory** [15] and **normative social influence** [14] which describe how users tend to perform actions which are socially safer, thus consuming and spreading information items that agree with the norms established in the community.

Users are also exposed to an abundance of information which exceeds their capacity to consume it (**finite attention**) resulting in **information overload** where the information dynamics are driven by an economy of attention [225, 261].

All these factors are related to a certain extent to the well-known **echo chamber** effect, which gives rise to the formation of homogeneous clusters where individuals are like-minded people that share and discuss similar ideas. These groups are usually characterized by extremely polarized opinions as they are insulated from opposite views and contrary perspectives [178, 239, 240]. Apropos, it has been shown that these close-knit communities are the primary driver of misinformation diffusion [58].

Another peculiar aspect of the social technologies involved is the well-known **algorithmic bias** which exacerbates the aforementioned phenomenon, for these platforms promote personalized content based on the preferences of users with the unique goal of maximize engagement [78, 138].

Finally, it is worth mentioning the role of demographics as well: recent analyses conclude that higher levels of education imply more analytical skills and, consequently, a more accurate perception of information [12, 185].

2.1.4 Effects on the real world

We can explain the explosive growth of attention on so-called **fake news** in light of a series of striking effects that the world has recently experienced. Politics indeed accounts for most of the attention on false news.

The 2016 US presidential elections have officially popularized the term *fake news* to the degree that it has been suggested that Donald Trump may not have been elected president were it not for the effects of false news (and the alleged interference of Russian trolls) [12]. Likewise, recent studies have shown that false news have also impacted on 2016 UK Brexit referendum [115] and the 2017 France presidential elections [79].

Over and above we may recall the finance stock crisis caused by a false tweet concerning president Obama [201], the shootout occurred in a restaurant as a consequence of the "Pizzagate" conspiracy theory and the diffused mistrust towards vaccines during Ebola and Zika epidemics [81, 157].

Recently, in the COVID-19 pandemic scenario, a number of studies have shown that widely shared misinformation includes false claims that vaccines genetically manipulate the population or contain microchips that interact with 5G networks [62]. Such exposure to online misinformation has been linked to increased health risks and vaccine hesitancy [88, 144]. Still, gaps remain in our understanding of how health-related misinformation is linked to unsafe health behaviour [258].

2.1.5 Challenges

We mention here a few challenges which characterize the fight against false information on social media, as highlighted by recent research on the subject.

Firstly, false news are deliberately created to deceive the readers and to mimic traditional news outlets, resulting in an adversarial scenario where it is very hard to distinguish true news articles from false ones [225].

Secondly, the rate and the volumes at which fake articles are produced overturn the possibility to fact-check and verify all items in a rigorous way, i.e. by sending articles to human expert for verification [225]. This also raises concern on developing tools for the early detection of fake news as to prevent them from spreading in the network [143]. Consequently, existing techniques only work in a supervised yet limited manner whereas a more powerful unsupervised fashion is sought to tackle the problem efficiently.

Finally, social media platforms impose limitations on the collection of public data and as of today the community has produced very limited datasets which do not always include all the possible levels of information available in social networks. We refer the interestead reader to [179] for a list of concrete examples of studies that misinformation researchers could conduct, if the community had better access to platforms' data and

processes.

2.1.6 Fundamentals of network science

Networks are a powerful tool to model real complex systems. In this thesis, we often employ networks to represent interactions between individuals on social media platforms such as Twitter and Facebook. In the following we provide the reader with some fundamental notions of network science which are necessary to understand results provided in later sections. We refer the reader to [22, 30, 274] for an exhaustive review of the topic.

Historically, networks have been studied in the domain of graph theory, a branch of discrete mathematics, and they were also popular in the domain of social sciences, whereas in the last few decades more attention has been devoted to the field of complex networks, inspired by findings on real networks such as the World Wide Web or power grids [30].

Basic definitions

We can represent a network as a *graph* $G = (N, L)$ which consists of a set $N = \{n_1, n_2, ..., n_N\}$ of *nodes* (or vertices) and a set $L = \{l_1, l_2, ..., l_K\}$ of *links* (or edges), which are pairs of vertices $l_1 = (i, j)$ indicating that they are connected.

We say that a network is *undirected* if the order of the nodes in a link is not important, whereas in a *directed* network $l_1 = (i, j)$ indicates a link from node i to node j. l_1 is said to be *incident* to nodes i and j, and nodes i and j are *adjacent*. A link (i, i) is called a *self-loop*. When multiple edges between the same pair of nodes are allowed, the graph is called a *multi-graph*.

We define a *walk* from node i to node j as an alternating sequence of adjacent nodes and edges that starts at i and ends in j. A *trail* is a walk in which no edge is repeated, whereas a *path* is a walk in which a node is visited only once. A *shortest path* is the walk of minimal length between two nodes. In an unweighted graph, the length of the walk is the number of edges in the sequence, whereas in a weighted graph, the length of a walk is the sum of the weights associated to edges in the walk. Finally, in a directed graph a walk follows the direction of edges to move from one node to another.

We say that a graph is *connected* if, for every distinct pair of nodes i and j, there exists a walk from i to j, otherwise it is said to be *disconnected*. We say that a directed graph is *weakly* connected if the above condition is satisfied when the direction of edges is not considered, whereas it is *strongly* connected when it is satisfied also when considering the direction of edges.

A graph can be completely described by its *adjacency* (or connectivity) matrix A, a $N \times N$ square matrix where each entry $a_{i,j} = 1$ if a link between i and j exists, and 0

otherwise. We can also associate a set of weights $W = \{w_1, w_2, ..., w_K\}$ to each link, and the network is said to be *weighted*. The weighted adjacency matrix of a graph is thus a $N \times N$ square matrix where each entry $w_{i,j}$ is the weight associated to each link if $a_{i,j} = 1$, and 0 otherwise.

Node centralities

In many applications, it is important to assess the influence of each node in a network. There are many definitions of node centralities in the literature which can be used to measure such influence.

The *degree* centrality (or simply degree) k_i of a node i is the total number of edges incident to the node:

$$k_i = \sum_{j \in N} a_{i,j}$$

in a directed graph we have two contributions: the number of outgoing links $k_i^{out} = \sum_j a_{i,j}$ and the number of ingoing links $k_i^{in} = \sum_j a_{j,i}$. We then call $k_i = k_i^{out} + k_i^{in}$ the total degree of node i.

In a weighted graph, we can define the *strength* centrality (or simply strength) s_i of a node i as the sum of the weights of the edges incident to the node:

$$s_i = \sum_{j \in N} w_{i,j}$$

similarly, in a directed graph we can decompose the strenght in two contributions: the out-strength $s_i^{out} = \sum_j w_{i,j}$ and in-strength $s_i^{in} = \sum_j w_{j,i}$. The total strength is $s_i = s_i^{out} + s_i^{in}$.

The *betweenness* centrality b_i of a node i is defined as:

$$b_i = \sum_{j,k \in N, j \neq k} \frac{n_{jk}(i)}{n_{jk}}$$

where n_{jk} is the total number of shortest paths connecting j and k, and $n_{jk}(i)$ is the number of shortest paths connecting j and k that pass through node i. It was traditionally employed to measure the importance of an individual in social networks. It can be easily generalized to the case of weighted networks by considering weights of edges when computing shortest paths.

The *clustering coefficient* c_i of a node (also known as local clustering coefficient) in an undirected network is defined as:

$$c_i = \frac{\sum_{j,m} a_{ij} a_{jm} a_{mi}}{k_i(k_i - 1)}$$

which is the fraction of *triangles*, triples of nodes connected with each other, in the

subgraph of neighbors of node i, i.e. a graph G' consisting of all nodes adjacent to node i. The clustering coefficient C of a graph is given by the average local clustering coefficient $C = \frac{1}{N}\sum_{i \in N} c_i$, and by definition $0 \leq C \leq 1$. A generalization of the clustering coefficient to the case of weighted networks can be found in [24]

The *coreness* of a node i is the highest value of k for which node i belongs to the k-core of a graph [26]. The k-core of a graph G is the maximal connected sub-graph of G in which all vertices have degree at least k. Given the k-core, recursively removing all nodes with degree k allows to extract the $(k+1)$-core; the main core is the non-empty graph with maximum value of k.

Considering the *eigenstructure* of the adjacency matrix A of a graph we can define the so-called *eigenvector* node centrality [215], which is the sum of the centrality values of the nodes to which it is connected. Thus, the eigenvector centrality is defined as the components of the leading eigenvector, i.e., the eigenvector associated to the largest eigenvalue of the adjacency matrix A:

$$x_i = \lambda_1^{-1} \sum_j a_{i,j} x_i$$

This measure is closely connected to spreading processes in networks, as they relate to the spectra of the adjacency matrix.

The Google *PageRank* centrality of a node [176], which was defined to measure the importance of web pages in a search engine, is calculated as $\pi^T = \pi^T \mathbf{G}$, where $\mathbf{G}$ is the Google matrix:

$$\mathbf{G} = \kappa(\mathbf{P} + \frac{de^T}{N}) + \frac{(1-\kappa)}{N} uu^T$$

$k = 0.85$ in the original formulation [176], d is the binary vector called dangling node vector (d_i is equal to one if i is a dangling node and 0 otherwise), u is a vector with unitary elements and $\mathbf{P}_{ij} = \frac{a_{ij}}{k_j}$ is the transition probability matrix of the respective network.

Community detection

The task of *community detection* in networks (or graph clustering) is to find communities (also called clusters or modules), i.e., groups of vertices which share common properties and/or play similar roles within the graph. The problem is ill posed, as there is no universal definition of community, and usually it depends on the specific application or system under analysis [83]. Communities may be overlapping (*soft clustering*) and share some vertices, or non-overlapping (*hard clustering*). Sometimes the generic term *clustering* is used to indicate both types of community detection. We refer the reader to [83] for an extensive review of the topic.

Given a subgraph C, with N_C nodes, of a graph G with N nodes and adjacency

matrix A, we can define a set of community variables based on internal connectedness, namely:

- *(internal degree)*: $k_C^{int} = \sum_{i,j \in C} a_{ij}$
- *(average internal degree)*: $k_C^{avg-int} = \frac{k_C^{int}}{N_C}$
- *(internal edge density)*: $\delta_C^{int} = \frac{k_C^{int}}{N_C(N_C-1)}$

and a set of community variables based on external connectedness:

- *(external degree or cut)*: $k_C^{ext} = \sum_{i \in C, j \notin C} a_{ij}$
- *(average external degree)*: $k_C^{avg-ext} = \frac{k_C^{ext}}{N_C}$
- *(external edge density)*: $\delta_C^{ext} = \frac{k_C^{ext}}{N_C(N-N_C)}$

and hybrid variables:

- *(total degree or volume)* $k_C = k^{int} + k^{ext} = \sum_{i \in C} a_{ij}$
- *(average degree)* $k_C^{avg} = \frac{k_C}{N_C}$
- *(conductance)* $C_C = \frac{k_C^{ext}}{k_C}$

These definitions hold for both undirected and unweighted networks, but can be generalized to weighted and directed networks easily, as it suffices to replace the "number of links" with the sum of the weights of each link.

There are many algorithms to detect communities in graphs. However, just a few popular algorithms are employed in most applications. Optimisation techniques, for instance, find an extremum of a function which indicates the quality of a clustering, over the space of all possible clusterings. This function can indicate the goodness of a partition or of single clusters. Modularity is one of the most popular quality functions [91]:

$$Q = \frac{1}{2m} \sum_{i,j} (a_{i,j} - p_{i,j}) \delta(C_i, C_j)$$

where *m* is the number of links in the network, $a_{i,j}$ is the element of the adjacency matrix, $p_{i,j}$ is the *null model* term and the Kronecker delta δ indicates the community of nodes i and j. The null model term indicates the average adjacency matrix of an ensemble of networks obtained by randomising the original graph, which preserve some of its features. Therefore, modularity measures how different the original graph is from randomisations. The idea is that randomisation destroys community structures, and modularity can measure how non-random groups are in the original network. A standard choice for the null model is $p_{i,j} = \frac{k_i k_j}{2m}$ where k_i and k_j are the degrees of

nodes i and j. It corresponds to the expected number of edges joining the two vertices if the degree of all vertices is preserved in the rewirings, on average. We can thus rewrite the modularity function as:

$$Q = \sum_{C} [\frac{l_C}{m} - \frac{k_c}{2m}]$$

where l_C is the number of edges incident to nodes belonging to cluster C and k_c is the total degree of the cluster, because the only contributions to the sum come from vertex pairs belonging to the same cluster, so we can rewrite it as a sum over the clusters. Modularity optimisation is NP-hard, and it is also not a perfect measure, as there are highly modular partitions even in graphs without structure [83].

The Louvain's algorithm [29] is one of the most popular modularity-based techniques for community detection. It performs a greedy optimisation of Q in a hierarchical manner, by assigning each node to the community of their neighbours yielding the largest Q, and creating a smaller weighted super-network whose vertices are the cluster found in the previous step. Partitions found on this super-network are clusters which include the ones found earlier, and represent a higher hierarchical level of clustering. The procedure is repeated until one reaches the level with the largest modularity.

Another set of techniques for detecting communities is based on running dynamical processes on a network, such as diffusion or synchronization. Random walk dynamics is by far the most exploited in community detection. If communities have high internal edge density and are well-separated from each other, random walkers would be trapped in each cluster for quite some time, before finding a way out and migrating to another cluster. The most popular of such methods is Walktrap [197], which defines a similarity between vertices i and j as the probability that a random walker moves from i to j in a fixed number of steps t. If communities are pronounced, pairs of nodes in the same cluster will be much more easily reachable by a random walk than pairs of nodes in different clusters, and the vertex similarity can be used to retrieve clusters with a hierarchical partitioning technique.

In order to validate a clustering algorithm , i.e., checking how precisely it can recover the communities, benchmark networks whose community structure is known are employed. These can be generated through a model, or they can be actual networks whose group structure is known via non-topological features (metadata). However, given the lack of an universal definition of community, benchmarks are often arbitrary.

2.1.7 Supervised machine learning classification

In this thesis, we employ supervised machine learning to develop a classifier that can automatically detect online disinformation (see Chapter 6). Therefore, in the following we provide the reader with some basic definitions of machine learning, with a focus on

supervised classification. We refer to [27, 161] for a broader view of the topic.

Basic definitions

Machine learning can be broadly defined as the task of using computational methods to improve the accuracy of predictions or performance in general, based on *experience*, i.e., some data collected by the learner. Specifically, the goal is to build an algorithm which allows to do so, and learning algorithms have unnumbered applications – spam detection, speech recognition, fraud detection, medical diagnosis, etc. We can define a few major classes of problems:

- **Classification**: assign a class to each input item, usually with a small number of classes.
- **Clustering**: divide input items into homogeneous groups.
- **Dimensionality reduction**: transform the input data with a different representation which has a lower-dimension and preserves some of the properties of the initial data.
- **Ranking**: order input items according to some criterion.
- **Regression**: predict a real value for each input item.

Here is a list of basic definitions to describe a typical machine learning setting:

- *Examples*: items that are used for developing and evaluating a learning algorithm.
- *Features*: the attributes of each example, often represented as vectors.
- *Labels*: values or categories assigned to each example, e.g. classes in classification problems or real values in regression problems.
- *Training sample*: examples that are used to train a learning algorithm.
- *Validation sample*: examples that are used to tune the parameters of a learning algorithm when labeled data are involved.
- *Test sample*: examples that are used to evaluate the performance of a learning algorithm. Usually this sample is not involved in the learning phase.
- *Loss function*: a function that measures the difference, or loss, between predicted labels and true labels.
- *Hypothesis set*: a set of mapping functions which take an input and produce a label.

There are many different machine learning scenarios which differ in the type of training and test data used to design a learning algorithm. We mention them briefly and refer the reader to [27, 161] for an exhaustive review. *Supervised learning* consists in labeled examples used as training data and the goal is to make prediction for unseen points; this is the most common scenario of classification and regression problems. *Unsupervised learning* consists in having unlabeled training data and making predictions for unseen points; common examples are clustering and dimensionality reduction. *Semi-supervised learning* is a scenario where the learning receives training samples which are both labeled and unlabeled. *Transductive inference* is similar to semi-supervised learning, but the goal is to predict labels for a specified test set. *On-line learning* involves multiple rounds of training and testing phases. *Reinforcement learning* is similar to on-line learning, but the learner has to actively interacts with the environment to receive new data. Finally, in *active learning* the learner can decide to interactively collect new training examples, with the goal of achieving performance similar to a supervised setting but with fewer labeled examples.

Tipically, the learning procedure consists in randomly partitioning the data into a training sample, a validation sample and a test sample. The size of each sample depends on the specifics of the algorithm, as they are used to tune its parameter and obtain the best performance. Features are then associated to each example, and they are used to fix different values of the free parameters of a given learning algorithm, selecting a hypothesis out of the hypothesis set which results in the best performance on the validation sample (*model selection*). Finally, we use it to predict labels of the examples in the test sample and we use the loss function to assess the performance.

Labeled examples are often very hard to obtain in most practical applications, and a widely adopted technique known as *k-fold cross-validation* is used to exploit labeled data for both model selection and training. The idea is to randomly partition a sample S of m labeled examples into k subsamples (or folds) of equal size. Then, for $i \in [1, k]$, the learning algorithm is trained on all folds but the ith fold, and the performance of the obtained hypothesis h_i is tested on the ith fold. Let θ be the free parameters (or hyperparameters) of the algorithm, $L(\cdot)$ a loss function and $((x_{i1}, y_{i1}, ..., (x_{im_i}, y_{im_i}))$ a labeled sample of size m_i in the ith fold, where x and y represent examples and labels. We can define a *cross-validation* error:

$$R_{CV}(\theta) = \frac{1}{k} \sum_{i=1}^{k} \frac{1}{m_i} \sum_{j=1}^{m_i} L(h_i(x_{ij}), y_{ij})$$

With large k the method tends to have a small bias but a large variance, and viceversa with smaller values of k it has a smaller variance but a larger bias. In many applications k is typically chosen to be 5 or 10; the special case where $k = m$ is called *leave-one-*

out cross-validation. In addition to model selection, cross validation is also employed for performance evaluation, i.e., by divided the full labeled sample in k folds with no distinction between training and test samples.

Classifiers

Here we briefly introduce a few supervised binary classifiers which are employed in later sections of this thesis, namely Support Vector Machines (SVM), Logistic Regression and K-Nearest Neighbors (k-NN).

Support Vector Machines. Given a set of n real-valued examples $\mathbf{x}$ with binary labels $y_i \in [-1, 1]$, the linear SVM algorithm proposed by Vapnik in 1963 aims to find the "maximum-margin hyperplane" that divides examples for which $y_i = 1$ from those for which $y_i = -1$, defined such that the distance between the hyperplane and the nearest point x_i from either group is maximized. Any hyperplane can be written as:

$$\mathbf{w}^T\mathbf{x} - b = 0$$

where $\mathbf{w}$ is the (not necessarily normalized) normal vector to the hyperplane. If the training examples are linearly separable (*hard margin*), we can select two parallel hyperplanes that separate the two classes of data, so that the distance between them, i.e., the margin, is as large as possible. When the data are normalized or standardized (Z-scoring), these two hyperplanes can be defined as:

$$\mathbf{w}^T\mathbf{x} - b = 1$$

where examples on or above this boundary have label $y_i = 1$, and

$$\mathbf{w}^T\mathbf{x} - b = -1$$

where examples on or below this boundary have label $y_i = -1$. The distance between the two hyperplanes is $\frac{2}{\mathbf{w}}$, so we want to minimize $\mathbf{w}$. The SVM algorithm puts together this and the constraints that each example lies on the correct side of the margin: $y_i(\mathbf{w}^T\mathbf{x_i} - b) \geq 1, \forall i = 1, ..., n$. The classifier is then given by $\text{sgn}(\mathbf{w}^T\mathbf{x_i} - b)$ where $\text{sgn}(\cdot)$ is the sign function. Interestingly, the max-margin hyperplane is completely determined by those examples which lie nearest to it, which are in fact called *support vectors*.

When examples are not linearly separable (*soft margin*), the *hinge loss* function is employed:

$$\max(0, 1 - y_i(\mathbf{w}^T\mathbf{x_i} - b))$$

For data on the wrong side of the margin, the function's value is proportional to the

distance from the margin, and the optimization problem becomes:

$$\lambda||\mathbf{w}||^2 + [\frac{1}{n}\sum_{i}^{n}\max(0, 1 - y_i(\mathbf{w}^T\mathbf{x_i} - b))]$$

where $\lambda > 0$ is a "hyperparameter" that determines the trade-off between increasing the margin size and having all the examples lying on the correct side of the margin. For sufficiently small values of λ we obtain the hard-margin SVM.
This linear algorithm – because it employs a linear combination of the characteristic of the features – can be adapted to a non-linear fashion by using the "kernel trick", i.e. using a nonlinear kernel function to fit the maximum-margin hyperplane in a transformed feature space [32]; however, although the classifier is a hyperplane in the transformed feature space, it may be nonlinear in the original input space.
Logistic Regression. Despite its name, logistic (or logit) regression is a classification algorithm which consists in estimating the logarithm of the odds (log-odds) for one class as a linear combination of the features of the input examples. In the two-class problem with classes 0 and 1, the algorithm outputs for each example a probability between 0 (certainly class 0) and 1 (certainly class 1), using the logistic function to convert log-odds to probability. The standard logistic function σ is defined as:

$$\sigma(x) = \frac{e^x}{1 + e^x} = \frac{1}{1 + e^{-x}}$$

and it's called *sigmoid* because it is a S-shaped curve. Assuming that the label variable follows a Bernoulli distribution, let p be the probability that an example belongs to class 1. We can write the following linear relationship for a given example $X_i = \{x_1, x_2\}$ with two features:

$$\log_b \frac{p}{1 - p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

where β_i are the parameters of the model. With a bit of algebraic manipulation we can obtain:

$$p = S_b(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$$

where S_b is the sigmoid function with base b.
In order to fit a logistic regression model, a maximum likelihood approach is employed using optimization techniques such as gradient descent. Let Y be the probability of a random variable to be 0 or 1 given input data X. We consider the generalized linear hypothesis parameterized by θ:

$$h_\theta(X) = \frac{1}{1 + e^{-\theta^T X}} = \mathbf{Pr}(Y = 1|X; \theta)$$

and $\Pr(Y = 0|X;\theta) = 1 - h_\theta(X)$. Since $Y \in 0, 1$ we have $\Pr(y|X;\theta) = h_\theta(X)^y(1 - h_\theta(X))^{(1-y)}$, and assuming that Y is Bernoulli distributed in the input data we can calculate the likelihood function:

$$L(\theta|y;x) = \Pr(Y = 0|X;\theta) = \prod_i h_\theta(X)^y(1 - h_\theta(X))^{(1-y)}$$

After fitting the model, one can examine the contribution of each feature x_i by looking at the corresponding coefficient, which represents the change in the logit (logarithm of the odds) for each unit change in the feature variable.

K-Nearest Neighbors. The k-Nearest Neighbors (k-NN) is a non-parametric algorithm which can be used for both classification and regression. It simply assigns to each example a label based on the most common label among its k nearest neighbors (usually k is a small positive integer). One can use different metrics to compute the distance between examples, usually the Euclidean distance for examples with continuous features and the overlap metric (or Hamming distance) for examples with discrete features. One can also assign a weight to different neighbors, e.g. the ith nearest neighbor has a weight w_{ni} such that $\sum_i^n w_{ni} = 1$. In the two-class setting, the k-NN with $k = 1$, which assigns to each example the label of its nearest neighbor, is guaranteed to yield an errore rate no worse than twice the Bayes error rate, i.e., the minimum achievable error rate given the distribution of the data. One drawback of this algorithm is that it needs to store the entire dataset in order to compute distance, and this can be expensive when dealing with large datasets. Usually, dimension reduction is performed to experimental data before applying the k-NN algorithm, in order to avoid the effect of the curse of dimensionality.

Evaluation metrics

In a binary classification setting, we can define several evaluation metrics which can be used to assess the performance of a classification algorithm.

In particular, given a set of labeled examples and considering in turn one class as "positive" with P examples and the other as "negative" with N examples, we denote: TP=true positives (when the actual label of an example and the predicted one are both positive), FP=false positives (when the actual label is negative but it is predicted as positive), FN=false negatives (when the actual label is positive and the predicted one is positive), TN = true negatives (when both actual and predicted labels are negative). The most popular metrics are:

1. **Accuracy** = $\frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}$, which is "symmetric" and doesn't depend on the choice of the positive class.
2. **Precision** = $\frac{\text{TP}}{\text{TP}+\text{FP}}$, the ability of a classifier not to label as positive a negative

sample.

3. **Recall** = $\frac{TP}{TP+FN}$, the ability of a classifier to retrieve all positive samples.

4. **F1-score** = $2\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}$, the harmonic average of Precision and Recall.

5. **Area Under the Receiver Operating Characteristic curve (AUROC)**: the Receiver Operating Characteristic (ROC) curve [77], which plots the TP rate versus the FP rate, shows the ability of a classifier to discriminate positive samples from negative ones as its threshold is varied; the AUROC value is in the range $[0, 1]$, with the random baseline classifier holding AUROC$= 0.5$ and the ideal perfect classifier AUROC$= 1$; thus larger AUROC values (and steeper ROCs) correspond to better classifiers.

6. **Area Under the Precision-Recall curve (AUPRC)**: similar to the ROC, one can plots the Precision versus the Recall of a classifier when varying the threshold, and compute the area under the curve. The random baseline classifer holds AUPRC$=$ $\frac{\text{P}}{\text{P}+\text{N}}$, and this measure is particularly suitable when the two classes are extremely imbalanced.

Given a metric, one might be interested in considering only one label as positive, or we may compute different averages of the metric considering in turn each label as positive. A *macro-average* consists in computing the metric independently for each label and then taking the average (hence treating all classes equally), whereas a *weighted-average* calculates metrics for each label, and finds their average weighted by the number of true instances for each label. Finally the *micro-average* is calculated by summing all the TPs, TNs, FPs, FNs of the two labels and the computing the metric globally.

2.2 Automatic detection of online disinformation

We start our literature review by considering a variegated landscape of research contributions which focus on the **detection** of false information spreading on social media. Their taxonomy, presented in Table 2.1, is based on two aspects: employed technique and considered features.

The problem has been traditionally formulated as a supervised binary classification problem, starting with datasets consisting of labeled news articles, related tweets and Facebook posts which allow to capture different features, from content based ones (text, image, video) to those pertaining to the social context (diffusion networks, users' profile, metadata) and, in some cases, to external knowledge bases (Wikipedia, Google News). Labels carrying the classification into true and false news are typically obtained via fact-checking organizations or by manual verification of researchers themselves.

	Machine Learning	Deep Learning	Other techniques
Content features	Wang et al (2017) [256] Horne et al. (2017) [111] Perez-Rosas et al. (2018) [186] Potthast et al. (2018) [199] Fairbanks et al. (2018) [75]	Baird et al. (2017) [21] Hanselowski et al. (2017) [105] Riedel et al. (2017) [204] Wang et al (2017) [256] Popat et al. (2018) [198]	Fairbanks et al. (2018) [75] Hosseini et al. (2018) [112]
Context features	Tacchini et al. (2017) [241]	Volkova et al. (2017) [252] Wang et al. (2018) [257] Wu et al. (2018) [267] Liu et al. (2018) [143]	Tacchini et al. (2017) [267] Wang et al. (2018) [77] Yang et al. (2019) [273]
Content and context features	Shu et al. (2019) [230] Volkova et al. (2018) [251]	Ruchansky et al. (2017) [214] Volkova et al. (2018) [251]	Shu et al. (2019) [230]

Table 2.1: *Comparative description of twenty studies for false news detection, in terms of method and considered features.*

Section 2.2.5 comparatively describes the datasets used as ground truth for false news classification.

For what concerns the classification method, a wide range of techniques are used, from traditional machine learning (Logistic Regression, Support Vector Machines, Random Forest) to deep learning (Convolutional and Recurrent Neural Networks) and to other models (Matrix Factorization, Bayesian Inference).

We approach these methods by starting from those contributions which focus only on content-based features; we next describe contributions which consider only the social context and finally those that consider both aspects.

2.2.1 Focus of the review

Aside from a few works appeared in 2015 and 2016 [50, 212, 213], we build our literature review with a focus on 2017 and 2018, when the research community first witnessed a sudden increase in the number of scientific contributions on the topic. This review is part of a paper published in 2019 [190], and we thus refer the reader to [279] for an overview of more recent literature.

Issues concerning false news on **collaborative platforms** such as Wikipedia and Yelp (namely fake reviews, spam detection, etc.) are out of the scope of this review; we thus refer the reader to [133] for an overview of related research. We suggest [147] for a comprehensive review of the research that focuses, instead, on **rumors detection and resolution**, as we observed that many aspects are shared with our subject. **Automated fact-checking** is another related topic; it deals with verification rather than search of false news on social media, and we refer the interested reader to [243]. We also suggest [81] to the readers who may be interested in the research on **social bots**.

2.2.2 Content-based techniques

In this section we consider research contributions which are content-based, meaning that they analyze solely the textual content of news articles, e.g. body, title, source.

Stance detection as a helpful first step towards fake news detection was introduced during the 2017 Fake News Challenge Stage 1[1] (FNC-1) organized by *D. Pomerleau et al. (2017)* [196]. The goal was to classify the stance of an entire news article relative to its headline, i.e. document-level stance detection. Neural networks are employed by three top-performing systems, respectively Talos (*Baird et al. (2017)* [21]), Athene (*Hanselowski et al. (2017)* [105]) and UCL Machine Reading (*Baird et al. (2017)* [204]). These models rely on a combination of lexical features, including Bag-of-Words, topic modeling and word similarity features. An extensive analysis of these approaches, with experiments on their ability to generalize on unseen data, is provided by *Hanselowski et al. (2018)* [106].

Wang et al. (2017) [256] consider a multi-label classification task on the Liar dataset, one of the first datasets introduced in the literature. This includes several textual and metadata features, such as the speaker affiliation or the source newspaper, and labels are based on the six degrees of truth provided by the PolitiFact[2] fact-checking organization. They solve the classification problem by considering several machine learning and deep learning methods, from logistic regression to convolutional and recurrent neural networks.

A deep textual analysis is carried out in *Horne at al. (2017)* [111], where authors examine the body and title of different categories of news articles (true, false and satire), extracting complexity, psychological and stylistic features. They highlight the relevance of each aspect in distinct classification tasks, using a linear Support Vector Machine (SVM), finally inferring that real news are substantially different from false news in title whereas satire and false news are similar in content. They also apply the Elaboration Likelihood Model [173] to news categories, and suggest that consuming false news requires little energy and cognition, making them more appealing to the readers.

A neural network model is also presented by *Popat et al. (2018)* [198], who build a framework to classify true and false claims, and also provide self-evidence for the credibility assessment. They evaluate their model against some state-of-the-art techniques on different collections of news articles and they show examples of explainable results enabled by the *attention mechanism* embedded in the model, which highlights the words in the text that are more relevant for the classification outcome.

Perez-Rosas et al. (2018) [186] produce a dataset of false and true news articles and consider different sets of linguistic features (extracted from the body of news articles) namely ngrams, LIWC [183], punctuation, syntax and readability. On top of these features they train a linear SVM classifier, showing different performances depending on the considered feature. They suggest that computational linguistics can effectively aide in the process of automatic detection of false news.

[1] `http://www.fakenewschallenge.org`
[2] `https://www.politifact.com/`

The goal of *Potthast et al. (2018)* [199] is to assess the style similarity of several categories of news, notably hyper-partisan, mainstream, satire and false. The proposed methodology employs an algorithm called *unmasking* [131], which is a meta learning approach originally intended for authorship verification. They carry out several experiments comparing topic and style-based features with a Random Forest classifier and they conclude that, while hyper-partisan, satire and mainstream news are well distinguished, a style-based analysis alone is not effective for detecting false news.

Fairbanks et al. (2018) [75] also aim to classify false and true news, using a collection of articles gathered from GDELT[3]); labels are manually crawled from a fact-checking website[4]. They compare two different models, a content-based one which uses a classifier on traditional textual features and a structural method that applies loopy belief propagation [165] on a graph built from the link structure of news articles. The conclusions indicate that by modeling just the text content of articles it is possible to detect bias, but it not possible to identify false news.

Hosseini et al. (2018) [112] tackle the problem of distinguishing different categories of false news (from satire to junk news), based only on the news content. They employ the Kaggle dataset, where they consider up to six different labels. Their approach involves a tensor decomposition of documents which aims to capture latent relationships between articles and terms and the spatial/contextual relations between terms. They further use an ensemble method to leverage multiple decompositions in order to discover classes with higher homogeneity and lower outlier diversity. They outperform other state-of-the-art clustering techniques and are able to correctly identify all categories of fake news.

2.2.3 Context-based techniques

Here we describe research contributions which are (social) context-based in the sense that they focus on information derived from social interactions between users, e.g. likes, comment and (re)tweets, as to detect fake content.

Tacchini et al. (2017) [241] propose a technique to identify false news on the basis of users who *liked* them on Facebook. They collect a set of posts and users from both conspiracy theories and scientific pages and they build a dataset where each feature vector represents the set of users who *liked* a page. They eventually compare logistic regression with a (boolean crowdsourcing) harmonic algorithm for showing that they are able to achieve high accuracy with a little percentage of the entire training data.

Volkova et al. (2017) [252] address the problem of predicting four sub-types of suspicious news: satire, hoaxes, click-bait and propaganda. They start from a (manually constructed) list of trusted and suspicious Twitter news accounts and they collect a set

[3] https://www.gdeltproject.org/
[4] https://mediabiasfactcheck.com/

of tweets in the period of Brussels bombing in 2016. They incorporate tweet text, several linguistic cues (bias, subjectivity, moral foundations) and user interactions in a *fused* neural network model which is compared against ad-hoc baselines trained on the same features. They qualitatively analyze the characteristics of different categories of news observing the performances of the model.

Wang et al. (2018) [257] propose a multi-modal neural network model which extracts both textual and visual features from Twitter and Weibo conversations in order to detect false news items. Inspired by adversarial settings [96] they couple it with an *event discriminator*, which they claim is able to remove event-specific features and generalize to unseen scenarios, where the number of events is specified as a parameter. They evaluate the model on two custom datasets, but they compare it with ad-hoc baselines which are not conceived for false news detection.

Wu et al. (2018) [267] instead concentrate on modelling the propagation of messages carrying malicious items in social networks. Therefore they build a custom dataset, reflecting both true and false news, by leveraging the Twitter API and the fact-checking website Snopes[5]. They first infer embeddings for users from the social graph and in turn use a neural network model to classify news items. To this extent they provide a new model to embed a social network graph in a low-dimensional space and they construct a sequence classifier, using *Long Short-Term Memory* (LSTM) networks [110] to analyze propagation pathways of messages. They show that their model performs better than other state-of-the-art embedding techniques.

Propagation of news items is also taken into account by *Yu et al. (2018)* [143], who use a combination of convolutional and *Gated Recurrent Units* (GRU) [46] to model diffusion pathways as multivariate time series, where each point corresponds to the characteristics of the user retweeting the news, and perform early detection of false news. The method is evaluated on two real-world datasets of sharing cascades showing better performances than other state-of-the-art-techniques, which were nonetheless originally conceived for rumor resolution.

The first unsupervised approach to false news detection is provided in *Yang et al. (2019a)* [273], where veracity of news and users' credibility are treated as latent random variables in a Bayesian network model, and the inference problem is solved by means of collapsed Gibbs sampling approach [208]. The method is evaluated on LIAR and BuzzFeedNews datasets, performing better than other general truth discovery algorithms, not explicitly designed for false news detection.

2.2.4 Content and Context-based techniques

In this section we describe research contributions which consider both news content and the associated social (context) interactions as to detect malicious information items.

[5] https://www.snopes.com/

The contribution of *Ruchansky et al. (2017)* [214] is a neural network model which incorporates the text of (false and true) news articles, the responses they receive in social networks and the source users that promote them. The model is tested on Twitter and Weibo sharing cascades datasets and it is evaluated against other techniques conceived for rumor detection. They finally present an analysis of users behaviours in terms of lag and activity showing that the source is a promising feature for the detection.

In *Shu et al. (2017)* [230] a tri-relationship among publishers, news items and users is employed in order to detect false news. Overall, user-news interactions and publisher-news relations are embedded using non-negative matrix factorization [181] and users credibility scores. Several different classifiers are built on top of the resulting features and performances are evaluated on the FakeNewsNet dataset against other state-of-the-art information credibility algorithms. Results show that the social context could effectively be exploited to improve false news detection.

Volkova et al. (2018) [251] focus on inferring different deceptive strategies (misleading, falsification) and different types of deceptive news (propaganda, disinformation, hoaxes). Extending their previous work [252], they collect summaries, news pages and social media content (from Twitter) that refer to confirmed cases of disinformation. Besides traditional content-based features (syntax and style) they employ psycho-linguistic signals, e.g. biased language markers, moral foundations and connotations, to train different classifiers (from Random Forests to neural networks) in a multi-classification setting. Final results show that falsification strategies are easier to identify than misleading and that disinformation is harder to predict than propaganda or hoaxes.

2.2.5 Datasets

The research community has produced a rich but heterogeneous ensemble of data collections for fact checking, often conceived for similar objectives and for slightly different tasks. We first introduce the datasets which are referenced in the aforementioned literature review along with a short description, the source and the main references; their features are summarized in Table 2.2. Next, we present some other interesting datasets. Finally, we refer the interested reader to [70] for a more recent review of existing data collections.

BuzzFeedNews. BuzzFeed[6] News journalists have produced different collections of verified false and true news, shared by both hyperpartisan and mainstream news media on Facebook in 2016 and 2017; two of them, introduced by *Silverman (2016)* [231], consist of title and source of news items and they are used in [111, 217, 273]

BuzzFeed-Webis. This collection extends the previous one as it also contains the full content of shared articles with attached multimedia; it is employed in [199].

[6] https://www.buzzfeed.com

	Content Features	Social Context Features	Size	Labeling	Platform	Reference
BuzzFeedNews	Article title and source	Engagement ratings	10^2	BuzzFeed	Facebook	[231]
BuzzFeedWebis	Full Article	-	10^3	BuzzFeed	Facebook	[199]
DeClare	Fact-checking post	-	10^5	NewsTrust PolitiFact Snopes	-	[198]
FakeNewsAMT	Article text only	-	10^3	Manual GossipCop	-	[186]
FakeNewsChallenge	Full article	-	10^3	Manual	-	[196]
FakeNewsNet	Full article	Users metadata	10^3	BuzzFeed PolitiFact	Twitter	[229]
Hoaxy	Full article	Diffusion network Temporal trends Bot score (for users)	$> 10^6$	-	Twitter	[224]
Kaggle	Article text and metadata	-	10^4	BS Detector	-	[207]
Liar	Short statement	-	10^4	PolitiFact	-	[256]
SemEval-2017 Task8	Full article Wikipedia articles	Threads (tweets, replies)	10^4	Manual	Twitter	[61]
Rumors	Fact-checking title	Diffusion network (Twitter) Original message, replies (Weibo)	10^4	Snopes Weibo	Twitter Sina Weibo	[147]

Table 2.2: *Comparative description of the datasets referenced in this survey.*

DeClare. This dataset contains several articles from Snopes, PolitiFact and NewsTrust [164] corresponding to both true and false claims; it is proposed in [198] and used for false news detection.

FakeNewsAMT. This collection contains some legitimate articles from mainstream news, some false news generated by Amazon Mechanical Turk workers and some false and true claims from GossipCop[7] (a celebrity fact-checking website); it is introduced in [186] for false news detection.

FakeNewsChallenge. This dataset was proposed for the 2017 Fake News Challenge Stage 1 [196]; it contains thousands of headlines and documents which have to be classified in a document-based stance detection task using 4 different labels (Agree, Discuss, Disagree, Unrelated). It was inspired by [82] where stance detection is instead applied at the level of single sentences. It is employed in [21, 105, 204]; an additional analysis is provided in [106].

FakeNewsNet. This dataset contains both news content (source, body, multimedia) and social context information (user profile, followers/followee) regarding false and true articles, collected from Snopes and BuzzFeed and shared on Twitter; it was

[7] `https://www.gossipcop.com`

presented in [229] and employed in [230].

Hoaxy. The Hoaxy platform[8] has been first introduced in [224] and employed in several studies [225, 226, 250] for different goals; it is continuously monitoring the diffusion network (on Twitter, since 2016) of news articles from both disinformation and fact-checking websites and it allows to generate custom data collections.

Kaggle. This dataset was conceived for a Kaggle false news detection competition [207] which contains text and metadata from websites indicated in the BS Detector[9]; it is employed in [112].

Liar. This is a collection of short labeled statements from political contexts, collected from PolitiFact, which serve for false news classification; it first appeared in [256] and it is employed in [273].

SemEval-2017 Task8. This data collection, composed of tweets and replies which form specific *conversations*, was designed for the specific tasks of stance and veracity resolution of social media content on Twitter; it is described in [61] and used in [198].

Rumors. This dataset was originally conceived for rumor detection and resolution in Twitter and Sina Weibo; introduced in [147], it contains retweet and discussion cascades corresponding to rumors/non-rumors and it is employed for false news detection and mitigation in [127, 143, 214].

Others. **BuzzFace** is a novel data collection composed of annotated news stories that appeared on Facebook during September 2016; it extends previous BuzzFeed dataset(s) with comments and the web-page content associated to each news article; it is introduced in [217]. As a complement to Hoaxy, **JunkNewsAggregator** is a platform that tracks the spread of disinformation on Facebook pages; it is described in [142]. Other datasets point to relevant organizations in the context of false news: [247] contains a list of false news outlets as indicated by different fact-checking organizations, whereas the list of signatories[10] of the International Fact Checking Network's code of principles is a collector of the main fact-checking organizations which operate in different countries. Finally, [12] provides a set of the most shared false articles identified on Facebook during 2016 US elections.

2.3 Characterizing the spread of online disinformation

In this section we review a series of contributions which have shed light on the mechanisms behind the spread of online disinformation, from the role of social bots and echo chambers to the demographics of users who are more vulnerable to deceptive content.

A first large-scale study on online misinformation is provided by *Del Vicario et al.* [58], who carry out a quantitative analysis on news consumption relatively to scientific

[8] https://hoaxy.iuni.iu.edu
[9] https://github.com/bs-detector/bs-detector
[10] https://ifcncodeofprinciples.poynter.org/signatories

and conspiracy theories news outlets on Facebook. They leverage Facebook Graph API[11] in order to collect a 5-year span of all the posts (and user interactions) which belong to the aforementioned categories. They analyze cascades (or *sharing trees*) in terms of lifetime, size and edge homogeneity (i.e. an indicator of the polarization of users involved) and they show that 1) the consumption patterns differ in the two categories and that 2) the *echo chambers* (or communities of interest) appear as the preferential drivers for the diffusion of content. On top of these results, they build a data driven percolation model which accounts for homogeneity and polarization and they simulate it in a small-world network reproducing the observed dynamics with high accuracy.

Similarly, a notable contribution is provided in [254], where the entire Twitter universe is explored in order to track the diffusion of false and true news. Authors build a collection of links to fact-checking articles (from six different organizations) which correspond to true, false and mixed news stories and they accordingly investigate how these rumors spread on the Twitter network by gathering only tweets that explicitly contain the URLs of the articles. The resulting dataset contains approx. 126,000 stories tweeted by 3 million users more than 4.5 million times. A series of measurements are carried out including statistical and structural indicators of the retweeting networks along with sentiment analysis, topic distribution and novelty estimation of the different categories of news. Final results show that overall falsehood spread significantly faster, deeper, farther and broader than the truth in all categories of information, with a prominent weight on political news. Moreover, they observe that false news usually convey a higher degree of novelty and that novel information is more likely to be shared by users (although they cannot claim this is the only reason behind the "success" of misinformation).

A slightly diverse analysis is issued in [226], where authors study the structural and dynamic characteristics of the core of the diffusion network on Twitter before and after the 2016 US Presidential Elections. They first illustrate the implementation and deployment of the Hoaxy platform [224] which is then employed to gather the data required for their analysis. They build different datasets (relative to a few months before and after the elections) which correspond to fact-checking and misinformation articles, i.e. the retweeting network of users that share URLs for related news items, and they perform a k-core decomposition analysis to investigate the role of both narratives in the network. They show that low-credibility articles prevail in the core, whereas fact-checking is almost relegated to the periphery of the network. They also carry out a network robustness analysis in order to analyze the role of most central nodes and guide possible different interventions of social platforms.

Same authors largely extend previous results in [225], as they carry out a huge anal-

[11] This is no longer available.

ysis on Twitter in a period of ten months in 2016 and 2017. They aim to find evidence of the considerable role of social bots in spreading low-credibility news articles. The Hoaxy [224] platform is leveraged once again and more than 14 million tweets, including fact-checking and misinformation sources, are collected. *Botometer* algorithm [56] is used to assess the presence of social bots among Twitter users. Results show that bots are active especially in the first phase of the diffusion, i.e. a few seconds after articles are published, and that although the majority of false articles goes unnoticed, a significant fraction tends to become viral. They also corroborate, to a certain extent, results provided by *Vosoughi et al. (2018)* [254]. Moving on, they highlight bot strategies for amplifying the impact of false news and they analyze the structural role of social bots in the network by means of a network dismantling procedure [9]. They finally conclude that curbing bots would be an effective strategy to reduce misinformation; using CAPTCHAs [253] is a simple tool to distinguish bots from humans, but with undesirable effects to the user experience of a platform.

A study of the agenda-setting [153] power of false news is instead accomplished in *Vargo et al. (2018)* [247], where authors focus on the online mediascape from 2014 to 2016. They leverage a few different agenda-setting models with a computational approach (collecting data from GDELT) in order to examine, among other targets, the influence of false news on real news reports, i.e. whether and to which extent false news have shifted journalistic attention in mainstream, partisan and fact-checking organizations. To this extent they gather news articles corresponding to partisan and mainstream news outlets as well as fact-checking organizations and false news websites; they refer to diverse references in the literature in order to manually construct the list. A network of different events and themes (as identified in the GDELT database) is built to relate distinct media and to model time series of (eigenvector) centrality scores [215] in order to carry out Granger causality tests and highlight potential correlations. Besides other results, they show that partisan media indeed appeared to be susceptible to the agendas of false news (probably because of the elections), but the agenda setting power of false news–the influence on mainstream and partisan outlets–is declining.

Bovet et al. [33] studied the influence of fake news on Twitter during 2016 US presidential elections. Leveraging a dataset of over 171 M tweets, they considered the full political spectrum of news sources shared by users, from extreme left to extreme right, plus those websites which notably shared fabricated information. Using a combination of network science techniques and a Granger causality analysis to characterize the information flow, they find that liberal news supporters influence the activity of most active users leaning towards Clinton, whereas Trump most active supporters influence the activity of fake news superspreaders.

In the same setting, authors of [100] analyze the activity of over 10k Twitter users which are linked to U.S. registered voters, during the elections. Using a source-based

approach similar to aforementioned work to estimate the prevalence of news articles coming from reliable vs unreliable outlets, they find that fake news are extremely concentrated in a minority of users. Also, they find that individuals most likely to engage with misleading content were conservative leaning, older, and highly engaged with political news.

2.4 Mitigation of online disinformation

A few potential interventions have been proposed for reducing the spread of misinformation on social platforms, from curbing most active (and likely to be bots) users [225] to leveraging the users' flagging activity in coordination with fact-checking organizations. The latter approach is proposed as a first practical mitigation technique in [127] and [246], where the goal is to reduce the spread of misinformation leveraging users' flagging activity on Facebook.

Kim et al. (2018) [127] develop CURB, an algorithm to select the most effective stories to send for fact-checking as to efficiently reduce the spreading of non-credible news with theoretical guarantees; they formulate the problem in the context of temporal point processes [3] and stochastic differential equations and they use the Rumors datasets to evaluate it in terms of precision and misinformation reduction (i.e. the fraction of prevented unverified exposures). They show that the algorithm accuracy is very sensitive to the ability of the crowd at spotting misinformation.

Tschiatschek et al. (2018) [246] also aim to select a small subset of news to send for verification and prevent misinformation from spreading; however, as they remark, with a few differences from the previous method respectively 1) they learn the accuracy of individual users rather than considering all of them equally reliable and 2) they develop an algorithm which is agnostic to the actual propagation of news in the network. Moreover, they carry out their experiments in a simulated Facebook environment where false and true news are generated by users in a probabilistic manner. They show that they are able at once to learn users' flagging behaviour and consider possible adversarial behaviour of spammer users who want to promote false news.

A different contribution is issued by *Vo et al. (2018)* [250], who are the first to examine active Twitter users who share fact-checking information in order to correct false news in online discussions. They incidentally propose a URL recommendation model to encourage these *guardians* (users) to engage in the spreading of credible information as to reduce the negative effects of misinformation. They use Hoaxy [224] to collect a large number of tweets referring to fact-checking organizations and they analyze several characteristics of the users involved (activity, profile, topics discussed, etc). Finally, they compare their recommendation model, which takes into account the social structure, against state-of-the-art collaborative filtering algorithms.

Main social networking platforms, from Facebook to Twitter, have recently provided to their users tools to combat disinformation [127], an approach which seems reasonable enough to tackle the problem of disinformation without raising censorship alerts. Resorting to the *wisdom of the crowd*, as discussed above, can be effective at identifying malicious news items and prevent from misinformation spreading on social networks.

2.5 The COVID-19 infodemic

The impact of the COVID-19 pandemic has been felt globally, with almost 220 million detected cases and 4.55 million deaths as of September 2021 (`coronavirus.jhu.edu/map.html`). Epidemiological strategies to combat the virus require collective behavioral changes. To this end, it is important that people receive coherent and accurate information from media sources that they trust. Within this context, the spread of false narratives in our information environment can have acutely negative repercussions on public health and safety. For example, misinformation about masks greatly contributed to low adoption rates and increased disease transmission [146]. The problem is not going away any time soon: false vaccine narratives [144] will drive hesitancy, making it difficult to reach herd immunity and prevent future outbreaks.

It is concerning that many people believe, and many more have been exposed to, misinformation about the pandemic [160, 168, 209, 220]. The spread of this misinformation has been termed the *Infodemic* [276]. Social media play a strong role in propagating misinformation because of peer-to-peer transmission [254]. There is also evidence that social media are manipulated [225, 236] and used to spread COVID-19 misinformation [80]. It is therefore important to better understand how users disseminate misinformation across social media networks.

Concerns regarding online health-related misinformation existed before the advent of online social media. Studies mostly focused on evaluating the quality of information on the web [73], and a new research field emerged, namely "infodemiology," to assess health-related information on the Internet and address the gap between expert knowledge and public perception [72]. We refer the interested reader to [258] for a deeper review of the existing literature on the subject.

With the wide adoption of online social media, the information ecosystem has seen large changes. Peer-to-peer communication can greatly amplify fake or misleading messages by any individual [254]. Many studies reported on the presence of misinformation on social media during the time of epidemics such as Ebola [85, 121, 180, 222] and Zika [31, 223, 228, 265]. Misinformation surrounding vaccines has been particularly persistent and is likely to reoccur whenever the topic comes into public focus [20, 62, 67, 149, 177, 221].

These studies focused on specific social media platforms including Twitter [149, 265], Facebook [221, 228], Instagram [223], and YouTube [31, 67]. The most common approach was content-based analysis of sampled social media posts, images, and videos to gauge the topics of online discussions and estimate the prevalence of misinformation. Unfortunately, the datasets analysed in these studies were usually small (at a scale of hundreds or thousands of items) due to difficulties in accessing and manually annotating large scale collections.

Unsurprisingly, the COVID-19 pandemic has inspired a new wave of health misinformation studies. In addition to traditional approaches like qualitative analyses of social media content [8,69,132,140,156,200] and survey studies [160,168,209], quantitative studies on the prevalence of links to low-credibility websites at scale have gained popularity in light of the recent development of computational methods [38,47,88,102, 233, 234, 271].

Many of these studies aimed to assess the prevalence of, and exposure to, COVID-19 misinformation on online social media [45]. However, different approaches yielded disparate estimates of misinformation prevalence levels ranging from as little as 1% to as much as 70%. These widely varying statistics indicate that different approaches to experimental design, including uneven access to data on different platforms and inconsistent definitions of misinformation, can generate inconclusive or misleading results.

2.5.1 Health-related disinformation in the Italian context

In this subsection, we zoom in a few contributions which investigate the spread of health-related disinformation in the Italian scenario.

Authors of [13] explored the relationships between Measles, mumps, and rubella (MMR) vaccination coverage in Italy and online search trends and social network activity from 2010 to 2015. Using a set of keywords related to the controversial link between MMR vaccines and autism, originated from a discredited 1998 paper, authors analyzed Google (search) Trends as well as the activity of Facebook pages and Twitter users on the same subject. They reported a significant negative correlation with the evolution of vaccination coverage in Italy (which decreased from 90% to 85% during the period of observation). They also identified real-world triggering events which most likely drove vaccine hesitancy, i.e. Court of Justice sentences that ruled in favor of a possible link between MMR vaccine and autism.

Authors of [67] provide a quantitative analysis of the Italian videos published on YouTube, from 2007 to 2017, about the link between vaccines and autism or other serious side effects in children. They showed that videos with a negative tone were more prevalent and got more views than those with a positive attitude. However, they did not inspect how videos were treating the link between vaccines and autism.

In [206], authors analyzed the Italian vaccine-related environment on Twitter in cor-

respondence with the child vaccination mandatory law promulgated in 2017. Using a keyword-based data collection similar to ours, the author showed that the strong "politicization" of the debate was associated with an increase in the amount of problematic information, such as conspiracy theories, anti-vax narratives, and false news, shared by online users.

Finally, authors of [52] also analyzed the debate about vaccinations in Italy on Twitter, following the mandatory law promulgated in 2017. They inspected the network of interactions between users, and they identified two main communities of people classified as "vaccine advocates" and "vaccine skeptics", in which they find evidence of echo chamber effects. Besides, they proposed a methodology to predicting the community in which a neutral user would fall, based on a content-based analysis of the tweets shared by users in the two groups.

CHAPTER 3

Investigating Italian disinformation spreading on Twitter

In this chapter we provide results from our investigation of Italian language disinformation spreading on Twitter. We used a combination of network-science, text mining and other data science techniques to study the phenomenon in the run-up to the 2019 European Parliament elections. The text in this chapter is thus taken and adapted from [189].

3.1 Background

As the European Union (EU) struggled to counter the financial crisis which took place since the end of 2009 (following 2008 financial crisis in the US), populist and anti-establishment movements emerged as new electoral forces in Europe [108].

After the 2016 Brexit Referendum, anti-Europeans parties spread across the continent defining national identities in terms of ethnicity and religion and supporting tighter immigration controls [60].

As Europeans were called to elect their new representatives at the European Parliament–between the 23^{rd} and the 26^{th} of May 2019–populist and nationalist parties contrasted more traditional ones, such as European People's Party (EPP), Socialists and Democrats (S&D) and Alliance of Liberals and Democrats for Europe (ALDE), generally engaged in the defense of fundamental values associated with the EU project.

Eventually, the pro-European side prevailed on aforementioned disruptive forces in most countries, but not in Italy where “Lega” amplified its electoral consensus (33%) and instead “Movimento 5 Stelle” declined (18%). Outside of our scope, a change of the Italian government occurred during the Summer of 2019.

For what concerns misbehavior on social platforms in European countries, research has highlighted the impact and the influence of social bots and online disinformation in different circumstances, including 2016 Brexit [25], 2017 French Presidential Elections [79, 114] and 2017 Catalan referendum [236]. A significant presence of disinformation in online conversations concerning 2019 European elections has been reported across several countries [107, 114, 130, 151]. The European Commission has itself raised concerns–since 2015 [48]–about the large exposure of citizens to disinformation, promoting an action plan to build capabilities and enforce cooperation between different member states. In anticipation of 2019 European Parliament elections, they sponsored an ad-hoc fact-checking portal (`www.factcheck.eu`) to debunk false claims relative to political topics, aggregating reports from several agencies across different countries.

For what concerns Italy, according to Reuters [167], trust in news is particularly low as of 2019 (40% of people trust overall news most of the time, 23% trust news in social media most of the time), as result of a long-standing trend which is mainly due to the political polarization of mainstream news organizations and of the resulting partisan nature of Italian journalism.

Previous research on online news consumption highlighted the existence of segregated communities [59] and explored the characteristics of polarizing and controversial topics which are traditionally prone to misinformation [249].

Remarkable exposure to online disinformation was highlighted by authors of [90], who exhaustively investigated online media coverage in the run-up to 2018 Italian General elections; in particular, the study observed a rising trend in the spread of malicious information, with a peak of interactions in correspondence with the Italian elections. This result was later substantiated in a report of the Italian Authority for Communications Guarantees (AGCOM) [5]. Another recent work [42] collected electoral and socio-demographic data, relative to Trentino and South Tyrol regions, as to directly estimate the impact of false news on the 2018 electoral outcomes, with a focus on the populist vote; this study argues that malicious information had a negligible and non-significant effect on the vote. Furthermore, a recent investigation by Avaaz [16] revealed the existence of a network of Facebook pages and fake accounts which spread low-credibility and inflammatory content–reaching over a million interactions–in explicit support of "Lega", "Movimento 5 Stelle" about controversial themes such as immigration, national safety and anti-establishment. Those pages were eventually shut down by Facebook as violating the platform’s terms of use.

3.2 Research questions and contributions

In our work we focus on the 5-month period preceding 2019 European elections; we use a consolidated setting, described in [116, 224, 226], for investigating the presence (and the impact) of disinformation in the Italian Twittersphere.

To collect relevant data, we manually curated a list of websites which have been flagged by fact-checking agencies for fabricating and spreading a variety of malicious information, namely inaccurate and misleading news reports, hyper-partisan and propaganda stories, hoaxes and conspiracy theories. Differently from [226], satire was excluded from the analysis. Following literature on the subject [33, 100, 138, 225], we used a Distant supervision approach, and assumed that all articles published on aforementioned outlets indeed produced deceptive information; nonetheless, we are aware that this might not be always true and reported cases of misinformation on mainstream outlets are not rare [138].

We recognize that our analysis has a few inherent limitations: first, according to Reuters [167] Twitter in Italy is overtaken by far by other social platforms, accounting for only 8% of total users (with a decreasing trend) when it comes to consume news online compared to Instagram (13%), YouTube (25%), WhatsApp (27%) and Facebook (54%), which exhibit instead a rising trend. Second, these differences are even more accentuated when comparing with the U.S. scenario [5], the focus of most of recent research. However, other aforementioned social media offer today little opportunities to researchers to conveniently analyze the spread of online information, given the limitations they impose on the acquisition of data and the different user experiences they offer. Our study sheds light on the Italian mechanisms of disinformation spreading, and thus the outcomes of the analysis indicate directions for future research in the field.

Our analysis is thus driven by the following research questions:

RQ1: What was the reach of disinformation which circulated on Twitter in the run-up to European Parliament elections? How active and strong is the community of users sharing disinformation?

RQ2: What were the most debated themes of disinformation? How much were they influenced by national vs European-scale topics?

RQ3: Who are the most influential spreaders of disinformation? Do they exhibit precise political affiliations? How could we dismantle the disinformation network?

RQ4: Do disinformation outlets share deceptive content in a coordinated manner? Can we identify connections with websites from other countries?

We shall first describe the data collection and the methodology employed to perform our analysis, then we discuss each of the aforementioned research questions, and finally

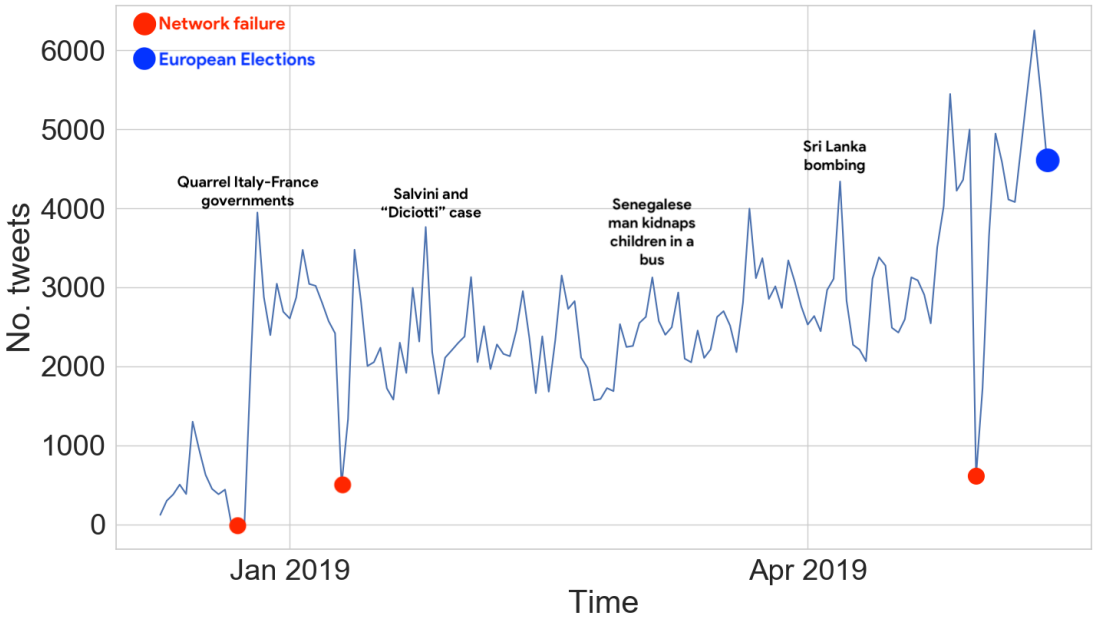


Figure 3.1: *Time series for the number of tweets, containing links to disinformation articles, collected in the period from 07/01/2019 to 27/05/2019. We annotated it with some events of interest; network failures indicate when the collection tool went down*

.

we summarize our findings.

3.3 Methods

3.3.1 Data Collection

Following a consolidated strategy [116, 224–226], we leveraged Twitter Streaming API in order to collect tweets containing an explicit Uniform Resource Locator (URL) associated to news articles shared on a set of Italian disinformation websites. As a matter of fact, using the standard streaming endpoint allows to gather 100% of shared tweets matching the defined query [116, 224, 226].

To this aim we manually compiled a list of 63 disinformation websites that were still active in January 2019. We relied on blacklists curated by local fact-checking organizations (such as "butac.net", "bufale.net" and "pagellapolitica.it"); these include websites and blogs which share hyper-partisan and conspiratorial news, hoaxes, pseudo-science and satire. We initially started with only a dozen of websites, and we successively added other sources; this did not alter the overall collection procedure.

For sake of comparison, we also included four Italian fact-checking and debunking

agencies, namely "lavoce.info", "pagellapolitica.it", "butac.net", "bufale.org".

In accordance with current literature [33, 100, 116, 224, 254] we use a Distant supervision approach: we do not verify each news article manually but we assign the *disinformation* label to all items published on websites labeled as such (the same holds for *fact-checking* articles).

In order to filter relevant tweets, we used all domains as query *filter* parameters (dropping "www", "https", etc) in the form "*byoblu com OR voxnews info OR ...*" as suggested by Twitter Developers guide (`https://developer.twitter.com`). We built a crawler to visit these websites and parse URLs as to extract article text and other metadata (published date, author, hyperlinks, etc). We handled URL duplicates by directly visiting hyperlinks and comparing the associated HTML content. We also extracted profile information and Twitter timelines for all users using Twitter API.

The collection of tweets containing disinformation (see Fig. 3.1) and fact-checking articles was carried out continuously from January 1st (2019) to May 27th, the day after EU elections in Italy. We collected 16,867 disinformation articles shared over 354,746 tweets by 23,243 unique users, and 1,743 fact-checking posts shared over 23,215 tweets by 9,814 unique users.

We can observe that, in general, articles devoted to debunk false claims were barely engaged, accounting only for 6% of the total volume of tweets spreading disinformation in the same period; such findings are comparable with the US scenario [226], and they are in accordance with the very low effectiveness of debunking strategies which is documented in [280]. We leave for future research an in-depth comparative analysis of diffusion networks pertaining to the two news domains.

The entire data is publicly available at: `https://doi.org/10.7910/DVN/OQHLAJ`

3.3.2 Comparison with Facebook

In order to perform a rough estimate of the different reach of disinformation on Twitter compared to Facebook, we collected data relative to the latter platform regarding two disinformation outlets, namely "byoblu.com" and "silenziefalsita.it", which have an associated Facebook page and are among Top-3 prolific and engaged sources of malicious information (see later section).

We used *netvizz* [205] to collect statistics on the number of daily shares of Facebook posts published by aforementioned outlets, and we compared with the traffic observed on Twitter.

As we can see in Fig. 3.2, disinformation has a stronger reach on Facebook than Twitter, for both sources, throughout the observation period; this is also shown in other research [5,16,90], coherently with the Italian consumption of social news. An in-depth

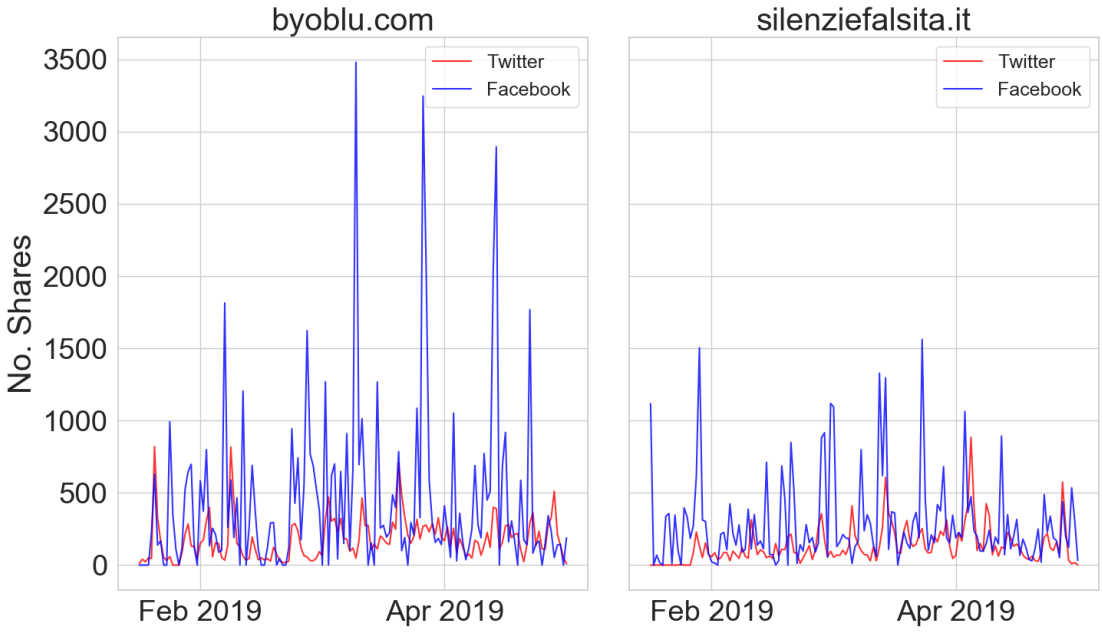


Figure 3.2: *Time series for the number of shares on both Twitter (red) and Facebook (blue) for two disinformation outlets, respectively "byoblu.com" (left) and "silenziefalsita.it" (right), in the period from 07/01/2019 to 27/05/2019.*

analysis of the Italian disinformation on Facebook would be required, but it needs suitable assistance from Facebook for what concerns the disinformation diffusion network.

3.3.3 Network analysis

Building Twitter diffusion network

We built Twitter global diffusion network–corresponding to the union of all sharing cascades associated to articles gathered in our dataset–following a consolidated strategy [225, 226]. We considered different Twitter social interactions altogether and for each tweet we add nodes and edges differently according to the action(s) performed by users:

- **Tweet**: a basic tweet corresponds to originally authored content, and it thus identifies a single node (author).
- **Mention**: whenever a tweet of user a contains a mention to user b, we build an edge from the author a of the tweet to the mentioned account b.
- **Reply**: when user a replies to user b we build an edge from a to b.
- **Retweet**: when user a retweets another account b, we build an edge from b to a.

- **Quote**: when user a quotes user b the edges goes from b to a.

When processing tweets, we add a new node for users involved in aforementioned interactions whenever they are not present in the network. As a remark, a single tweet can contain simultaneously several actions and thus it can generate multiple nodes and edges. Finally, we consider edges to be weighted, where the weight corresponds to the number of times two users interacted via actions mentioned beforehand.

Building the network of websites

In order to investigate existing inter-connections among different disinformation websites, and to understand the nature of external sources which are usually mentioned by deceptive outlets, we searched for URLs in all articles present in our dataset, i.e. which were shared at least once on Twitter. We accordingly built a graph where each node is a distinct Top-Level Domain–the highest level in the hierarchical Domain Name System (DNS) of the Internet–and an edge is built between two nodes a and b whenever a has published at least an article containing an URL belonging to b domain; the weight of an edge corresponds to the number of shared tweets carrying an URL with an hyperlink from a to b. The final result is a directed weighted network of approximately 5k nodes and 8k edges. We used *networkx* Python package [103] to handle the network.

Main core decomposition, centrality measures and community detection

In our analysis we employed several techniques coming from the network science toolbox [22], namely k-core decomposition, community detection algorithms and centrality measures. We used *networkx* Python package to perform all the computations.

The k-core [26] of a graph G is the maximal connected sub-graph of G in which all vertices have degree at least k. Given the k-core, recursively removing all nodes with degree k allows to extract the $(k+1)$-core; the main core is the non-empty graph with maximum value of k. k-core decomposition can be employed as to uncover influential nodes in a social network [226].

Community detection is the task of identifying *communities* in a network, i.e. dense sub-graphs which are well separated from each other [83]. In this work we consider Louvain's fast greedy algorithm [29], which is an iterative procedure that maximizes the Newman-Girvan *modularity* [91]; this measure is based on randomizations of the original graph as to check how non-random the group structure is.

A centrality measure is an indicator that allows to quantify the importance of a node in a network. In a weighted directed network we can define the *In-strength* of a node as the sum of the weights on the incoming edges, and the *Out-strength* as the sum of the weights on the out-going edges. *Betweenness* centrality [84] instead quantifies the probability for a node to act as a bridge along the shortest path between two other nodes;

it is computed as the sum of the fraction of all-pairs shortest paths that pass through the node. *PageRank* centrality [176] is traditionally used to rank webpages in search engine queries; it counts both the number and quality of links to a page to estimate the importance of a website, assuming that more important websites will likely receive more links from other websites.

3.3.4 Time series analysis

In our experiments, we carried out a trend analysis of time series concerning users' activity, topics contained in disinformation articles and the number of interconnections between different outlets.

In statistics, a trend analysis refers to the task of identifying a population characteristic changing with another variable, usually time or spatial location. Trends can be increasing, decreasing, or periodic (cyclic). We used the Mann-Kendall statistical test [124, 150] as to determine whether a given time series showed a monotonic trend. The test is non-parametric and distribution-free, e.g. it does not make any assumption on the distribution of the data. The null hypothesis H_0, no monotonic trend, is tested against the alternative hypothesis H_a that there is either an upward or downward monotonic trend, i.e. the variable consistently increases or decreases through time; the trend may or may not be linear. We used *mkt* Python package.

The multiple testing (or large-scale testing) problem occurs when observing simultaneously a set of test statistics, to decide which if any of the null hypotheses to reject [71]. In this case it is desirable to have confidence level for the whole family of simultaneous tests, e.g. requiring a stricter significance value for each individual test. For a collection of null hypotheses we define the family-wise error rate (FWER) as the probability of making at least one false rejection, (at least one type I error). We used the classical *Bonferroni* correction to control the FWER at $\leq \alpha$ by strengthening the threshold of each individual testing, i.e. for an overall significance level α and N simultaneous tests, we reject the individual null hypothesis at significance level α/N.

3.3.5 Limitations

As anticipated before hand, our methodology has some limitations which must be considered when assessing results.

First, we remark that we investigate disinformation spreading on a single social platform (Twitter) which has not a widespread usage in Italy, specifically if compared to other social networks such as Facebook, WhatsApp and Instagram – which, however, do not exhibit good APIs for data collection.

Second, we are subject to limitations of the Twitter Streaming API; [163] indicates that the API returns at most 1% of all the tweets produced on Twitter at a given time;

that source reports that once the number of tweets matching a given query exceeds 1% of the global daily volume of tweets, Twitter begins to sample the data returned to the user. In more recent documentation we found no mention of such limitation. Authors of [224–226] used the same approach as ours, and in an e-mail exchange they mentioned this Twitter policy as a potential limitation of their work. However, as the global volume of daily tweets exceeds $2 \cdot 10^8$ tweets (see `https://blog.twitter.com/official/en_us/a/2011/200-million-tweets-per-day.html`), most likely our data collection is not hindered by such limitation: in fact, we filter approximately $2 \cdot 10^3$ tweets per day, which are well below the 1% limit (which is roughly $2 \cdot 10^6$ tweets per day).

Third, we are collecting a specific typology of disinformation content originated from a limited set of sources, i.e. news articles published on websites which have been repeatedly flagged by journalists and fact-checkers as disinformation outlets. In line with findings from [226], we believe that we are drawing a consistent picture of Italian disinformation spreading on Twitter. However we miss photos and videos which may contain misleading or malicious content, and that can't be captured in a straightforward way. Besides, we are not verifying any of these shared items and at the same time we are not monitoring any unverified and misleading content which might be published on traditional and reliable news outlets.

Finally, for what concerns connections between disinformation outlets (see later sections) we remark that, when we observe out-going hyperlinks from Italian sources to disinformation outlets of other countries, we just show that outlets sharing disinformation news often refer to similar sources and tend to deliver similar stories; we cannot prove actual coordination between different outlets and/or countries.

3.4 Results and discussion

3.4.1 Assessing the reach of Italian disinformation

Sources of disinformation

To understand the reach of different disinformation outlets, we first computed the distribution of the number of articles and tweets per source. We observed, as shown in Fig. 3.3, that a few websites dominate on the remaining ones both in terms of activity and social audience.

In particular, with approximately 200k tweets (over 50% of the total volume) and 6k articles (about $1/3$ of the total number), "voxnews.info" stands out on all other sources; this outlet spreads disinformation spanning several subjects, from immigration to health-care and conspiratorial theories, and it runs campaigns against fact-checkers as well as labeling its articles with false "fact-checking" labels as to deceive readers.

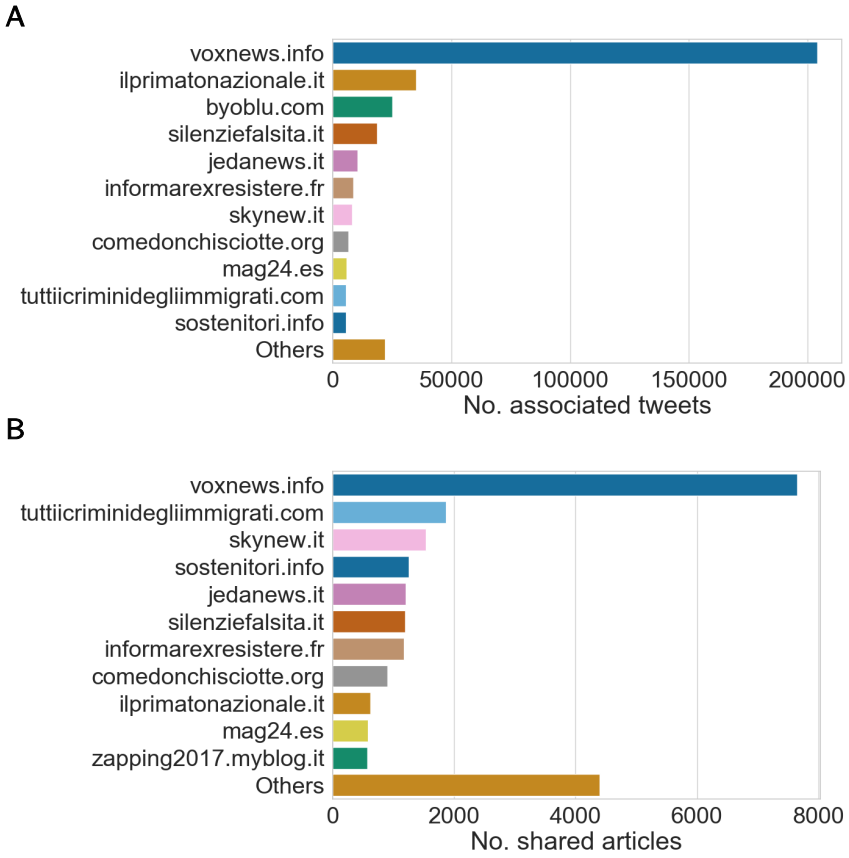


Figure 3.3: ***A.*** *The distribution of the total number of shared articles per website.* ***B.*** *The distribution of the total number of associated tweets per website. We show Top-11 (which account for over 95% of the total volume of tweets), and we aggregate remaining sources as "Others".*

Interestingly, two other uppermost prolific sources such as "skynew.it" and "tuttiicriminidegliimmigrati.com" do not receive the same reception on the platform; the former has stopped its activity on March and the latter is literally–it translates as "All the immigrants crimes"–a repository of true, false and mixed statements about immigrants who committed crimes in Italy.

We can also recognize three websites associated to public Facebook pages that have been recently banned after the investigation of Avaaz NGO, namely "jedanews.it", "catenaumana.it" and "mag24.es", as they were "regularly spreading fake news and hate speech in Italy" violating the platform's terms of use [16].

We further computed the distribution of the daily engagement (the ratio *no.articles published/no.tweets shared* per day) per each source, noticing that a few sources exhibit a considerable number of social interactions in spite of fewer associated tweets, compared to uppermost "voxnews.info". We show the time series for the daily engagement of Top-10 sources, which account for over 95% of total tweets, in Fig. 3.4. We can notice in particular that "byoblu.com" exhibits remarkable spikes of engagement w.r.t

to a very small number of total tweets compared to other outlets, whereas "mag24.es" shows a suspiciously large number of shares in the month preceding the elections (and after the release of Avaaz report).

We excluded "ilprimatonazionale.it" from this analysis as it was added only at the end of April (we collected around 30k associated tweets and less than 1,000 articles); official magazine of "CasaPound" (former) neo-fascist party–with style and agenda-setting that remind of Breitbart News–it exhibits a daily engagement of over 200 tweets, exceeding all other websites .

As elections approached, we were interested to understand whether there were particular trends in the daily reception of different sources. Focusing on Top-10 sources (except "ilprimatonazionale.it") we performed a Mann-Kendall test to assess the presence of an upward or downward monotonic trend in the time series of (a) daily shared tweets and (b) daily engagement. Taking into account Bonferroni's correction, the test was rejected at $\alpha = 0.05/10 = 0.005$; both (a) and (b) exhibit an upward trend for "byoblu.com" alone, whereas the remaining sources are either stationary or monotonically decreasing. As this outlet strongly supported euro-skeptical positions (and often gave visibility to many Italian representatives of such arguments) we argue that in the run-up to the European elections its agenda became slightly more captivating for the social audience.

User activity

For what concerns the underlying community of users sharing disinformation, we first computed the distribution of the number of shared tweets and unique URLs shared per number of users, noticing that a restricted community of users is responsible for spreading most of the online disinformation. In fact, approximately 20% of the community ($\sim$4k users) accounts for more than 90% of total tweets ($\sim$330k), in accordance with similar findings elsewhere [33,100,226]. Among them, we identified accounts officially associated to 18 different outlets (we manually looked at users' profile description and usernames); they overall shared 8,310 tweets.

We also distinguished five classes of users based on their generic activity, i.e. the number of shared tweets containing an URL to disinformation articles: *Rare* (about 9.5k users) with only 1 tweet; *Low* (about 8k users) with more than 1 tweet and less than 10; *Medium* (about 3k users) with a number of tweets between 11 and 100; *High* (about 500 users) with more than 100 tweets but less than 1,000; *Extreme* (exactly 20 users) with more than 1,000 shared tweets. About 1 user out of 5 shared more than 10 disinformation articles in five months.

As shown in Fig. 3.5A, we can notice that a minority of very active users (the ensemble with *High* and *Extreme* activity) accounts for half of the deceptive stories that were shared, and over 3/4 of the total number of tweets was shared by less than 4

Figure 3.4: *Daily engagement for Top-10 sources (ranked according to the total number of shared tweets). The Mann-Kendall test (upward trend at significance level 0.005) was accepted only for "byoblu.com".*

thousand users (*Medium*, *High* and *Extreme* activity).

We overall report 21,124 active (20 of which are also verified), 800 deleted, 124 protected and 112 suspended accounts. Verified accounts were altogether involved in 5761 tweets, only 18 of which in an "active" way, i.e. a verified account actually authored the tweet. We observed that they were mostly called in with the intent to mislead their followers, adding deceptive content on top of quoted statuses or replies.

Next we inspected the distribution of the number of users concerning their re-tweeting activity, i.e. the fraction of re-tweets compared to the number of pure tweets; as shown in Fig. 3.5B this is strongly bi-modal, and it reveals that users sharing disinformation are mostly "re-tweeters": more than 60% of the accounts exhibit a re-tweeting activity larger than 0.95 and less than 30% have a re-tweeting activity smaller than 0.05. This shows that a restricted group of accounts is presumably responsible for conveying in the first place disinformation articles on the platform, which are propagated afterwards by the rest of the community.

We computed the distribution of some user profile features, namely the count of followers and friends, the number of statuses authored by users and the age on the social platform (in number of months passed since the creation date to May 2019).

We report that users sharing disinformation tend to be quite "old" and active on the platform–with an average age of 3 years and more than a thousand authored statuses. We were able to gather information via Twitter API only for active and non-protected users.

We further inspected recently created accounts, noticing that approximately a thousand user was registered during the collection period, i.e. the last six months; they show similar distributions of aforementioned features compared to older users. Overall (see Fig. 3.5B) they mostly pertain to active classes (*Medium* and *High*) and they account for 15% (around 18k tweets) of the total volume of tweets considered–which lowers to approximately 288k tweets excluding those authored by non-active, suspended and protected accounts. Furthermore, about a hundred exhibit abnormal activities, producing more than 10k (generic) tweets in the period preceding the elections and directly sharing more than 10 disinformation stories each. We performed a Mann-Kendall test to the time series of daily tweets shared by such users (see Fig. 3.5C), assessing the presence of a monotonically increasing trend (at significance level $\alpha = 0.05$). The main referenced source of disinformation is "voxnews.info" with more than 60% (circa 12k tweets) of the total number of shared stories. An activity of this kind is quite suspicious and could be further investigated as to detect the presence of "cyber-troops" (bots, cyborgs or trolls) that either attempted to drive public opinion in light of upcoming elections (via so-called "astroturfing" campaigns [202]) or simply redirected traffic as to generate online revenue through advertisement [12, 138, 190].

3.4.2 The agenda-setting of disinformation

Theme analysis

For what concerns the main themes covered by different disinformation outlets, relative to the resulting audience on Twitter, we based our analysis on the first level of agenda-setting theory [154], which states that news media set the public importance for objects based on the frequency in which these are mentioned and covered. In the case of disinformation news an agenda-setting effect could occur as a result of the rise in the coverage, even if some audience members are aware that news are false [247]. We focused on the prevalence of titles, which were shared at least once, as they usually pack a lot of information about their claims in simple and repetitive structures [111]; besides, the exposure (such as the presence alone of misleading titles on users' timelines) could affect ordinary beliefs and result in resistance to opposite arguments [280] and an increased perceived accuracy of the content, irrespective of its credibility [275].

We avoided automatic topic modeling algorithms [28] as they are not suitable for small texts, and we employed a dictionary-based text-analysis, an approach which is largely used for testing communication theories such as agenda setting and selective

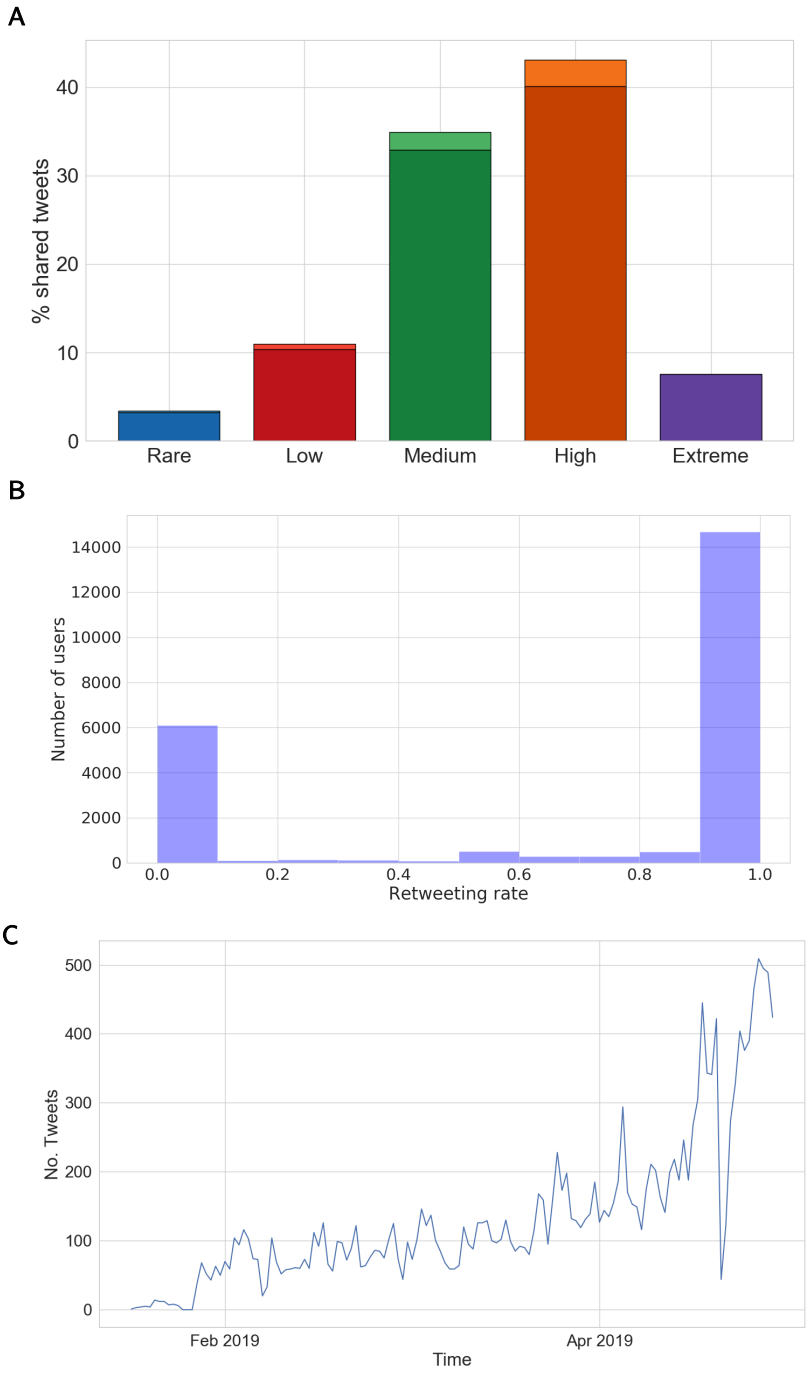


Figure 3.5: ***A.*** *A breakdown of the total volume of tweets according to the activity of users. Fractions of users created in the six months before the elections are indicated with lighter shades; these account respectively for 0.18% (Rare), 0.6% (Low), 2.04% (Medium) and 2.98% (High) of total tweets.* ***B.*** *The distribution of the number of users per retweeting activity.* ***C.*** *The distribution of daily tweets shared by recently created users.*

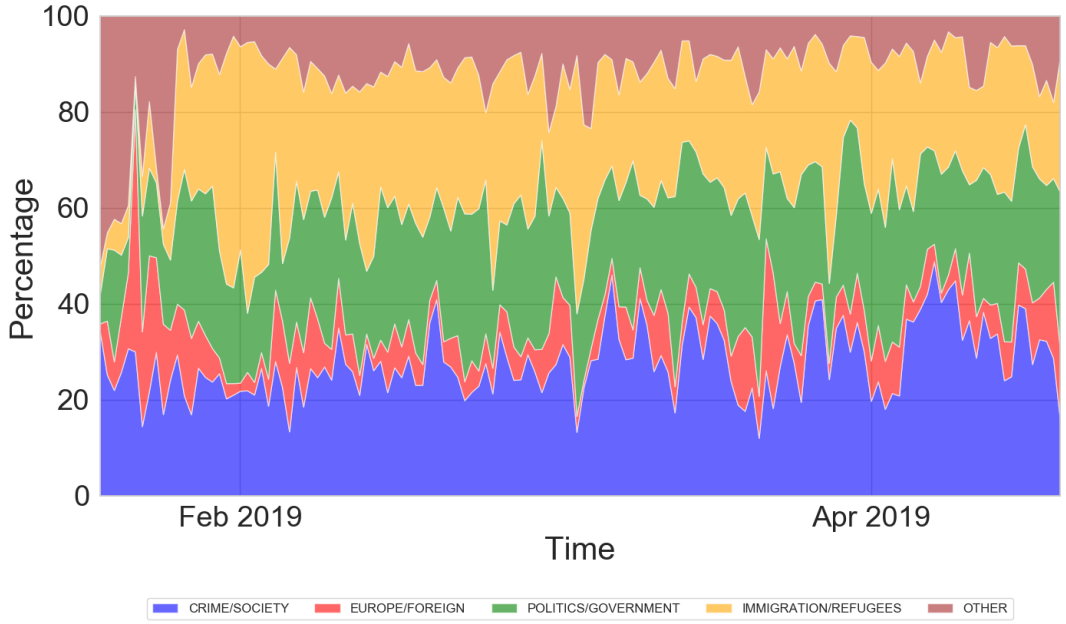


Figure 3.6: *A stacked-area chart showing the distribution of different topics over the collection period. The daily coverage on themes related to Immigration/Refugees and Europe/Foreign is stationary, whereas focus on subjects related to Crime/Society and Politics/Government is monotonically increasing towards the elections (end of May 2019).*

exposure in big social media data [248]. Therefore we manually compiled a list of keywords associated to five distinct topics namely: Politics/Government (PG), Immigration/Refugees (IR), Crime/Society (CS), Europe/Foreign (EF), Other (OT). Keywords were obtained with a data-driven approach, i.e. inspecting Top-500 most frequent words appearing in the titles, and taking into account relevant events that occurred in the last months. We provide Top-20 keywords for each topic in Table 3.1.

In particular, PG refers to main political parties and state government as well as the main political themes of debate. IR includes references to immigration, refugees and hospitality whereas CS includes terms mostly referring to crime, minorities and national security. Finally EF contains direct references to European elections and foreign countries. It is worth mentioning that the most frequent keyword was "video", suggesting that a remarkable fraction of disinformation was shared as multimedia content [255].

We computed the relative presence of each topic in each article by counting the number of keywords appearing in the title and accordingly assessed their distribution across tweets over different months. We can observe in Fig. 3.6 that the discussion was stable on controversial topics such immigration, refugees, crime and government, whereas

Politics/Government	Immigration/Refugees	Europe/Foreign	Crime/Society	Other
salvini	immigrati	euro	rom	video
italia	profughi	europa	milano	anni
pd	clandestini	ue	casa	contro
italiani	profugo	fusaro	bergoglio	foto
m5s	ong	diego	morti	vuole
italiana	porti	meluzzi	mafia	può
italiano	migranti	libia	bambini	vogliono
milioni	africani	macron	roma	parla
lega	immigrato	soros	donne	byoblu
sinistra	islamici	francia	bruciato	via
casapound	imam	francesi	confessa	niccolò
maio	seawatch	gilet	falsi	casal
soldi	nigeriani	gialli	bus	vero
guerra	nigeriana	europee	choc	ufficiale
cittadinanza	nigeriano	germania	figli	bufala
prima	islamica	tedesca	case	anti
raggi	africano	mondo	chiesa	sta
governo	stranieri	notre	famiglia	grazie
renzi	chiusi	dame	magistrato	casarini
zingaretti	sea	francese	polizia	farli

Table 3.1: *Top-20 keywords associated with each topic.*

focus on European elections and foreign affairs was quite negligible throughout the period, with only a single spike of interest at the beginning of January corresponding to the quarrel between Italian and France prime ministers. We also performed Mann-Kendall test to assess the presence of any monotonic trends in the daily distribution of different topics; we rejected the test for $\alpha = 0.05/5 = 0.01$ for IR and EF whereas we accepted it for the remaining topics, detecting the presence of an upward monotonic trend in CS and PG, and a downward monotonic trend in OT.

In the observation period, the disinformation agenda was well settled on main arguments supported by leading parties, namely "Lega" and "Movimento 5 Stelle", since 2018 general elections; this suggests that they might have profited from and directly exploited hoaxes and misleading reports as to support their populist and nationalist views (whereas "Partito Democratico" appeared among main targets of misinformation campaigns); empirical evidence for this phenomenon has been also widely reported elsewhere [42, 90]. However, the electoral outcome confirmed the decreasing trend of "Movimento 5 Stelle" electoral consensus in favor of "Lega", which was rewarded with an unprecedented success.

Differently from 2018 [90] we in fact observed one main cited leader: Matteo Salvini ("Lega" party). This is consistent with a recent report on online hate speech [118], contributed by Amnesty International, which has shown that his activity (and reception) on Twitter and Facebook is 5 times higher than Luigi Di Maio (leader of "Movimento 5 Stelle"); not surprisingly, his main agenda focuses (negatively) on immi-

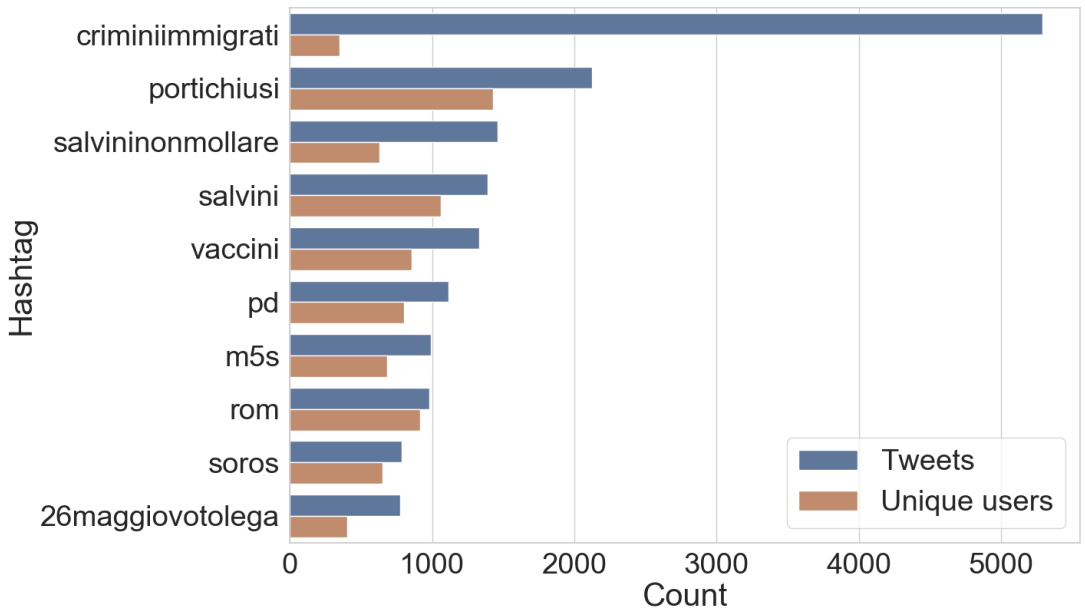


Figure 3.7: *Top-10 hashtags per number of shared tweets (blue) and unique users (orange).*

gration, refugees and Islam (which generated most of online interactions in 2018 [90]), which are also the main objects of hate speech and controversy in online conversations of Italian political representatives overall.

It appears that mainstream news actually disregarded European elections in the months preceding them, focusing on arguments of national debate [51]; this trend was also observed in other European countries according to FactCheckEU [74], claiming that misinformation was not prominent in online conversations mainly because European elections are not particularly polarized and are seen as less important compared to national elections. We believe that this might have affected the agenda of disinformation outlets, which are in general susceptible to traditional media coverage [153], thus explaining the focus on different targets in their deceptive strategies.

Usage of hashtags

Among most relevant hashtags shared along with tweets–in terms of number of tweets and unique users who used them (see Fig. 3.7)–a few indicate main political parties (cf. "m5s", "pd", "lega") and others convey supporting messages for precise factions, mostly "Lega" (cf. "salvininonmollare", "26maggiovotolega"); some hashtags manifest instead active engagement in public debates which ignited on polarizing and controversial topics (such as immigrants hospitality, vaccines, the Romani community and George Soros). We also found explicit references to (former) far-right party "CasaPound" and the associated "Altaforte" publishing house, as well as some disinformation websites (with a remarkable polarization on "criminiimmigrati" which was shared

Figure 3.8: *The cloud of words for Top-50 most frequent hashtags embedded in the users' profile description.*

more than 5,000 times by only a few hundred accounts).

We also extracted hashtags directly embedded in the profile description of users collected in our data, for which we provide a cloud of words in Fig. 3.8. The majority of them expresses extreme positions in matter of Europe and immigration: beside explicit references to "Lega" and "Movimento 5 Stelle", we primarily notice euro-skeptical (cf. "italexit", "noue"), anti-Islam (cf. "noislam") and anti-immigration positions (cf. "noiussoli", "chiudiamo i porti") and, surprisingly enough, also a few (alleged) Trump followers (cf. "maga" and "kag"). The latter finding is odd but somehow reflects the vicinity of Matteo Salvini and Donald Trump on several political matters (such as refugees and national security). On the other hand, we also notice "facciamorete", which refers to a Twitter grassroots anti-fascist and anti-racist movement that was born on December 2018, as a reaction to the recent policies in matter of immigration and national security of the Italian establishment.

3.4.3 Principal spreaders of disinformation

Central users in the main core

In order to identify most influential nodes in the diffusion network, we computed the value of several centrality measures for each account. We show in Table 3.2 the list of Top-10 users according to each centrality measure, and we also indicate whether they belong or not to the main K-core of the network [26]; this corresponds to the sub-graph of neighboring nodes with degree greater or equal than $k = 47$, which is shown in Fig

Table 3.2: *List of Top-10 users according to different centrality measures, namely In-strength, Out-Strength, Betweenness and PageRank; we indicate with a cross nodes that do not belong to the main K-core (k=47) of the network.*

Rank	In-Strength	Out-Strength	Betweenness	PageRank
1	*napolinordsud* ×	*Filomen30847137*	*IlPrimatoN*	*IlPrimatoN*
2	*RobertoPer1964*	*POPOLOdiTWITTER*	*matteosalvinimi*	*matteosalvinimi*
3	*razorblack66*	*laperlaneranera*	*Filomen30847137*	*Sostenitori1* ×
4	*polizianuovanaz* ×	*byoblu*	*byoblu*	*armidmar*
5	*Giulia46489464*	*IlPrimatoN*	*a_meluzzi*	*Conox_it* ×
6	*geokawa*	*petra_romano*	*AdryWebber*	*lauraboldrini* ×
7	*Gianmar26145917*	*araldoiustitia*	*claudioerpiu*	*pdnetwork* ×
8	*pasqualedimaria* ×	*max_ronchi*	*razorblack66*	*libreidee* ×
9	*il_brigante07*	*Fabio38437290*	*armidmar*	*byoblu*
10	*AngelaAnpoche*	*claudioerpiu*	*Sostenitori1* ×	*Pontifex_it* ×

9. We color nodes according to the communities identified by the Louvain modularity-based community algorithm [29] run on the original diffusion network (over 20k nodes and 100k edges).

Although we expect centrality measures to display small differences in their ranking, we can notice that the majority of nodes with highest values of In-Strength, Out-Strength and Betweenness centralities also belong to the main K-core of the network; the same does not hold for users which have a large PageRank centrality value. A few users strike the eye:

1. *matteosalvinimi* is Matteo Salvini, leader of the far-right wing "Lega" party; he is not an active spreader of disinformation, being responsible for just one (true) story coming from disinformation outlet "lettoquotidiano.com" (available at `https://twitter.com/matteosalvinimi/status/1102654128944308225`), which was shared over 1,800 times. He is generally passively involved in deceptive strategies of malicious users who attempt to "lure" his followers by attaching disinformation links in replies/re-tweets/mentions to his account.

2. *a_meluzzi* is Alessandro Meluzzi, a former representative of centre-right wing "Forza Italia" party (whose leader is Silvio Berlusconi); he is a well-known supporter of conspiracy theories and a very active user in the disinformation network, with approximately 400 deceptive stories shared overall.

3. Accounts associated to disinformation outlets, namely *IlPrimatoN* with "ilprimatonazionale.it", *byoblu* with "byoblu.com", *libreidee* with "libreidee.org", *Sostenitori1* with "sostenitori.info" and *Conox_it* with "conoscenzealconfine.it".

A manual inspection revealed that most of the influential users are indeed actively involved in the spread of disinformation, with the only exception of *matteosalvinimi* who is rather manipulated by other users, via mentions/retweets/replies, as to mislead his

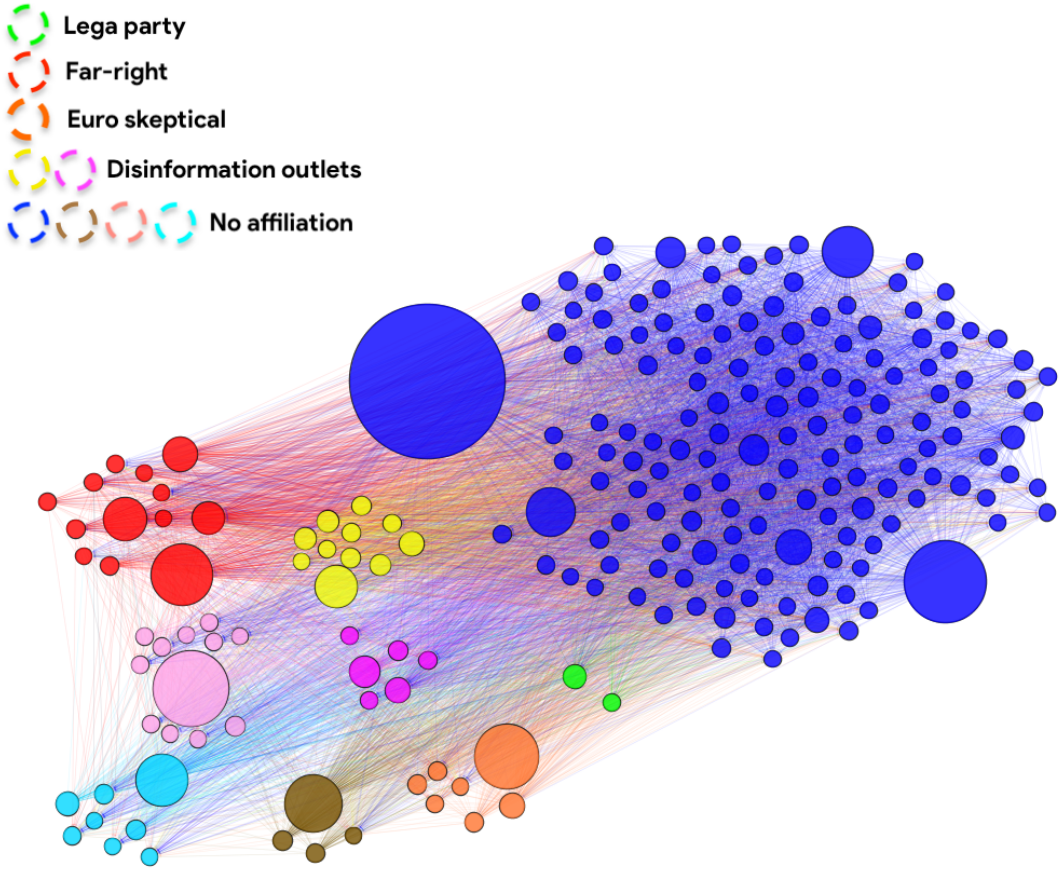


Figure 3.9: *The main K-core ($k = 47$) of the re-tweeting diffusion network. Colors correspond to different communities identified with the Louvain's algorithm. Node size depends on the total Strength (In + Out) and edge color is determined by the source node.*

huge community of followers (more than 2 millions). The story shared by Matteo Salvini underlines a common strategy of disinformation outlets identified in this analysis: they often publish simple true and factual news as to bait users and expose them to other harmful and misleading content present on the same website.

Besides, we notice in the ranking a few users who are (or have been in the past) target of several disinformation campaigns, such as *lauraboldrini* (Laura Boldrini), *pdnetwork* ("Partito Democratico" party) and *Pontifex_it* (Papa Francesco). We also report a suspended account (*polizianuovanaz*), a protected one (*Giulia46489464*) and a deleted user (*pasqualedimaria*).

In addition, we investigated communities of users in the main K-core–which con-

tains 218 nodes (see Fig. 3.9)–and we noticed systematic interactions between distinct accounts. We manually inspected usernames, most frequent hashtags and referenced sources, deriving the following qualitative characterizations:

1. the **Green** community corresponds to "Lega" party official accounts: *matteosalvinimi* and *legasalvini*, whereas the third account, *noipersalvini*, belongs to the same community but does not appear in the core.

2. the **Red** community represents Italian far-right supporters, with several representatives of CasaPound (former) party (including his secretary *distefanoTW* who does not appear in the core), who obviously refer to "ilprimatonazionale.it" news outlet.

3. the **Yellow** community is strongly associated to two disinformation outlets, namely "silenziefalsita.it" (*SilenzieFalsita*) and "jedanews.it" (*jedasupport*); the latter was one of the pages identified in Avaaz report [16] and deleted by Facebook.

4. the **Orange** community is associated to the euro-skeptical and conspiratory outlet "byoblu.com" (*byoblu*), and it also features Antonio Maria Rinaldi (*a_rinaldi*), a well-known euro-skeptic economist who has just been elected with "Lega" in the European Parliament.

5. the **Purple** community corresponds to the community associated to "tuttiicriminidegliimmigrati.com" (*TuttICrimin*) disinformation outlet.

6. the remaining **black** (*Filomen30847137*), **Light-blue** (*araldoiustitia*) and **Brown** communities (*petra_romano*) represent different groups of very active "loose cannons" who do not exhibit a clear affiliation.

Eventually, we employed Botometer algorithm [56] as to detect the presence of social bots among users in the main core of the network. We set a threshold of 50% on the Complete Automation Probability (CAP)–i.e. the probability of an account to be completely automated–which, according to the authors, is a more conservative measure that takes into account an estimate of the overall presence of bots on the network; besides, we computed the CAP value based on the language independent features only, as the model includes also some features conceived for English-language users. We only detected two bot-like accounts, namely *simonemassetti* and *jedanews*, respectively with probabilities 58% and 64%, that belong to the same Purple community. A manual check confirmed that the former habitually shares random news content (also mainstream news) in an automatic flavour whereas the latter is the official spammer account of "jedanews.it" disinformation outlet. We argue that the impact of automated accounts in the diffusion of malicious information is quite negligible compared to findings re-

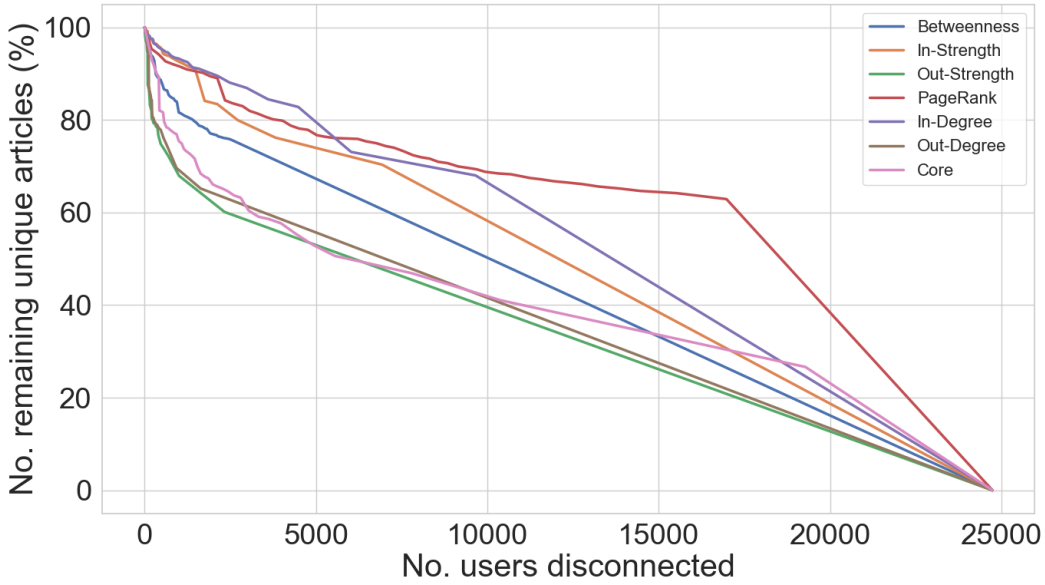


Figure 3.10: *Results of different network dismantling strategies w.r.t to remaining unique disinformation articles in the network. The x-axis indicates the number of disconnected accounts and the y-axis the fraction of remaining items in the network.*

ported in [226], where about 25% of accounts in the main core of the US disinformation diffusion network were classified as bots.

Dismantling the disinformation network on Twitter

Similar to [226], we performed an exercise of network dismantling analysis using different centrality measures, as to investigate possible intervention strategies that could prevent disinformation from spreading with the greatest effectiveness.

We first ranked nodes in decreasing order w.r.t to each metric, plus the core number–the largest k for which the node is present in the corresponding k-core–and the In and Out-degree, which exhibited the same Top 10 ranking as their weighted formulation (Strengths), but they do entail different results at dismantling the network. Next we delete them one by one while tracking the resulting fraction of remaining edges, tweets and unique articles in the network.

We observed that eliminating a few hundred nodes with largest values of Out-Degree promptly disconnects the network; in fact these users alone account for 90% of the total number of interactions between users. For what concerns the number of tweets sharing disinformation articles, the best strategy would be to target users with largest values of In-Strength who, according to our network representation, are likely to be users with a high re-tweeting activity; in fact, confirming previous observations, a few thousand nodes account for more than 75% of the total number of tweets shared in the five months

before the elections. However, as shown in Fig. 3.10, it is more challenging to prevent users to be exposed from even a tiny fraction of disinformation articles, as the network exhibits an almost linear relationship between the number of users disconnected and the corresponding number of remaining stories; as such the spread of malicious information would be completely prevented only blocking the entire network.

3.4.4 Interconnections of deceptive agents

To investigate existing connections between different disinformation outlets and other external sources, we first analyzed the network of websites with a core decomposition [26], obtaining a main core ($k = 14$) which contains 35 nodes as a result of over 75,000 external re-directions via hyperlinks (shown in Fig. 3.11A). Over 99% of the articles includes a hyperlink in the body. We may first notice frequent connections between distinct disinformation outlets, suggesting the presence of shared agendas and presumably coordinated deceptive tactics, as well as frequent mentions to reputable news websites; among them we distinguish "IlFattoQuotidiano", which is a historical supporter of "Movimento 5 Stelle", and conservative outlets such as "IlGiornale" and "LiberoQuotidiano" which lean instead towards "Lega". We also observe that most of the external re-directions point to social networks (Facebook and Twitter) and video sharing websites (Youtube); this is no wonder given that disinformation is often shared on social networks as multimedia content [138, 190]. In addition, we inspected nodes with the largest number of incoming edges (In-degree) in the original network, discovering among uppermost 20 nodes a few misleading reports originated on dubious websites (such as "neoingegneria.com"), flagged by fact-checkers but that were not included in any blacklist. We believe that a more detailed network analysis could reveal additional relevant connections and we leave it for future research.

Furthermore, we focused on the sub-graph composed of three particular classes of nodes, namely Russian (RU) sources, EU/US disinformation websites and our list of Italian (IT) outlets; we manually identified notable Russian sources ("RussiaToday" and "SputnikNews" networks) and we resorted to notable blacklists to spotlight other EU/US disinformation websites–namely "opensources.co", "décodex.fr", the list compiled by Hoaxy [224] and references to junk news in latest data memos by COMPROP research group [107, 114, 130, 151].

The resulting bipartite network–we filtered out intra-edges between IT sources to better visualize connections with the "outside" world–contains over 60 foreign websites (RU, US and EU) and it is shown in Fig. 3.11B.

We observe a considerable number of external connections (over 500 distinct hyperlinks present in articles shared more than 5 thousand times) with other countries sources, which were primarily included within "voxnews.info", "ilprimatonazionale.it" and "jedanews.it". Among foreign sources we encounter several well-known US sources

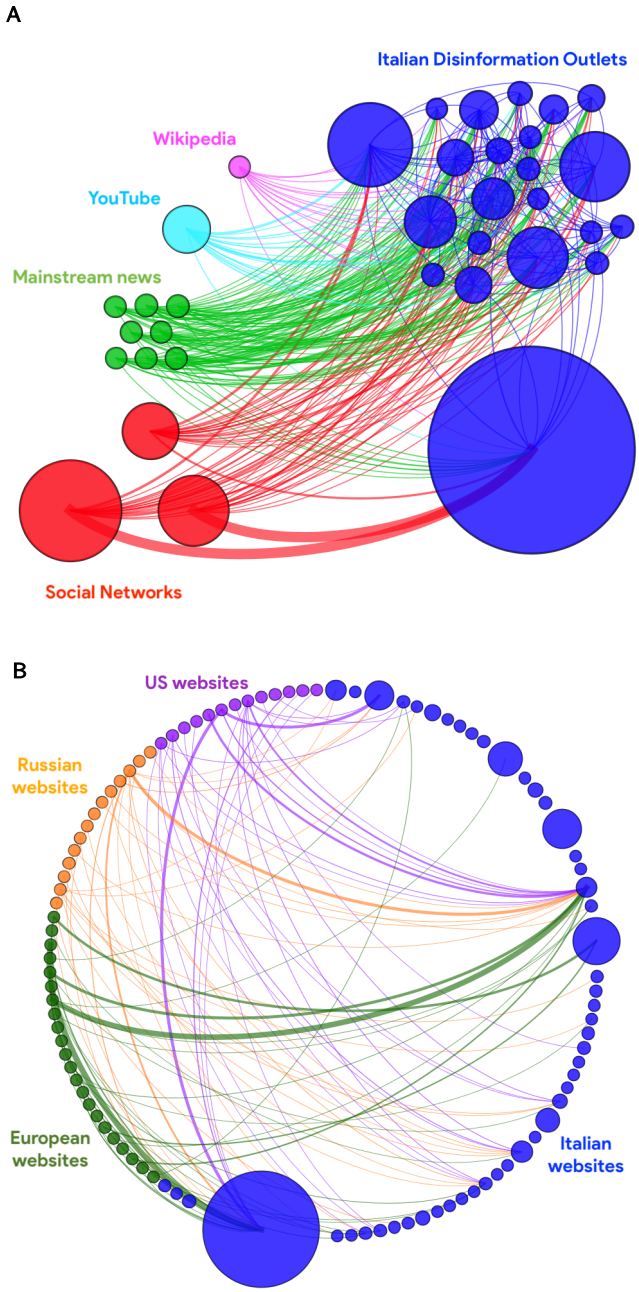


Figure 3.11: *Two different views of the network of websites; the size of each node is adjusted w.r.t to the Out-strength, the color of edges is determined by the target node and the thickness depends on the weight (i.e. the number of shared tweets containing an article with that hyperlink).* ***A (Left).*** *The main core of the network ($k = 14$); blue nodes are Italian disinformation websites, green ones are Italian traditional news outlets, red nodes are social networks, the sky-blue node is a video sharing website and the pink one is an online encyclopedia.* ***B (Right).*** *The sub-graph of Russian (orange), EU (olive green), US (violet) and Italian (blue) disinformation outlets.*

("breitbart.com", "naturalnews.com" and "infowars.com" to mention a few) as well as RU ("rt.com", "sputniknews.com" and associated networks in several countries), but

we also find interesting connections with notable disinformation outlets from France ("fdesouche.com" and "breizh-info.com"), Germany ("tagesstimme.com), Spain ("latribunadeespana.com") and even Sweden ("nyheteridag.se" and "samnytt.se"). Besides, a manual inspection of a few articles revealed that stories often originated in one country were immediately translated and promoted from outlets in different countries (see Fig. 3.12). Such findings suggest the existence of inter-connected deceptive strategies which span across several countries, consistently with claims in latest report by Avaaz [16] which revealed the existence of a network of far-right and anti-EU websites, leading to the shutdown of hundreds of Facebook pages with more than 500 million views just ahead of the elections. Far-right disinformation tactics comprised the massive usage of fake and duplicate accounts, recycling followers and bait and switch of pages covering topics of popular interest (e.g. sport, fitness, beauty).

It is interesting that Facebook decided on the basis of external insights to shutdown pages delivering misleading content and hate speech; differently from the recent past [138, 225, 226] it might signal that social media are more willing to take action against the spread of deceptive information in coordination with findings from third-party researchers. Nevertheless, we argue that closing malicious pages is not sufficient and more proactive strategies should be followed [16, 138].

In order to check the relevance of inter-connections with websites of different countries, we applied a simple degree preserving randomization [152] to the network depicted in Fig. 3.11B and tested whether the percentages of links respectively towards EU, US and RU were significantly different from the mean value observed in the random ensemble (obtained re-wiring the network for 1000 times). We thus performed a Z-test at $\alpha = 0.05/3$, rejecting the null hypothesis in all cases; in particular the number of RU and US connections are higher than expected whereas the number of EU connections is lower.

Finally, we performed a Mann-Kendall test to see whether there was an increasing trend, towards the elections, in the number of external connections with US and RU disinformation websites; we rejected it at $\alpha = 0.05/2 = 0.0025$.

3.5 Conclusions

We studied the reach of Italian disinformation on Twitter for a period of five months immediately preceding the European elections **(RQ1)** by analyzing the content production of websites producing disinformation, and the characteristics of users sharing malicious items on the social platform. Overall, thousands of articles–which included hoaxes, propaganda, hyper-partisan and conspiratorial news–were shared in the period preceding the elections. We observed that a few outlets accounted for most of the deceptive information circulating on Twitter; among them, we also encountered a

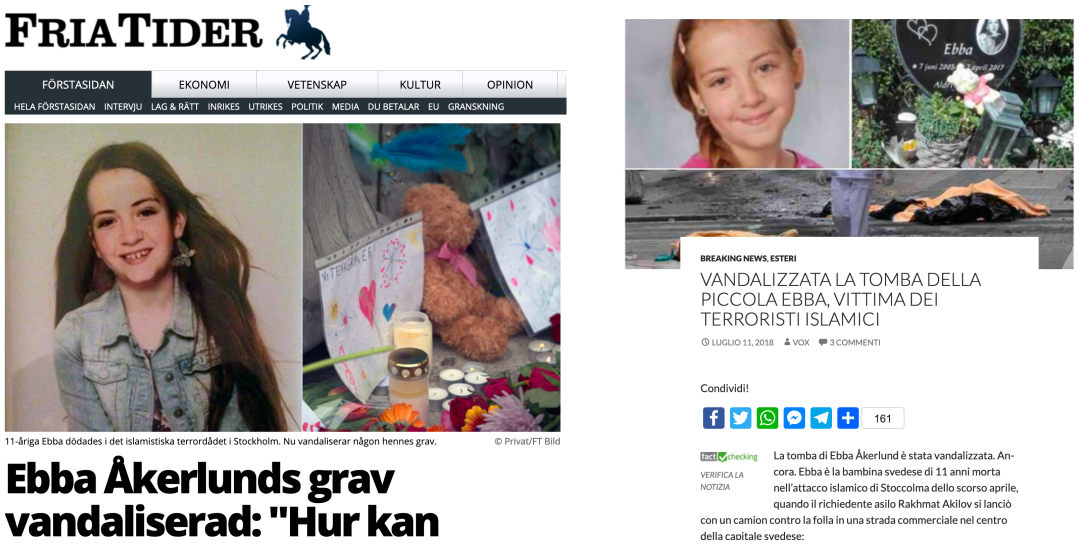


Figure 3.12: *An example of disinformation story who was published on a Swedish website ("friatider.se") and then reported by an Italian outlet ("voxnews.info"). Interestingly, this news is old (July 2018) but it was diffused again in the first months of 2019.*

few websites which were recently banned from Facebook after violating the platform's terms of use. We identified a heterogeneous yet limited community of thousands of users who were responsible for sharing disinformation. The majority of the accounts (more than 75%) occasionally engaged with malicious content, sharing less than 10 stories each, whereas only a few hundred accounts were responsible for (the spreading) of thousands of articles (see Fig 5).

We singled out the most debated topics of disinformation **(RQ2) by inspecting news items and Twitter hashtags.** We observed that they mostly concern polarizing and controversial arguments of the local political debate such as immigration, crime and national safety, whereas discussion around the topics of Europe global management had a negligible presence throughout the collection period; the lack of European topics was also reported in the agendaof mainstream media.

Then we identified the most influential accounts in the diffusion network resulting from users sharing disinformation articles on Twitter **(RQ3)**, so as to detect the presence of active groups with precise political affiliations. We discovered strong ties with the Italian far-right and conservative community, in particular with "Lega" party, as most of the users manifested explicit support to the party agenda through the use of keywords and hashtags. Besides, a common deceptive strategy was to passively involve his leader Matteo Salvini via mentions, quotes and replies as to potentially mislead his audience of million of followers. We found limited evidence of bot activity in the main core, and we observed that disabling a limited number of central users in the network would considerably reduce the spread of disinformation circulating on Twitter, but it would immediately raise censorship concerns.

Finally, we investigated inter-connections within different deceptive agents **(RQ4)**, thereby observing that they repeatedly linked to each other websites during the period preceding the elections. Moreover we discovered many cases where the same (or similar) stories were shared in different languages across different European countries.

This analysis confirms that disinformation is present on Twitter and that its spread shows some peculiarities in terms of themesbeing discussed and of political affiliation of the key members of the information spreading community. We are aware that disinformation news in Italy have a higher share on Facebook than Twitter and that the use of Twitter in Italy as a social channel is limited compared to other social platforms such as Facebook, WhatsApp or Instagram. Therefore similar studies on other social media platforms will be needed and beneficial to our understanding of the spread of disinformation.

CHAPTER 4

Understanding the COVID-19 infodemic on Twitter and Facebook

In this chapter we provide results from a systematic comparison of English language COVID-19 related posts on Twitter and Facebook throughout 2020. We refer the reader to the **Related Work** section for the related literature on the topic. The text is based and adapted from and [270].

4.1 Research contributions

We analyze the prevalence and diffusion of links to low-credibility content about the pandemic across two major social media platforms, Twitter and Facebook. We characterize cross-platform similarities and differences in popular sources, diffusion patterns, influencers, coordination, and automation.

Comparing the two platforms, we find divergence among the prevalence of popular low-credibility sources and suspicious videos. A minority of accounts and pages exert a strong influence on each platform. These misinformation “superspreaders” are often associated with the low-credibility sources and tend to be verified by the platforms. On both platforms, there is evidence of coordinated sharing of "infodemic" content.

The overt nature of this manipulation points to the need for societal-level solutions in addition to mitigation strategies within the platforms. However, we highlight limits

imposed by inconsistent data-access policies on our capability to study harmful manipulations of information ecosystems.

The main contributions of this study stem from exploring three sets of research questions:

- **RQ1:** What is the prevalence of low-credibility content on Twitter and Facebook? Are there similarities in how sources are shared over time? How does this activity compare to that of popular high-credibility sources? Are the same suspicious sources and YouTube videos shared in similar volumes across the two platforms?
- **RQ2:** Is the sharing of misinformation concentrated around a few active accounts? Do a few influential accounts dominate the resharing of popular misinformation? What is the role of verified accounts and those associated with the low-credibility sources on the two platforms?
- **RQ3:** Is there evidence of inauthentic coordinated behavior in sharing low-credibility content? Can we identify clusters of users, pages, or groups with suspiciously similar sharing patterns? Is low-credibility content amplified by Twitter bots more prevalent on Twitter as compared to Facebook?

We extract website links from social media posts that include COVID-19 related keywords. We identify a link with low-credibility content in one of two ways. First, we follow the convention of classifying misinformation at the source rather than the article level [138]. We do this by matching links to an independently-generated corpus of low-credibility website *domains* (or *sources*). Second, in the case of links to YouTube, we label videos as suspicious if they have been banned by the site or are otherwise unavailable to the public. This enables us to quantify the prevalence of individual uploads likely to propagate COVID-19 misinformation and the different ways in which they are shared on Twitter and Facebook.

4.2 Methodology

In this section we describe in detail the methodology employed in our analyses, allowing other researchers to replicate our approach. The outline is as follows: we collect social media data from Twitter and Facebook using the same keywords list. We then identify low- and high-credibility content from the tweets and posts automatically by tracking the URLs linking to the domains in a pre-defined list. Finally, we identify suspicious YouTube videos by their availability status.

4.2.1 Identification of low-credibility information

We focus on news articles linked in social media posts and identify those pertaining to low-credibility domains by matching the URLs to sources, following a corpus of

Table 4.1: *List of high-credibility sources.*

huffpost.com	newyorker.com
msnbc.com	newsweek.com
cnn.com	nytimes.com
economist.com	time.com
washingtonpost.com	reuters.com
apnews.com	npr.org
usatoday.com	wsj.com
foxnews.com	marketwatch.com
nypost.com	dailycaller.com
theblaze.com	dailywire.com
cdc.gov	who.int

literature [33, 100, 138, 184, 225]. We define our list of low-credibility domains based on information provided by the Media Bias/Fact Check website (MBFC[1]), an independent organization that reviews and rates the reliability of news sources. We gather the sources labeled by MBFC as having a "Very Low" or "Low" factual-reporting level. We then add "Questionable" or "Conspiracy-Pseudoscience" sources and we leave out those with factual-reporting levels of "Mostly-Factual," "High," or "Very High." We remark that although many websites exhibit specific political leanings, these do not affect inclusion in the list. The list has 674 low-credibility domains [269].

4.2.2 High-credibility sources

As a benchmark for interpreting the prevalence of low-credibility content, we also curate a list of 20 more credible information sources. We start from the list provided in a recent Pew Research Center report [159] and used in a few studies on online disinformation [193, 194], and we select popular news outlets that cover the full U.S. political spectrum. These sources have a MBFC factual-reporting level of "Mixed" or higher. In addition, we include the websites of two organizations that acted as authoritative sources of COVID-19 related information, namely the Centers for Disease Control and Prevention (CDC) and World Health Organization (WHO). For simplicity we refer to the full list in Table 4.1 as high-credibility sources.

4.2.3 Data collection

We collect data related to COVID-19 from both Twitter and Facebook. To provide a general and unbiased view of the discussion, we chose the following generic query terms: `coronavirus`, `covid` (to capture keywords like `covid19` and `covid-19`), and `sars` (to capture `sars-cov-2` and related variations).

[1] `mediabiasfactcheck.com`

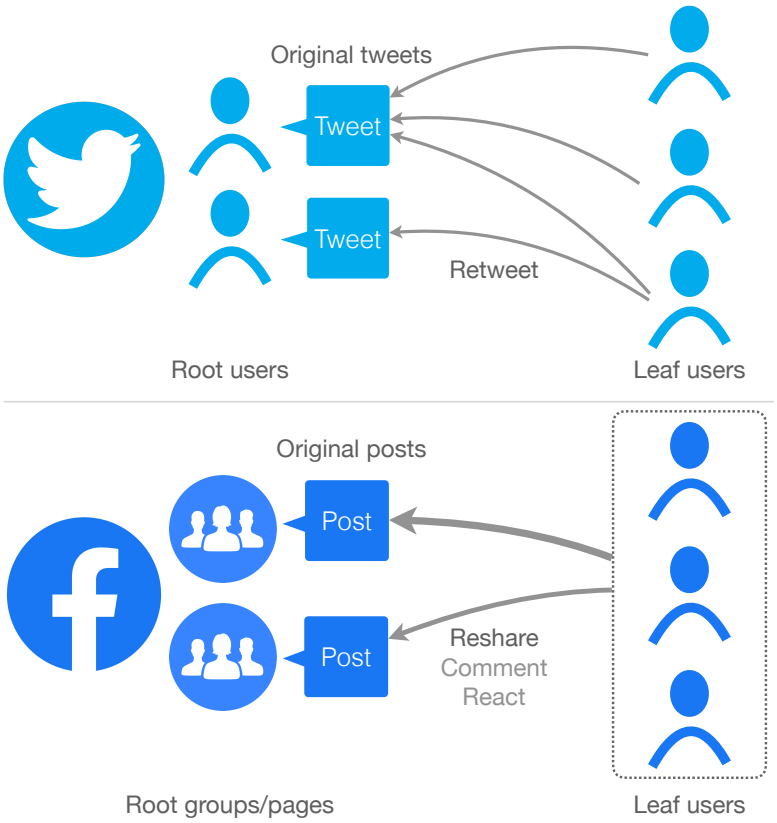


Figure 4.1: *Structure of the data collected from Twitter and Facebook. On Twitter, we have the information about original tweets, retweets, and all the accounts involved. On Facebook, we have information about original posts and public groups/pages that posted them. For each post, we also have aggregate numbers of reshares, comments, and reactions, with no information about the users responsible for those interactions.*

Twitter data.

Our Twitter data was collected using an API from the Observatory on Social Media [55], which allows to search tweets from the Decahose, a 10% random sample of public tweets. We searched for English tweets containing the keywords between Jan. 1 and Oct. 31, 2020, resulting in over 53M tweets posted by about 12M users. Note that since the Decahose samples tweets and not users, the sample of users in our Twitter dataset is biased toward more active users.

Our collection contains two types of tweets, namely original tweets and retweets. The content of original tweets is published by users directly, while retweets are generally used to endorse/amplify original tweets by others (no quoted tweets are included). We refer to authors of original tweets as "root" users, and to authors of retweets as "leaf" users (see Fig. 4.1).

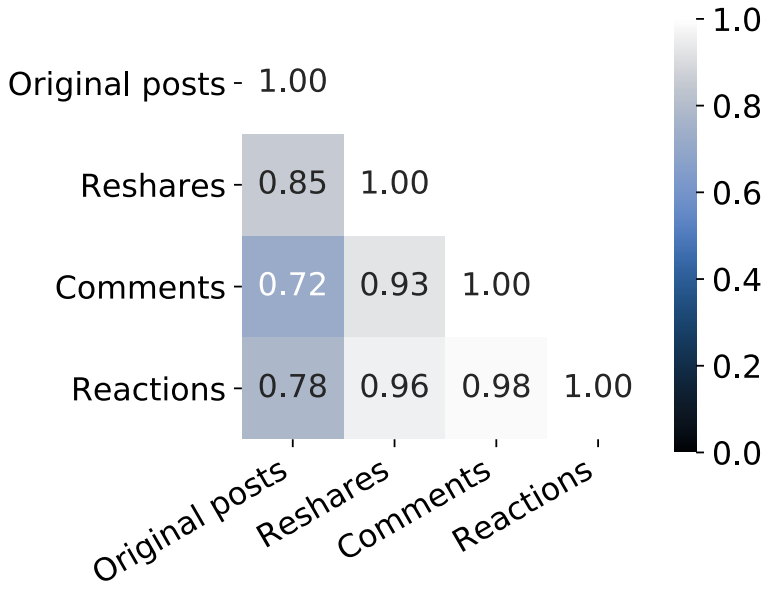


Figure 4.2: *Pearson correlation coefficients between Facebook metrics aggregated at the domain level for low-credibility domains. A reaction can be a "like," "love," "wow," "haha," "sad," "angry," or "care." All correlations are significant ($p < 0.01$).*

Facebook data.

We used the *posts/search* endpoint of the CrowdTangle API [54] to collect data from Facebook. We filtered the entire set of English posts published by public pages and groups in the period from Jan. 1 to Oct. 31, 2020 using the above list of keywords, resulting in over 37M posts by over 140k public pages/groups.

Our Facebook data collection is limited by the coverage of pages and groups in CrowdTangle, a public tool owned and operated by Facebook. CrowdTangle includes over 6M Facebook pages and groups: all those with at least 100k followers/members, U.S. based public groups with at least 2k members, and a very small subset of verified profiles that can be followed like public pages. We include these public accounts among pages and groups. In addition, some pages and groups with fewer followers and members are also included by CrowdTangle upon request from users. This might bias the dataset in ways that are hard to gauge. For example, requests from researchers interested in monitoring low-credibility pages and groups might lead to over-representation of such content.

As shown in Fig. 4.2, the collected data contains information about original Facebook posts and the pages/groups that published these posts. For each post, we also have access to *aggregate* statistics such as the number of reshares, comments, and reactions (e.g., "likes") by Facebook users. The numbers of comments and reactions are highly correlated with reshares (Fig. 4.2), so we focus on reshares in this study.

Similarly to Twitter, Facebook pages and groups that publish posts are referred to

as "roots" and users who reshare them are "leaves." However, in contrast to Twitter, we don't have access to any information about leaf users on Facebook. We refer generically to Twitter users and Facebook pages and groups as "accounts." To compare Facebook and Twitter in a meaningful way, we compare root users with root pages/-groups, original tweets with original posts, and retweet counts with reshare counts. We define *prevalence* as the sum of original tweets and retweets on Twitter, and as the sum of original posts and reshares on Facebook.

YouTube data.

We observed a high prevalence of links pointing to `youtube.com` on both platforms — over 64k videos on Twitter and 204k on Facebook. Therefore, we also provide an analysis of popular videos published on Facebook and Twitter. Specifically, we focus on popular YouTube videos that are likely to contain low-credibility content. An approach analogous to the way we label links to websites would be to identify sources that upload low-credibility videos and then label every video from those sources as misinformation. However, this approach is infeasible because the list of YouTube channels would be huge and fluid. To circumvent this difficulty, we use removal of videos by YouTube as a proxy to label low-credibility content. We additionally consider private videos to be suspicious, since this can be used as a tactic to evade the platform's sanctions when violating terms of service.

To identify the most popular and suspicious YouTube content, we first select the 16,669 videos shared at least once on both platforms. We then query the YouTube API *Videos:list* endpoint to collect their metadata and focus on the 1,828 (11%) videos that had been removed or made private. To validate this approach for identifying low credibility YouTube content, we follow a two-step manual inspection process for a sample of about 3% of the unavailable videos, comprising a mix of randomly selected and popular ones. We first search for the deleted video IDs in other YouTube videos and web pages. When these references contain the deleted videos' titles, we search for these titles on `bitchute.com` to find copies of the original videos. This process allows us to identify the narratives of 40 deleted videos, 90% of which contain misinformation. A similar approach was also adopted by [129] in their recent study of COVID-19 misinformation on YouTube.

4.2.4 Link extraction

Estimating the prevalence of low-credibility information requires matching URLs, extracted from tweets and Facebook metadata, against our lists of low- and high-credibility websites. As shortened links are very common, we also identified 49 link shortening services that appear at least 50 times in our datasets (Table 4.2) and expanded shortened

Table 4.2: *List of URL shortening services.*

bit.ly	dlvr.it	liicr.nl	tinyurl.com
goo.gl	ift.tt	ow.ly	fxn.ws
buff.ly	back.ly	amzn.to	nyti.ms
nyp.st	dailysign.al	j.mp	wapo.st
reut.rs	drudge.tw	shar.es	sumo.ly
rebrand.ly	covfefe.bz	trib.al	yhoo.it
t.co	shr.lc	po.st	dld.bz
bitly.com	crfrm.us	flip.it	mf.tt
wp.me	voat.co	zurl.co	fw.to
mol.im	read.bi	disq.us	tmsnrt.rs
usat.ly	aje.io	sc.mp	gop.cm
crwd.fr	zpr.io	scq.io	trib.in
owl.li			

Table 4.3: *Breakdown of Facebook and Twitter posts/tweets matched to low- and high-credibility domains.*

	Low-credibility	High-credibility
Facebook		
Original posts	303,119	1,194,634
Reshares	20,462,035	98,415,973
Twitter		
Original tweets	245,620	734,409
Retweets	653,415	2,184,050

URLs referring to these services through HTTP requests to obtain the actual domains. We finally match the extracted and expanded links against the lists of low- and high-credibility domains. A breakdown of matched posts/tweets is shown in Table 4.3. For low-credibility content, the ratio of retweets to tweets is 2.7:1, while the ratio of reshares to posts is 68:1. This large discrepancy is due to various factors: the difference in traffic on the two platforms, the fact that we only have a 10% sample of tweets, and the bias toward popular pages and groups on Facebook.

4.3 Results

4.3.1 Infodemic prevalence trends

We plot the daily prevalence of links to low-credibility sources on Twitter and Facebook in Fig. 4.3(a). The two time series are strongly correlated (Pearson $r = 0.87$, $p < 0.01$). They both experience a drastic growth during March, when the number of COVID-19 cases was growing worldwide. Towards summer, the prevalence of low-credibility information decreases to a relatively low level and then becomes more stable.

To analyze the Infodemic surge with respect to the pandemic's development and public awareness, Fig. 4.3(b) shows the worldwide hospitalization rate and the overall

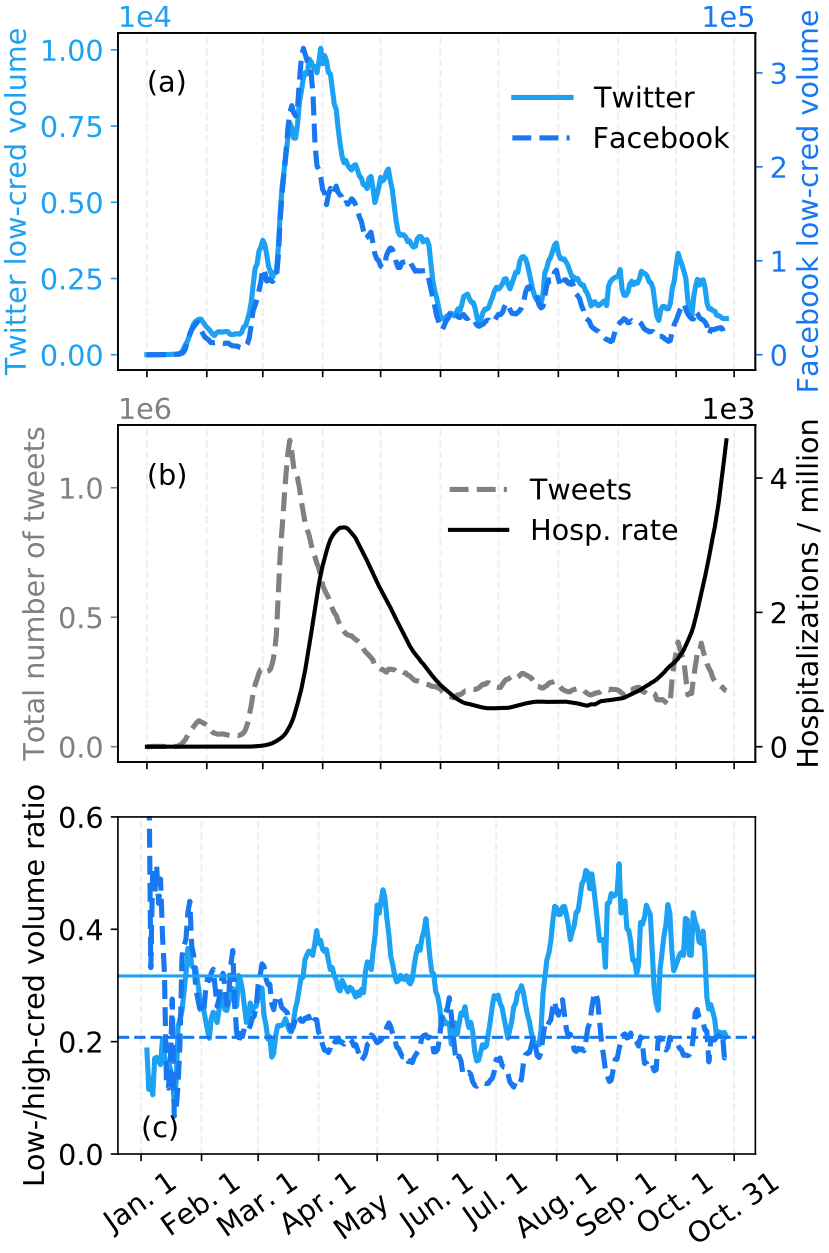


Figure 4.3: *Infodemic content surge on both platforms around the COVID-19 pandemic waves, from Jan. 1 to Oct. 31, 2020. All curves are smoothed via 7-day moving averages.* ***(a)*** *Daily volume of posts/tweets linking to low-credibility domains on Twitter and Facebook. Left and right axes have different scales and correspond to Twitter and Facebook, respectively.* ***(b)*** *Overall daily volume of pandemic-related tweets and worldwide COVID-19 hospitalization rates (data source: Johns Hopkins University).* ***(c)*** *Daily ratio of volume of low-credibility links to volume of high-credibility links on Twitter and Facebook. The noise fluctuations in early January are due to low volume. The horizontal lines indicate averages across the period starting Feb. 1.*

volume of tweets in our collection. The Infodemic surge roughly coincides with the general attention given to the pandemic, captured by the overall Twitter volume. The

peak in hospitalizations trails by a few weeks. A similar delay was recently reported between peaks of exposure to Infodemic tweets and of COVID-19 cases in different countries [88]. This plot suggests that the delay is related to general attention toward the pandemic rather than specifically toward misinformation.

To further explore whether the decrease in low-credibility information is organic or due to platform interventions, we also compare the prevalence of low-credibility content to that of links to credible sources. As shown in Fig. 4.3(c), the ratios are relatively stable across the observation period. These results suggest that the prevalence of low-credibility content is mostly driven by the public attention to the pandemic in general, which progressively decreases after the initial outbreak. We finally observe that Twitter exhibits a higher ratio of low-credibility information than Facebook (32% vs. 21% on average).

4.3.2 Infodemic prevalence of specific domains

We use the high-credibility domains as a benchmark to assess the prevalence of low-credibility domains on each platform. As shown in Fig. 4.4, we notice that the low-credibility sources exhibit disparate levels of prevalence. Low-credibility content as a whole reaches considerable volume on both platforms, with prevalence surpassing every single high-credibility domain considered in this study. On the other hand, low-credibility domains generally exhibit much lower prevalence compared to high-credibility ones (with a few exceptions, notably `thegatewaypundit.com` and `breitbart.com`).

4.3.3 Source popularity comparison

As shown in Fig. 4.4, we observe that low-credibility websites may have different prevalence on the two platforms. To further contrast their prevalence levels on Twitter and Facebook, we measure the popularity of websites on each platform by ranking them by prevalence, and then compare the resulting ranks in Fig. 4.5. The ranks on the two platforms are not strongly correlated (Spearman $r = 0.57$, $p < 0.01$). A few domains are much more popular or only appear on one of the platforms (see annotations in Fig. 4.5(a)). We also show the domains that are very popular on both platforms in Fig. 4.5(b). They are dominated by right-wing and state sources, such as `breitbart.com`, `washingtontimes.com`, `thegatewaypundit.com`, `oann.com`, and `rt.com`.

4.3.4 YouTube Infodemic content

Thus far, we examined the prevalence of links to low-credibility web page sources. However, a significant portion of the links shared on Twitter and Facebook point to YouTube videos, which can also carry COVID-19 misinformation. Previous work has

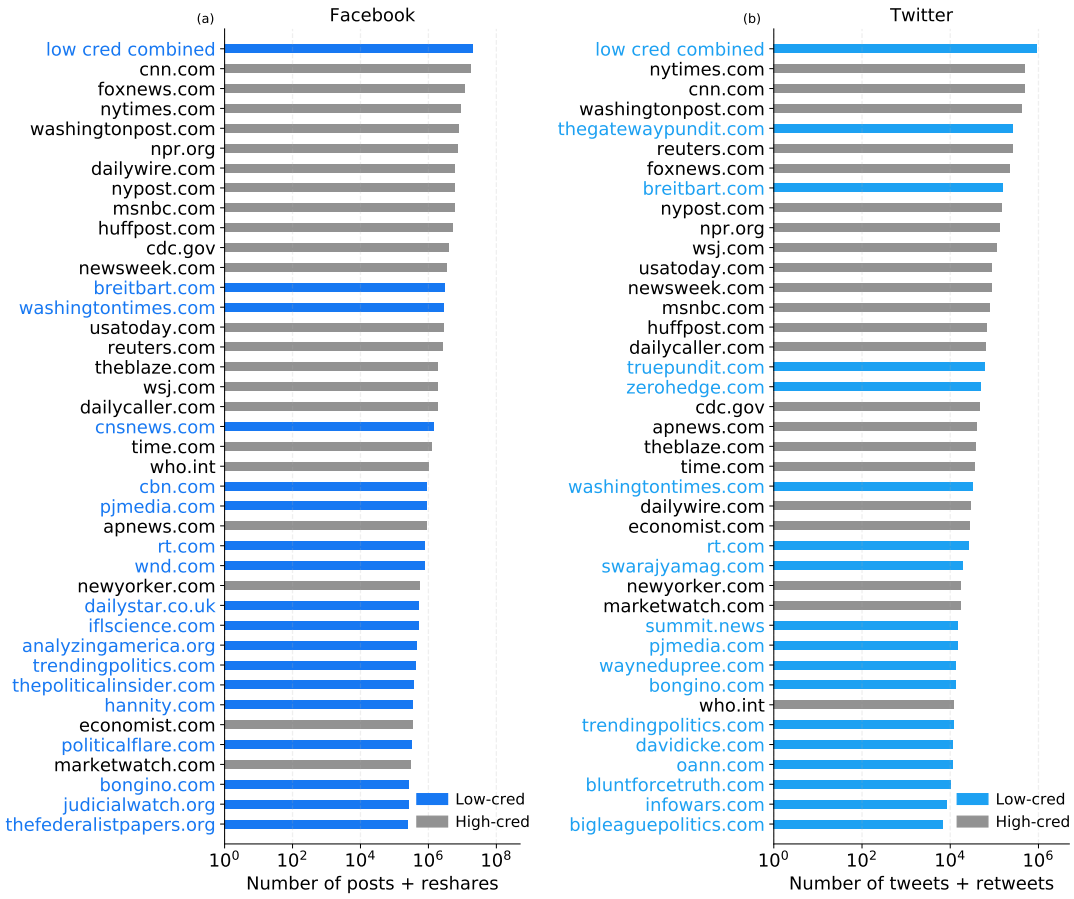


Figure 4.4: *Total prevalence of links to low- and high-credibility domains on both* ***(a)*** *Facebook and* ***(b)*** *Twitter. Due to space limitation, we only show the 40 most frequent domains on the two platforms. The high-credibility domains are all within the top 40. We also show low-credibility information as a whole (cf. "low cred combined").*

shown that bad actors utilize YouTube in this manner for their campaigns [263]. Specifically, anti-scientific narratives on YouTube about vaccines, Idiopathic Pulmonary Fibrosis, and the COVID-19 pandemic have been documented [67, 69, 95, 129].

To measure the prevalence of Infodemic content introduced from YouTube, we consider the unavailability (deletion or private status) of videos as an indicator of suspicious content, as explained in the Methods section. Fig. 4.6 compares the prevalence rankings on Twitter and Facebook for unavailable videos ranked within the top 500 on both platforms. These videos are linked between 6 and 980 times on Twitter and between 39 and 64,257 times on Facebook. While we cannot apply standard rank correlation measures due to the exclusion of low-prevalence videos, we do not observe a correlation in the cross-platform popularity of suspicious content from a qualitative inspection of the figure. A caveat to this analysis is that the same video content (sometimes re-edited) can be recycled within many video uploads, each having a unique video ID. Some of

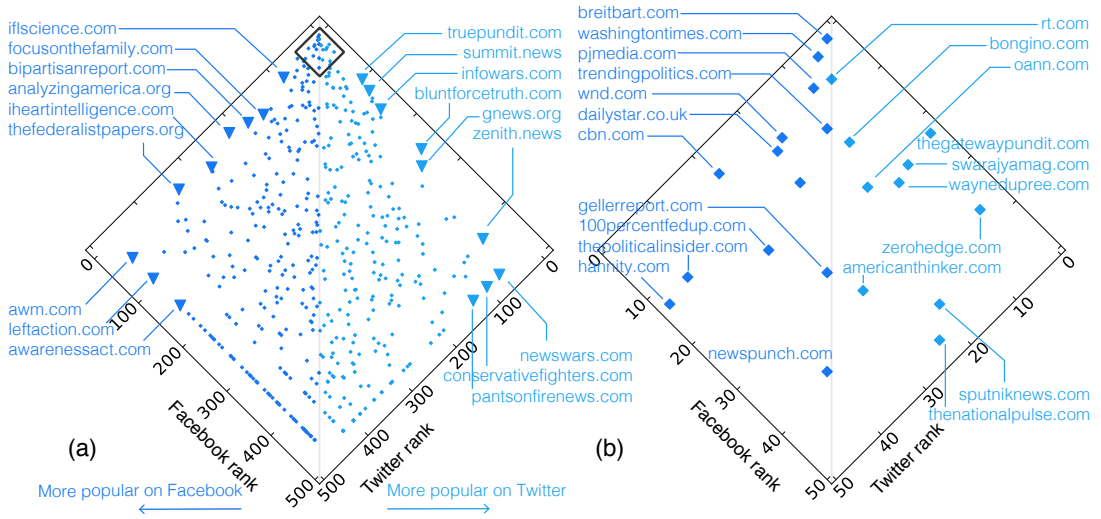


Figure 4.5: ***(a)*** *Rank comparison of low-credibility sources on Facebook and Twitter. Each dot in the figure represents a low-credibility domain. The most popular domain ranks first. Domains close to the vertical line have similar ranks on the two platforms. Domains close to the edges are much more popular on one platform or the other. We annotate a few selected domains that exhibit high rank discrepancy.* ***(b)*** *A zoom-in on the sources ranked among the top 50 on both platforms (highlighted square in (a)).*

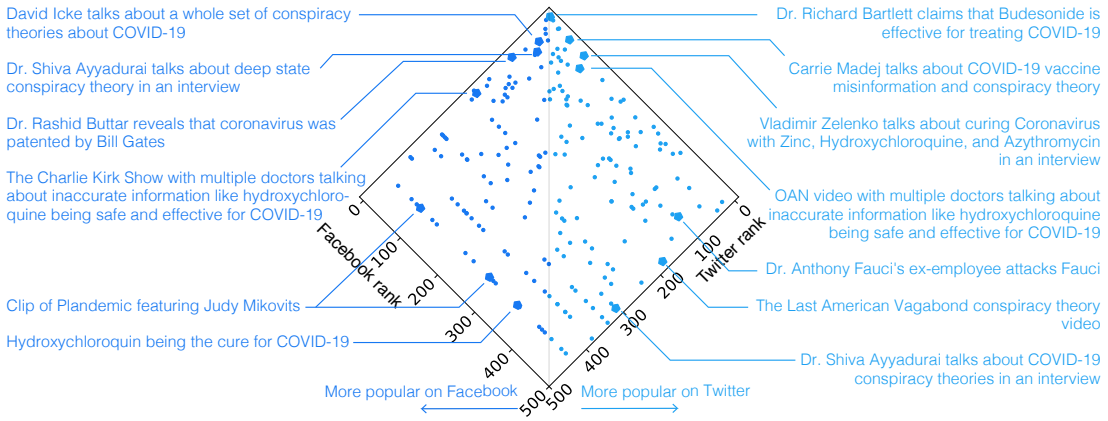


Figure 4.6: *Rank comparison of suspicious YouTube videos within the top 500 on both Facebook and Twitter. The most popular video ranks first. Each dot in the figure represents a suspicious video. Videos close to the vertical line have similar ranks on both platforms. Videos close to the edges are more popular on one platform or the other. We annotated a few selected videos with their narratives extracted from their copies on* `bitchute.com` *or other web pages.*

these videos are promptly removed while others are not. Therefore, the lack of correlation could partly be driven by YouTube's efforts to remove Infodemic content in conjunction with attempts by uploaders to counter those efforts [129].

Having looked at the prevalence of suspicious content from YouTube, we wish to explore the question from another angle: are videos that are popular on Facebook or Twitter more likely to be flagged as suspicious? Fig. 4.7 shows this to be the case on both platforms: a larger portion of videos with higher prevalence are unavailable,

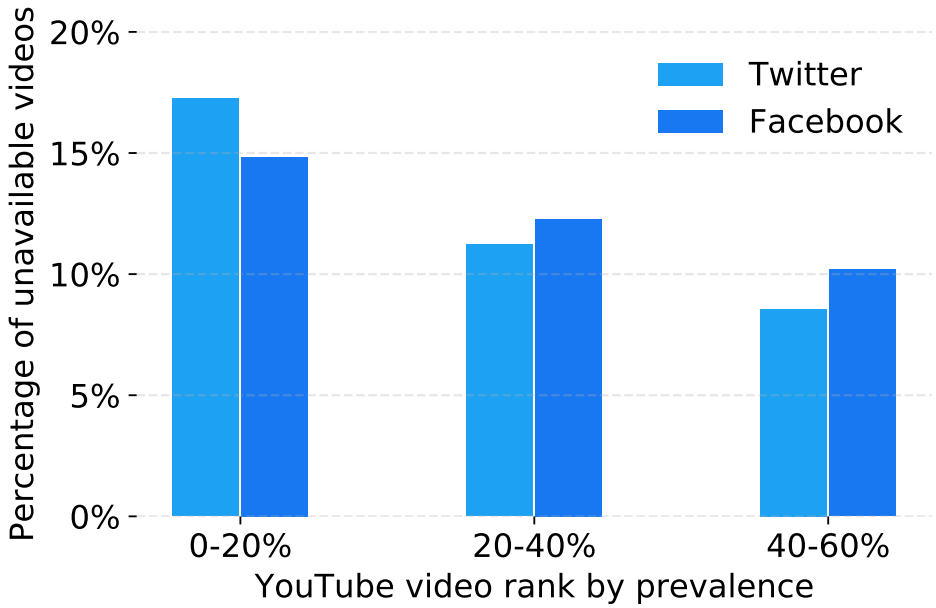


Figure 4.7: *Percentages of suspicious YouTube videos against their percent rank among all videos linked from pandemic-related tweets/posts on both Twitter and Facebook.*

but the trend is stronger on Twitter than on Facebook. The overall trend suggests that YouTube may have a bias toward moderating videos that attract more attention. This may be a function of the fact that an Infodemic video that is spreading virally on Twitter/Facebook may receive more abuse reports through YouTube's reporting mechanism. The fact that this trend is greater on Twitter may be explained by the differences between each platform's demographics. Survey data cited in the Discussion section shows that Twitter users are younger and more educated; it is therefore plausible that the average Twitter user is more likely to report unreliable content.

4.3.5 Infodemic spreaders

Links to low-credibility sources are published on social media through original tweets and posts first, then retweeted and reshared by leaf users. In this section, we study this dissemination process on Twitter and Facebook with a focus on the top spreaders, or "superspreaders": those disproportionately responsible for the distribution of Infodemic content.

Concentration of influence

We wish to measure whether low-credibility content originates from a wide range of users, or can be attributed to a few influential actors. For example, a source published 100 times could owe its popularity to 100 distinct users, or to a single root whose post is

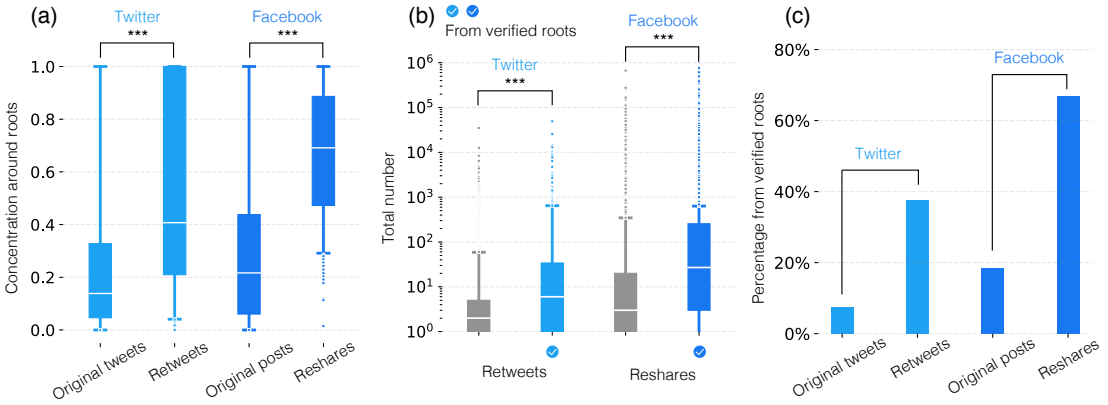


Figure 4.8: *Evidence of Infodemic superspreaders. Boxplots show the median (white line), 25th–75th percentiles (boxes), 5th–95th percentiles (whiskers), and outliers (dots). Significance of statistical tests is indicated by *** ($p < 0.001$). **(a)** Distributions of the concentration of original tweets, retweets, original posts, and reshares linking to low-credibility domains around root accounts. Each domain corresponds to one observation. **(b)** Distributions of the total number of retweets and reshares of low-credibility content posted by verified and unverified accounts. Each account corresponds to one observation. **(c)** Fractions of original tweets, retweets, original posts, and reshares by verified accounts.*

republished by 99 leaf users. To quantify the concentration of original posts/tweets or reshares/retweets for a source s, we use the inverse normalized entropy [170], defined as:

$$\mathcal{C}_s = 1 + \sum_{r=1}^{N_s} \frac{P_r(s) \log P_r(s)}{\log N_s},$$

where r represents a root user/group/page linking to source s, $P_r(s)$ stands for the fraction of posts/tweets or reshares/retweets linking to s and associated with r, and N_s is the total number of roots linking to s. Entropy measures how evenly quantities of content are distributed across roots; it is normalized to account for the varying numbers of roots in different cases. Its inverse captures concentration, and is defined in the unit interval. It is maximized at 1 when the content originates from a single root user/group/page, and minimized at 0 when each root makes an equal contribution. We set $\mathcal{C}_s = 1$ when $N_s = 1$.

Let us gauge the *concentration of activity* around root accounts through their numbers of original tweets/posts for each source. Similarly, we calculate the *concentration of popularity* around the root accounts using their numbers of retweets/reshares for each source. We show the distributions of these concentration variables in Fig. 4.8(a). On both platforms, we find that popularity is significantly more concentrated around root accounts compared to their activity ($p < 0.001$ for paired sample t-tests). This suggests the existence of superspreaders: despite the diversity of root accounts publishing links to low-credibility content on both platforms, only messages from a small group of influential accounts are shared extensively.

Who are the Infodemic superspreaders?

Both Twitter and Facebook provide verification of accounts and embed such information in the metadata. Although the verification processes differ, we wish to explore the hypothesis that verified accounts on either platform play an important role as top spreaders of low-credibility content. Fig. 4.8(b) compares the popularity of verified accounts to unverified ones on a per-account basis. We find that verified accounts tend to receive a significantly higher number of retweets/reshares on both platforms ($p < 0.001$ for Mann-Whitney U-tests).

We further compute the proportion of original tweets/posts and retweets/reshares that correspond to verified accounts on both platforms. Verified accounts are a small minority compared to unverified ones, i.e., 1.9% on Twitter and 4.5% on Facebook, among root accounts involved in publishing low-credibility content. Despite this, Fig. 4.8(c) shows that verified accounts yield almost 40% of low-credibility retweets on Twitter and almost 70% of reshares on Facebook.

These results suggest that verified accounts play an outsize role in the spread of Infodemic content. Are superspreaders all verified? To answer this question, let us analyse superspreader accounts separately for each low-credibility source. We extract the top user/page/group (i.e., the account with most retweets/reshares) for each source, and find that 19% and 21% of them are verified on Twitter and Facebook, respectively. While these values are much higher than the percentages of verified accounts among all roots, they show that not all superspreaders are verified.

Who are the top spreaders of Infodemic content? Table 4.4 answers this question for the 23 top low-credibility sources in Fig. 4.5(b). We find that the top spreader for each source tends to be the corresponding official account. For instance, about 20% of the retweets containing links to `thegatewaypundit.com` pertain to `@gatewaypundit`, the official handle associated with *The Gateway Pundit* website, on Twitter. (The `@gatewaypundit` account was suspended by Twitter in February 2021.) The remaining retweets have 10,410 different root users. Similarly, on Facebook, among all 2,821 pages/groups that post links to `thegatewaypundit.com`, the official page `@gatewaypundit` accounts for 68% of the reshares. We observe in Table 4.4 that most of the top low-credibility sources have official accounts on both Twitter and Facebook, which tend to be verified (71.4% on Twitter and 65.2% on Facebook). They are also the top spreaders of those domains in 16 out of 21 cases (76.2%) on Twitter and 18 out of 23 (78.3%) on Facebook.

4.3.6 Infodemic manipulation

Here we consider two types of inauthentic behaviors that can be used to spread and amplify COVID-19 misinformation: coordinated networks and automated accounts.

Table 4.4: *Official social media handles for the 23 top low-credibility sources from Fig. 4.5(b). Accounts with a checkmark (✓) are verified. Accounts with an asterisk (*) are the top spreaders for the corresponding domains. Accounts with a dagger (†) were suspended as of February 2021.*

Domain	Twitter handle	Facebook page/group
thegatewaypundit.com	@gatewaypundit ✓* †	@gatewaypundit ✓*
breitbart.com	@BreitbartNews ✓*	@Breitbart ✓*
zerohedge.com	@zerohedge *	@ZeroHedge *
washingtontimes.com	@WashTimes ✓*	@TheWashingtonTimes ✓
rt.com	@RT_com ✓*	@RTnews ✓*
swarajyamag.com	@SwarajyaMag ✓*	@swarajyamag ✓*
pjmedia.com	@PJMedia_com *	@PJMedia ✓*
waynedupree.com	@WayneDupreeShow ✓*	@WayneDupreeShow ✓*
bongino.com	@dbongino ✓*	@dan.bongino ✓*
trendingpolitics.com	–	@trendingpoliticsdotcom
oann.com	@OANN ✓*	@OneAmericaNewsNetwork ✓*
wnd.com	@worldnetdaily ✓*	@WNDNews *
sputniknews.com	@SputnikInt ✓*	@SputnikNews ✓*
dailystar.co.uk	@dailystar ✓*	@thedailystar ✓*
politicalflare.com	@nicolejames	@nicolejameswriter
thenationalpulse.com	@RaheemKassam ✓*	@thenationalpulse *
americanthinker.com	@AmericanThinker	@AmericanThinker *
gellerreport.com	@PamelaGeller ✓†	@pamelageller ✓
cbn.com	@CBNOnline ✓	@cbnonline ✓
100percentfedup.com	@100PercFEDUP *	@100PercentFEDUp *
newspunch.com	–	@thepeoplesvoicetv *
thepoliticalinsider.com	@TPInsidr	@ThePoliticalInsider ✓*
hannity.com	@seanhannity ✓*	@SeanHannity ✓*

Coordinated amplification of low-credibility content

Social media accounts can act in a coordinated fashion (possibly controlled by a single entity) to increase influence and evade detection [171, 175, 227]. We apply the framework proposed by [175] to identify coordinated efforts in promoting low-credibility information, both on Twitter and Facebook.

The idea is to build a network of accounts where the weights of edges represent how often two accounts link to the same domains. A high weight on an edge means that there is an unusually high number of domains shared by the two accounts. We first construct a bipartite graph between accounts and low-credibility domains linked in tweets/posts. The edges of the bipartite graph are weighted using TF-IDF [235] to discount the contributions of popular sources. Each account is therefore represented as a TF-IDF vector of domains. A projected co-domain network is finally constructed, with edges weighted by the cosine similarity between the account vectors.

We apply two filters to focus on active accounts and highly similar pairs. On Twitter, the users must have at least 10 tweets containing low-credibility links, and we retain edges with similarity above 0.99. On Facebook, the pages/groups must have at least 5 posts containing links, and we retain edges with similarity above 0.95. These thresholds are selected by manually inspecting the outputs.

Fig. 4.9 shows densely connected components in the co-domain networks for Twitter and Facebook. These clusters of accounts share suspiciously similar sets of sources. They likely act in a coordinated fashion to amplify Infodemic messages, and are possibly controlled by the same entity or organization. We highlight the fact that using a more stringent threshold on the Twitter dataset yields a higher number of clusters than a more lax threshold on the Facebook dataset. However, this does not necessarily imply a higher level of abuse on Twitter; it could be due to the difference in the units of analysis. On Facebook, we only have access to public groups and pages with a bias toward high popularity, and not to all accounts as on Twitter.

An examination of the sources shared by the suspicious clusters on both platforms shows that they are predominantly right-leaning and mostly U.S.-centric. The list of domains shared by likely coordinated accounts on Twitter is mostly concentrated on the leading low-credibility sources, such as `breitbart.com` and `thegatewaypundit.com`, while likely coordinated groups and pages on Facebook link to a more varied list of sources. Some of the amplified websites feature polarized rhetoric, such as defending against "attack by enemies" (see `www.frontpagemag.com/about`) or accusations of "liberal bias" (`cnsnews.com/about-us`), among others. Additionally, there are clusters on both platforms that share Russian state-affiliated media such as `rt.com` and an Indian right-wing magazine (`swarajyamag.com`).

In terms of the composition of the clusters, they mostly consist of users, pages, and

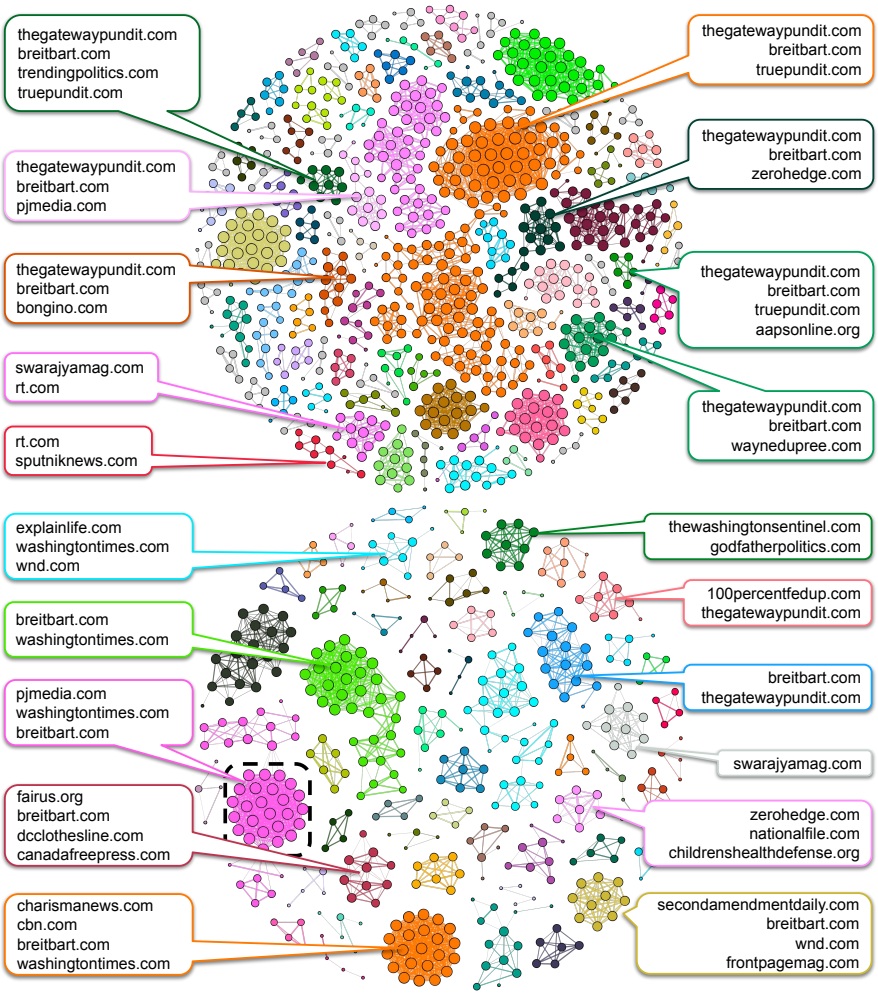


Figure 4.9: *Networks showing clusters that share suspiciously similar sets of sources on* ***(top)*** *Twitter and* ***(bottom)*** *Facebook. Nodes represent Twitter users or Facebook pages/groups. The size of the each node is proportional to its degree. Edges are drawn between pairs of nodes that share an unlikely high number of the same low-credibility domains. The edge weight represents the number of co-shared domains. The most shared sources are annotated for some of the clusters. Facebook pages associated with Salem Media Group radio stations are highlighted by a dashed box.*

groups that mention President Trump or his campaign slogans. Some of the Facebook clusters are notable because they consist of groups or pages that are owned by organizations with a wider reach beyond the platform, or that are given an appearance of credibility by being verified. Examples of the former are the pages associated with *The Answer* radio stations (highlighted in Fig. 4.9). These are among 100 stations owned by the publicly traded Salem Media Group, which also airs content on 3,100 affiliate stations. Examples of verified pages engaged in likely coordinated behavior are those affiliated with the non-profit Media Research Center, some of which have millions of followers. On Twitter, some of the clusters include accounts with tens of thousands of followers. Many of the suspicious accounts in Fig. 4.9 no longer exist.

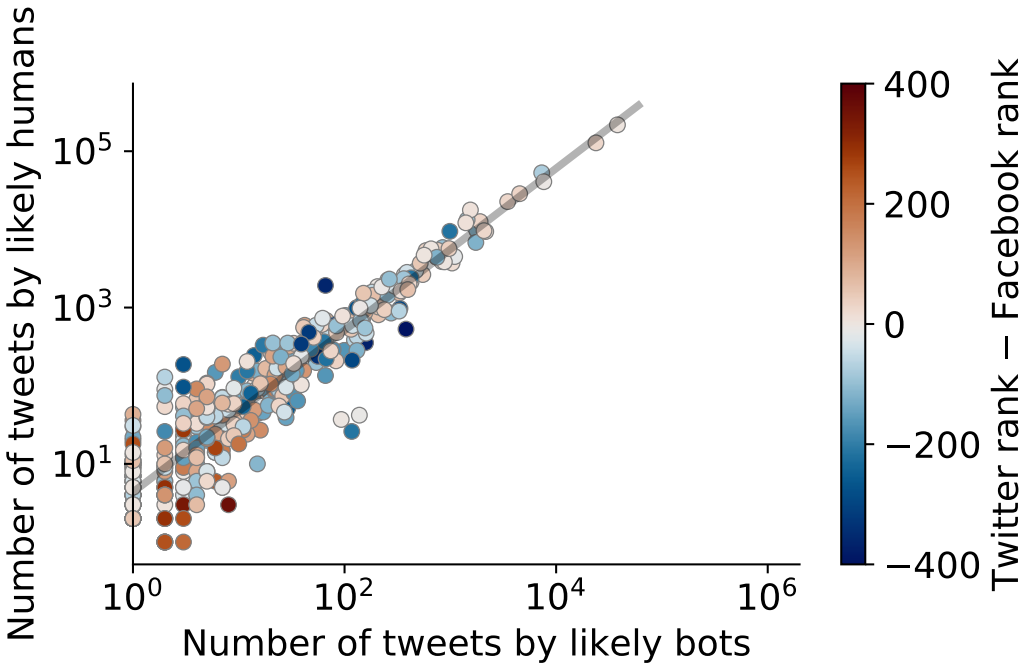


Figure 4.10: *Total number of tweets with links posted by likely humans vs. likely bots for each low-credibility source. The slope of the fitted line is 1.04. The color of each source represents the difference between its popularity rank on the two platforms. Red means more popular on Facebook, blue more popular on Twitter.*

Role of social bots

We are interested in revealing the role of inauthentic actors in spreading low-credibility information on social media. One type of inauthentic behavior stems from accounts controlled in part by algorithms, known as *social bots* [81]. Malicious bots are known to spread low-credibility information [225] and in particular create confusion in the online debate around health-related topics like vaccination [37].

We adopt BotometerLite (`rapidapi.com/OSoMe/api/botometer-pro`), a publicly-available tool that allows efficient bot detection on Twitter [272]. BotometerLite generates a bot score between 0 and 1 for each Twitter account; higher scores indicate bot-like profiles. To the best of our knowledge, there are no similar techniques designed for Facebook because insufficient training data is available. Therefore we limit this analysis to Twitter.

When applying BotometerLite to our Twitter dataset, we use 0.5 as the threshold to categorize accounts as likely humans or likely bots. For each domain, we calculate the total number of original tweets plus retweets authored by likely humans (n_h) and bots (n_b). We plot the relationship between the two in Fig. 4.10. The linear trend on the log-log plot signifies a power law $n_h \sim n_b^\gamma$ with exponent $\gamma \approx 1.04$, suggesting a weak

level of bot amplification (4%) [225].

While we are unable to perform automation detection on Facebook groups and pages, the ranks of the low-credibility sources on both platforms allow us to investigate whether sources with more Twitter bot activity are more prevalent on Twitter or Facebook. For each domain, we calculate the difference of its ranks on Twitter and Facebook and use the value of the difference to color the dots in Fig. 4.10. The results show that sources with more bot activity on Twitter are equally shared on both platforms.

4.4 Discussion

In this study, we provide the first comparison between the prevalence of low-credibility content related to the COVID-19 pandemic on two major social media platforms, namely Twitter and Facebook. Our results indicate that the primary drivers of low-credibility information tend to be high-profile, official, and verified accounts. We also find evidence of coordination among accounts spreading Infodemic content on both platforms, including many controlled by influential organizations. Since automated accounts do not appear to play a strong role in amplifying content, these results indicate that the COVID-19 Infodemic is an overt, rather than a covert, phenomenon.

We find that low-credibility content, as a whole, has higher prevalence than content from any single high-credibility source. However, there is evidence of differences in the misinformation ecosystems of the two platforms, with many low-credibility websites and suspicious YouTube videos at higher prevalence on one platform when compared to the other. Such a discrepancy might be due to a combination of the supply and demand factors. On the supply side, the official accounts associated with specific low-credibility websites are not symmetrically present on both platforms. On the demand side, the two platforms have very different user demographics. According to recent surveys, 69% of adults in the U.S. say they use Facebook, but only 22% of adults are on Twitter. Further, while Facebook usage is relatively common across a range of demographic groups, Twitter users tend to be younger, more educated, and have higher than average income. Finally, Facebook is a pathway to consuming online news for around 43% of U.S. adults, while the same number for Twitter is 12 [187, 264].

During the first months of the pandemic, we observe similar surges of low-credibility content on both platforms. The strong correlation between the timelines of low- and high-credibility content volume reveals that these peaks were likely driven by public attention to the crisis rather than by bursts of malicious content.

Our results provide us with a way to assess how effective the two platforms have been at combating the Infodemic. The ratio of low- to high-credibility information on Facebook is lower than on Twitter, suggesting that Facebook may be more effective.

On the other hand, we also find that verified accounts played a stronger role on Facebook than Twitter in spreading low-credibility content. However, the accuracy of these comparisons is subject to the different data collection biases. Suspicious YouTube uploads also exhibit an asymmetric prevalence between Facebook and Twitter. As stated previously, this may be partly a result of uploaders recycling sections of videos and uploading the content with a new video ID. Having such duplicates can mean that one version becomes popular on Facebook and another on Twitter, each potentially shared by a different demographic. This asymmetry might also be driven by Twitter users being more likely to flag videos. YouTube may then quickly remove reported videos before Facebook users have a chance to share them.

There are a number of limitations to our work. As we have remarked throughout the paper, differences between platform data availability and biases in sampling and selection make direct and fair comparisons impossible in many cases. The content collected from the Twitter Decahose is biased toward active users due to being sampled on a per-tweet basis. The Facebook accounts provided by CrowdTangle are biased toward popular pages and public groups, and data availability is also based upon requests made by other researchers. The small set of keywords driving our data collection pipeline may have introduced additional biases in the analyses. This is an inevitable limitation of any collection system, including the Twitter COVID-19 stream (`developer.twitter.com/en/docs/labs/covid19-stream/filtering-rules`). The use of source-level rather than article-level labels for selecting low-credibility content is necessary [138], but not ideal; some links from low-credibility sources may point to credible information. In addition, the list of low-credibility sources was not specifically tailored to our subject of inquiry. Finally, we do not have access to many deleted Twitter and Facebook posts, which may lead to an underestimation of the Infodemic's prevalence. All of these limitations highlight the need for cross-platform, privacy-sensitive protocols for sharing data with researchers [179].

Low-credibility information on the pandemic is an ongoing concern for society. Our study raises a number of questions. For example, user demographics might strongly affect the consumption of low-credibility information on social media: how do users in distinct demographic groups interact with different information sources? The answer to this question can lead to a better understanding of the Infodemic and more effective moderation strategies, but will require methods that scale with the nature of big data from social media. Another critical question is how social media platforms are handling the flow of information and allowing dangerous content to spread. Regrettably, since we find that high-status accounts play an important role, addressing this problem will prove difficult. As Twitter and Facebook have increased their moderation of COVID-19 misinformation, they have been accused of political bias. While there are many legal and ethical considerations around free speech and censorship, our work suggests that

these questions cannot be avoided and are an important part of the debate around how we can improve our information ecosystem.

CHAPTER 5

The impact of vaccine-related disinformation

In this chapter we provide results from two on-going projects where we investigate the interplay between vaccine-related disinformation spreading in online social networks and the nation-wide vaccination programs, respectively in the U.S. and in Italy. The former, in collaboration with the Observatory on Social Media of the Indiana University, is called CoVaxxy, and a public dashboard with visualizations of results is available at `osome.iu.edu/tools/covaxxy`. The latter is called VaccinItaly, and a dashboard is also available at `genomic.elet.polimi.it/vaccinitaly/`. The text in this chapter is mostly based on [62, 192, 195].

5.1 Background

The COVID-19 pandemic has killed over 4.55 million people and infected 219 million worldwide as of September 2021 [2]. Vaccination is the lynchpin of the global strategy to fight the SARS-CoV-2 coronavirus [126, 174]. Surveys conducted during February and March 2021 found high levels of vaccine hesitancy with around 40-47% of American adults were hesitant to take the COVID-19 vaccine [1, 86]. However, populations must reach a threshold vaccination rate to achieve herd immunity (i.e., 60-70%) [6, 94, 145]. Evidence of uneven distributions of vaccinations [43] raises the possibility of geographical clusters of non-vaccinated people [216]. In early July 2021, increased rates of the highly transmissible SARS-CoV-2 Delta variant were recorded

in several poorly vaccinated U.S. states [43]. These localised outbreaks will preclude eradication of the virus and may exacerbate racial, ethnic, and socioeconomic health disparities.

Vaccine hesitancy covers a spectrum of intentions, from delaying vaccination to outright refusal to be vaccinated [148]. Some factors are linked to COVID-19 vaccine hesitancy, with rates in the U.S. highest among three groups: African Americans, women, and conservatives [41]. Other predictors, including education, employment, and income are also associated with hesitancy [125]. Targeted messaging can be used to build confidence and address complacency in target groups [148], but these strategies are undermined by exposure to misinformation.

A number of studies discuss the spread of vaccine misinformation on social media [37] and argue that such campaigns have driven negative opinions about vaccines and even contributed to the resurgence of measles [40, 262]. In the COVID-19 pandemic scenario, widely shared misinformation includes false claims that vaccines genetically manipulate the population or contain microchips that interact with 5G networks [62, 113]. Exposure to online misinformation has been linked to increased health risks [88] and vaccine hesitancy [144]. Gaps remain in our understanding of how vaccine misinformation is linked to broad-scale patterns of COVID-19 vaccine uptake rates.

A possible driver for vaccine hesitancy is the anti-vaccination movement. This movement has been on the rise in the U.S. for two decades, beginning with unfounded fears over a Measles, Mumps and Rubella (MMR) vaccine [117]. The vocal online presence of the anti-vaccination movement has undermined confidence in vaccines. Worse, resistance to the COVID-19 vaccines is currently much more prevalent than resistance to the MMR vaccine. Since COVID-19 vaccine hesitancy and its drivers remains understudied, a goal of our project is to help address this gap.

There is a growing body of evidence linking social media and the antivaccination movement to vaccine hesitancy [37, 40, 122]. Studies show that vaccine hesitancy in one's peer group is associated with future vaccine refusal [39], and that misinformation spread on social networks is linked to poor compliance with public health guidance about COVID-19 [210].

Based on these findings, the core hypothesis behind our analyses is that the social spread of vaccine mis/disinformation and vaccine hesitancy will impact public health outcomes such as vaccine uptake and COVID mortality rates.

5.2 CoVaxxy

The goal of *CoVaxxy* is to track English-language discourse about COVID-19 vaccines on Twitter. We provide public access[1] to the data allowing researchers to study vaccine

[1] https://github.com/osome-iu/CoVaxxy

misinformation and hesitancy, and their relationship to public health outcomes. A public dashboard associated to the project is available at `https://osome.iu.edu/tools/covaxxy`.

We present results in that direction, currently under review [192], in the following section **Association between online misinformation and vaccine outcomes in the U.S.**.

The long-term aim of this project is to tackle the ambitious challenge of linking social media observations directly to public health. We hope that researchers will be able to leverage the *CoVaxxy* dataset to obtain a clearer picture of how vaccine hesitancy and misinformation affect health outcomes. In turn, such insight might enable public health officials to design better strategies for confronting vaccine hesitancy and refusal.

5.2.1 Identifying COVID-19 vaccines content on Twitter

To create as complete a set of Twitter posts related to COVID-19 vaccines as possible, we carefully select a list of keywords through a snowball sampling technique [49, 268]. We start with the two most relevant keywords, i.e., *covid* and *vaccine*, as our initial seeds. Keywords also match hashtags, URLs, and substrings. For example, *covid* matches "cnn.com/covid" and "#covid." Next, we gather tweets utilizing the filtered stream endpoint of the Twitter API[2] for three hours. From these gathered tweets, we then identify potential keywords that frequently co-occur with the seeds. These keywords are separately reviewed by two authors and added to the seed list if both agree that a keyword is related to our topic. This process was repeated six times between Dec. 15, 2020 and Jan. 2, 2021 with each iteration's data collection taking place at different times of the day to capture tweets from different geographic areas and demographics. The seed list serves as our initial keyword list.

We further refine the keyword list by manually combining certain keywords into composites, leveraging the query syntax of Twitter's filtered stream API. For example, using *covid19 pfizer* as a single composite matching phrase will capture tweets that contain *both* "covid19" *and* "pfizer." On the other hand, including *covid19* and *pfizer* as separate keywords will capture tweets that contain "covid19" *or* "pfizer," which we consider as too broad for our analysis. The final keyword list includes 76 (single or composite) keywords. Constructing various composites of relevant keywords in this way ensures the dataset is broad enough to include most relevant conversations while excluding tweets that are not related to the vaccine discussion.

[2] `https://developer.twitter.com/en/docs/twitter-api/v1/tweets/filter-realtime/overview`

5.2.2 Content Coverage

To demonstrate the effectiveness of the snowball sampling technique introduced above, we calculate the popularity of each keyword in the final list by the number of unique tweets and unique users associated with it.

Figure 5.1, where keywords are ranked by popularity, shows that additional keywords beyond the 60 most popular ones tend to capture very small numbers of users and tweets, relative to other keywords in the collection. This suggests that including more keywords in the seed list described above is not likely to alter the size and structure of the dataset significantly. In fact, the inclusion of additional keywords could be redundant, due to the co-occurrence of multiple keywords and hashtags in a single tweet, especially for the most popular terms. Thus, we believe that our set of keywords provides reasonable coverage and is representative of tweets communicating about COVID-19 vaccines.

As the collection of tweets is intended to persist over time, new relevant keywords may emerge. To ensure that the keyword list remains comprehensive throughout the entire data collection period, our team will continue to monitor the ongoing public discussion related to COVID-19 vaccinations and update the list with important emerging keywords, if necessary.

5.2.3 Data Collection Architecture

Our server architecture (Figure 5.2) is designed to collect and process large quantities of data. This infrastructure is hosted by Extreme Science and Engineering Discovery Environment (XSEDE) Jetstream virtual machines (VMs) [237, 244]. To maintain the integrity of our tweet streaming pipeline, we have incorporated redundancy. We maintain two *streamer* (stream collection) VMs in different U.S. states so that if one suffers a fault we can use data from the other. These servers connect to Twitter's filtered stream API to collect tweets that match any of the keywords in real time. We use the language metadata to filter out non-English tweets.

The data from the two streamers is collated on a general purpose server VM where we run data analysis. The server VM is also linked to Indiana University's high performance computing infrastructure for running advanced analyses.

We upload new data files to a public data repository [63] each day[3] and will continue to do so as long as the topic of COVID-19 vaccinations remains relevant in public discourse. This repository also includes our list of keywords. In compliance with Twitter's Terms, we are only able to share tweet IDs with the public. One can rehydrate the dataset by querying the Twitter API or using tools like Hydrator[4] or twarc[5].

[3] https://doi.org/10.5281/zenodo.4526494
[4] https://github.com/DocNow/hydrator
[5] https://github.com/DocNow/twarc

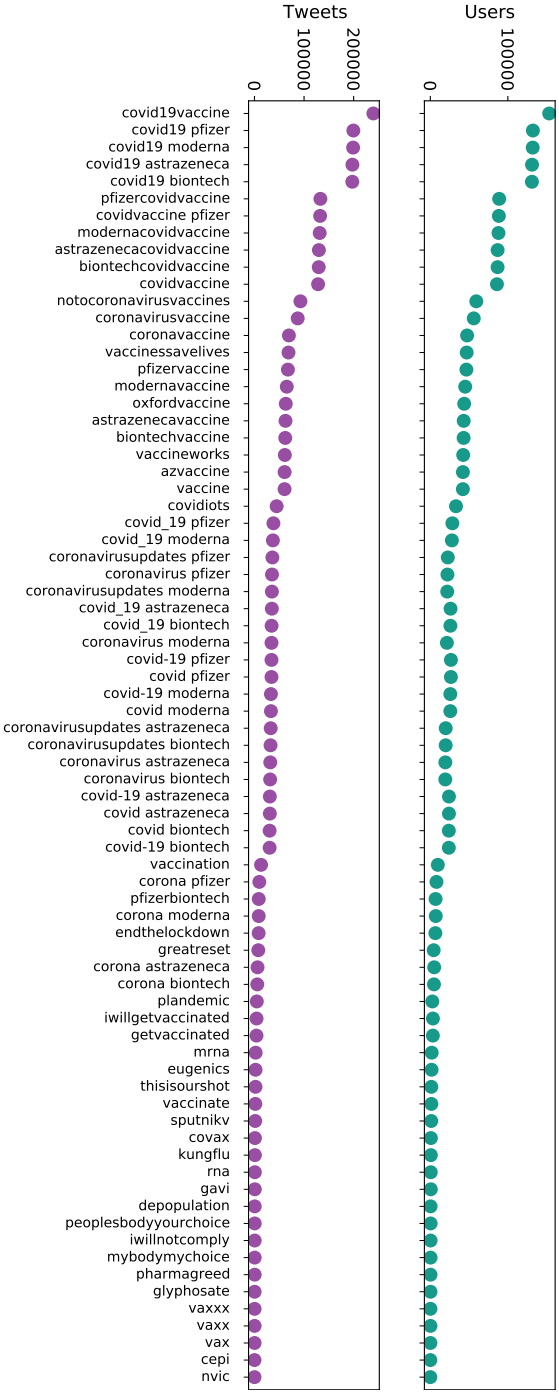


Figure 5.1: *Number of tweets (purple, left) and users (green, right) captured by each keyword/phrase in the final list (ranked by popularity) between January 4–11, 2021.*

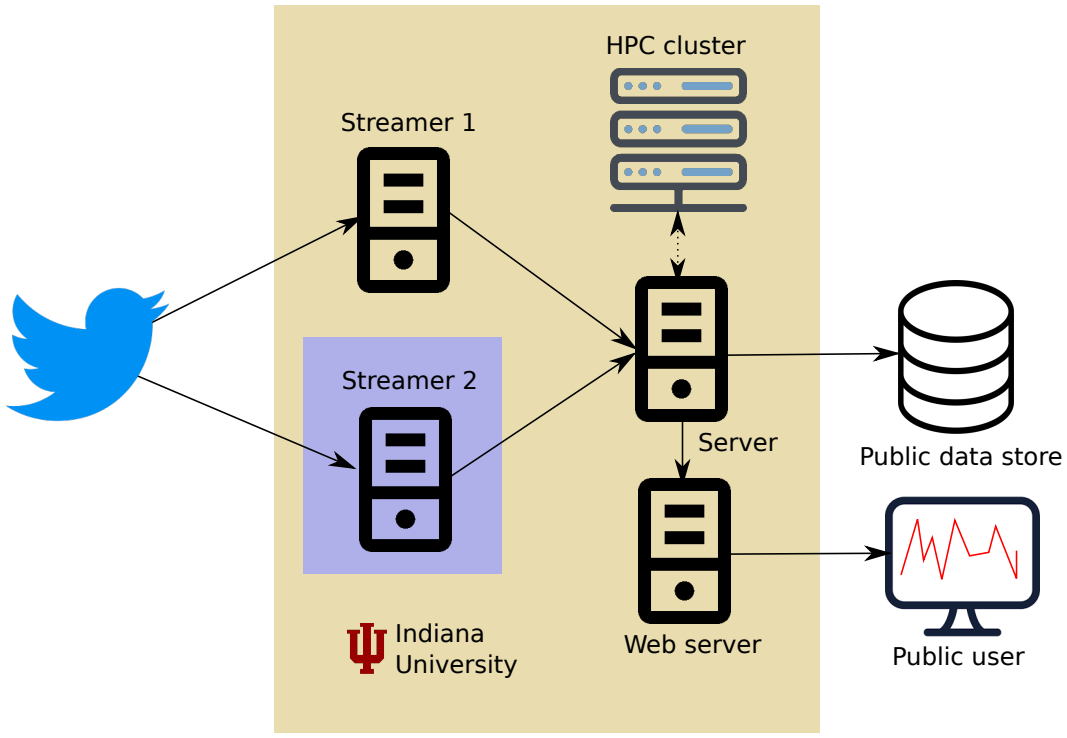


Figure 5.2: *The VM server architecture for the CoVaxxy project. Data flows in the direction of the arrows. Machines in the larger yellow box are hosted by Indiana University. The VM "Streamer 2," in the embedded blue box, is hosted by the Texas Advanced Computing Center.*

Finally, a web server provides access to the data on the server VM through the interactive *CoVaxxy* dashboard, described next.

5.2.4 Dashboard

Existing COVID-19 visualization tools include those by Johns Hopkins University [66] and The Atlantic.[6] These trackers address hospitalization and mortality. Another dashboard from the Fondazione Bruno Kessler reports on the proportions of misinformation and epidemic-related statistics (e.g., confirmed cases and deaths) per country.[7] Finally, the Our World in Data COVID-19 vaccination dataset publishes vaccine uptake information by country.[8]

We are not aware of any tools that concurrently explore the relationships between COVID-19 vaccine conversations, vaccine uptake, and epidemic trends. Consequently, we have created a web-based dashboard to fill this void. The *CoVaxxy* dashboard will track and quantify credible information and misinformation narratives over time, as well as their sources and related popular keywords.[9] Although we collect English tweets related to vaccines globally, the dashboard provides state-level statistics in the

[6] https://covidtracking.com/
[7] https://covid19obs.fbk.eu
[8] https://ourworldindata.org/covid-vaccinations
[9] https://osome.iu.edu/tools/covaxxy

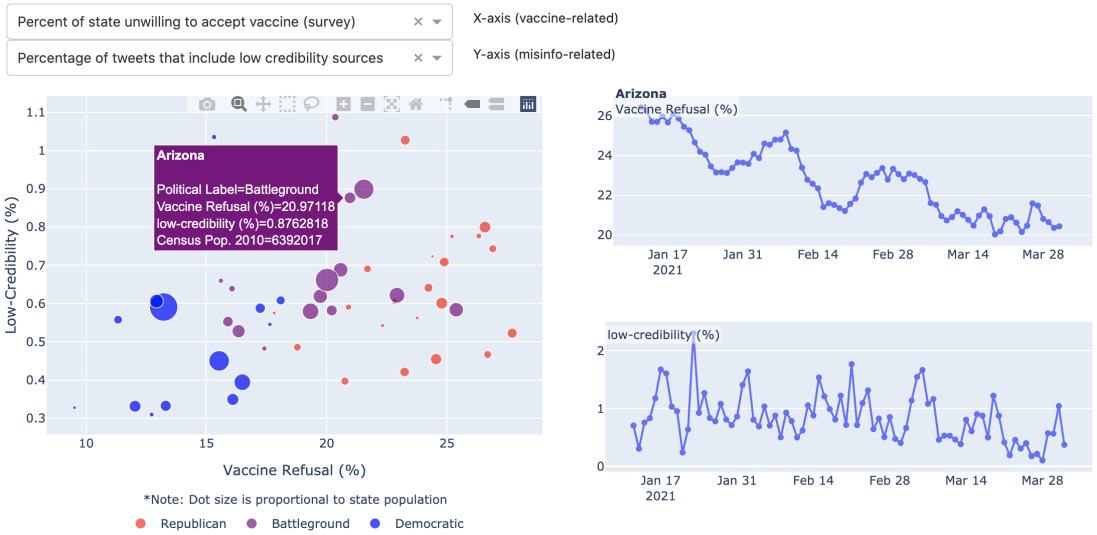


Figure 5.3: *Example visualization from the CoVaxxy web dashboard. This visualization lets users plot relationships (at the state-level) between vaccine-related and misinformation-related data. The left figure's axes are selected from the dropdowns, displaying the aggregate relationship. The two figures on the right illustrate the same relationship from a temporal perspective for an individual state. The user chooses which state to visualize in the figures on the right by hovering over a dot within the left figure.*

United States. Additionally, it shows global hashtag and domain sharing trends. It is updated daily. Figure 5.3 illustrates one example of an interactive visualization that lets users visualize the relationship between various misinformation-related and COVID-19 pandemic data. This data will be displayed alongside COVID-19 pandemic and vaccine trends. By highlighting the connection between misinformation and public health actions and outcomes, we hope to encourage the public to be more vigilant about the information they consume from their daily social media feeds in the fight against COVID-19.

5.2.5 Limitations

This dataset has some key limitations. Critically, Twitter users are not a representative sample of the population, nor are their posts a representative sample of public opinions [264]. Additionally, filtering our stream to include only English-language tweets comes at the price of occasionally excluding some variants of this language. This is because our stream gathers tweets that have been marked as containing English by Twitter's automatic language identification system, which may not capture some tweets by minority dialect speakers and multilingual speakers [123].

The Twitter Filtered Stream API imposes a rate limitation of 1% of all public tweets, which could limit our ability to capture all relevant content in the future. Fortunately, if this happens, Twitter provides the number of tweets not delivered within our stream.

Another potential source of bias is the keyword sampling procedure used to identify and collect COVID-19 vaccine related content, which involved evaluation of keywords to determine what was relevant. We are unable to fully exclude irrelevant content using only keyword-based filtering. However, further filtering is possible at a later stage. Other researchers may also refine the data to properly address their own topics of interest.

Given the large-scale, real-time nature of our data collection infrastructure, users do not have the ability to opt-out. This raises important ethical concerns related to anonymity. To address this concern, we note that (1) our dashboard only displays aggregate data, obfuscating the ability of users to identify those captured within our data; and (2) should a user delete a tweet or account, the related information will not be returned by Twitter during the re-hydration process.

5.3 Association between online misinformation and vaccine outcomes in the U.S.

The Pfizer-BioNTec COVID-19 vaccine was the first to be given U.S. Food and Drug Administration approval on December 10th 2020 [4]. Since then, two other vaccines have been approved in the U.S. Until recently, vaccines have been selectively administered with nationwide priority being given to more vulnerable cohorts such as the more elderly members of the population. As vaccines become available to the entire adult population [53], adoption will be driven by limits in demand rather than in supply. It is therefore important to study the variability in uptake across U.S. states and counties, as reflected in recent surveys [76, 137].

In this work we study relationships between vaccine uptake, vaccine hesitancy and online misinformation. We measure vaccine uptake from the daily vaccination rates recorded by the Centers for Disease Control and Prevention (CDC) [43] for each U.S. state averaged over the week of March 19 to 25, 2021, when variability across U.S. states became apparent [53]. Vaccine hesitancy is likely to affect uptake rates, so we specify a longer time window to measure that variable, Jan 4th to March 25th, 2021, and likewise for online misinformation. We leverage over 22 M individual responses to surveys administered on Facebook to assess vaccine hesitancy rates [76], and we identify online misinformation by focusing on low-credibility sources shared on Twitter [33,100,137,225] by over 1.67M users geolocated within U.S. regions (see next section for details on the methodology).

For statistical analysis, we use multivariate regression models adjusting for socioeconomic, demographic and political confounding factors. The variables are recorded at group level, which makes drawing inference at the individual level problematic; however, we account for likely issues using interaction variables, logarithmic transforms,

heteroskedasticity tests, clustering at multiple levels (county and state), and uncertainty weighting of variables. Finally, to investigate whether there is evidence for a directional effect from misinformation onto vaccine hesitancy, we perform a Granger causality analysis. Supplementary material is available in Appendix B.

5.3.1 Methods

Our key independent variable is the mean percentage of vaccine-related misinformation shared via Twitter at the U.S. state or county level. We used 55 M tweets from the CoVaxxy dataset [64], which were collected between Jan 4th and March 25th using the Twitter filtered stream API using a comprehensive list of keywords related to vaccines (see previous section). We leveraged the carmen library [68] to geolocate almost 1.67 M users residing in 50 U.S. states, and a subset of approximately 1.15 M users residing in over 1,300 counties. The larger set of users accounts for a total of 11 M shared tweets. Following a consolidated approach in the literature [33, 100, 138, 225], we identified misinformation by considering tweets that contained links to news articles from a list of low-credibility websites compiled by a politically neutral third-party (see Supplementary Information in Appendix B). We measured the prevalence of misinformation about vaccines in each region by (i) calculating the proportion of vaccine-related misinformation tweets shared by each geo-located account; and (ii) taking the average of this proportion across accounts within a specific region. The Twitter data collection was evaluated and deemed exempt from review by the Indiana University IRB (protocol 1102004860).

Our dependent variables include vaccination uptake rates at the state level and vaccine hesitancy at the state and county levels. Vaccination uptake is measured from the number of daily vaccinations administered in each state during the week 19th-25th March 2021, and measurements are derived from the CDC [43]. Vaccine hesitancy rates are based on Facebook Symptom Surveys provided by the Delphi Group [76] at Carnegie Mellon University in the period Jan 4th-March 25th 2021. We computed hesitancy by taking the complementary proportion of individuals "who either have already received a COVID vaccine or would definitely or probably choose to get vaccinated, if a vaccine were offered to them today." See Supplementary Information in Appendix B for further details.

There are no missing vaccine-hesitancy survey data at the state level. Observations are missing at the county level because Facebook survey data are available only when the number of respondents is at least 100. We use the same threshold on the minimum number of Twitter accounts geolocated in each county, resulting in a sample size of $N = 548$ counties.

Our multivariate regression models adjust for six potential confounding factors. These include percent of the population below the poverty line, percent aged 65+, per-

cent of residents in each racial and ethnic group (Asian, Black, Native American, and Hispanic; White non-Hispanic is omitted), rural-urban continuum code (RUCC, county level only), number of COVID-19 deaths per thousand, and percent republican vote (in 10 percent units). Other covariates (listed in supplementary table S9 in Appendix B) were considered but dropped due to non-significance and/or multicollinearity (i.e., high variance inflation factors).

We also conduct a large number of sensitivity analyses, including different specifications of the misinformation variable (with a restricted set of keywords and different thresholds for the inclusion of Twitter accounts) as well as logged versions of misinformation (to correct positive skew). These results are presented in Supplementary Information (Tables S3-S8, see Appendix B).

We conduct multiple regression models predicting vaccination rate and vaccine hesitancy. Both dependent variables are normally distributed, making weighted least squares regression the appropriate model. Data are observed (aggregated) at the state or county level rather than at the individual level. Analytic weights are applied to give more influence to observations calculated over larger samples. The weights are inversely proportional to the variance of an observation such that the variance of the j-th observation is assumed to be $\frac{\sigma^2}{w_j}$ where w_j is the weight. The weights are set equal to the size of the sample from which the average is calculated. We estimate weighted regression with the *aweights* command in Stata 16. In addition, because counties are nested hierarchically in states, we use cluster robust standard errors to correct for lack of independence between county-level observations.

We investigate Granger causality between vaccine hesitancy and misinformation by comparing two auto-regressive models. The first considers daily vaccine hesitancy rates x at time t in geographical region r (state or county):

$$x_{t,r} = \sum_{i}^{n} a_i x_{t-i,r} + \epsilon_{t,r}$$

where n is the length of the time window. The second model adds daily misinformation rates per account as an exogenous variable y:

$$x_{t,r} = \sum_{i}^{n} (a_i x_{t-i,r} + b_i y_{t-i,r}) + \epsilon'_{t,r}$$

The variable y is said to be Granger causal [99, 104] on x if, in statistically significant terms, it reduces the error term $\epsilon_{t,r}$, i.e., if $E_{a,b} = \sum_{t,r} \epsilon^2_{t,r} - \epsilon'^2_{t,r} > 0$, meaning that misinformation rates y help forecast hesitancy rates x. We assume geographical regions to have equivalence and independence in terms of the way misinformation influences vaccine attitudes. Thus, we use the same parameters for a_i and b_i across all regions.

We employ Ordinary Least Squares (using the Python `statsmodels` package version 0.11.1) linear regression to fit a and b, standardizing the two variables and removing trends in the time series of each region. We select the value of the time window n which maximises $E_{a,b}$. For both counties and states, this was $n = 6$ days. We use data points with at least 1 tweet and at least 100 survey responses for every day in the time window for the specified region.

The traditional statistic used to assess the significance of Granger Causality is the F-statistic [99]. However, in our case, there are several reasons why this is not appropriate. First, we have missing time-windows in some of our regions. Second, our assumptions of equivalence and independence for regions may not be accurate. For these reasons, we use a bootstrap method to estimate the expected random distribution of $E_{a,b}$ with the time signal removed. This is done by generating trial surrogates for y by randomly shuffling the data points. The reduction in error (which we call $E*_{a,b}$) is recalculated for each trial. The significance or P-value of our Granger Causality analysis is then given by the proportion of trials ($N = 10,000$) for which $E*_{a,b} > E_{a,b}$.

5.3.2 Results

Looking across U.S. states, we observe a negative association between vaccination uptake rates and online misinformation (Pearson R = –0.49, P < .001). Investigating covariates known to be associated with vaccine uptake or hesitancy, we find that an increase in the mean amount of online misinformation is significantly associated with a decrease in daily vaccination rates per million (b = –3518.00, P = .009, Fig.5.4a, and see Methods and Table S1 in Supplementary Information). Political partisanship (a 10% increase in GOP vote) is also strongly associated with vaccination rate (b = –640.32, P = .004). These two factors alone explain nearly half the variation in state-level vaccination rates, and are themselves moderately correlated (Fig. S1 and Table S1 in the Supplementary Information in Appendix B), consistent with prior research [169]. Remaining covariates, including religiosity, unemployment rate, and population density, are non-significant and/or collinear with other variables and thus dropped for parsimony.

To investigate vaccine hesitancy, we leverage over 22 M individual responses to daily survey data provided by Facebook [76] (see Methods). Reports of vaccine hesitancy are aggregated at the state level (i.e., percent hesitant) and weighted by sample size. We find a strong negative correlation between vaccine uptake and hesitancy across U.S. states (Pearson R = –0.71, P < .001, Fig. S1 in Supplementary Information), suggesting that daily vaccination rates largely reflect demand for vaccines rather than supply. Taking into account the same set of potential confounding factors in a weighted regression model, we find a significant positive association between misinformation (b = 6.88, P = .007) and state-level vaccine hesitancy, and between political partisanship (b

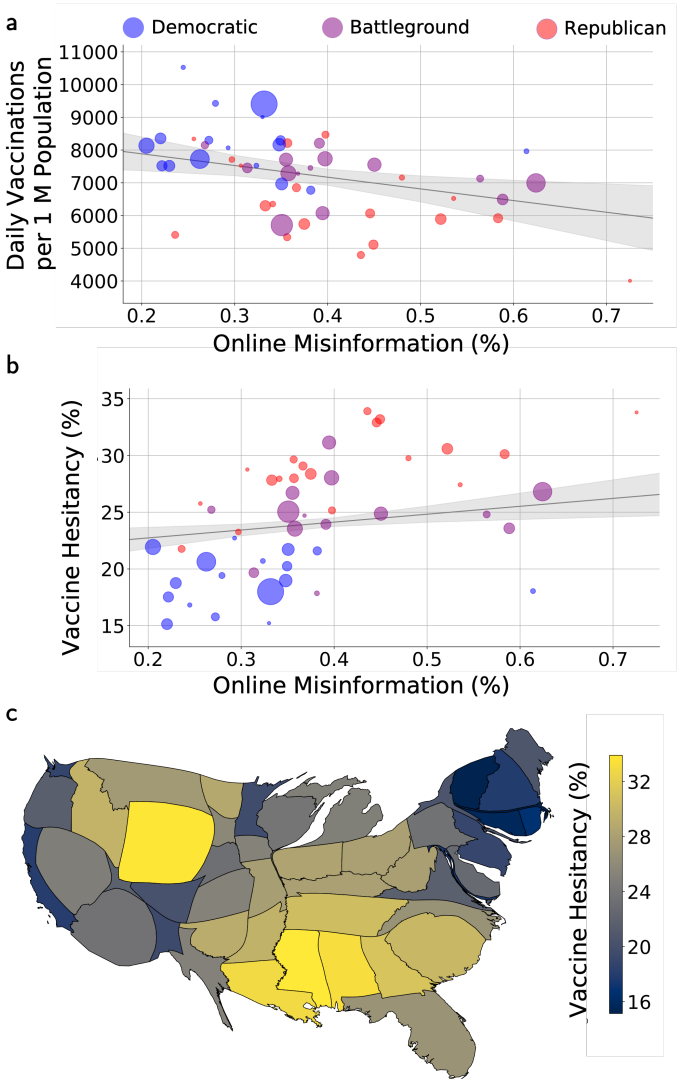


Figure 5.4: *Online misinformation is associated with vaccination uptake and hesitancy at the state level. (a) State-level mean daily vaccinations per million population during the period from March 19 to 25, 2021, against the average proportion of vaccine misinformation tweets shared by geolocated users on Twitter during the period from Jan 4th to March 25th, 2021. (b) Levels of state-wide vaccine hesitancy, computed as the fraction of individuals who would not get vaccinated according to Facebook daily surveys administered in the period from January 4th to March 25th, 2021, and misinformation about vaccines shared on Twitter. Each dot represents a U.S. state and is colored according to the share of Republican voters (battleground states have a share between 45% and 55%) and sized according to population. Grey lines show the partial correlation between the two variables after adjusting for socioeconomic, demographic, and political factors in a weighted multiple linear regression model (shaded areas correspond to 95% C.I.). (c) Cartogram [89] of the U.S. in which the area of each state is proportional to the average number of misinformation links shared by geolocated users, and the color is mapped to the vaccine hesitancy rate, with lighter colors corresponding to higher hesitancy.*

= 2.96, P < .001) and hesitancy (see Fig. 5.4b and Fig. S1 in Supplementary Information). Fig. 5.4c provides an illustration of the correlation between misinformation and hesitancy. For example, the large size and yellow color of Wyoming indicate it is the

state with the highest level of misinformation and hesitancy. Among other variables, we find that the percent of Black residents is positively related to reports of hesitancy (b = 0.12, P = .001), while percent Hispanic or Latinx is negatively associated (b = –0.07, P = .021). The percent of residents below the poverty line is also positively associated with vaccine hesitancy (b = 0.53, P = .001).

To test the robustness of these results, we also consider a more granular level of information by examining county data. Similar to previous analyses, we compute online misinformation shared by almost 1.15 M Twitter users geolocated in over 1,300 U.S. counties. We measure vaccine hesitancy rates by leveraging over 17 M daily responses to the Facebook survey for over 700 distinct counties. The total number of observations (i.e. counties) for which we are able to measure both variables is $N = 548$ (see Methods). Political partisanship and misinformation are both significantly correlated with county-level vaccine hesitancy, net covariates (Table S4, Fig. S2 in Supplementary Information in Appendix B). Using a weighted multiple linear regression model, we find a significant interaction between political partisanship and misinformation. Specifically, as levels of misinformation increase, democratic and republican counties converge to the same level of vaccine hesitancy (Fig. 5.5).

Our results so far demonstrate an association between online misinformation and vaccine hesitancy. We investigate evidence for directionality in this association by performing a Granger Causality analysis [99, 104]. We find that misinformation helps forecast vaccine hesitancy, weakly at state level (P = .0519) and strongly at county level (P < .001; see Methods and Tables S10, S11 in the Supplementary Information in Appendix B). Analysis of the significant lagged coefficients (see Table S10 in Appendix B) indicates that there is a lag of around 2-6 days from misinformation posted in a county to a corresponding increase in vaccine hesitancy in the same county.

5.3.3 Discussion

Our results provide evidence for the problem of geographical regions with lower levels of COVID-19 vaccine uptake, which may be driven by online misinformation. Considering variability across regions with low and high levels of misinformation, the best estimates from our data predict a 20% decrease in vaccine uptake between states, and a 67% increase in hesitancy rates across democratic counties, across the full range of misinformation prevalence. At these levels of impact on vaccine uptake, the data predict SARS-CoV-2 will remain endemic in many U.S. regions. Vaccine-hesitant individuals are potentially more likely to post vaccine misinformation, and our data cannot demonstrate a causal relationship between misinformation and vaccine refusal. However, we find evidence of a directional relationship from misinformation to vaccine hesitancy, consistent with another study that used controlled circumstances [144]. This evidence suggests a need to counter misinformation, and the beliefs associated with misinforma-

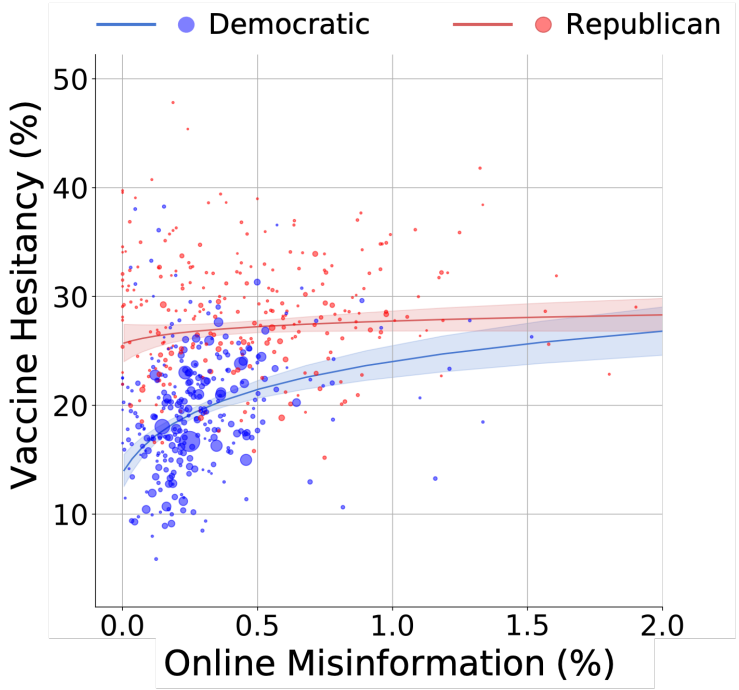


Figure 5.5: *Associations of online misinformation and political partisanship with vaccination hesitancy at the U.S. county level. Each dot represents a U.S. county, with size and color indicating population size and political majority, respectively. The average proportion of misinformation shared on Twitter by geolocated users was fitted on a log scale due to non-normality (i.e., positive skew) at the county level. The two lines show predicted values of vaccine hesitancy as a function of misinformation for majority Democratic and Republican counties, adjusting for county-level confounding factors (see Methods). Shaded area corresponds to 95% C.I.*

tion, to promote vaccine uptake.

Public opinion is very sensitive to the information ecosystem and sensational posts tend to spread widely and quickly [138]. Our results indicate that there is a geographical component to this spread, with opinion on vaccines spreading at a local scale. While social media users are not representative of the general public, existing evidence suggests that vaccine hesitancy flows across social networks [39], providing a mechanism for the lateral spread of misinformation offline among those connected directly or indirectly to misinformation spreading online. More broadly, our results provide additional insight into the effects of information diffusion on human behavior and the spread of infectious diseases [87].

A limitation of our findings is that they are based on data averaged over geographical regions, which does not provide evidence at an individual level. However, to account for group-level effects we present a number of sensitivity analyses, and note that our findings are consistent over two geographical scales. Our results are also limited to a snapshot in time. Vaccination hesitancy levels will potentially change over time due to novel factors, including changes in COVID-19 infection and death rates, as well as

legitimate reports about vaccine safety, among other factors [136].

Associations between online misinformation and detrimental offline effects, like the results presented here, call for better moderation of our information ecosystem. COVID-19 misinformation is shared overtly by known entities on major social media platforms [270]. While people have a constitutional right to free speech, it is important to maintain an environment where individuals have access to good information that benefits public health.

5.4 Vaccinitaly

Analogously to CoVaxxy, VaccinItaly is a project to monitor Italian conversations around vaccines on multiple social media (Twitter, Facebook) with the aim of understanding the interplay between online public discourse and the vaccine roll-out campaign in Italy. A public dashboard associated to the project is available at `http://genomic.elet.polimi.it/vaccinitaly/`.

5.4.1 Twitter data collection

Starting on December 20th, 2020, we use Twitter Filter[10] API to collect tweets matching the set of keywords in our repository, in real-time. We routinely check for trending hashtags and relevant events to add new peculiar keywords, e.g. "#novaccinoainovax" and "#iononsonounacavia" were hashtags trending on specific days and consequently they were added to the list of keywords. The latter refers to vaccine advocates stating that no-vax should not be vaccinated, and the former indicates vaccine skeptics who "do not want to be guinea pigs for vaccines". The overall data up to September 2021 comprises approximately 9 M tweets shared by over 500 k unique users.

5.4.2 Facebook data collection

We used the *posts/search* endpoint of the CrowdTangle API [54] to collect public posts shared by pages and groups which matched the list of keywords previously defined, resulting in over 50 M posts published by over 100 k public pages and groups, and re-shared over 300 M times, as of September, 2021. In the following, we will use the number of shares to compare Facebook with Twitter.

A limitation to our collection of Facebook is the coverage of pages and groups, whose data can be retrieved using the API. The tool includes over 6M Facebook pages and groups: all those with at least 100k followers/members and a very small subset of verified profiles that can be followed like public pages. Besides, some pages and groups with fewer followers and members can be included by CrowdTangle upon request from

[10] `https://developer.twitter.com/en/docs/twitter-api/v1/tweets/filter-realtime/api-reference/post-statuses-filter`

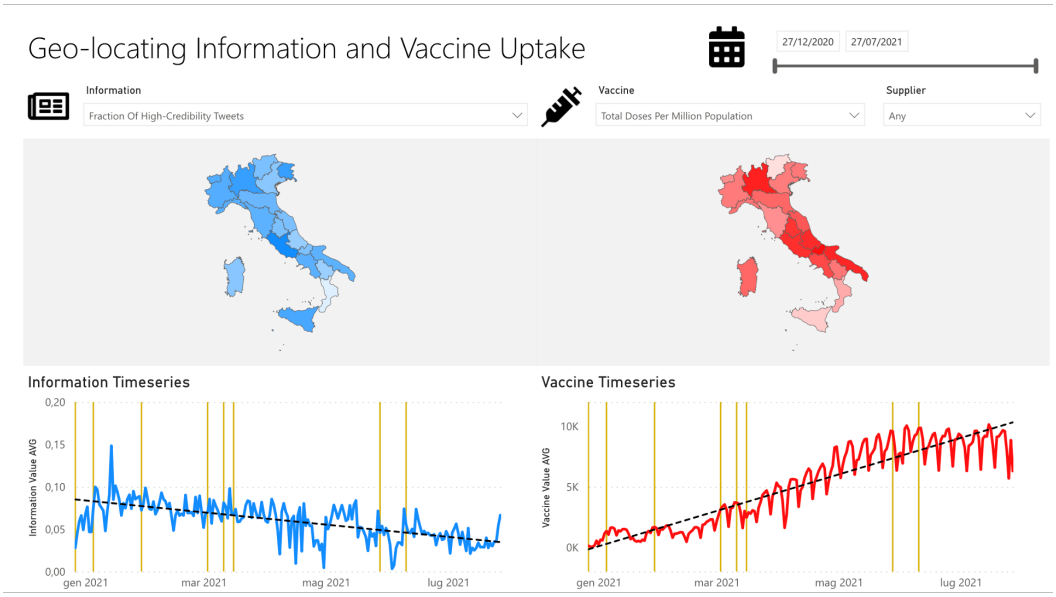


Figure 5.6: *Statistics about information spreading on Twitter **(left)** and the vaccination program **(right)** for each Italian region. We geolocalize Twitter users and we average, for each region, the mean number (Mean) and the fraction (Fraction) of tweets with Low/High/Fact-checking news articles shared by users as well as Pro and Anti vaccine hashtags. Vertical yellow lines highlight some relevant events (e.g. start date of the vaccination campaign, Astrazeneca blood clots, etc).*

users. This might bias the data as, for instance, researchers and journalists might be interested in monitoring pages and groups sharing low-credibility thus leading to an over-representation of such content.

5.4.3 Sources of low- and high-credibility information

We extract URLs contained in tweets and Facebook posts to understand the prevalence of low- and high-credibility information shared in vaccine-based conversations [270]. We use a consolidated source-based approach to label news articles [62, 88, 100, 138, 188, 189, 193, 194, 225] depending on the reliability of the source, referring to two lists of Italian low- and high credibility news websites. The former corresponds to websites flagged by Italian fact-checkers for publishing false news, hoaxes and conspiracy theories[11]); the latter corresponds to Italian traditional and most popular news websites [249], and it is used as a reference to understand the prevalence of misleading and (potentially) harmful information. Lists are available in our repository[12].

Again, we are aware that this approach, widely adopted in the research community, is not 100% accurate, as cases of misinformation on mainstream websites are not rare and, similarly, low-credibility websites do not publish solely "fake news". However, to date, it is the most reliable and scalable way to study misleading and harmful infor-

[11] See `www.pagellapolitica.it`, `www.facta.news` and `www.butac.it`
[12] `https://github.com/frapierri/VaccinItaly`

mation. Another limitation to our estimates is that our lists might not fully capture the amount of low- and high-credibility information circulating on Twitter. Besides, we do not consider different typologies of content such as photos, videos, memes, etc.

5.4.4 Geolocating Twitter users

We attempt to geolocate Twitter users by using a naive string matching algorithm, i.e. checking whether they have a "location" field disclosed in their profile and matching it against a list of Italian municipalities, provinces, and regions[13]. In the case of multiple matches, we retain the longest one. We matched circa 16 k unique locations and, among over 135 k users putting a "location" in their profile, we accordingly geolocated 73 k users to either an Italian municipality or region. These shared over 1.3 M tweets. The number of accounts mapped to each Italian region is significantly positively correlated with the actual population (Pearson R$= 0.89$, PVAL< 0.001). However, it is known that the Twitter sample of users might not be fully representative of the Italian population, and this is a limitation to analyses that infer demographics from Twitter [10].

These results are still preliminary, as the methodology presents several limitations and needs further assessment, e.g., how to handle multiple locations appearing in the "location" field of user profiles or when false places match with Italian municipalities with misleading names (e.g. "Paese" which translates as "village"). We plan to carry out a rigorous evaluation of our methodology and compare it to existing ones [68, 155].

5.4.5 Dashboard

Similar to CoVaxxy, we deployed a public dashboard to show visualizations of preliminary results, which is available at `http://genomic.elet.polimi.it/vaccinitaly/`. An example of visualization is available in Figure 5.6.

In addition to CoVaxxy, we also keep track of pro and anti vaccine hashtags, collected manually with a snowball sampling approach and evaluated through independent annotators.

The main goal is to find statistical associations between online signals, e.g. the prevalence of disinformation and anti vaccine hashtags, and vaccine outcomes. In Figure 5.7 we provide a scatterplot where the user can visualize the correlation between different variables. Currently, we are working on building a multiple linear regression model (similar to the work described in previous section) to take into account other confounding factors.

[13] Taken from the Italian National Institute of Statistics and available at `https://www.istat.it`.

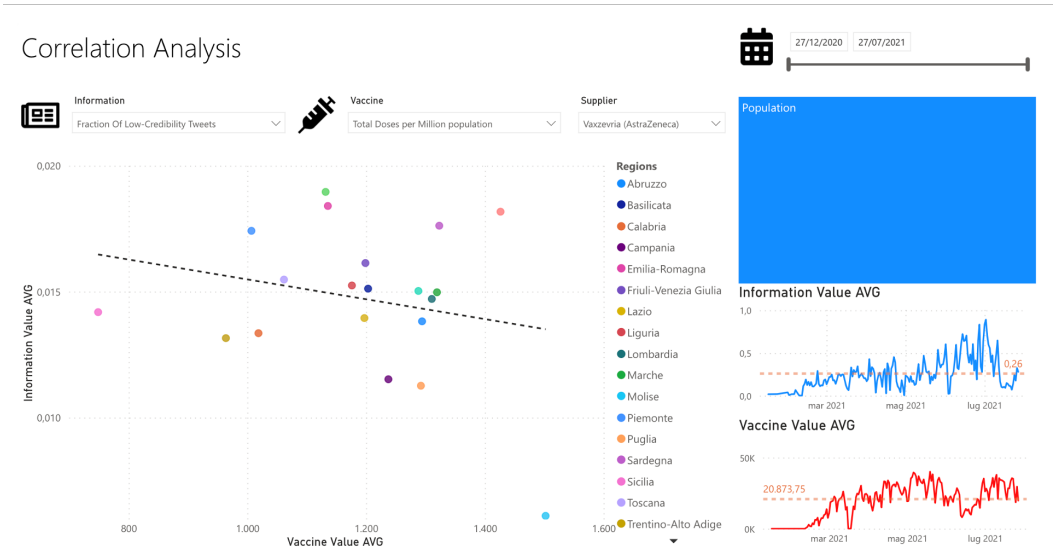


Figure 5.7: *Correlation between a variable measured from Twitter **(Y-axis)** and a variable measured from vaccine data **(X-axis)**. Each point represents an Italian region, and both X and Y values are computed as the average over the time period chosen by the user (see slider on top-right). The dashed line represents a linear fit.*

5.4.6 Extension to European countries

There is a number of research directions that we are currently pursuing, which involve collecting data from multiple countries.

We set-up a procedure to collect tweets about vaccines in multiple countries, namely France, Netherlands and Germany, and to build a classifier to automatically detect pro and anti vaccine tweets. Again, the overall goal is to find associations between online disinformation and negative opinions about vaccines expressed on Twitter, and the evolution of the vaccination programs.

To this aim, we used a snowball sampling approach to compile a list of relevant keywords for gathering Twitter data (similar to CoVaxxy and VaccinItaly), with the help of native speakers[14]. They also identified hashtags which were clearly associated with either positive or negative views about vaccines. Finally, they independently labeled tweets according to whether they were pro, anti, neutral or unrelated to vaccines. We are currently working on building a machine learning classifier that can identifies the stance of tweets in multiple languages. We refer the reader to [57] for more details on the methodology.

In addition, we are going to integrate data from multiple countries in the current dashboard, which will be called VaccinEU, so that users can visualize and compare statistics about online conversations and vaccine outcomes in different countries. We are also planning to leverage data from Newsguard[15] to obtain more accurate labels for

[14] We resorted to other partners of the H2020 Periscope.
[15] https://www.newsguardtech.com

news websites.

CHAPTER 6

A network-based approach to detect online disinformation on Twitter

In this section we shall describe a methodology to classify online disinformation spreading on Twitter, which is solely based on the interactions between users when sharing news articles. We present results from two different approaches: one that considers a single-layer representation of Twitter diffusion networks, and an extension which employs a multi-layer representation of Twitter diffusion networks. The chapter is based on [194] and [193].

6.1 Context and Problem Formulation

As discussed in the introductory sections of this thesis, a few recent studies, featuring large-scale analyses of social media data, have produced deeper knowledge about the phenomenon, showing that: false news spread faster and more broadly than the truth on social media [254]; social bots play an important role as "super-spreaders" in the core of diffusion networks [225]; echo chambers are primary drivers for the diffusion of true and false content [58]; the majority of fake news circulating on-line is accounted by a limited community of online users, who tend to be older, conservative and very active in politics [100].

In the work presented in this section, we focus on analyzing the diffusion of disin-

formation news along the direction pointed by these studies.

As we saw in Chapter 2, the problem of automatically detecting online disinformation news has been typically formulated as a binary classification task (i.e. credible vs non-credible articles), and tackled with a variety of different techniques, based on traditional machine learning and/or deep learning, which mainly differ in the dataset and the features they employ to perform the classification. We may distinguish three approaches: those built on content-based features, those based on features extracted from the social context, and those which combine both aspects.

Leveraging the sole diffusion network allows to by-pass the intricate information related to individual news articles–such as content, style, editorship, audience, etc–and to capture the overall diffusion properties for two distinct news domains: reputable outlets that produce mainstream, reliable and objective information, opposed to sources which notably fabricate and spread different kinds of disinformation articles. We consider any article published on the former domain as a proxy for credible and factual information (although it might not be true in all cases) and all news published on the latter domain as proxies for disinformation and/or inaccurate information (we do not investigate whether disinformation news can be accurately distinguished also from factual but non-mainstream news originated from niche outlets).

In the research described in the following sections, we are driven by two broad research questions:

- **RQ1**: Can we accurately classify disinformation versus mainstream news articles solely based on their diffusion patterns on Twitter?

- **RQ2**: Does a multi-layer representation of Twitter diffusion networks yield a significant advance in terms of classification accuracy over a conventional single-layer diffusion network?

To this aim, we collect thousands of Twitter diffusion networks pertaining to disinformation and mainstream news domains and we carry out an extensive network comparison using several alignment-free approaches. These include training a classifier on top of global network properties and centrality measures distributions, as well as computing network distances. We also disentangle different Twitter actions, e.g. tweets, retweets, replies, etc, to build a multi-layer representation of Twitter diffusion networks, and train binary classifiers accordingly.

We perform classification experiments with off-the-shelf classification models on two different datasets of mainstream and disinformation news shared on Twitter respectively in the United States and in Italy during 2019. In the former case we also perform multiple disaggregated tests to control for political biases inherent to different news sources, referring to the procedure proposed in [33] to label different outlets.

Overall we show that we are able to classify credible vs non-credible diffusion networks (and consequently news articles) with high accuracy (AUROC up to 94%), also when controlling for the political bias of sources (e.g., training only on left-biased or right-biased articles). We observe that the layer of mentions alone conveys useful information for the classification, denoting a different usage of this functionality when sharing news belonging to the two news domains.

We also show that the most discriminative features, which are relative to the breadth and depth of the largest cascades in different layers, are the same across the two countries.

As our datasets are collected in different countries, we also investigate whether disinformation can be detected independently from the country where it originates. Cross-country experiments show that our methodology fails to distinguish reliable vs non-reliable news regardless of where it originates from. We argue that this might be due either to the high imbalance of data or to the class discrepancies which are country specific. It emerges that a classifier based on our methodology should be trained in a country-wise fashion.

6.1.1 Existing techniques for network-based detection of online disinformation

We briefly report here a few contributions which tackle the problem of classifying false and true news articles specifically based on the propagation of URLs on Twitter.

A deep learning framework for detecting fake news cascades is proposed in [162], where the authors refer to [254] in order to collect Twitter cascades pertaining to verified false and true rumors. They employ *geometric* deep learning, a novel paradigm for graph-based structures, to classify cascades based on four categories of features, such as user profile, user activity, network and spreading, and content. They also observe that a few hours of propagation are sufficient to distinguish false news from true news with high accuracy.

Drawing on the same data [254], authors of [211] (the paper was published contemporary to our own contributions [193, 194]) make use of graph kernels, wherein graphs are embedded in a vector space, with dimensions corresponding to attributes of motif-like substructures. They represent a single Twitter rumor cascade as a graph, and they employ Weisfeiler-Lehman graph kernels to train classifiers that are able to tell whether a cascade pertains to true or false rumor.

Finally, diffusion cascades on Weibo and Twitter are analyzed in [278], where the authors focus on highlighting different topological properties, such as the number of *hops* from the source or the heterogeneity of the network, to show that fake news shape diffusion networks which are highly different from credible news, even at early stages of propagation.

6.2 Methodology

6.2.1 Mainstream versus Disinformation

As highlighted by recent research on the subject [33, 100, 138, 225, 254], there is not a general consensus on a definition for malicious and deceptive information, e.g authors in [259] define disinformation as information at the intersection between false information and information intended to harm whereas Wikipedia defines it as "false information spread deliberately to deceive"; consequently, to assess whether a news outlet is spreading unreliable or objective information is a controversial matter, subject to imprecision and individual judgment.

The consolidated strategy in the literature–which we follow in this work–consists of building a classification of websites, based on multiple sources (e.g. reputable third-party news and fact-checking organizations).

Along this approach, we characterize a list of websites that notably produce disinformation, i.e. low-credibility content, false and/or hyper-partisan news as well as hoaxes, conspiracy theories, click-bait and satire. We oppose to these malicious sources a set of traditional news outlets (defined as in [33]) which deliver *mainstream* reliable news, i.e. factual, objective and credible information. We are aware that this might not be always true as reported cases of misinformation on mainstream outlets are not rare [138], yet we adopt this approach as it is currently the best available proxy for a correct classification.

We formulate our classification problem as follows: given two classes of news articles, respectively D (*disinformation*) and M (*mainstream*), a set of news articles A_i and associated class labels $C_i \in \{D, M\}$, and, for each article A_i, a set of tweets $\Pi_i = \{T_i^1, T_i^2, ...\}$ each containing an Uniform Resource Locator (URL) pointing explicitly to article A_i, predict the class C_i of each article A_i.

6.2.2 U.S. data collection

We collected all tweets containing a Uniform Resource Locator (URL) pointing to websites (specified next) which belong either to a *disinformation* or *mainstream* domain. Following the approach described in [33, 100, 138, 224–226] we assume that article labels are associated with the label of their source, i.e. all items published on a disinformation (mainstream) website are disinformation (mainstream) articles. We took into account censoring effects described in [92], by retaining only diffusion cascades relative to articles that were published after the beginning of the collection process (*left censoring*), and observing each of them for at least one week (*right censoring*).

For what concerns disinformation sources we referred to the curated list of 100+ news outlets provided by [224–226], which contains websites featured also in [33, 100,

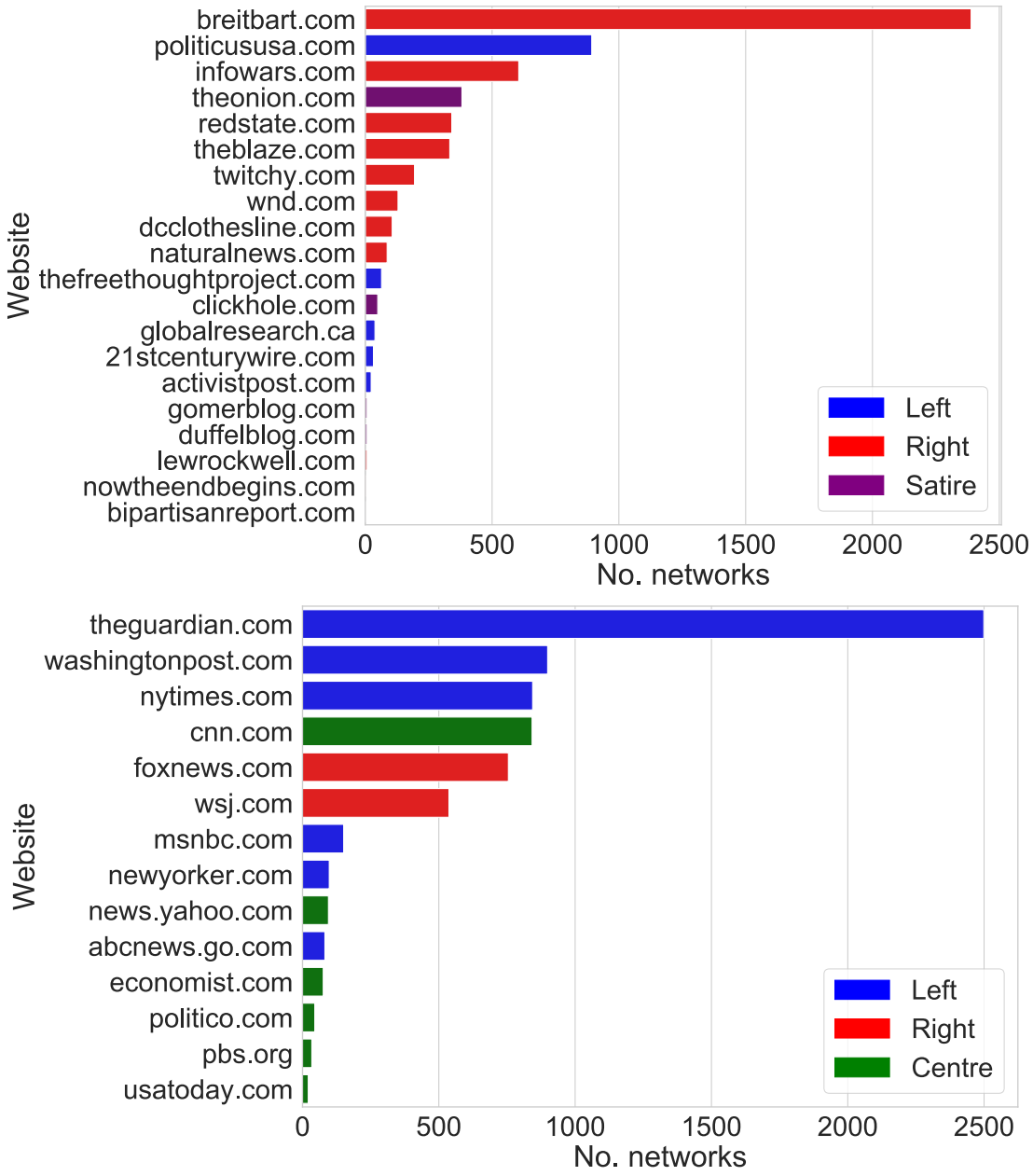


Figure 6.1: *Distribution of the number of networks per each source for disinformation **(top)** and mainstream **(bottom)** outlets; colors indicate different political bias labels as specified in the legend.*

254]. Leveraging Hoaxy API, we obtained tweets pertaining to news items published in the period from Jan, 1st 2019 to March, 15th 2019, filtering articles with less than 50 associated tweets. The final collection comprises 5,775 diffusion networks. In Figure 6.1 we show the distribution of the number of networks per each source.

We replicated the collection procedure described in [224, 226] in order to gather reliable news articles by using the Twitter Streaming API. We referred to U.S. most trusted news sources described in [159]; this includes websites described also in [33,

100] and employed in other research also described in this thesis [270]. We associated tweets to a given article after canonicalization of the attached URL(s), using tracking parameters as in [224, 226], to handle duplicated hyperlinks. We collected the tweets during a window of three weeks, from February 25th 2019 to March 18th 2019; we restricted the period w.r.t the disinformation collection in order to obtain a balanced dataset of the two news domains. At the end of the collection, we excluded articles for which the number of associated tweets was less than 50, obtaining 6,978 diffusion networks; we show in Figure 6.1 the distribution of the number of network per each source. A different classification approach using several data sampling strategies on networks in the same 3-week period is available in the Supplementary Information. The number of disinformation networks is only $\sim$1,200, resulting in an imbalanced dataset with disinformation/mainstream proportion 1 to 5. Results are nonetheless in accordance with those provided in the main paper.

Furthermore, we assigned a *political bias* label to sources in both news domains, as to perform binary classification experiments considering separately *left*-biased and *right*-biased outlets. We derived labels following the procedure outlined in [33]. Overall, we obtained 4,573 left-biased, 1,079 centre leaning and 1,292 right-biased mainstream diffusion networks; on the other side, we counted 1,052 left-biased, 444 satire and 4,194 right-biased disinformation diffusion networks. Labels for each source in both news domains are shown in Figure 6.1, and a more detailed description is provided in the Supplementary Information available in Appendix A.

Eventually, mainstream news generated $\sim$1.7 million tweets, corresponding to $\sim$400k independent cascades, $\sim$680k unique users and $\sim$ 1.2 million edges; disinformation news generated $\sim$1.6 million tweets, $\sim$210k independent cascades, $\sim$420k unique users and $\sim$1.4 million edges.

6.2.3 Italian data collection

For what concerns the Italian scenario we first collected tweets with the Streaming API in a 3-week period (April 19th, 2019-May 5th, 2019), filtering those containing URLs pointing to Italian official newspapers websites as described in [188, 249]; these correspond to the list provided by the association for the verification of newspaper circulation in Italy (Accertamenti Diffusione Stampa)[1]. We instead referred to the dataset described in Chapter 3 for what concerns Italian disinformation. In order to get balanced classes, we retained data collected in a longer period w.r.t to mainstream news (April 5th, 2019-May 5th, 2019). In both cases we filtered out articles with less than 50 tweets; overall this dataset contains $\sim$160k mainstream tweets, corresponding to 227 news articles, and $\sim$100k disinformation tweets, corresponding to 237 news articles. We provide in Figure 6.2 the distribution of articles according to distinct sources

[1] `http://www.adsnotizie.it`. Accessed: April 18th, 2019.

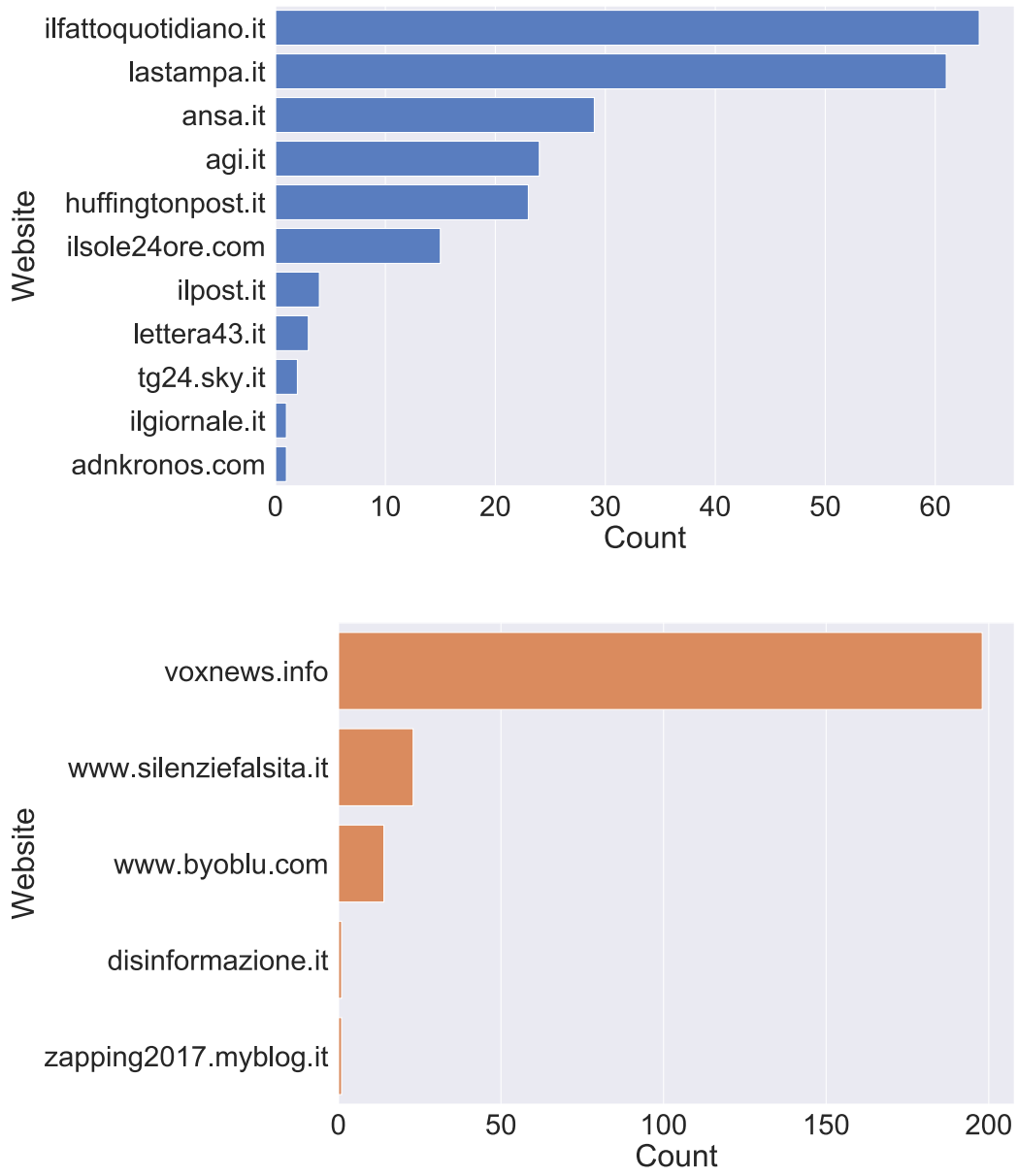


Figure 6.2: *Distribution of the number of articles per source for Italian* ***(top)*** *mainstream and* ***(bottom)*** *disinformation news.*

for both news domains. As in the US dataset, we took into account censoring effects [92] by excluding tweets published before (*left-censoring*) or after two weeks (*right-censoring*) from the beginning of the collection process.

The different volumes of news shared on Twitter in the two countries are due both to the different population size of US and Italy (320 vs 60 million) but also to the different usage of Twitter platform (and social media in general) for news consumption [167].

6.2.4 Twitter diffusion networks

Single-layer representation. We represent Twitter sharing diffusion networks as directed, unweighted graphs following [225, 226]: for each unique URL we process all tweets containing that hyperlink and build a graph where each node represents a unique user and a directed edge is built between two nodes whenever a user re-tweets/quotes, mentions or replies to another user. Edges between nodes are built only once and they all have weight equal to 1. Isolated nodes correspond to users who authored tweets which were never re-tweeted nor replied/quoted/mentioned.

An intrinsic yet unavoidable limitation in our methodology is that, as pointed out in [92, 226, 254], it is impossible on Twitter to retrieve *true* diffusion cascades because the re-tweeting functionality makes any re-tweet pointing to the original content, losing intermediate re-tweeting users. As such, the majority of Twitter cascades often end up in *star* topologies. In contrast to [254], we consider as a single diffusion network the union of several cascades generated from different users which shared the same news article on the social network; thus such network is not necessarily a single connected component. Notice that our approach, although yielding a description of diffusion cascades which might be partial, is the only viable approach based on publicly available Twitter information.

Multi-layer representation. Using the notation described in [128] we employ a multi-layer representation for Twitter diffusion networks. Sociologists have indeed recognized decades ago that it is crucial to study social systems by constructing multiple social networks where different types of ties among the same individuals are used [260]. Therefore, for each news article we build a multi-layer diffusion network composed of four different layers, one for each type of social interaction on Twitter platform, namely retweet (RT), reply (R), quote (Q) and mention (M), as shown in Figure 6.3. These networks are not necessarily *node-aligned*, i.e. users might be missing in some layers. We do not insert "dummy" nodes to represent users not active in a given layer as it would have severe impact on the global network properties (e.g. number of weakly connected components). Alternatively one may look at each multi-layer diffusion network as an ensemble of individual graphs [128]; since global network properties are computed separately for each layer, they are not affected by the presence of *inter-layer* edges, which nonetheless allow the diffusion of information across layers.

In our multi-layer representation, each layer is a directed graph where we add edges and nodes for each tweet of the layer type. While the direction of information flow - thus the edge direction - is unambiguous for some layers, e.g. RT, the same is not true for others. Here we follow the conventional approach described e.g. in [49, 202, 225, 226] to define the direction of edges. For the RT layer: whenever user a retweets account b we first add nodes a and b if not already present in the RT layer, then we build

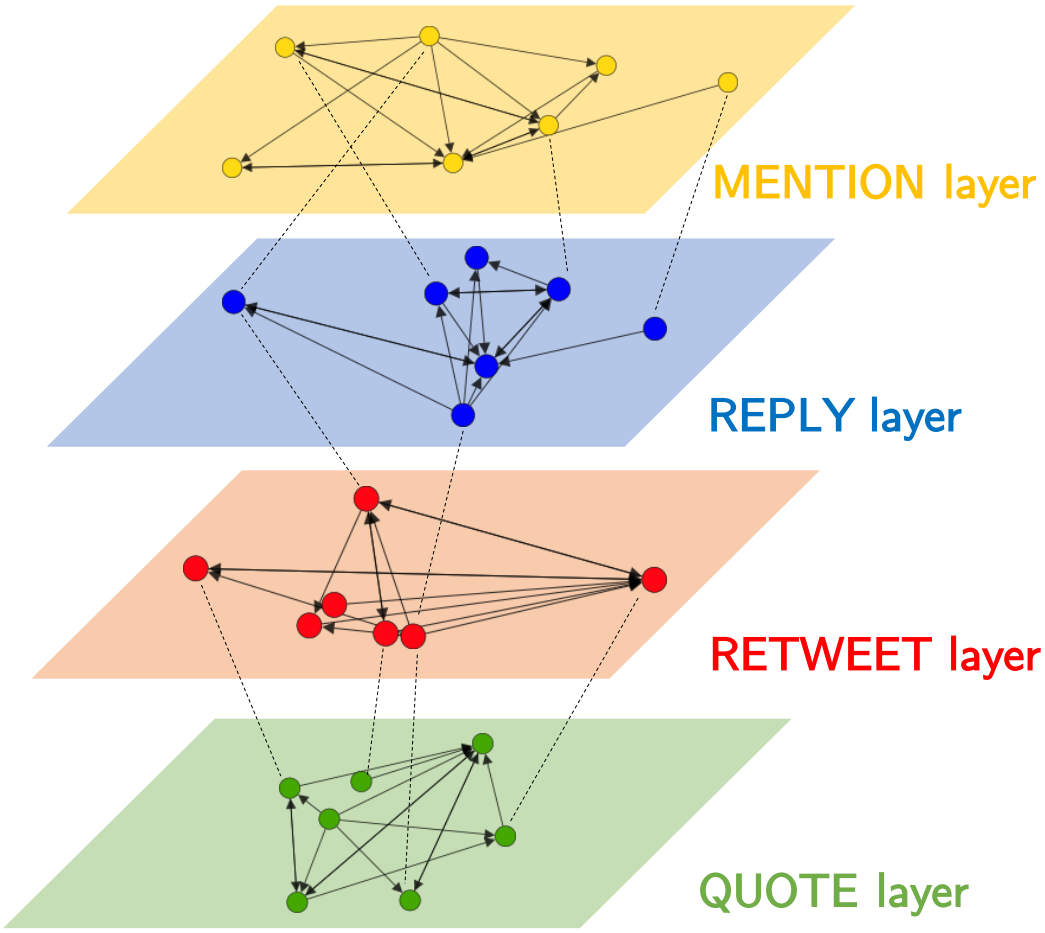


Figure 6.3: *A visualization of a Twitter multi-layer diffusion network with four layers.*

an edge that goes from b to a if it does not exist. Similarly for the other layers: for the R layer edges go from user a (who replies) to user b, for the Q layer edges go from user b (who is quoted by) to user a and for the M layer edges go from user a (who mentions) to user b. Note that, by construction, our layers do not include isolated nodes; they correspond to "isolated tweets", i.e. tweets which have not originated any interactions with other users. However, they are present in our dataset, and their number is exploited for classification, as described below.

6.2.5 Breakdown of Twitter interactions

We disentangle different social interactions on Twitter according to five categories:

Mention (M): Including in a tweet another account's Twitter user name, preceded by the "@" symbol;

Reply (R): Responding to another account's tweet;

Retweet (RT): Re-posting a tweet;

Country	Class	Mentions	Replies	Retweets	Quotes	Tweets
United States	Mainstream	87,183	30,745	1,482,261	29,365	409,544
	Disinformation	123,047	22,599	1,207,243	94,027	220,891
Italy	Mainstream	1,578	473	18,794	1,378	4,832
	Disinformation	929	186	35,323	3,192	5,302

Table 6.1: *Breakdown of US and IT datasets in terms of different Twitter interactions.*

Quote (Q): Retweeting with the addition of a comment;

Tweet (T): Posting a tweet containing an article URL.

We show in Table 6.1 the breakdown of our datasets for what concerns cardinalities of different Twitter interactions across news domains. We notice that news sharing mostly involves retweeting and tweets in both countries and for both classes of news articles.

We notice that news sharing mostly involves retweeting and tweets in both countries and for both classes of news articles.

For what concerns different Twitter actions, users primarily interact with each other using retweets and mentions [49]. The former are the main engagement activity and act as a form of endorsement, allowing users to rebroadcast content generated by other users [35]. Besides, when node B retweets node A we have an implicit confirmation that information from A appeared in B's Twitter feed [202]. Quotes are simply a special case of retweets with comments. Mentions usually include personal conversations as they allow someone to address a specific user or to refer to an individual in the third person; in the first case they are located at the beginning of a tweet and they are known as replies, otherwise they are put in the body of a tweet [49]. The network of mentions is usually seen as a stronger version of interactions between Twitter users, compared to the traditional graph of follower/following relationships [98].

6.2.6 Global network properties

We used a set of global network indicators which encode each network layer by a tuple of features. Then we simply concatenated tuples as to represent each multi-layer network with a single feature vector. We used the following global network properties:

1. **Number of Strongly Connected Components (SCC)**: a Strongly Connected Component of a directed graph is a maximal (sub)graph where for each pair of vertices u, v there is a path in each direction ($u \to v$, $v \to u$).

2. **Size of the Largest Strongly Connected Component (LSCC)**: the number of nodes in the largest strongly connected component of a given graph.

3. **Number of Weakly Connected Components (WCC)**: a Weakly Connected Component of a directed graph is a maximal (sub)graph where for each pair of vertices

(u, v) there is a path $u \leftrightarrow v$ ignoring edge directions.

4. **Size of the Largest Weakly Connected Component (LWCC)**: the number of nodes in the largest weakly connected component of a given graph.

5. **Diameter of the Largest Weakly Connected Component (DWCC)**: the largest distance (number of edges of the shortest path) between two nodes in the (undirected version of) largest weakly connected component of a graph.

6. **Average Clustering Coefficient (CC)**: the average of the local clustering coefficients of all nodes in a graph; the local clustering coefficient of a node quantifies how close its neighbourhood is to being a complete graph (or a clique). It is computed according to [219].

7. **Main K-core Number (KC)**: a K-core [26] of a graph is a maximal sub-graph that contains nodes of internal degree K or more; the main K-core number is the highest value of K (in directed graphs the total degree is considered).

8. **Density (d)**: the density for directed graphs is $d = \frac{|E|}{|V||V-1|}$, where $|E|$ is the number of edges and $|V|$ is the number of vertices in the graph; the density equals 0 for a graph without edges and 1 for a complete graph.

9. **Structural virality of the largest weakly connected component (SV)**: this measure is defined in [92] as the average distance between all pairs of nodes in a cascade tree or, equivalently, as the average depth of nodes, averaged over all nodes in turn acting as a root; for $|V| > 1$ vertices, $SV = \frac{1}{|V||V-1|} \sum_i \sum_j d_{ij}$ where d_{ij} denotes the length of the shortest path between nodes i and j. This is equivalent to compute the Wiener's index [65] of the graph and multiply it by a factor $\frac{1}{|V||V-1|}$. In our case we computed it for the undirected equivalent graph of the largest weakly connected component, setting it to 0 whenever $|V| = 1$.

We used `networkx` Python package [103] to compute all features.

We remark that in the single-layer setting [194] we only employed features 1-7, while in the multi-layer setting [193] we introduced feature 8 and 9. Specifically, in the multi-layer setting, whenever a layer is empty, we simply set to 0 all its features. In addition, we added two indicators for encoding information about isolated tweets, namely the number **T** of isolated tweets (containing URLs to a given news article) and the number **U** of unique users authoring those tweets. Therefore, a diffusion network for a given article is represented by a vector with $9 \cdot 4 + 2 = 38$ entries.

6.2.7 Interpretation of network features and layers

The aforementioned network properties can be qualitatively explained in terms of social footprints as follows (see the illustrative examples in Figure 6.4): in this specific class

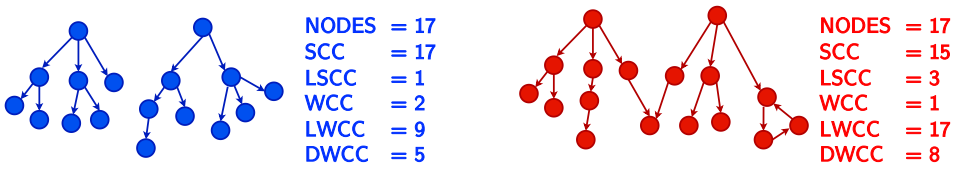


Figure 6.4: *Two illustrative examples of diffusion layers. Left: The same news spreads, in a pure top-down broadcast manner, along two distinct cascades. Thus, SCC equals the number of nodes, since each strongly connected component is a single node, while WCC is the number of distinct cascades. Right: The two cascades merge in a common node (thus WCC=1) and, additionally, mono-directionality is broken by a loop (thus SCC is less than the number of nodes).*

of networks, **SCC** correlates with the size (i.e. number of nodes) of the diffusion layer, as the propagation of news occurs in a broadcast manner in most cases, i.e. re-tweets dominate on other interactions, while **LSCC** allows to distinguish cases where such mono-directionality is somehow broken. **WCC** equals (approximately) the number of distinct diffusion cascades pertaining to each news article, with exceptions corresponding to those cases where some cascades merge together via Twitter interactions such as mentions, quotes and replies, and accordingly **LWCC** and **DWCC** equals the size and the depth of the largest cascade. **CC** corresponds to the level of connectedness of neighboring users in a given diffusion network whereas **KC** identifies the set of most influential users in a network [226]. Finally, **d** describes the proportions of potential connections between users which are actually activated and **SV** indicates whether a news item has gained popularity with a single and large broadcast or in a more viral fashion through multiple generations [92].

We observed what follows: **(a)** is highly correlated with the size of the network (see Supplementary Information), as the diffusion flow of news mostly occurs in a broadcast manner, i.e. edges almost consist of re-tweets, and **(b)** allows to capture cases where the mono-directionality of the information diffusion is broken; **(c)** indicates approximately the number of distinct cascades, with exceptions corresponding to cases where two or more cascades are merged together via mentions/quotes/replies on Twitter; **(d)** and **(e)** represent respectively the size and the depth of the largest cascade of a given news article; **(f)** indicates the degree to which users in diffusion networks tend to form local cliques whereas **(g)** is commonly employed in social networks to identify influential users and to describe the efficiency of information spreading [226].

6.2.8 Network distances

In our first contribution [194], in addition to encode networks using a vector of topological features, we considered two alignment-free network distances that are commonly used in the literature to assess the topological similarity of networks, namely the Di-

rected Graphlet Correlation Distance (DGCD) and the Portrait Divergence (PD).

The first distance [218] is based on directed graphlets [119]. These are used to catch specific topological information and to build graph similarity measures; depending on the graphlets (and the orbits) considered, different DGCD can be obtained, e.g. DGCD-13 is the one that we employed in this work. Among all graphlet-based distances, which often yield a prohibitive computational cost to compute graphlets, DGCD has been demonstrated as the most effective at classifying networks from different domains.

The second distance [19], which was recently defined, is based on the network portrait [18], a graph invariant measure which yields the same value for all graph isomorphisms. This distance is purely topological, as it involves comparing, via Jensen-Shannon divergence, the distribution of all shortest path lengths of two graphs; moreover, it can handle disconnected networks and it is computationally efficient.

We also conducted experiments on several centrality measures distributions–such as total degree, eigenvector and betweenness centrality–and results are available in the Supplementary Information in Appendix A. They overall perform worse than the above methods, in accordance with current literature on network comparison techniques [242].

6.2.9 Dataset splitting

As we expect networks to exhibit different topological properties within different ranges of node sizes (see also Supplementary Information), prior to our analyses, we partitioned the original collection of networks into subsets of similar sizes. This simple heuristic criterion produced a splitting of the dataset into three subsets according to specific ranges of cardinalities (see Tables 6.2, 6.3); we also considered the entire original dataset for comparison. Splitting proved effective for improving the classification and also for highlighting interesting properties of diffusion networks.

6.2.10 Performance evaluation

We used the following evaluation metrics to assess the performance of different classifiers (TP=true positives, FP=false positives, FN=false negatives):

1. **Precision** = $\frac{TP}{TP+FP}$, the ability of a classifier not to label as positive a negative sample.
2. **Recall** = $\frac{TP}{TP+FN}$, the ability of a classifier to retrieve all positive samples.
3. **F1-score** = $2\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}$, the harmonic average of Precision and Recall.
4. **Area Under the Receiver Operating Characteristic curve (AUROC)**: the Receiver Operating Characteristic (ROC) curve [77], which plots the TP rate versus

	No. Mainstream			No. Disinformation		
Size Class	Left	Right	Tot.	Left	Right	Tot.
$[0, 100)$	774	2746	4177	379	2086	2640
$[100, 1000)$	1712	464	2605	654	1946	2900
$[1000, +\infty)$	115	54	196	19	162	235
$[0, +\infty)$	4573	1292	6978	1052	4194	5575

Table 6.2: *Composition of the US dataset according to domain (mainstream vs disinformation), size class (number of unique users who interact with a given news) and political bias.*

Size Class	**No. Mainstream**	**No. Disinformation**
$[0, 100)$	165	79
$[100, 1000)$	61	158
$[0, +\infty)$	226	237

Table 6.3: *Composition of the Italian dataset according to domain (mainstream vs disinformation) and size class (number of unique users who interact with a given news).*

the FP rate, shows the ability of a classifier to discriminate positive samples from negative ones as its threshold is varied; the AUROC value is in the range $[0, 1]$, with the random baseline classifier holding AUROC$= 0.5$ and the ideal perfect classifier AUROC$= 1$; thus larger AUROC values (and steeper ROCs) correspond to better classifiers.

In particular we computed so-called *macro* average–simple unweighted mean–of these metrics evaluated considering both labels (*disinformation* and *mainstream*). We employed stratified shuffle split cross validation (with 10 folds) to evaluate performance.

6.2.11 Limitations

As mentioned beforehand, we use a coarse approach to label articles at the source level relying on a huge corpus of literature on the subject. We believe that this is currently the most reliable classification approach, although it entails obvious limitations, as disinformation outlets may also publish true stories and likewise misinformation is sometimes reported on mainstream media [138]. Also, given the choice of news sources, we cannot test whether our methodology is able to classify disinformation vs factual but not mainstream news which are published on niche, non-disinformation outlets.

Another crucial aspect in our approach is the capability to fully capturing sharing cascades on Twitter associated to news articles. It has been reported [163] that the Twitter streaming endpoint filters out tweets matching a given query if they exceed 1% of the global daily volume[2] of shared tweets, which nowadays is approximately $5 \cdot 10^8$; however, as we always collected less than 10^6 tweets per day, we did not incur in this issue and we thus gathered 100% of tweets matching our query.

[2] `https://www.internetlivestats.com/twitter-statistics/`

Classifier	Dataset	Recall	Precision	F1-Score
Logistic Regression	D_{all}	0.71 (sd 0.02)	0.74 (sd 0.02)	0.71 (sd 0.02)
	$D_{[0,100)}$	0.65 (sd 0.01)	0.70 (sd 0.01)	0.65 (sd 0.01)
	$D_{[100,1000)}$	0.75 (sd 0.02)	0.76 (sd 0.01)	0.74 (sd 0.02)
	$D_{[1000,+\infty)}$	0.85 (sd 0.06)	0.86 (sd 0.06)	0.85 (sd 0.06)
K-NN (k=10)	D_{all}	0.70 (sd 0.01)	0.72 (sd 0.02)	0.70 (sd 0.01)
	$D_{[0,100)}$	0.67 (sd 0.02)	0.70 (sd 0.01)	0.67 (sd 0.02)
	$D_{[100,1000)}$	0.76 (sd 0.02)	0.76 (sd 0.02)	0.76 (sd 0.02)
	$D_{[1000,+\infty)}$	0.84 (sd 0.04)	0.84 (sd 0.04)	0.84 (sd 0.04)

Table 6.4: *Evaluation metrics for Logistic Regression and Random Forest classifiers, built using global network properties. We report average values and standard deviations (in brackets) from 10-fold cross validation.*

We built Twitter diffusion networks using an approach widely adopted in the literature [33,225,226]. We remark that there is an unavoidable limitation in Twitter Streaming API, which does not allow to retrieve *true* re-tweeting cascades because re-tweets always point to the original source and not to intermediate re-tweeting users [92, 254]; thus we adopt the only viable approach based on Twitter's public availability of data. However, by disentangling different interactions with multiple layers we potentially reduce the impact of this limitation on the global network properties compared to the approach used in our baseline.

Finally, a limitation of the present work is the lack of a direct comparison of our methodology with other techniques, an exercise which boils down to assessing several classification metrics on the same dataset(s). As thoroughly discussed in [36], the problem of reliably comparing fake-news classifiers is open and faces many types of challenges that go out the scope of this work. We just mention that the performance of our classification framework is quantitatively comparable (in terms of AUROC value) to that of state-of-the-art deep learning models for fake news detection [162, 277]. However, this result is only indicative, because obtained on different datasets and, in one case [277], with different focus of the classification task.

6.3 Results of the single-layer approach

In this section we show and discuss results of the single-layer approach as presented in [194].

6.3.1 Experiments

Before evaluating global network properties in a classification task, we employed a non-parametric statistical test, Kolmogorov-Smirnov (KS) test, to verify the null hypothesis (each individual feature has the same distribution in the two classes). Hypothesis is rejected ($\alpha = 0.05$) for all indicators in all data subsets, with a few exceptions on

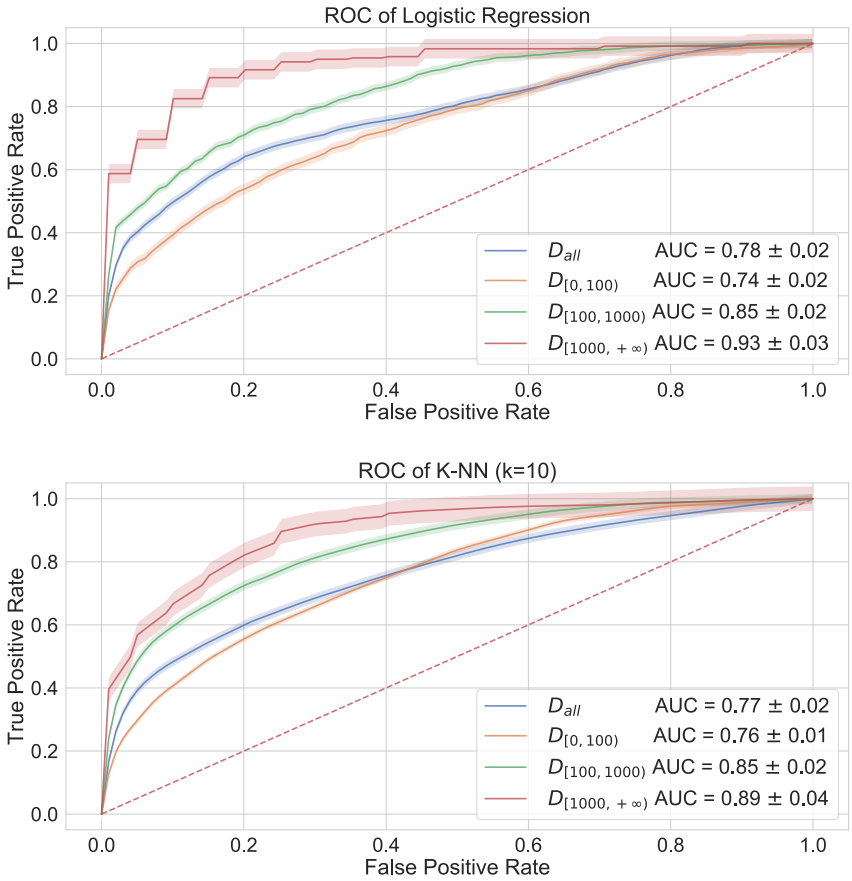


Figure 6.5: *ROC curves for Logistic Regression and K-NN (with $k = 10$) classifiers evaluated using global network properties. The dashed line corresponds to the ROC of a random classifier baseline with AUC=0.5.*

networks of larger size. More details are available in the Supplementary Information in Appendix A.

We then employed these features to train two traditional classifiers, namely Logistic Regression (LR) and K-Nearest Neighbors (K-NN) (with different choices of the number k of neighbors). Experiments on other state-of-the-art classifiers, which exhibit comparable results, are described in the Supplementary Information in Appendix A. Before training each model, we applied feature normalization, as commonly required in standard machine learning frameworks [7]. Finally we evaluated performances of both classifiers using a 10-fold stratified-shuffle-split cross validation approach, with 90% of the samples as training set and 10% as test set in each fold. In Table 6.4 we show values for Precision, Recall and F1-score, and in Figure 6.5 we show the resulting

Receiver Operating Characteristic curve (ROC) for both classifiers with corresponding Area Under the Curve (AUC) values. The performance is in all cases much better than that of a random classifier.

Next, we considered two specific classifiers, Support Vector Machines (SVM) and K-NN, applied to the network similarity matrix computed considering network distances (DGCD and PD). In Figure 6.6, we report the Area Under the Receiver Operating Characteristic curve (AUROC) values for the K-NN classifier, trained on top of PD and DGCD-13 similarity matrix; we excluded SVM as it was considerably outperformed (results are available in the Supplementary Information in Appendix A). DGCD-13 was evaluated only on networks with less than 1000 nodes (which still account for over 95% of the data) as the computational cost for larger networks was prohibitive. We carried out the same cross validation procedure as previously described. Again, the performance of the classifiers is in all instances much better than the baseline random classifier value.

Finally, we carried out several tests to assess the robustness of our classification framework when taking into account the political bias of sources, by computing global network properties and evaluating the performances of several classifiers (including balanced versions of Random Forest and Adaboost classifiers using `imblearn` Python package [139]). We first classified networks altogether excluding two specific sources of disinformation articles, namely "breitbart.com" and "politicususa.com", one at a time and both at the same time; we carried out these tests as these are very prolific sources (the former has by far the largest number of networks among right-biased sources, which is 4 times larger than "infowars.com", the second uppermost right-biased source; similarly, the latter has 10 times the number of networks of "activistpost.com", the second left-biased source). We then evaluated classifiers performance in two different scenarios, i.e. including in training data only mainstream and disinformation networks with a specific bias (in turn left and right) and testing on the entire set of sources; in Figure 6.7 we show the resulting Receiver Operating Characteristic curve (ROC) with corresponding Area Under the Curve (AUC) values. Results are in all cases better than those of a random classifier and in agreement with the result of the classifier developed without excluding any source from the training and test sets; a more detailed description of aforementioned classification results is available in the Supplementary Information in Appendix A.

6.3.2 Discussion

In a nutshell, we demonstrated that our choice of basic global network properties provides an accurate classification of news articles based solely on their Twitter diffusion networks–AUROC in the range 0.75-0.93 with basic K-NN and LR, and comparable or better performances with other state-of-the-art classifiers (see Supplementary Informa-

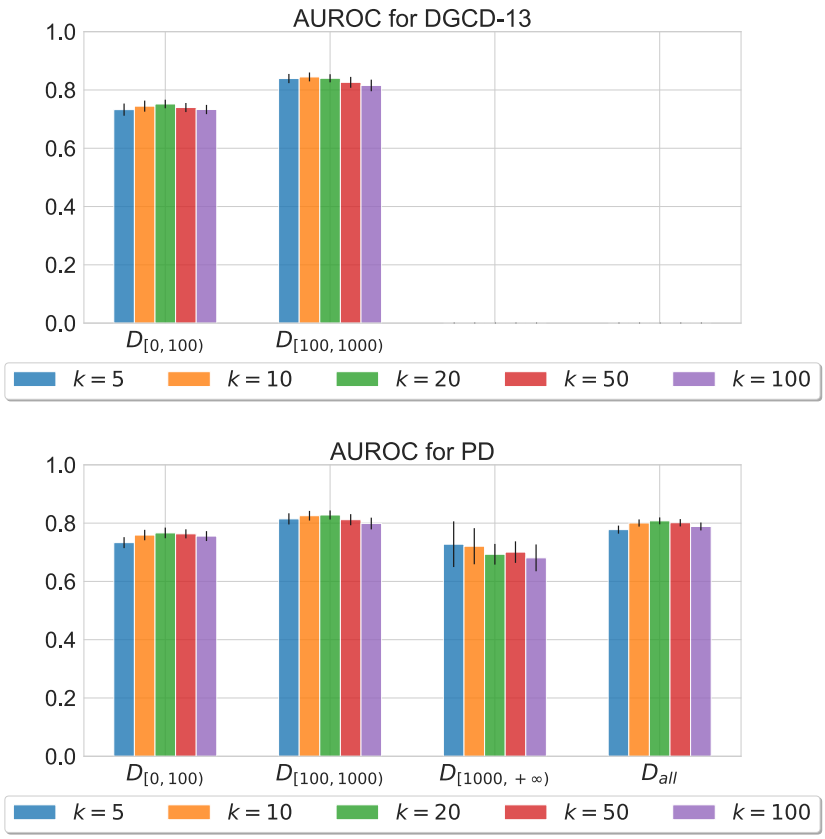


Figure 6.6: *AUROC values for K-NN classifiers (with different choices of k) using PD and DGCD-13 distances.*

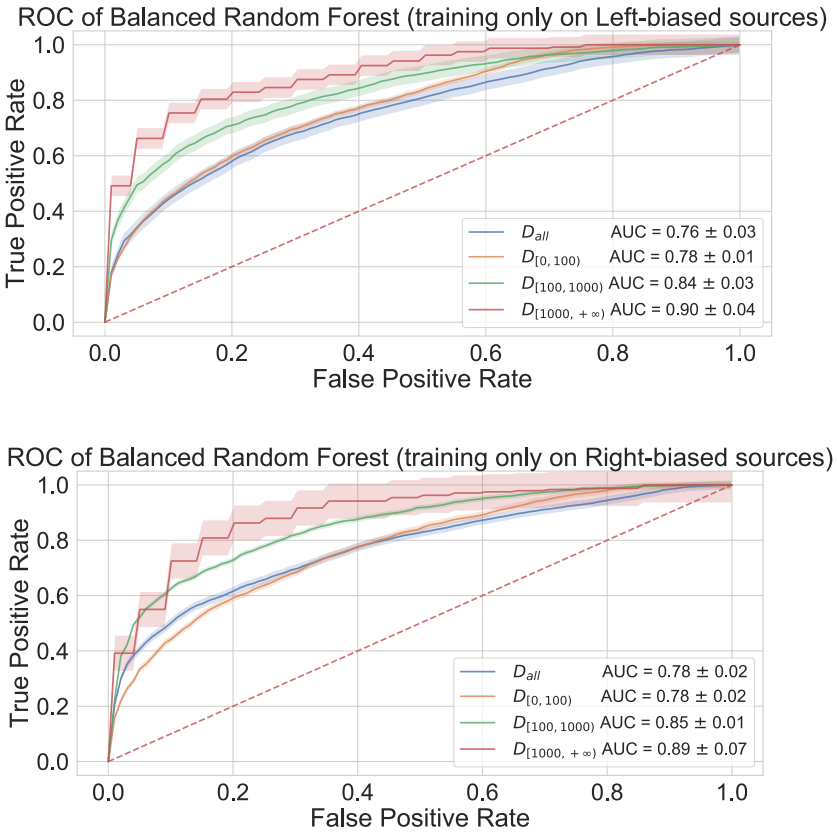


Figure 6.7: *ROC curves for a balanced Random Forest classifier, evaluated using global network properties, training only on left-biased (top) or right-biased (bottom) sources and testing using all sources. The dashed line corresponds to the ROC of a random classifier baseline with AUC*=0.5.

tion); these results hold also when considering news based on the political bias of their sources. The use of more sophisticated network distances confirms the result, which is altogether in accordance with prior work on the detection of online political *astroturfing* campaigns [202], and two more recent network-based contributions on *fake news* detection [162, 278]. However, we remark that, given the composition of our dataset, we did not test whether our methodology allows to accurately classify disinformation information vs factual but non-mainstream news which is produced by niche outlets.

For what concerns global network properties, comparing networks with similar sizes turned out to be the right choice, yielding a general increase in all classification metrics (see Supplementary Information in Appendix A for more details). We experienced the worst performances when classifying networks with smaller sizes (with less than 100 nodes); we argue that small diffusion networks appear more similar and that differences across news domains emerge particularly when their size increases.

For what concerns network distances, they overall exhibit a similar trend in classification performances, with worst results on networks with less than 100 nodes and a slight improvement when considering the entire dataset; accuracy in classifying networks with more that 1000 nodes is lower, perhaps due to data scarcity. DCGD and PD distances appear equivalent in our specific classification task; the former is generally used in biology to efficiently cluster together similar networks and identify associated biological functions [218]. They reinforce the results of our more naive approach involving a manual selection of the input features. Understanding classification results in terms of input features is notably a controversial problem in machine learning [141]. In the following we give our own qualitative interpretation of the results in terms of global network properties.

For networks with less than 1000 nodes, we observed that disinformation networks exhibit higher values of both size and diameter of the largest weakly connected components; recalling that the largest weakly connected component corresponds to the largest cascade, this result is in accordance with [254] where it is shown that false rumor cascades spread deeper and broader than true ones.

For networks with more than 100 nodes, we noticed higher values of both size of the largest strongly connected component and clustering coefficient in disinformation networks compared to mainstream ones. This denotes that communities of users sharing disinformation news tend to be more connected and clustered, with stronger interaction between users, whereas mainstream articles are shared in a more broadcast manner with less discussion between users. A similar result was reported in [120] where a sample of most shared news was inspected in the context of 2016 US presidential elections. Conversely, mainstream networks manifest a much larger number of weakly connected components (or cascades). This is not surprising since traditional outlets have a larger audience than websites sharing disinformation news. [33, 100].

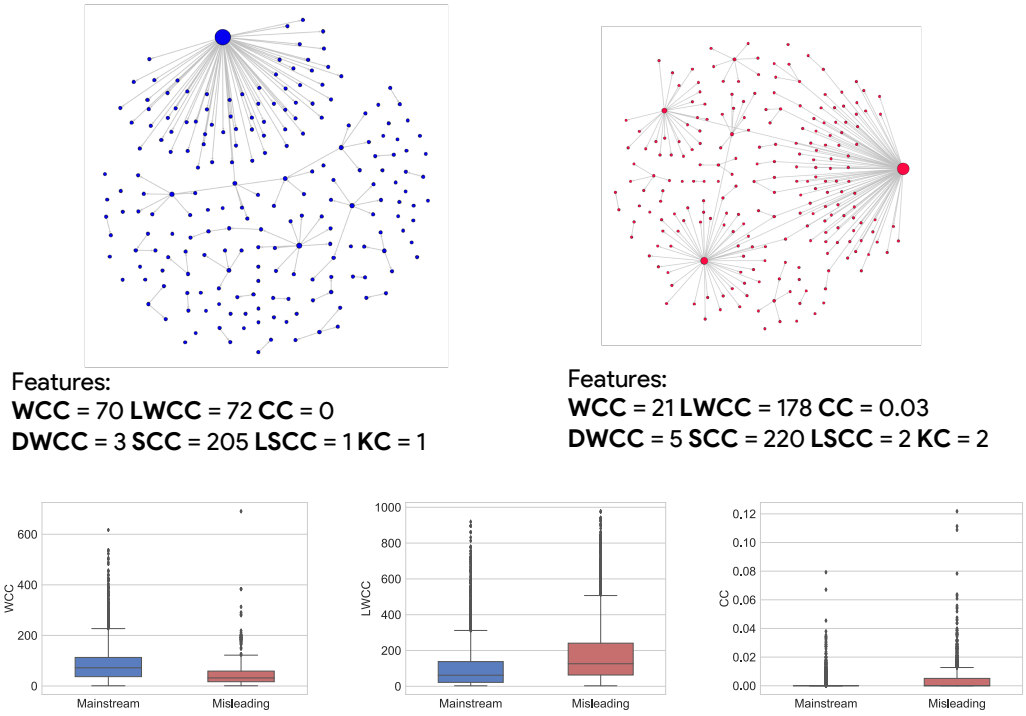


Figure 6.8: *Top. Prototypical examples (the nearest individuals) of two diffusion networks in the subset $D_{[100,1000)}$ of mainstream (left) and disinformation (right) networks. The size of nodes is adjusted according to their degree centrality, i.e. the higher the degree value the larger the node.*
*Middle. Feature values corresponding to the two examples (**WCC** = Number of Weakly Connected Components; **LWCC** = Size of the Largest Weakly Connected Component; **CC** = Average Clustering Coefficient; **DWCC** = Diameter of the Largest Weakly Connected Components; **SCC** = Number of Strongly Connected Components; **LSCC** = Size of the Largest Strongly Connected Component; **KC** = Main K-Core Number).*
Bottom. Box-plots of values of the three most significant features–WCC, LWCC, CC–highlighting different distributions in the $D_{[100,1000)}$ subset of the two news domains.

Finally, we observed that the main K-core number takes higher values indisinformation networks rather than in mainstream ones. This result confirms considerations from [226] where authors perform a K-core decomposition of a massive diffusion network produced on Twitter in the period of 2016 US presidential elections; they show that low-credibility content proliferates in the core of the network. More details on differences between news domains, according to the size of diffusion networks, are available in the Supplementary Information in Appendix A.

A pictorial representation of these properties is provided in Figure 6.8, where we display two networks, with comparable size, which represent the *nearest* individuals pertaining to both news domains in the $D_{[100,1000)}$ subset, i.e. the network with the smallest Euclidean distance–in the feature space of global network properties–from all other individuals in the same domain. Although they may appear similar at first

sight, they actually exhibit different global properties. In particular we observe that the disinformation network has a non-zero clustering coefficient, and higher value of size and diameter of the largest weakly connected component, but a smaller number of weakly connected components w.r.t to the mainstream network. Additional examples relative to other subsets are available in the Supplementary Information in Appendix A.

Coordinated activity and political bias

As far as the well known deceptive efforts to manipulate the regular spread of information are concerned (see for instance the documented activity of Russian trolls [17, 238] and the influence of automated accounts in general [225]), we argue that such hidden forces might indeed play to accentuate the discrepancies between the diffusion patterns of disinformation and mainstream news thus enhancing the effectiveness of our methodology.

As far as the political bias of sources is concerned, a few contributions [23, 34, 49] report differences in how conservatives and liberals socially react to relevant political events. For instance, Conovet et al. [49] report discrepancies in a few network indicators (e.g. average degree and clustering coefficient) among three specific pairs of diffusion networks (one for right-leaning users and one for left-leaning users), namely those described by follower/followee relationships between users, re-tweets and mentions, which however differ from the URL-based diffusion networks which we analyzed in this work. In addition they carried out their analysis in a different experimental setting w.r.t to ours: they used Twitter Gardenhose API (which collects a random 10% sample of tweets) and filtered tweets based on political hashtags. Although they found that right-leaning users shared hyperlinks (not necessarily news) 43% of the times compared to 36.5% of left-leaning users (percentages that grow to 51% and 62.5% in case of tweets classified as political), their findings are not commensurable to our topology comparison. Similarly, other work [23, 33, 34] used different experimental settings w.r.t to our data collection, they did not specifically focus on URL-wise diffusion networks, and they carried out analyses with approaches not comparable to ours. Nevertheless, despite any possible influence of the political leaning on some features of the diffusion patterns, our methodology proves to be insensitive to the presence of political biases in news sources, as we are capable of accurately distinguishing disinformation and mainstream news in several different experimental scenarios. The presence of specific outlets of disinformation articles which outweigh the others in terms of data samples (respectively "breitbart.com" for right-biased networks and "politicususa.com" for left-biased networks) does not affect the classification, as results are similar even when excluding articles from these sources. Also, when we considered only left-biased (or right-biased) mainstream and disinformation articles in the training data while including all sources in the test set, we observed results which are in accordance with our

Size Class	AUROC	Precision	Recall	F1-score
(US) $[0, 100)$	0.87 ± 0.01	0.79 ± 0.01	0.77 ± 0.01	0.78 ± 0.01
(US) $[100, 1000)$	0.93 ± 0.01	0.87 ± 0.01	0.87 ± 0.01	0.87 ± 0.01
(US) $[1000, +\infty)$	0.94 ± 0.02	0.86 ± 0.05	0.86 ± 0.05	0.86 ± 0.05
(US) $[0, +\infty)$	0.88 ± 0.01	0.81 ± 0.01	0.80 ± 0.01	0.80 ± 0.01
(IT) $[0, 100)$	0.89 ± 0.06	0.81 ± 0.11	0.82 ± 0.11	0.81 ± 0.11
(IT) $[100, 1000)$	0.86 ± 0.07	0.83 ± 0.08	0.78 ± 0.06	0.80 ± 0.06
(IT) $[0, +\infty)$	0.90 ± 0.02	0.81 ± 0.05	0.81 ± 0.05	0.81 ± 0.05

Table 6.5: *Performance of the LR classifier (using a multi-layer approach) evaluated on different size classes on both the US (top rows) and the Italian (bottom rows) dataset.*

general aforementioned findings, for what concerns both classification performances and features distributions. Overall, our results show that mainstream news, regardless of their political bias, can be accurately distinguished from disinformation news.

6.4 Results of the multi-layer approach

In this section we show and discuss results of the multi-layer approach as presented in [193].

6.4.1 Experiments

Using the single-layer setting as baseline, we performed classification experiments using a basic off-the-shelf classifier, namely Logistic Regression (LR) with L2 penalty. We applied a standardization of the features and we used the default configuration for parameters as described in `scikit-learn` package [182]. We also tested other classifiers (such as K-Nearest Neighbors, Support Vector Machines and Random Forest) but we omit results as they give comparable performance. We remark that our goal is to show that a very simple machine learning framework, with no parameter tuning and optimization, allows for accurate results with our network-based approach.

Finally, we partitioned networks according to the total number of unique users involved in the sharing, i.e. the number of nodes in the aggregated network represented with a single-layer representation considering together all layers and also isolated tweets.

6.4.2 Classification performance

In Table 6.5 we first provide classification performance on the US dataset for the LR classifier evaluated on the size classes described in Table 6.1. We can observe that in all instances our methodology performs much better than a random classifier (50% AUROC), with AUROC values above 85% in all cases.

For what concerns political biases, as the classes of mainstream and disinformation networks are not balanced (e.g., 1,292 mainstream and 4,149 disinformation networks

Size Class	Single-layer	Multi-layer
(US) $[0, 100)$	0.74 ± 0.02	0.87 ± 0.01
(US) $[100, 1000)$	0.85 ± 0.02	0.93 ± 0.01
(US) $[1000, +\infty)$	0.93 ± 0.03	0.94 ± 0.02
(US) $[0, +\infty)$	0.78 ± 0.02	0.88 ± 0.01
(IT) $[0, 100)$	0.77 ± 0.08	0.89 ± 0.06
(IT) $[100, 1000)$	0.66 ± 0.14	0.86 ± 0.07
(IT) $[0, +\infty)$	0.74 ± 0.12	0.90 ± 0.02

Table 6.6: *Comparison of performance of our multi-layer approach vs the baseline (single-layer). We show AUROC values for the LR classifier evaluated on different size classes of both US and IT datasets.*

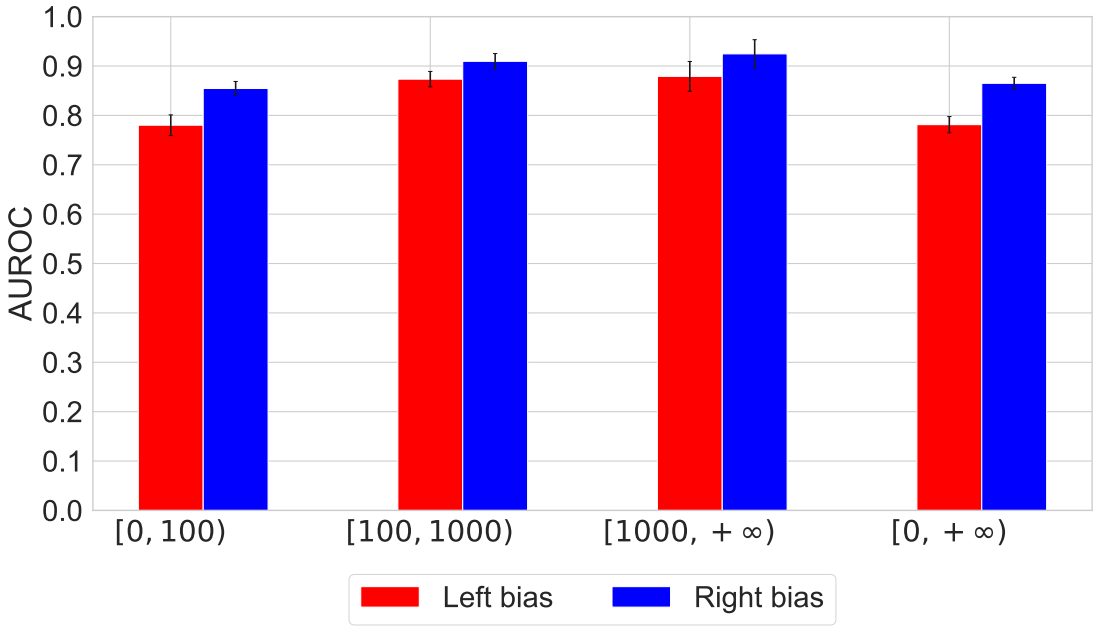


Figure 6.9: *AUROC values for the Balanced Random Forest classifier trained on left-biased (red) and right-biased (blue) news articles in the US dataset, and tested on the entire dataset. Error bars indicate the standard deviation of AUROC values over different folds of the cross validation.*

with right bias) we employ a Balanced Random Forest with default parameters (as provided in `imblearn` Python package [139]). In order to test the robustness of our methodology, we trained only on left-biased networks or right-biased networks and tested on the entire set of sources (relative to the US dataset); we provide a comparison of AUROC values for both biases in Figure 6.9. We can notice that our multi-layer approach still entails significant results, thus showing that it can accurately distinguish mainstream news from disinformation regardless of the political bias. We further corroborated this result with additional classification experiments, that yield similar performance, in which we excluded from the training/test set two specific sources (one at a time and both at the same time) that outweigh the others in terms of data samples– respectively "breitbart.com" for right-biased sources and "politicususa.com" for left-

biased ones [36].

We performed classification experiments on the Italian dataset using the LR classifier and different size classes (notice that $[1000, +\infty)$ is empty for the Italian dataset); we show results for different evaluation metrics in Table 6.5. We can see that despite the limited number of samples (one order of magnitude smaller than the US dataset) the performance is overall in accordance with the US scenario.

As shown in Table 6.6, we obtain results which are much better than our baseline in all size classes:

- In the US dataset our multi-layer methodology performs much better in all size classes except for large networks ($[1000, +\infty)$ size class), reaching up to 13% improvement on smaller networks ($[0, 100)$ size class);
- In the IT dataset our multi-layer methodology outperforms the baseline in all size classes, with the maximum performance gain (20%) on medium networks ($[100, 1000)$ size class); the baseline generally reaches worst performance compared to the US scenario.

6.4.3 Layer importance analysis

In order to understand the impact of each layer on the performance of classifiers, we performed additional experiments considering separately each layer (we ignored **T** and **U** features relative to isolated tweets).

In Table 6.7 we show metrics for each layer and all size classes, computed with a 10-fold stratified shuffle split cross validation, evaluated on the US dataset; in Figure 6.10 we show AUROC values for each layer compared with the general multi-layer approach. We can notice that both Q and M layers alone capture adequately most of discrepancies of the two distinct news domains in the United States as they obtain good results with AUROC values in the range 75%-86%; these are comparable with those of the multi-layer approach which, nevertheless, outperforms them across all size classes.

In the Italian dataset we observe that the M layer obtains comparable performance w.r.t the multi-layer approach for what concerns small networks and the dataset altogether, whereas the RT layer performs better on large networks (see Table 6.8 and Figure 6.11). We also notice higher values in standard deviations of performance metrics which are likely due to the limited size of the training/test data.

6.4.4 Feature importance analysis and cross-country experiments

We further investigated the importance of each feature by performing a χ^2 test, with 10-fold stratified shuffle split cross validation, considering the entire range of network sizes $[0, +\infty)$. We show the Top-5 most discriminative features for each country in Table 6.9.

Size Class	Metric	Quotes	Retweets	Mentions	Replies
$[0, 100)$	AUROC	$\mathbf{0.75 \pm 0.02}$	0.63 ± 0.02	$\mathbf{0.75 \pm 0.02}$	0.61 ± 0.02
	Precision	$\mathbf{0.71 \pm 0.02}$	0.59 ± 0.02	0.70 ± 0.02	0.60 ± 0.04
	Recall	0.66 ± 0.01	0.55 ± 0.01	$\mathbf{0.67 \pm 0.01}$	0.54 ± 0.02
	F1-score	0.66 ± 0.02	0.53 ± 0.02	$\mathbf{0.68 \pm 0.02}$	0.50 ± 0.06
$[100, 1000)$	AUROC	$\mathbf{0.81 \pm 0.02}$	0.63 ± 0.02	$\mathbf{0.81 \pm 0.02}$	0.65 ± 0.03
	Precision	0.73 ± 0.02	0.61 ± 0.02	$\mathbf{0.75 \pm 0.02}$	0.65 ± 0.02
	Recall	0.73 ± 0.02	0.60 ± 0.02	$\mathbf{0.75 \pm 0.02}$	0.62 ± 0.02
	F1-score	0.73 ± 0.02	0.60 ± 0.02	$\mathbf{0.75 \pm 0.02}$	0.60 ± 0.02
$[1000, +\infty)$	AUROC	$\mathbf{0.85 \pm 0.08}$	0.62 ± 0.08	0.84 ± 0.04	0.66 ± 0.06
	Precision	$\mathbf{0.80 \pm 0.08}$	0.61 ± 0.08	0.75 ± 0.06	0.61 ± 0.10
	Recall	$\mathbf{0.80 \pm 0.08}$	0.60 ± 0.07	0.75 ± 0.06	0.59 ± 0.07
	F1-score	$\mathbf{0.79 \pm 0.08}$	0.59 ± 0.08	0.75 ± 0.06	0.58 ± 0.09
$[0, +\infty)$	AUROC	0.76 ± 0.01	0.62 ± 0.01	$\mathbf{0.77 \pm 0.01}$	0.59 ± 0.04
	Precision	0.70 ± 0.01	0.58 ± 0.01	$\mathbf{0.73 \pm 0.01}$	0.59 ± 0.05
	Recall	0.69 ± 0.01	0.56 ± 0.01	$\mathbf{0.71 \pm 0.01}$	0.55 ± 0.03
	F1-score	0.69 ± 0.01	0.53 ± 0.01	$\mathbf{0.71 \pm 0.01}$	0.52 ± 0.05

Table 6.7: *Different evaluations metrics for LR classifier evaluated on different size classes of the US dataset and trained using features separately for each layer. Best scores for each row are written in bold.*

Size Class	Metric	Quotes	Retweets	Mentions	Replies
$[0, 100)$	AUROC	0.49 ± 0.12	0.73 ± 0.08	$\mathbf{0.74 \pm 0.06}$	0.49 ± 0.09
	Precision	0.34 ± 0.00	$\mathbf{0.61 \pm 0.15}$	0.58 ± 0.08	0.34 ± 0.00
	Recall	0.50 ± 0.00	$\mathbf{0.63 \pm 0.13}$	0.57 ± 0.07	0.50 ± 0.00
	F1-score	0.40 ± 0.00	$\mathbf{0.61 \pm 0.13}$	0.57 ± 0.07	0.40 ± 0.00
$[100, 1000)$	AUROC	0.64 ± 0.10	$\mathbf{0.80 \pm 0.07}$	0.62 ± 0.11	0.51 ± 0.07
	Precision	0.59 ± 0.18	$\mathbf{0.77 \pm 0.13}$	0.74 ± 0.15	0.66 ± 0.18
	Recall	0.56 ± 0.08	$\mathbf{0.67 \pm 0.10}$	0.64 ± 0.08	0.56 ± 0.07
	F1-score	0.55 ± 0.11	$\mathbf{0.67 \pm 0.11}$	0.65 ± 0.10	0.56 ± 0.08
$[0, +\infty)$	AUROC	0.72 ± 0.08	0.72 ± 0.06	$\mathbf{0.82 \pm 0.07}$	0.51 ± 0.05
	Precision	0.66 ± 0.09	0.75 ± 0.06	$\mathbf{0.76 \pm 0.06}$	0.53 ± 0.06
	Recall	0.66 ± 0.09	0.70 ± 0.04	$\mathbf{0.75 \pm 0.06}$	0.51 ± 0.03
	F1-score	0.66 ± 0.09	0.70 ± 0.04	$\mathbf{0.75 \pm 0.06}$	0.47 ± 0.04

Table 6.8: *Different evaluations metrics for LR classifier evaluated on different size classes of the IT dataset and trained using features separately for each layer. Best scores for each row are written in bold.*

We find exactly the same set of features (with different rankings in the Top-3) in both countries; these correspond to two global network properties–LWCC, which indicates the size of the largest cascade in the layer, and SCC, which correlates with the size of the layer ($\rho \approx 0.99$, with $p \approx 0$ in all cases)–associated to the same set of layers (Q, RT and M).

We further performed a χ^2 test to highlight the most discriminative features in the M layer of both countries, which performed equally well in the classification task as previously highlighted; also in this case we focused on the entire range of network sizes $[0, +\infty)$. Interestingly, we discovered exactly the same set of Top-3 features in both countries, namely LWCC, SCC and DWCC (which indicates the depth of the

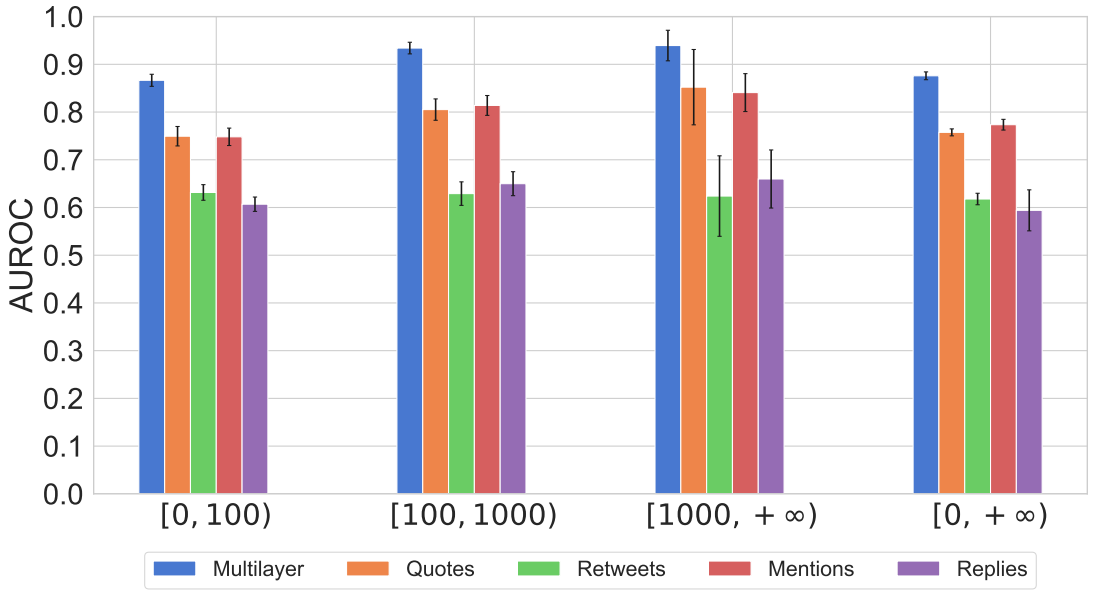


Figure 6.10: *AUROC values for the LR classifier (evaluated on different size classes of the US dataset) trained using different layers separately and together (our multi-layer approach). Error bars indicate the standard deviation of AUROC values over different folds of the cross validation.*

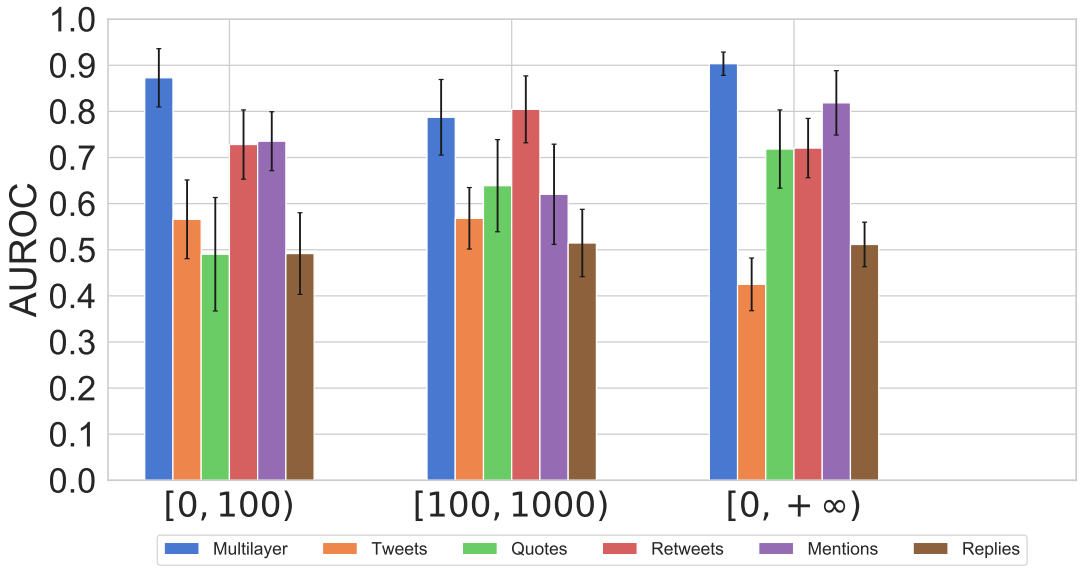


Figure 6.11: *AUROC values for the LR classifier (evaluated on different size classes of the IT dataset) trained using different layers separately and together (our multi-layer approach). Error bars indicate the standard deviation of AUROC values over different folds of the cross validation.*

largest cascade in the layer). We performed a Kolmogorov-Smirnov two-sample test to assess whether distributions of these features are statistically equivalent across the two news domains; the hypothesis was rejected in all cases at $\alpha = 0.05$.

The similarities evidenced so far in both countries–i.e., classification performance of

Rank	**US**	**IT**
#1	SCC (Quotes)	LWCC (Retweets)
#2	LWCC (Retweets)	SCC (Retweets)
#3	SCC (Retweets)	SCC (Quotes)
#4	LWCC (Quotes)	LWCC (Quotes)
#5	LWCC (Mentions)	LWCC (Mentions)

Table 6.9: *Top-5 most discriminative features according to χ^2 test evaluated on both US and IT datasets (considering networks in the $[0, +\infty)$ size class).*

single layers and features importance–might suggest that the two news domains exhibit discrepancies which are geographic-independent. We further investigated this hypothesis by testing the performance of both LR and Balanced Random Forest classifiers in several cross-country settings, e.g. training on the US dataset and testing on the Italian (and viceversa), performing feature normalization either over the entire data or separately for training and test sets, to investigate whether we can classify disinformation vs mainstream news regardless of the country where they originate. Interestingly, performance is in all cases worse or equal than those of a random classifier (AUROC=50%); this might be due either to the high imbalance of data across the two countries, or most likely suggests that sharing patterns of the two news domains exhibit coupled dissimilarities which are very country specific.

6.5 Conclusions and future work

Following the latest insights on the characterization of disinformation news spreading on social media compared to more traditional news, we investigated the topological structure of Twitter diffusion networks pertaining to distinct domains. Leveraging different network comparison approaches, from manually selected global properties to more elaborated network distances, we corroborate what previous research has suggested so far: disinformation content spreads out differently from mainstream and reliable news, and dissimilarities can be remarkably exploited to classify the two classes of information using purely topological tools, i.e. basic global network indicators and standard machine learning.

We disentangled different types of interactions on Twitter to accordingly build a multi-layer representation of news diffusion networks, and we computed a set of global network properties–separately for each layer–in order to encode each network with a tuple of features. Our goal was to investigate whether a multi-layer representation performs better than an aggregated one, and to understand which of the features, observed at given layers, are most effective in the classification task.

Experiments with an off-the-shelf classifier such as Logistic Regression on datasets pertaining to two different media landscapes (US and Italy) yield very accurate classification results (AUROC up to 94%), also when controlling for the different political bias

of news sources, with improvements up to 20% w.r.t the single-layer baseline. Classification performance using individual layers shows that the layer of mentions alone entails better performance w.r.t. other layers in both countries, pointing in both cases to a peculiar usage of this type of Twitter interaction across the two domains.

We also highlighted the most discriminative features across different layers in both countries; we noticed the exact same set of features, suggesting, at first glance, that differences between the two news domains might be country-independent and rather due only to the typology of content shared. However, the two news domains exhibit coupled dissimilarities in sharing patterns which appear to be very country specific, and our methodology fails to detect disinformation regardless of where it originates.

Overall, our results prove that the topological features of multi-layer diffusion networks might be effectively exploited to detect online disinformation. Notice that we do not deny the presence of deceptive efforts to orchestrate the regular spread of information on social media via content amplification and manipulation [17, 238]. On the contrary, we postulate that such hidden forces might play to accentuate the discrepancies between the diffusion patterns of disinformation and mainstream news (and thus to make our methodology effective).

In the future we aim to further investigate two main directions: (1) employ temporal networks to represent news diffusion and apply classification techniques (e.g. recurrent neural networks) that take into account the sequential aspect of data; (2) leverage our network-based features in addition to state-of-the-art text-based approaches for "fake news" detection in order to deliver a real-world system to detect misleading and harmful information spreading on social media.

We believe that future research directions might successfully exploit these results to develop real world applications that could resolve and mitigate malicious information spreading on social media.

CHAPTER 7

Epilogue

Ever since the 2016 U.S. Presidential elections, the research community has focused its attention on understanding the role played by online social media in the diffusion and amplification of mis- and disinformation narratives. The problem of false information is not novel, but it has become crucial in the era of Internet and social media, as barriers to enter the media industry dropped and general trust towards mass media collapsed.

We discussed how most of the research tackled the problem of promptly detecting false news articles, with a variety of techniques and different approaches. However, the massive amount of information generated on social media hinders the possibility to build a universal classifier that can automatically identify "fake news"; to date, platforms mostly rely on human fact-checkers and crowd-sourcing efforts to identify malicious content and remove it.

A few studies addressed the problem from a different perspective, with the goal of understanding who are the actors involved in the spreading of harmful content, both actively and passively, and what are the mechanisms that exacerbate the phenomenon on online social media. They found that both human and algorithmic factors play a major role, but they also highlighted how malicious agents such as bots, cyborgs and trolls are deployed to manipulate and influence public opinion.

We also saw that there are many challenges that researchers face when tackling the problem of online disinformation. First and foremost, the "big data" that flood on social

media every day and, at the same time, the lack of access to relevant data on platforms for researchers, as companies rush to remove content which might impact their reputation and, thus, their business. Besides, researchers from different disciplines still struggle to find a common ground when it comes to formal definitions of the problem.

Overall, much remains unknown for what regards the vulnerabilities of individuals, institutions, and society to the spread of online mis- and disinformation, and their backlashes on the real world. Consequently, the research community has been promoting interdisciplinary research to address the problem from multiple perspectives.

Recently, the COVID-19 pandemic and the *infodemic* that followed, have urged the need for addressing the proliferation of low-credibility and inflammatory content that overrun online social media. Furthermore, the global crisis also exposed the weaknesses of platforms, and called for intervention by governments worldwide.

7.1 Summary of the contributions

I can summarize the contributions of the thesis as follows.

First, in Chapter 2, I provided the reader with the background required to follow the output of my own research, presented in later chapters. I started from an overview of the terminology and the problem formulation(s), and I presented existing literature on the detection, characterization and mitigation of mis- and disinformation spreading in online social networks. This review was part of my preliminary work during the first half of my first year as a Ph.D. student, as I started exploring research on the topic in order to formulate my own research questions, and it became part of a review paper published on ACM SIGMOD Record [190].

In Chapter 3, I described my contribution to the problem of understanding the spread of Italian language disinformation spreading on Twitter, which I tackled during the second half of my 1st year. The intuition for the research presented there [189], published on PLoS One, came from previous work [226] which analyzed the spread of English language online mis- and disinformation, using a source-based approach to track unreliable news articles. During my thorough literature review, I had noticed that most of previous research focused on the U.S., due to the attention raised by 2016 Presidential elections, and I was eager to investigate what was going on in my own country. I followed their methodology to a certain extent, but I also introduced additional techniques to gain further insights, from understanding the agenda setting effect to find inter-connections between deceptive websites. I had the possibility to co-supervise a M.Sc. student, Alessandro Artoni, and we eventually extended his thesis to write a scientific paper. We showed that Italian disinformation in the run-up to the 2019 European Parliament elections focused on controversial topics such as immigration and refugees, and not specifically on the elections. We also highlighted the role of far-right

communities as the main sources of disinformation narratives, and we found evidence of inter-connections between different websites across Europe. All in all, our results aligned with findings from other research which focused on the United States.

In Chapter 4, I described the results of an international collaboration with the Observatory on Social Media of the Indiana University, where I spent a year virtually and a few months in presence. They had published a few preliminary analyses of COVID-19 related disinformation on Twitter [271], and I suggested to extend those analyses to Facebook and cover the entire 2020 when I stumbled in a call for papers, focused on the COVID-19 infodemic, of the Big Data & Society journal [101]. As I had previous experience with Crowdtangle [54], a tool which allows to collect data from Facebook, I was responsible for gathering, pre-processing and analyzing posts from Facebook public pages and groups, with a focus on the prevalence of COVID-19 related disinformation. We carried out a thorough comparison between the two platforms from multiple perspectives, and we incidentally analyzed the role of YouTube as a potential source of misinformation. The main result is, indeed, the role of "superspreaders", i.e., influential users that account for most of the misleading and false information spreading on both Twitter and Facebook. We argued that suspending, or at least moderating these accounts, would promptly reduce the spread of disinformation, with positive and visible effects on the pandemic. Following our publication, the concept of superspreaders became mainstream when (data) journalists found out that it was easy to spot as much as 12 accounts responsible for over 65% of the vaccine-related disinformation circulating on Facebook, Twitter and Instagram[1].

In Chapter 5, I moved on with other output(s) of my collaboration with the Observatory on Social Media, and a research project which I started on my own with some Italian colleagues. Following the roll-out of vaccination campaigns worldwide, we decided to keep track of conversations about vaccines taking place in online social media, with a specific focus on the influence and the impact of mis- and disinformation. For what concerns the U.S., I am co-responsible for Covaxxy[2], which is thoroughly described in the chapter, whereas in Italy I am responsible for Vaccinitaly[3], which will be soon extended to consider multiple European countries (under the H2020 Periscope[4]). In both cases, the broad research question is to understand whether vaccine-related disinformation has an impact on the vaccination programs. Both projects were presented at the 2021 International Conference of Weblogs and Social Media of the AAAI society. In particular, we provide an aswer to that question in the United States, as I discuss in the central sections of the chapter. By leveraging data from both Twitter and Face-

[1] https://www.npr.org/2021/05/13/996570855/disinformation-dozen-test-facebooks-twitters-ability-to-curb-vaccine-hoaxes

[2] https://osome.iu.edu/tools/covaxxy

[3] http://genomic.elet.polimi.it/vaccinitaly/

[4] https://www.periscopeproject.eu/start

book, we built a multiple regression model that takes into account several confounding factors, from politics to demographics, and we found that online disinformation is significantly associated to both vaccine uptake and vaccine hesitancy rates. This work is currently under evaluation for publication, and I plan to carry out similar analyses in the European context.

Finally, in Chapter 6, I described the results of my collaboration with Carlo Piccardi and my advisor Stefano Ceri, who helped me developing another line of research which specifically deals with the automatic detection of online disinformation, by means of network science and machine learning. The intuition to focus on sharing patterns came from reviewing the literature, as I discovered that a few relevant contributions, published on top journals such as Science and Nature Communications, had shed light on the differences in news sharing patterns when consuming reliable versus unreliable information. Thus, I planned to test the hypothesis that false and true news are shared differently, i.e., users shape diffusion patterns which are topologically distinct and that could be classified by a machine learning algorithm. I had the opportunity to deepen my knowledge of network science under the supervision of Carlo Piccardi, and our collaboration was indeed successful, as discussed in the chapter. Overall, we show that encoding diffusion networks by mean of a single or multi-layer representation and some topological features (in the graph theory sense and not in the purely mathematical one) allows to train off-the-shelf classifiers that can accurately tell whether the source of a news article is reliable or not. To date, we published the output of our research in two international journals, respectively Nature Scientific Reports and EPJ Data Science, and we are currently working on a third "chapter" where we leverage a temporal representation of networks.

7.2 Outlook

Despite the vast amount of contributions discussed above, we believe that the detection of false and misleading information spreading in online social networks requires a deeper and more structured approach. Several contributions appear as academic exercises, not always compared to each other (and often not comparable). Mostly, they achieve good performance when applied to given input dataset, but they do not generalize to unseen data. From our analysis, it seems that methods purely based upon content analysis work within a limited scope, whereas context analysis addresses generic actions (such as liking, commenting, propagating) that generalize more easily. Leveraging data from social interactions drives our own research discussed in chapter 6, and we plan to extend our methodology from different perspectives, e.g., employ temporal networks to represent news diffusion and apply classification techniques (e.g. recurrent neural networks) that take into account the sequential aspect of data and leverage

our network-based features in addition to state-of-the-art text-based deep learning approaches, with the final goal of building a real-world system that can promptly detect misleading and harmful information spreading on social media.

From a characterization perspective, most of previous research, including our contributions, focused on Twitter and, with some limitations, Facebook. In recent times, the attention has shifted to other social media such as Reddit, YouTube, Instagram, Whatsapp, etc, but there are still a lot of unanswered questions regarding the spread of disinformation across multiple platforms. Studies which consider other social media platforms will be needed and beneficial to our understanding of the spread of disinformation. However, such research is hindered by the lack of data access to researchers from the majority of social platforms, and we hope that efforts are made in that direction to allow academics to study the phenomenon.

With the ongoing COVID-19 pandemic, addressing the spread of low-credibility information is crucial and it raises many questions. For instance, how user demographics affect the consumption of unreliable information on social media? It is also unclear how social media platforms are handling the flow of information and allowing dangerous content to spread. As Twitter and Facebook increased their moderation of COVID-19 misinformation, they have been accused of political bias. Indeed, there are many legal and ethical considerations around free speech and censorship, but these questions should not be avoided and it is important to maintain an environment where individuals have access to good information that benefits public health.

APPENDIX $\mathcal{A}$

Supplementary Information for "A network-based approach to detect online disinformation on Twitter"

In this chapter we provide supplementary material for Chapter 6.

A.1 Mainstream and misleading news

A.1.1 Collecting mainstream news on Twitter

In order to gather mainstream news diffusion networks we followed the same approach as in [226]: we first used Twitter Streaming API, via `tweepy`[1] package in Python, to filter all tweets containing an URL matching the domains of top trusted sources specified in [159]. Among sources described in the research report, we only selected most reliable sources listed in Table A.1 (the bias is derived according to [33]). In particular we specified domains URLs as *track* parameter to call Twitter Api, e.g. "wsj com OR nytimes com OR news yahoo com. OR ..." as suggested by Twitter Developers documentation[2]. The tweets collected in this way contained more than 300k unique

[1] `http://docs.tweepy.org/en/v3.5.0/`

[2] `https://developer.twitter.com/en/docs/tweets/filter-realtime/guides/basic-stream-parameters`

URL; about 12k of them were associated with at least 50 tweets, further reduced to 6978 after handling censoring effects described in [92].

Newspaper	Domain	Bias
The Wall Street Journal	wsj.com	Right
The New York Times	nytimes.com	Left
The Washington Post	washingtonpost.com	Left
USA Today	usatoday.com	Centre
CNN	cnn.com	Centre
ABCNews	abcnews.go.com	Left
Bloomberg	bloomberg.com	Centre
Fox News	foxnews.com	Right
PBS	pbs.org	Centre
NPR	npr.org	Centre
CBS News	cbsnews.org	Left
NBC News	nbcnews.org	Left
The Economist	economist.com	Centre
MSNBC	msnbc.com	Left
The Guardian	theguardian.com	Left
The New Yorker	newyorker.com	Left
Politico	politico.com	Centre
Yahoo News	news.yahoo.com	Centre

Table A.1: *List of monitored mainstream news domains.*

A.1.2 Misleading sources

In Table A.2 we provide bias labels for misleading news outlets; we indicate fewer sources w.r.t to the original list provided by [224] (and available at `https://docs.google.com/spreadsheets/d/1S5eDzOUEByRcHSwSNmSqjQMpaKcKXmUzYT6YlRy3UOg/edit#gid=1882442466`), i.e. sources with at least one news article in our dataset. Bias labels are obtained resorting to "allside.com" and "mediabiasfactcheck.com", as in [33], and we indicate missing labels with "-".

Outlet	Bias
breitbart.com	Right
politicususa.com	Left
redstate.com	Right
infowars.com	Right
theblaze.com	Right
activistpost.com	Left
dcclothesline.com	Right
theonion.com	Satire
thefreethoughtproject.com	Left
wnd.com	Right
lewrockwell.com	Right
beforeitsnews.com	-
naturalnews.com	Right
twitchy.com	Right
govtslaves.info	-
21stcenturywire.com	Left
globalresearch.ca	Left
worldtruth.tv	-
anonews.co	-
disclose.tv	-
realfarmacy.com	-
burrardstreetjournal.com	-
gomerblog.com	Satire
huzlers.com	-
coasttocoastam.com	-
geoengineeringwatch.org	-
worldnewsdailyreport.com	-
clickhole.com	Satire
duffelblog.com	Satire
bipartisanreport.com	Left
nowtheendbegins.com	Right
veteranstoday.com	-

Table A.2: *List of misleading outlets collected in our dataset.*

A.1.3 Composition of the dataset

We provide in Fig. A.1 and Fig. A.2 (respectively for *mainstream* and *misleading* news) barplots for the distribution of articles according to different sources and bias. Only 98.5% of misleading news networks is present as some sources do not have a bias label. We further provide in Table A.3 a concise breakdown of the dataset of network cascades according to both class and bias labels.

A.2 Network Comparison Approaches

A.2.1 Centrality Measures

We computed the following centrality measures as provided in `networkx` package [103]: *Betweenness*, *Clustering*, *Closeness*, *Degree*, *Eigenvector*, *In-Degree*, *Katz*,

Appendix A. Supplementary Information for "A network-based approach to detect online disinformation on Twitter"

Class	Bias	No. networks
Mainstream	Left	4573
Mainstream	Centre	1079
Mainstream	Right	1292
Misleading	Left	1052
Misleading	Satire	444
Misleading	Right	4194

Table A.3: *Breakdown of the dataset of networks in terms of class and bias labels.*

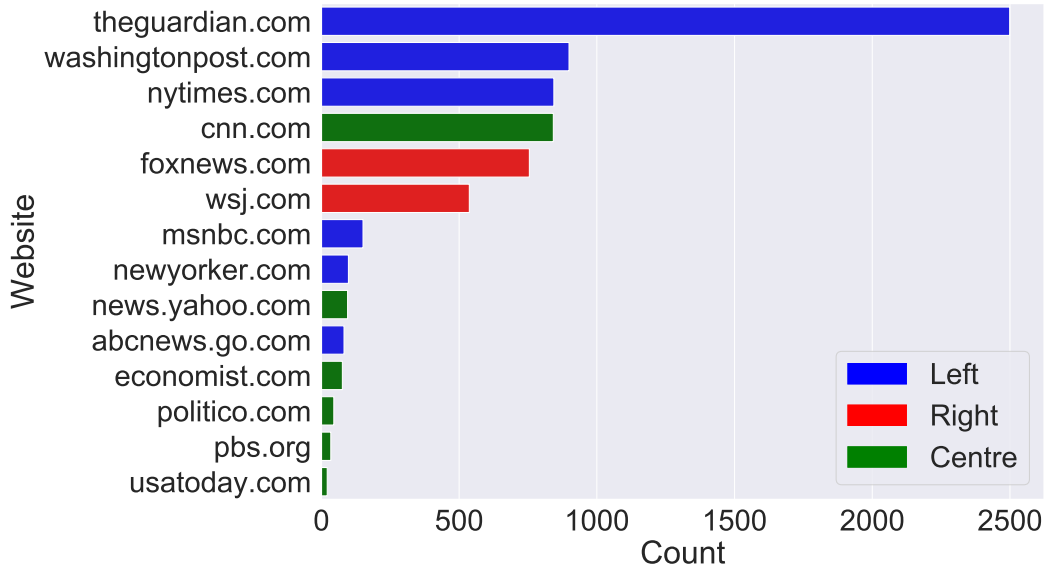


Figure A.1: *Distribution of the number of networks per mainstream source. Colors indicate the bias of the source as specified in the legend.*

Load (or *Newman's Betweenness*) and *Out-Degree*.

References and details on how these measures are computed are available in `networkx` documentation[3]. In particular *Degree*, *In-Degree* and *Out-Degree* distributions are normalized, e.g. for each node the corresponding centrality value is divided by $N-1$ (the maximum degree of a graph) where N is the number of nodes in the graph. For what concerns the similarity matrix, we computed Wasserstein Distance between empirical distributions using `scipy` Python package and `numpy` Python package to compute the empirical distribution function (with `no_bins`=100).

[3] `https://networkx.github.io/documentation/stable/reference/algorithms/centrality.html`

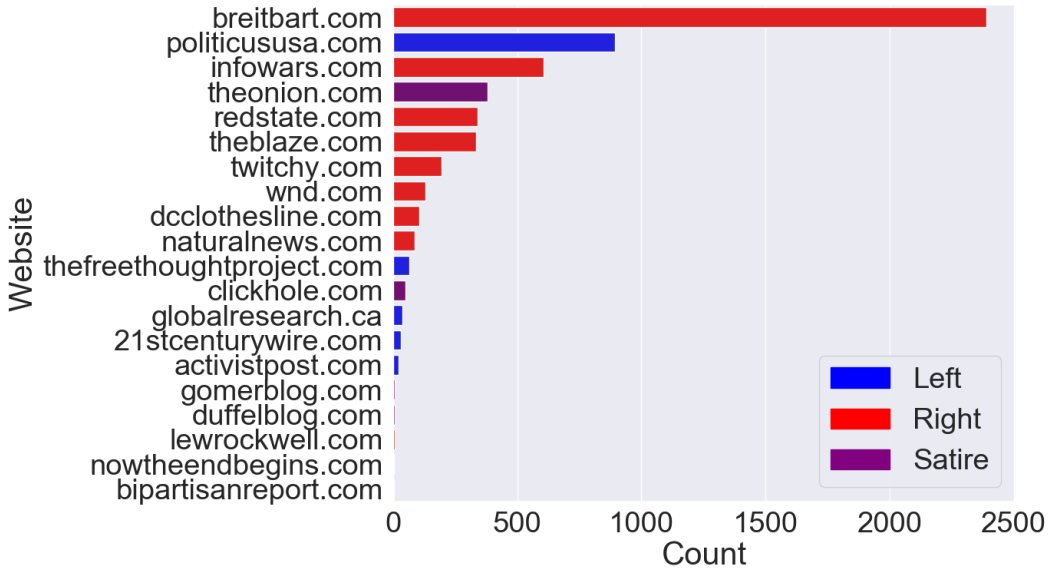


Figure A.2: *Distribution of the number of networks per misleading source. Colors indicate the bias of the source as specified in the legend.*

A.3 Analysis of Global Network Properties

A.3.1 Statistical Tests

In order to assess statistical differences between features distributions we used the non-parametric Kolmogorov Smirnov (KS) two-sample test, as provided by `scipy` Python package. In tables A.2 to A.5 we report KS statistic and associated p-value for all features (plus network size) in all subsets; as usual we reject the null hypothesis (a given feature has the same distribution in the two domains) when we observe p-value < $\alpha = 0.05$.

Kolmogorov Smirnov two-sample test for all features in $D_{[0,100)}$

Feature	KS statistic	KS p-value
Size	0.2716	3.0831e-104
SCC	0.2658	6.5331e-100
LSCC	0.1857	5.9470e-49
WCC	0.0987	3.7962e-14
LWCC	0.1483	2.1263e-31
DWCC	0.2687	4.4415e-102
CC	0.0654	1.8972e-06
KC	0.2189	7.1959e-68

Table A.4: *Kolmogorov-Smirnov statistic and p-value for all features in $D_{[0,100)}$.*

Kolmogorov Smirnov two-sample test for all features in $D_{[100,1000)}$

Feature	KS statistic	KS p-value
Size	0.0535	7.2564e-04
SCC	0.0512	1.3914e-03
LSCC	0.4363	1.4851e-229
WCC	0.3519	1.8491e-149
LWCC	0.2698	4.7916e-88
DWCC	0.3059	4.3610e-113
CC	0.2750	1.7796e-91
KC	0.4055	2.1375e-198

Table A.5: *Kolmogorov-Smirnov statistic and p-value for all features in $D_{[100,1000)}$.*

Kolmogorov Smirnov two-sample test for all features in $D_{[1000,+\infty)}$

Feature	KS statistic	KS p-value
Size	0.0606	8.1456e-01
SCC	0.0606	8.1456e-01
LSCC	0.5617	1.6478e-30
WCC	0.4854	6.8044e-23
LWCC	0.1908	6.7693e-04
DWCC	0.1351	3.6397e-02
CC	0.3997	1.1645e-15
KC	0.0916	3.1587e-01

Table A.6: *Kolmogorov-Smirnov statistic and p-value for all features in $D_{[1000,+\infty)}$.*

Kolmogorov Smirnov two-sample test for all features in D_{all}

Feature	KS statistic	KS p-value
Size	0.1862	5.2890e-96
SCC	0.1833	4.3347e-93
LSCC	0.3371	9.4902e-314
WCC	0.1202	3.1607e-40
LWCC	0.2468	2.3192e-168
DWCC	0.3102	1.2055e-265
CC	0.2037	7.8364e-115
KC	0.3481	0.0000e-00

Table A.7: *Kolmogorov-Smirnov statistic and p-value for all features in D_{all}.*

A.3.2 Box-plots for the distribution of features

In this section we provide box-plots in all subsets for the empirical distributions of all features: **SCC** = Number of Strongly Connected Components; **LSCC** = Size of the Largest Strongly Connected Component; **WCC** = Number of Weakly Connected

Components; **LWCC** = Size of the Largest Weakly Connected Component; **DWCC** = Diameter of the Largest Weakly Connected Components; **CC** = Average Clustering Coefficient; **KC** = Main K-Core Number.

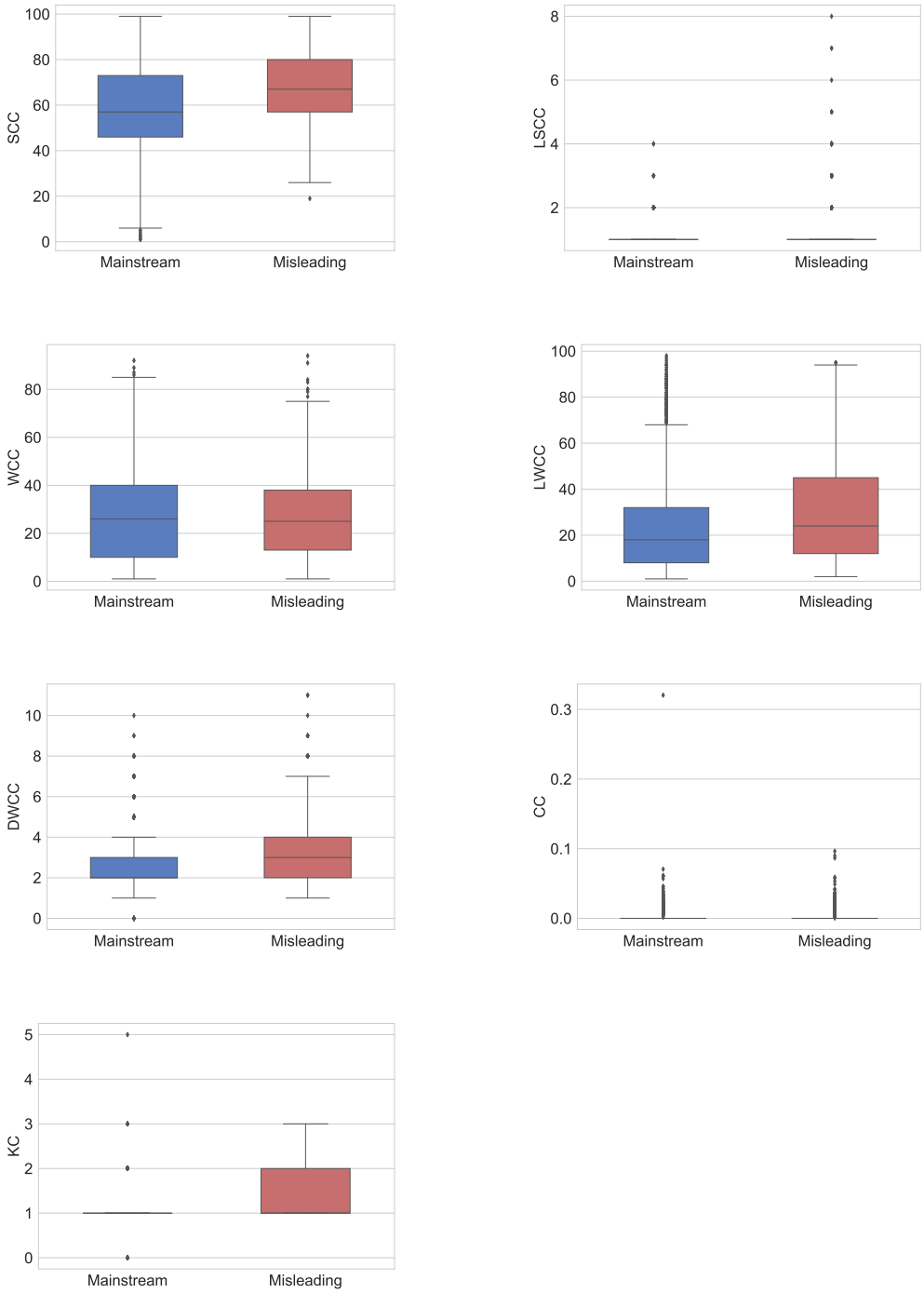


Figure A.3: *Box plots for all global network properties in* $D_{[0,100)}$

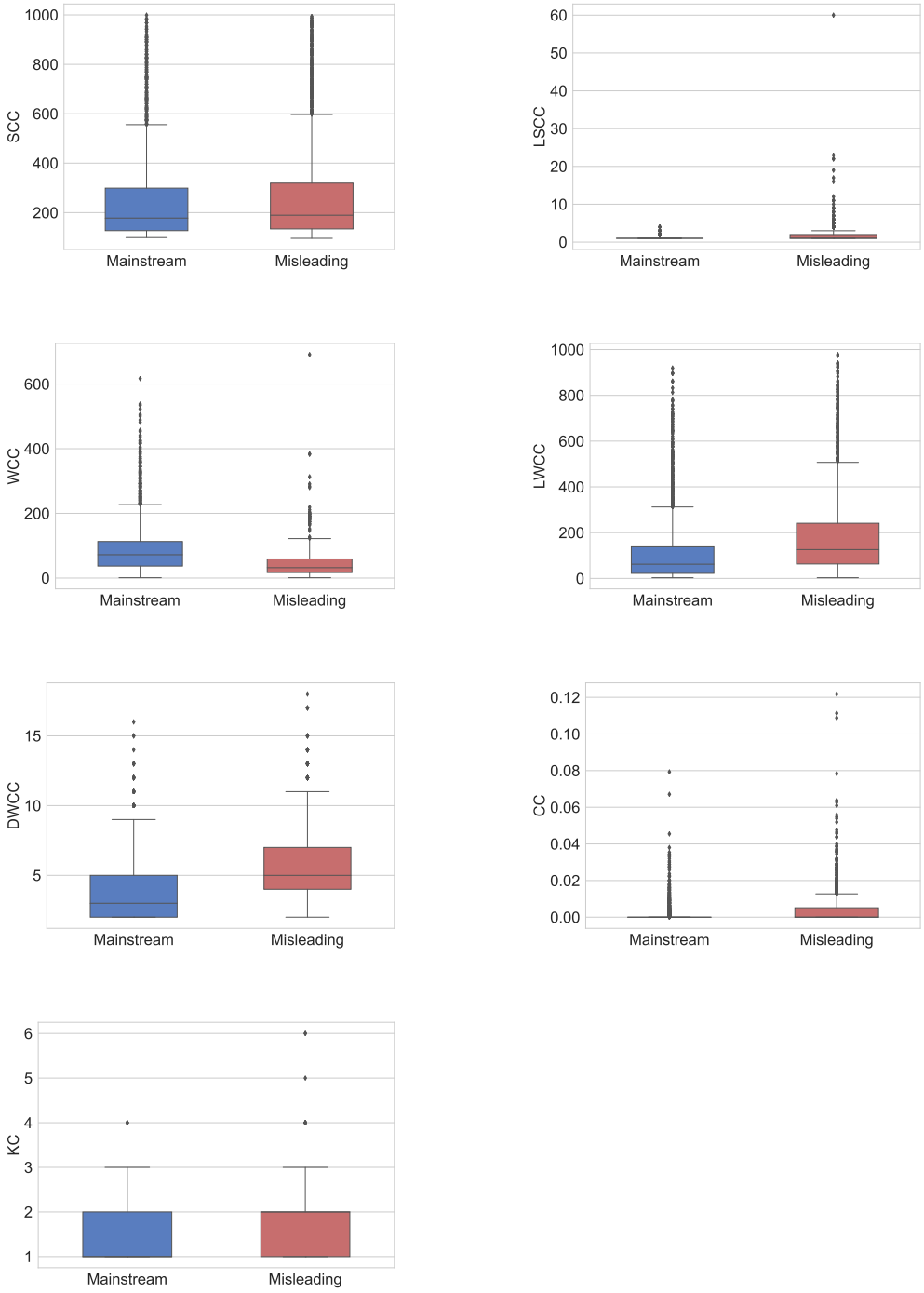


Figure A.4: *Box plots for all global network properties in* $D_{[100,1000)}$

A.3.3 Correlation Analysis

In this section we provide the Pearson pairwise correlation of all features in all subsets (including the size of networks) computed according to `pandas` Python package.

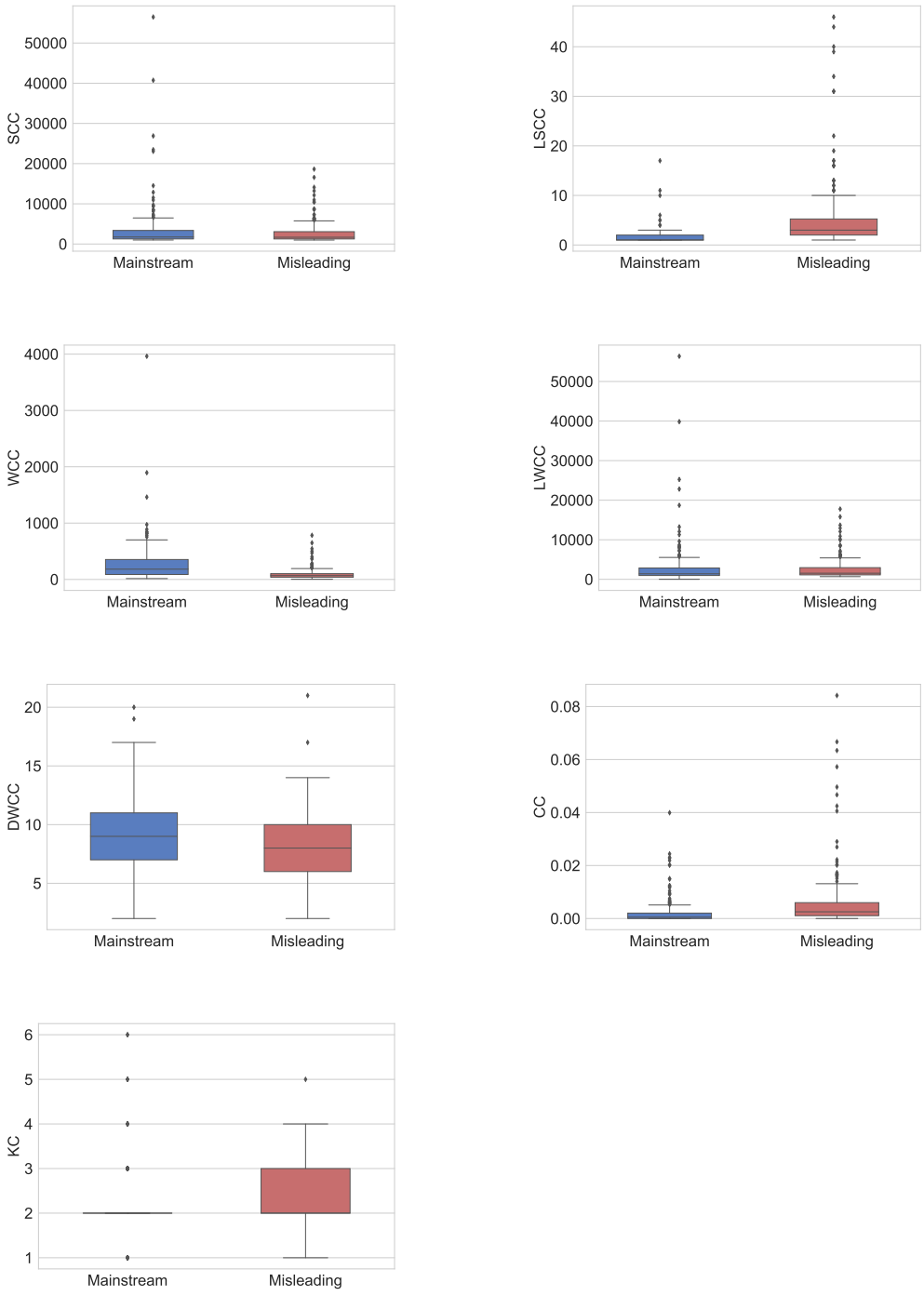


Figure A.5: *Box plots for all global network properties in* $D_{[1000,+\infty)}$

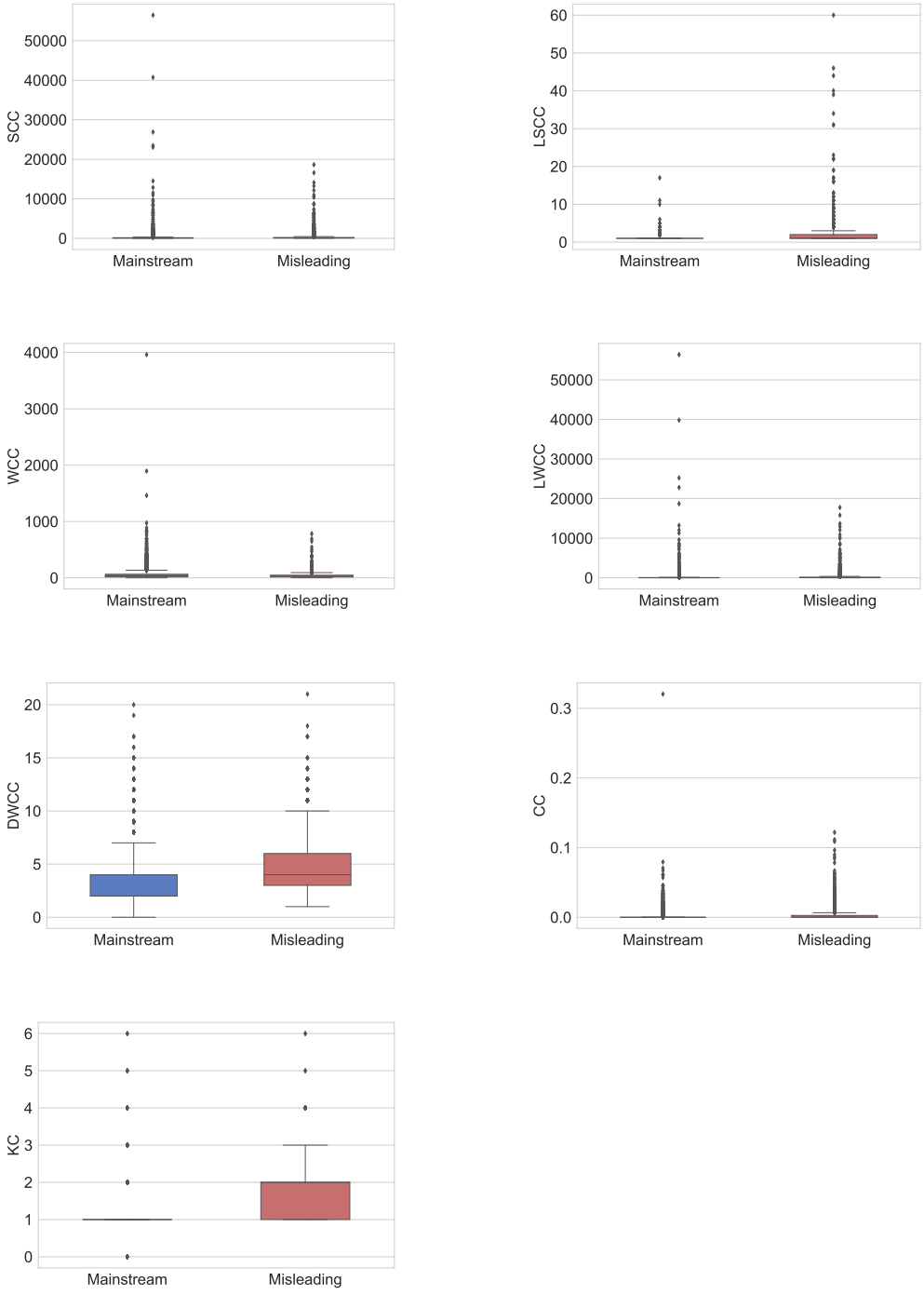


Figure A.6: *Box plots for all global network properties in* D_{all}

A.4 Classification

A.4.1 Classification results for Global Network Properties

In addition to classifiers specified in the main text, we also evaluated the following state-of-the-art classifiers: Support Vector Machine (SVM) with linear and RBF kernels,

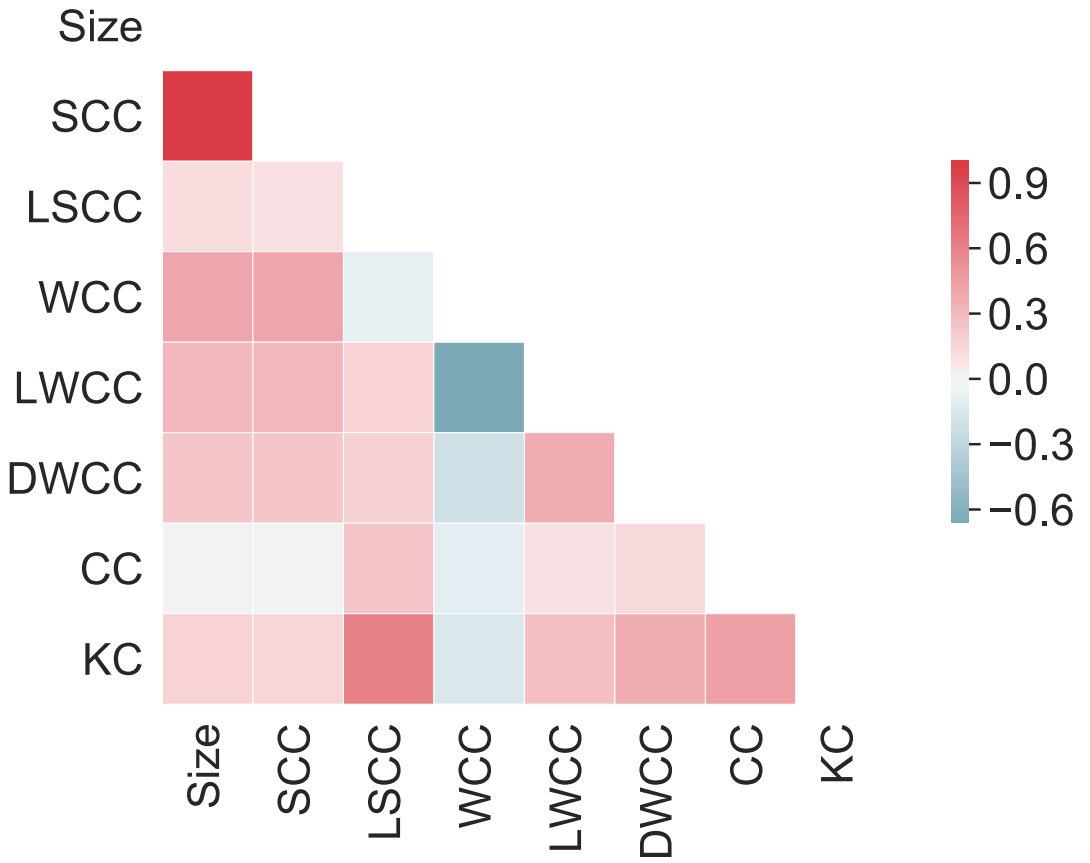


Figure A.7: *Correlation matrix for* $D_{[0,100)}$.

Gradient Boosting with exponential and deviance loss, Random Forest. We used the implementations available in the `sklearn` Python package with default parameters (and no hyperparameter tuning).

	Evaluation Metrics for $D_{[0,100)}$			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.63 (sd 0.01)	0.73 (sd 0.03)	0.63 (sd 0.01)	0.72 (sd 0.03)
SVC RBF	0.68 (sd 0.02)	0.71 (sd 0.02)	0.69 (sd 0.02)	0.74 (sd 0.02)
Logistic Regression	0.65 (sd 0.01)	0.70 (sd 0.01)	0.65 (sd 0.01)	0.74 (sd 0.02)
Random Forest	0.68 (sd 0.02)	0.68 (sd 0.02)	0.68 (sd 0.02)	0.75 (sd 0.02)
K-NN (k=5)	0.66 (sd 0.02)	0.67 (sd 0.01)	0.67 (sd 0.02)	0.74 (sd 0.01)
K-NN (k=10)	0.67 (sd 0.02)	0.70 (sd 0.01)	0.67 (sd 0.02)	0.76 (sd 0.01)
K-NN (k=20)	0.68 (sd 0.02)	0.71 (sd 0.02)	0.69 (sd 0.02)	0.77 (sd 0.01)
K-NN (k=50)	0.68 (sd 0.01)	0.72 (sd 0.02)	0.69 (sd 0.01)	0.78 (sd 0.01)
Gradient Boosting (exponential)	0.69 (sd 0.01)	0.72 (sd 0.01)	0.70 (sd 0.01)	0.78 (sd 0.01)
Gradient Boosting (deviance)	0.69 (sd 0.01)	0.72 (sd 0.01)	0.69 (sd 0.01)	0.78 (sd 0.01)

Table A.8: *Classification metrics for all classifiers evaluated using global network properties in* $D_{[0,100)}$.

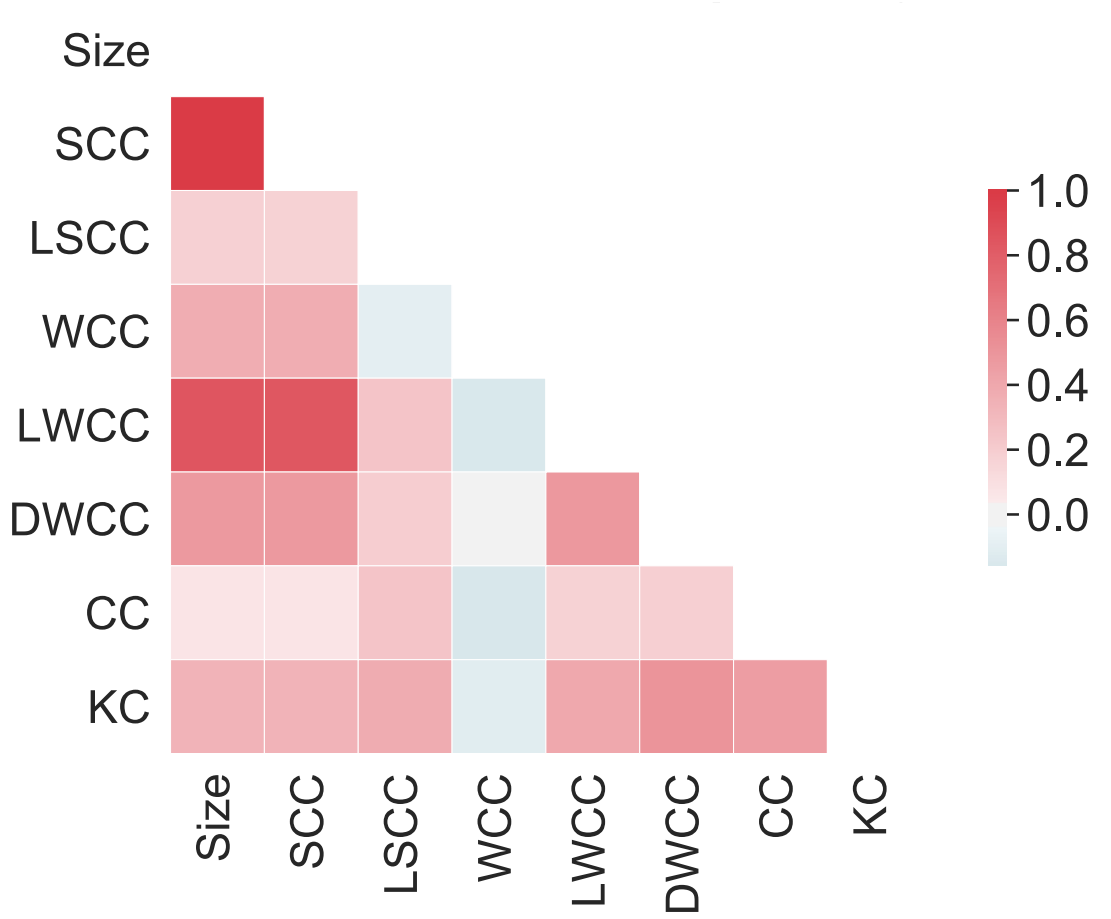


Figure A.8: *Correlation matrix for* $D_{[100,1000)}$.

	Evaluation Metrics for $D_{[100,1000)}$			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.75 (sd 0.01)	0.75 (sd 0.01)	0.75 (sd 0.01)	0.84 (sd 0.01)
SVC RBF	0.76 (sd 0.02)	0.76 (sd 0.02)	0.76 (sd 0.02)	0.82 (sd 0.02)
Logistic Regression	0.75 (sd 0.02)	0.76 (sd 0.01)	0.74 (sd 0.02)	0.85 (sd 0.02)
Random Forest	0.76 (sd 0.02)	0.76 (sd 0.02)	0.76 (sd 0.02)	0.85 (sd 0.02)
K-NN (k=5)	0.75 (sd 0.02)	0.75 (sd 0.02)	0.75 (sd 0.02)	0.83 (sd 0.02)
K-NN (k=10)	0.76 (sd 0.02)	0.76 (sd 0.02)	0.76 (sd 0.02)	0.85 (sd 0.02)
K-NN (k=20)	0.76 (sd 0.02)	0.76 (sd 0.02)	0.76 (sd 0.02)	0.85 (sd 0.01)
K-NN (k=50)	0.76 (sd 0.01)	0.76 (sd 0.01)	0.76 (sd 0.01)	0.86 (sd 0.01)
Gradient Boosting (exponential)	0.76 (sd 0.02)	0.76 (sd 0.02)	0.76 (sd 0.02)	0.86 (sd 0.02)
Gradient Boosting (deviance)	0.76 (sd 0.02)	0.76 (sd 0.02)	0.76 (sd 0.02)	0.86 (sd 0.02)

Table A.9: *Classification metrics for all classifiers evaluated using global network properties in* $D_{[100,1000)}$.

A.4.2 Classification results for Global Network Properties with Sampling

We collected misleading stories in the period from 25th February 2019 to 18th March 2019, which was used for mainstream news. As resulting misleading networks were

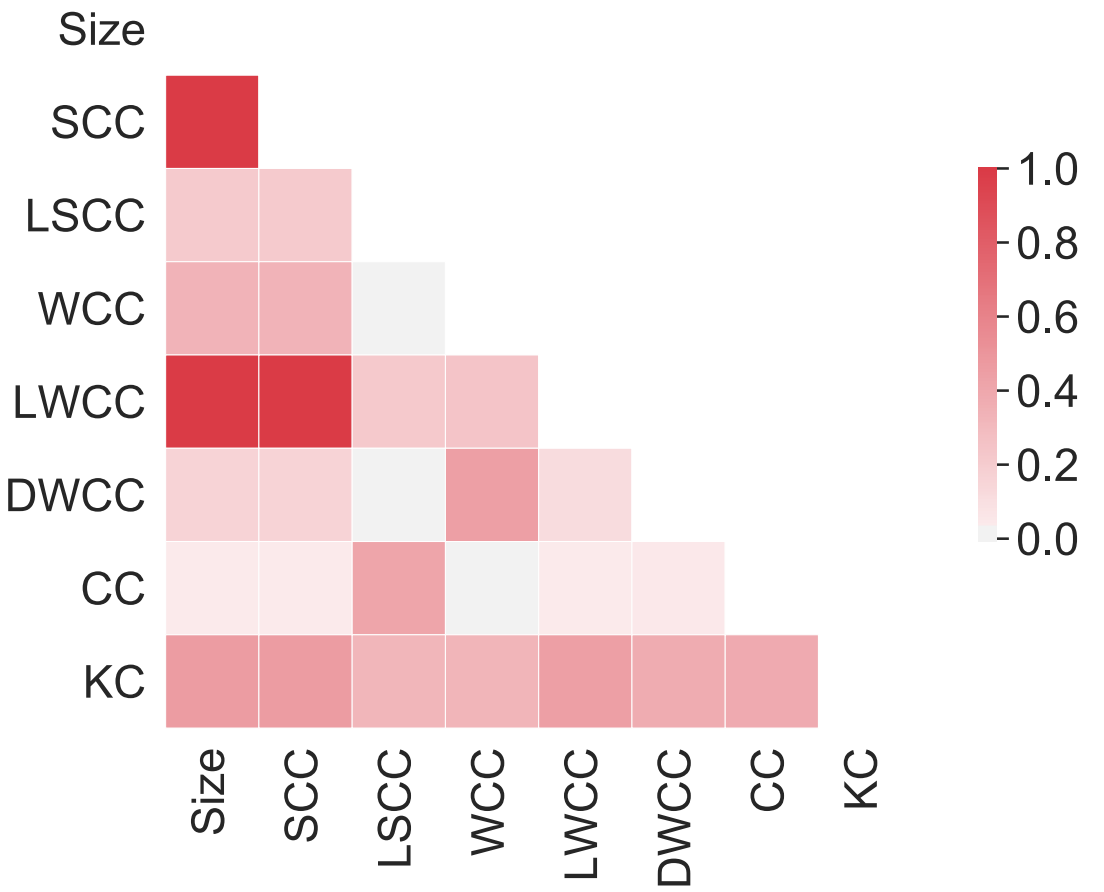


Figure A.9: *Correlation matrix for* $D_{[1000,+\infty)}$.

	Evaluation Metrics for $D_{[1000,+\infty)}$			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.78 (sd 0.04)	0.82 (sd 0.05)	0.78 (sd 0.05)	0.89 (sd 0.04)
SVC RBF	0.73 (sd 0.06)	0.79 (sd 0.06)	0.73 (sd 0.07)	0.86 (sd 0.06)
Logistic Regression	0.85 (sd 0.06)	0.86 (sd 0.06)	0.85 (sd 0.06)	0.93 (sd 0.03)
Random Forest	0.84 (sd 0.07)	0.85 (sd 0.07)	0.84 (sd 0.07)	0.91 (sd 0.04)
K-NN (k=5)	0.80 (sd 0.05)	0.81 (sd 0.04)	0.81 (sd 0.05)	0.87 (sd 0.04)
K-NN (k=10)	0.84 (sd 0.04)	0.84 (sd 0.04)	0.84 (sd 0.04)	0.89 (sd 0.04)
K-NN (k=20)	0.82 (sd 0.05)	0.83 (sd 0.05)	0.82 (sd 0.05)	0.89 (sd 0.03)
K-NN (k=50)	0.79 (sd 0.06)	0.81 (sd 0.06)	0.79 (sd 0.06)	0.87 (sd 0.04)
Gradient Boosting (exponential)	0.81 (sd 0.05)	0.84 (sd 0.06)	0.82 (sd 0.05)	0.90 (sd 0.04)
Gradient Boosting (deviance)	0.82 (sd 0.06)	0.83 (sd 0.07)	0.82 (sd 0.06)	0.89 (sd 0.04)

Table A.10: *Classification metrics for all classifiers evaluated using global network properties in* $D_{[1000,+\infty)}$.

strongly imbalanced (1157 misleading networks vs 6878 mainstream networks), we used `imblearn` Python package to apply three different sampling approaches:

1. Random Under Sampling: we uniformly randomly sampled individuals from the majority class (mainstream).

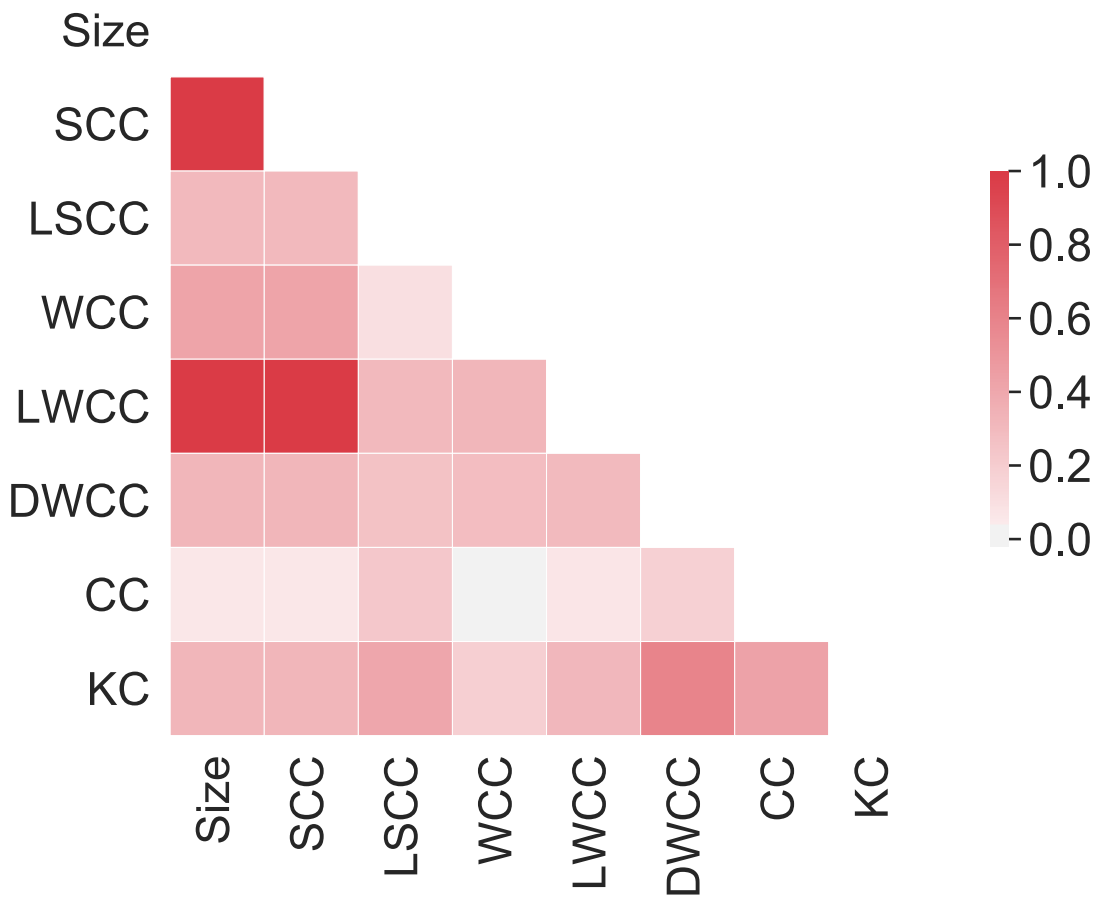


Figure A.10: *Correlation matrix for D_{all}.*

	Evaluation Metrics for D_{all}			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.69 (sd 0.01)	0.72 (sd 0.02)	0.69 (sd 0.02)	0.77 (sd 0.01)
SVC RBF	0.72 (sd 0.01)	0.74 (sd 0.01)	0.72 (sd 0.01)	0.75 (sd 0.03)
Logistic Regression	0.71 (sd 0.02)	0.74 (sd 0.02)	0.71 (sd 0.02)	0.78 (sd 0.02)
Random Forest	0.65 (sd 0.03)	0.65 (sd 0.03)	0.65 (sd 0.03)	0.73 (sd 0.04)
K-NN (k=5)	0.69 (sd 0.02)	0.70 (sd 0.02)	0.69 (sd 0.02)	0.75 (sd 0.02)
K-NN (k=10)	0.70 (sd 0.01)	0.72 (sd 0.02)	0.70 (sd 0.01)	0.77 (sd 0.02)
K-NN (k=20)	0.71 (sd 0.01)	0.73 (sd 0.02)	0.71 (sd 0.01)	0.78 (sd 0.01)
K-NN (k=50)	0.72 (sd 0.01)	0.74 (sd 0.02)	0.72 (sd 0.01)	0.79 (sd 0.01)
Gradient Boosting (exponential)	0.69 (sd 0.04)	0.70 (sd 0.04)	0.69 (sd 0.04)	0.76 (sd 0.03)
Gradient Boosting (deviance)	0.67 (sd 0.05)	0.67 (sd 0.05)	0.67 (sd 0.05)	0.73 (sd 0.06)

Table A.11: *Classification metrics for all classifiers evaluated using global network properties in D_{all}.*

2. Random Over Sampling: we uniformly randomly sampled individuals from the minority class (misleading).

3. Balanced Classifiers: the `imblearn` package also provides a way to train boosting classifiers, in particular we chose Random Forest and AdaBoost, trained on

samples which are balanced in the two classes.

We evaluated Precision, Recall, F1-Score and AUROC for all before mentioned classifiers on all subsets as it follows: for 1) and 2) we first split networks according to the number of nodes ($[0, 100)$, $[100, 1000)$, $[1000, +\infty)$, all) and then evaluated metrics with 10-fold stratified shuffle split on 100 different samples; we report accordingly the average value of metrics over all samples (tables A.10 to A.20). For what concerns 3) we simply evaluated both classifiers with 10-fold stratified shuffle split on the original collection.

	Evaluation Metrics with Random Under Sampling for $D_{[0,100)}$			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.68 (sd 0.00)	0.69 (sd 0.00)	0.68 (sd 0.00)	0.75 (sd 0.00)
SVC RBF	0.69 (sd 0.00)	0.69 (sd 0.00)	0.69 (sd 0.00)	0.75 (sd 0.00)
Logistic Regression	0.68 (sd 0.00)	0.68 (sd 0.00)	0.68 (sd 0.00)	0.75 (sd 0.00)
Random Forest	0.66 (sd 0.01)	0.66 (sd 0.01)	0.66 (sd 0.01)	0.74 (sd 0.00)
K-NN (N=5)	0.68 (sd 0.00)	0.68 (sd 0.00)	0.68 (sd 0.00)	0.74 (sd 0.00)
K-NN (N=10)	0.67 (sd 0.00)	0.67 (sd 0.00)	0.67 (sd 0.00)	0.75 (sd 0.00)
K-NN (N=20)	0.67 (sd 0.00)	0.67 (sd 0.00)	0.67 (sd 0.00)	0.76 (sd 0.00)
K-NN (N=50)	0.69 (sd 0.00)	0.69 (sd 0.00)	0.68 (sd 0.00)	0.78 (sd 0.00)
Gradient Boosting (exponential)	0.69 (sd 0.00)	0.69 (sd 0.00)	0.69 (sd 0.00)	0.77 (sd 0.00)
Gradient Boosting (deviance)	0.69 (sd 0.00)	0.69 (sd 0.00)	0.69 (sd 0.00)	0.76 (sd 0.00)

Table A.12: *Classification metrics for all classifiers evaluated using global network properties and random under sampling in $D_{[0,100)}$.*

	Evaluation Metrics with Random Under Sampling for $D_{[100,1000)}$			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.75 (sd 0.00)	0.76 (sd 0.00)	0.75 (sd 0.00)	0.83 (sd 0.00)
SVC RBF	0.74 (sd 0.00)	0.74 (sd 0.00)	0.74 (sd 0.00)	0.78 (sd 0.00)
Logistic Regression	0.76 (sd 0.00)	0.76 (sd 0.00)	0.76 (sd 0.00)	0.85 (sd 0.00)
Random Forest	0.75 (sd 0.01)	0.75 (sd 0.01)	0.74 (sd 0.01)	0.83 (sd 0.00)
K-NN (N=5)	0.75 (sd 0.00)	0.75 (sd 0.00)	0.75 (sd 0.00)	0.81 (sd 0.00)
K-NN (N=10)	0.76 (sd 0.00)	0.76 (sd 0.00)	0.76 (sd 0.00)	0.84 (sd 0.00)
K-NN (N=20)	0.76 (sd 0.00)	0.76 (sd 0.00)	0.76 (sd 0.00)	0.85 (sd 0.00)
K-NN (N=50)	0.75 (sd 0.00)	0.75 (sd 0.00)	0.75 (sd 0.00)	0.84 (sd 0.00)
Gradient Boosting (exponential)	0.74 (sd 0.00)	0.74 (sd 0.00)	0.73 (sd 0.00)	0.84 (sd 0.00)
Gradient Boosting (deviance)	0.74 (sd 0.00)	0.74 (sd 0.00)	0.74 (sd 0.00)	0.83 (sd 0.00)

Table A.13: *Classification metrics for all classifiers evaluated using global network properties and random under sampling in $D_{[100,1000)}$.*

A.5 Classification performances taking into account bias labels on sources

In this section we provide classification performance results when taking into account also the bias of news domains. Labels are obtained as in [**?**]. We show results, in terms of Precision, Recall, F1-Score and AUROC, concerning several combinations of training and test sets, as specified in the caption of tables A.24 to A.55.

Appendix A. Supplementary Information for "A network-based approach to detect online disinformation on Twitter"

	Evaluation Metrics with Random Under Sampling for $D_{[1000,+\infty)}$			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.64 (sd 0.00)	0.73 (sd 0.00)	0.59 (sd 0.00)	0.66 (sd 0.08)
SVC RBF	0.70 (sd 0.00)	0.72 (sd 0.00)	0.68 (sd 0.00)	0.79 (sd 0.00)
Logistic Regression	0.76 (sd 0.00)	0.77 (sd 0.00)	0.75 (sd 0.00)	0.82 (sd 0.00)
Random Forest	0.73 (sd 0.02)	0.74 (sd 0.02)	0.71 (sd 0.02)	0.79 (sd 0.01)
K-NN (N=5)	0.77 (sd 0.00)	0.80 (sd 0.00)	0.75 (sd 0.00)	0.84 (sd 0.00)
K-NN (N=10)	0.71 (sd 0.00)	0.77 (sd 0.00)	0.67 (sd 0.00)	0.83 (sd 0.00)
K-NN (N=20)	0.75 (sd 0.00)	0.80 (sd 0.00)	0.72 (sd 0.00)	0.85 (sd 0.00)
K-NN (N=50)	0.62 (sd 0.00)	0.66 (sd 0.00)	0.59 (sd 0.00)	0.62 (sd 0.00)
Gradient Boosting (exponential)	0.71 (sd 0.01)	0.72 (sd 0.01)	0.70 (sd 0.01)	0.75 (sd 0.01)
Gradient Boosting (deviance)	0.72 (sd 0.01)	0.74 (sd 0.01)	0.71 (sd 0.01)	0.76 (sd 0.01)

Table A.14: *Classification metrics for all classifiers evaluated using global network properties and random under sampling in $D_{[1000,+\infty)}$.*

	Evaluation Metrics with Random Under Sampling for D_{all}			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.68 (sd 0.00)	0.70 (sd 0.00)	0.67 (sd 0.00)	0.75 (sd 0.00)
SVC RBF	0.70 (sd 0.00)	0.70 (sd 0.00)	0.70 (sd 0.00)	0.73 (sd 0.00)
Logistic Regression	0.69 (sd 0.00)	0.71 (sd 0.00)	0.68 (sd 0.00)	0.76 (sd 0.00)
Random Forest	0.59 (sd 0.00)	0.60 (sd 0.01)	0.57 (sd 0.00)	0.63 (sd 0.01)
K-NN (N=5)	0.67 (sd 0.00)	0.68 (sd 0.00)	0.67 (sd 0.00)	0.73 (sd 0.00)
K-NN (N=10)	0.70 (sd 0.00)	0.71 (sd 0.00)	0.69 (sd 0.00)	0.76 (sd 0.00)
K-NN (N=20)	0.69 (sd 0.00)	0.70 (sd 0.00)	0.69 (sd 0.00)	0.77 (sd 0.00)
K-NN (N=50)	0.71 (sd 0.00)	0.72 (sd 0.00)	0.71 (sd 0.00)	0.78 (sd 0.00)
Gradient Boosting (exponential)	0.62 (sd 0.00)	0.63 (sd 0.00)	0.61 (sd 0.01)	0.68 (sd 0.00)
Gradient Boosting (deviance)	0.61 (sd 0.00)	0.62 (sd 0.00)	0.59 (sd 0.01)	0.66 (sd 0.00)

Table A.15: *Classification metrics for all classifiers evaluated using global network properties and random under sampling in D_{all}.*

	Evaluation Metrics with Random Over Sampling for $D_{[0,100)}$			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.67 (sd 0.00)	0.68 (sd 0.00)	0.67 (sd 0.00)	0.74 (sd 0.00)
SVC RBF	0.79 (sd 0.00)	0.79 (sd 0.00)	0.79 (sd 0.00)	0.86 (sd 0.00)
Logistic Regression	0.68 (sd 0.00)	0.68 (sd 0.00)	0.68 (sd 0.00)	0.74 (sd 0.00)
Random Forest	0.87 (sd 0.00)	0.87 (sd 0.00)	0.86 (sd 0.00)	0.94 (sd 0.00)
K-NN (N=5)	0.86 (sd 0.00)	0.88 (sd 0.00)	0.86 (sd 0.00)	0.92 (sd 0.00)
K-NN (N=10)	0.80 (sd 0.00)	0.81 (sd 0.00)	0.79 (sd 0.00)	0.88 (sd 0.00)
K-NN (N=20)	0.77 (sd 0.00)	0.77 (sd 0.00)	0.76 (sd 0.00)	0.85 (sd 0.00)
K-NN (N=50)	0.73 (sd 0.00)	0.73 (sd 0.00)	0.73 (sd 0.00)	0.82 (sd 0.00)
Gradient Boosting (exponential)	0.72 (sd 0.00)	0.72 (sd 0.00)	0.72 (sd 0.00)	0.82 (sd 0.00)
Gradient Boosting (deviance)	0.73 (sd 0.00)	0.73 (sd 0.00)	0.73 (sd 0.00)	0.82 (sd 0.00)

Table A.16: *Classification metrics for all classifiers evaluated using global network properties and random over sampling in $D_{[0,100)}$.*

They overall show similar classification performances compared to the general case discussed in the main text.

	Evaluation Metrics with Random Over Sampling for $D_{[100,1000)}$			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.76 (sd 0.00)	0.77 (sd 0.00)	0.76 (sd 0.00)	0.85 (sd 0.00)
SVC RBF	0.83 (sd 0.00)	0.84 (sd 0.00)	0.83 (sd 0.00)	0.89 (sd 0.00)
Logistic Regression	0.76 (sd 0.00)	0.76 (sd 0.00)	0.76 (sd 0.00)	0.86 (sd 0.00)
Random Forest	0.86 (sd 0.00)	0.87 (sd 0.00)	0.86 (sd 0.00)	0.93 (sd 0.00)
K-NN (N=5)	0.84 (sd 0.00)	0.84 (sd 0.00)	0.83 (sd 0.00)	0.90 (sd 0.00)
K-NN (N=10)	0.80 (sd 0.00)	0.80 (sd 0.00)	0.80 (sd 0.00)	0.89 (sd 0.00)
K-NN (N=20)	0.79 (sd 0.00)	0.79 (sd 0.00)	0.79 (sd 0.00)	0.88 (sd 0.00)
K-NN (N=50)	0.79 (sd 0.00)	0.79 (sd 0.00)	0.79 (sd 0.00)	0.87 (sd 0.00)
Gradient Boosting (exponential)	0.78 (sd 0.00)	0.79 (sd 0.00)	0.78 (sd 0.00)	0.88 (sd 0.00)
Gradient Boosting (deviance)	0.79 (sd 0.00)	0.79 (sd 0.00)	0.79 (sd 0.00)	0.88 (sd 0.00)

Table A.17: *Classification metrics for all classifiers evaluated using global network properties and random over sampling in* $D_{[100,1000)}$.

	Evaluation Metrics with Random Over Sampling for $D_{[1000,+\infty)}$			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.74 (sd 0.00)	0.74 (sd 0.00)	0.74 (sd 0.00)	0.82 (sd 0.00)
SVC RBF	0.72 (sd 0.00)	0.76 (sd 0.00)	0.70 (sd 0.00)	0.86 (sd 0.00)
Logistic Regression	0.76 (sd 0.00)	0.77 (sd 0.00)	0.76 (sd 0.00)	0.82 (sd 0.00)
Random Forest	0.73 (sd 0.01)	0.80 (sd 0.01)	0.71 (sd 0.02)	0.87 (sd 0.01)
K-NN (N=5)	0.80 (sd 0.00)	0.80 (sd 0.00)	0.79 (sd 0.00)	0.87 (sd 0.00)
K-NN (N=10)	0.74 (sd 0.00)	0.74 (sd 0.00)	0.74 (sd 0.00)	0.84 (sd 0.00)
K-NN (N=20)	0.72 (sd 0.00)	0.73 (sd 0.00)	0.72 (sd 0.00)	0.84 (sd 0.00)
K-NN (N=50)	0.72 (sd 0.00)	0.73 (sd 0.00)	0.72 (sd 0.00)	0.83 (sd 0.00)
Gradient Boosting (exponential)	0.73 (sd 0.00)	0.76 (sd 0.00)	0.72 (sd 0.00)	0.85 (sd 0.00)
Gradient Boosting (deviance)	0.72 (sd 0.00)	0.75 (sd 0.01)	0.71 (sd 0.00)	0.84 (sd 0.00)

Table A.18: *Classification metrics for all classifiers evaluated using global network properties and random over sampling in* $D_{[1000,+\infty)}$.

	Evaluation Metrics with Random Over Sampling for D_{all}			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.70 (sd 0.00)	0.72 (sd 0.00)	0.69 (sd 0.00)	0.76 (sd 0.00)
SVC RBF	0.73 (sd 0.00)	0.75 (sd 0.00)	0.73 (sd 0.00)	0.78 (sd 0.00)
Logistic Regression	0.71 (sd 0.00)	0.72 (sd 0.00)	0.70 (sd 0.00)	0.77 (sd 0.00)
Random Forest	0.65 (sd 0.00)	0.70 (sd 0.01)	0.62 (sd 0.00)	0.71 (sd 0.00)
K-NN (N=5)	0.71 (sd 0.00)	0.72 (sd 0.00)	0.71 (sd 0.00)	0.75 (sd 0.00)
K-NN (N=10)	0.71 (sd 0.00)	0.71 (sd 0.00)	0.71 (sd 0.00)	0.77 (sd 0.00)
K-NN (N=20)	0.71 (sd 0.00)	0.72 (sd 0.00)	0.71 (sd 0.00)	0.78 (sd 0.00)
K-NN (N=50)	0.72 (sd 0.00)	0.72 (sd 0.00)	0.71 (sd 0.00)	0.79 (sd 0.00)
Gradient Boosting (exponential)	0.67 (sd 0.00)	0.68 (sd 0.00)	0.67 (sd 0.00)	0.74 (sd 0.00)
Gradient Boosting (deviance)	0.67 (sd 0.00)	0.67 (sd 0.00)	0.67 (sd 0.00)	0.72 (sd 0.00)

Table A.19: *Classification metrics for all classifiers evaluated using global network properties and random over sampling in* D_{all}.

	Evaluation Metrics with Balanced Classifiers for $D_{[0,100)}$			
Classifier	Recall	Precision	F1-Score	AUROC
Balanced Random Forest	0.69 (sd 0.03)	0.59 (sd 0.01)	0.56 (sd 0.02)	0.78 (sd 0.03)
Balanced AdaBoost	0.69 (sd 0.05)	0.58 (sd 0.02)	0.57 (sd 0.03)	0.77 (sd 0.04)

Table A.20: *Classification metrics for all classifiers evaluated using global network properties and balanced classifiers in $D_{[0,100)}$.*

	Evaluation Metrics with Balanced Classifiers for $D_{[100,1000)}$			
Classifier	Recall	Precision	F1-Score	AUROC
Balanced Random Forest	0.77 (sd 0.02)	0.69 (sd 0.02)	0.71 (sd 0.02)	0.86 (sd 0.02)
Balanced AdaBoost	0.75 (sd 0.03)	0.69 (sd 0.02)	0.70 (sd 0.03)	0.84 (sd 0.02)

Table A.21: *Classification metrics for all classifiers evaluated using global network properties and balanced classifiers in $D_{[100,1000)}$.*

	Evaluation Metrics with Balanced Classifiers for $D_{[1000,+\infty)}$			
Classifier	Recall	Precision	F1-Score	AUROC
Balanced Random Forest	0.74 (sd 0.09)	0.69 (sd 0.12)	0.68 (sd 0.10)	0.82 (sd 0.11)
Balanced AdaBoost	0.73 (sd 0.11)	0.67 (sd 0.09)	0.68 (sd 0.10)	0.81 (sd 0.11)

Table A.22: *Classification metrics for all classifiers evaluated using global network properties and balanced classifiers in $D_{[1000,+\infty)}$.*

	Evaluation Metrics with Balanced Classifiers for D_{all}			
Classifier	Recall	Precision	F1-Score	AUROC
Balanced Random Forest	0.68 (sd 0.03)	0.61 (sd 0.01)	0.61 (sd 0.03)	0.75 (sd 0.05)
Balanced AdaBoost	0.69 (sd 0.04)	0.63 (sd 0.02)	0.63 (sd 0.02)	0.76 (sd 0.05)

Table A.23: *Classification metrics for all classifiers evaluated using global network properties and balanced classifiers in D_{all}.*

	Evaluation Metrics for Classifiers in D_{all}			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.66 (sd $1.73E-02$)	0.79 (sd $1.26E-02$)	0.68 (sd $1.97E-02$)	0.73 (sd $2.45E-02$)
SVC RBF	0.66 (sd $1.28E-02$)	0.79 (sd $1.48E-02$)	0.68 (sd $1.48E-02$)	0.67 (sd $3.28E-02$)
Logistic Regression	0.66 (sd $2.12E-02$)	0.78 (sd $2.45E-02$)	0.68 (sd $2.48E-02$)	0.77 (sd $1.62E-02$)
Random Forest	0.66 (sd $1.92E-02$)	0.74 (sd $2.64E-02$)	0.68 (sd $2.12E-02$)	0.74 (sd $1.84E-02$)
K-NN (N=5)	0.67 (sd $1.37E-02$)	0.74 (sd $1.67E-02$)	0.69 (sd $1.37E-02$)	0.72 (sd $1.53E-02$)
K-NN (N=10)	0.67 (sd $1.44E-02$)	0.78 (sd $1.06E-02$)	0.69 (sd $1.56E-02$)	0.74 (sd $1.21E-02$)
K-NN (N=20)	0.68 (sd $1.36E-02$)	0.79 (sd $1.58E-02$)	0.70 (sd $1.49E-02$)	0.76 (sd $1.09E-02$)
K-NN (N=50)	0.67 (sd $1.34E-02$)	0.79 (sd $1.44E-02$)	0.70 (sd $1.44E-02$)	0.78 (sd $1.14E-02$)
Gradient Boosting (exponential)	0.66 (sd $2.35E-02$)	0.80 (sd $1.89E-02$)	0.68 (sd $2.72E-02$)	0.75 (sd $2.39E-02$)
Gradient Boosting (deviance)	0.66 (sd $2.68E-02$)	0.80 (sd $1.67E-02$)	0.68 (sd $3.07E-02$)	0.76 (sd $2.57E-02$)
Balanced RF	0.68 (sd $2.04E-02$)	0.69 (sd $3.02E-02$)	0.68 (sd $2.46E-02$)	0.75 (sd $1.72E-02$)
Balanced ADABOOST	0.69 (sd $3.89E-02$)	0.69 (sd $2.13E-02$)	0.68 (sd $3.25E-02$)	0.77 (sd $2.70E-02$)

Table A.24: *Classification metrics for balanced classifiers evaluated using global network properties in D_{all}. Training and test data include all sources except Breitbart.com and Politicususa.com.*

A.6 Box-plots for the distribution of features taking into account bias of sources

In this section we provide box-plots in all subsets for the empirical distributions of all features per different bias of sources.

	Evaluation Metrics for Classifiers in $D_{[0,100]}$			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.59 (sd $1.84E-02$)	0.80 (sd $3.67E-02$)	0.60 (sd $2.70E-02$)	0.65 (sd $6.58E-02$)
SVC RBF	0.57 (sd $2.36E-02$)	0.70 (sd $4.02E-02$)	0.57 (sd $3.51E-02$)	0.69 (sd $3.31E-02$)
Logistic Regression	0.60 (sd $2.22E-02$)	0.76 (sd $5.53E-02$)	0.61 (sd $3.06E-02$)	0.75 (sd $2.71E-02$)
Random Forest	0.62 (sd $1.82E-02$)	0.69 (sd $1.89E-02$)	0.64 (sd $2.01E-02$)	0.75 (sd $2.18E-02$)
K-NN (N=5)	0.63 (sd $2.46E-02$)	0.70 (sd $2.90E-02$)	0.65 (sd $2.75E-02$)	0.71 (sd $2.63E-02$)
K-NN (N=10)	0.62 (sd $2.05E-02$)	0.75 (sd $3.78E-02$)	0.64 (sd $2.59E-02$)	0.74 (sd $2.13E-02$)
K-NN (N=20)	0.61 (sd $1.85E-02$)	0.75 (sd $3.53E-02$)	0.64 (sd $2.42E-02$)	0.76 (sd $2.42E-02$)
K-NN (N=50)	0.60 (sd $2.53E-02$)	0.76 (sd $4.04E-02$)	0.62 (sd $3.40E-02$)	0.77 (sd $2.43E-02$)
Gradient Boosting (exponential)	0.60 (sd $2.07E-02$)	0.75 (sd $3.01E-02$)	0.62 (sd $2.82E-02$)	0.77 (sd $1.78E-02$)
Gradient Boosting (deviance)	0.60 (sd $2.41E-02$)	0.74 (sd $4.27E-02$)	0.62 (sd $3.16E-02$)	0.77 (sd $1.90E-02$)
Balanced RF	0.69 (sd $2.23E-02$)	0.63 (sd $1.48E-02$)	0.63 (sd $1.61E-02$)	0.77 (sd $2.28E-02$)
Balanced ADABOOST	0.69 (sd $2.02E-02$)	0.64 (sd $1.36E-02$)	0.64 (sd $1.62E-02$)	0.77 (sd $1.59E-02$)

Table A.25: *Classification metrics for balanced classifiers evaluated using global network properties in $D_{[0,100]}$. Training and test data include all sources except Breitbart.com and Politicususa.com.*

	Evaluation Metrics for Classifiers in $D_{[100,1000]}$			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.70 (sd $1.50E-02$)	0.78 (sd $1.54E-02$)	0.72 (sd $1.60E-02$)	0.83 (sd $1.50E-02$)
SVC RBF	0.72 (sd $1.79E-02$)	0.78 (sd $2.39E-02$)	0.74 (sd $1.93E-02$)	0.79 (sd $1.55E-02$)
Logistic Regression	0.72 (sd $9.61E-03$)	0.79 (sd $1.92E-02$)	0.73 (sd $1.08E-02$)	0.85 (sd $2.02E-02$)
Random Forest	0.75 (sd $1.29E-02$)	0.78 (sd $1.81E-02$)	0.76 (sd $1.40E-02$)	0.85 (sd $1.24E-02$)
K-NN (N=5)	0.74 (sd $1.93E-02$)	0.76 (sd $2.15E-02$)	0.75 (sd $1.93E-02$)	0.81 (sd $1.68E-02$)
K-NN (N=10)	0.74 (sd $1.49E-02$)	0.78 (sd $2.24E-02$)	0.75 (sd $1.60E-02$)	0.84 (sd $1.67E-02$)
K-NN (N=20)	0.75 (sd $1.26E-02$)	0.80 (sd $1.98E-02$)	0.77 (sd $1.42E-02$)	0.84 (sd $1.63E-02$)
K-NN (N=50)	0.75 (sd $1.74E-02$)	0.79 (sd $2.45E-02$)	0.76 (sd $1.90E-02$)	0.85 (sd $2.10E-02$)
Gradient Boosting (exponential)	0.75 (sd $1.80E-02$)	0.81 (sd $2.84E-02$)	0.77 (sd $2.02E-02$)	0.86 (sd $1.59E-02$)
Gradient Boosting (deviance)	0.75 (sd $1.81E-02$)	0.81 (sd $2.65E-02$)	0.77 (sd $1.99E-02$)	0.86 (sd $1.37E-02$)
Balanced RF	0.77 (sd $1.41E-02$)	0.75 (sd $1.41E-02$)	0.76 (sd $1.41E-02$)	0.85 (sd $1.23E-02$)
Balanced ADABOOST	0.75 (sd $2.14E-02$)	0.73 (sd $2.84E-02$)	0.74 (sd $2.65E-02$)	0.84 (sd $1.29E-02$)

Table A.26: *Classification metrics for all classifiers evaluated using global network properties in $D_{[100,1000]}$. Training and test data include all sources except Breitbart.com and Politicususa.com.*

	Evaluation Metrics for Classifiers in $D_{[1000,+\infty]}$			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.63 (sd $7.04E-02$)	0.78 (sd $1.30E-01$)	0.63 (sd $9.24E-02$)	0.91 (sd $3.85E-02$)
SVC RBF	0.69 (sd $1.09E-01$)	0.78 (sd $1.10E-01$)	0.69 (sd $1.29E-01$)	0.88 (sd $5.82E-02$)
Logistic Regression	0.82 (sd $9.51E-02$)	0.85 (sd $7.86E-02$)	0.82 (sd $9.48E-02$)	0.92 (sd $4.72E-02$)
Random Forest	0.77 (sd $9.65E-02$)	0.81 (sd $9.39E-02$)	0.78 (sd $9.48E-02$)	0.90 (sd $6.17E-02$)
K-NN (N=5)	0.77 (sd $4.86E-02$)	0.79 (sd $5.68E-02$)	0.77 (sd $5.06E-02$)	0.86 (sd $5.65E-02$)
K-NN (N=10)	0.73 (sd $9.11E-02$)	0.78 (sd $7.48E-02$)	0.73 (sd $9.23E-02$)	0.87 (sd $4.83E-02$)
K-NN (N=20)	0.72 (sd $8.42E-02$)	0.77 (sd $6.48E-02$)	0.73 (sd $8.65E-02$)	0.86 (sd $4.45E-02$)
K-NN (N=50)	0.68 (sd $4.86E-02$)	0.80 (sd $7.21E-02$)	0.69 (sd $5.51E-02$)	0.83 (sd $4.50E-02$)
Gradient Boosting (exponential)	0.77 (sd $7.78E-02$)	0.77 (sd $7.86E-02$)	0.75 (sd $7.35E-02$)	0.88 (sd $7.61E-02$)
Gradient Boosting (deviance)	0.75 (sd $1.07E-01$)	0.75 (sd $9.24E-02$)	0.73 (sd $9.48E-02$)	0.87 (sd $8.85E-02$)
Balanced RF	0.81 (sd $7.91E-02$)	0.82 (sd $7.75E-02$)	0.81 (sd $7.37E-02$)	0.91 (sd $5.24E-02$)
Balanced ADABOOST	0.78 (sd $8.95E-02$)	0.80 (sd $1.05E-01$)	0.77 (sd $9.70E-02$)	0.86 (sd $7.23E-02$)

Table A.27: *Classification metrics for all classifiers evaluated using global network properties in $D_{[1000,+\infty]}$. Training and test data include all sources except Breitbart.com and Politicususa.com.*

A.7 Networks Plots

In this section we provide some example plots for networks belonging to both news domains. We used two different strategies to identify most appropriate individuals to plot:

1. *nearest* individual: the network with the smallest Euclidean distance from all other

Appendix A. Supplementary Information for "A network-based approach to detect online disinformation on Twitter"

	Evaluation Metrics for Classifiers in D_{all}			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.67 (sd $4.97E-03$)	0.78 (sd $1.04E-02$)	0.66 (sd $6.13E-03$)	0.76 (sd $8.18E-03$)
SVC RBF	0.72 (sd $1.05E-02$)	0.75 (sd $1.43E-02$)	0.72 (sd $1.12E-02$)	0.74 (sd $2.86E-02$)
Logistic Regression	0.70 (sd $1.07E-02$)	0.75 (sd $1.25E-02$)	0.70 (sd $1.20E-02$)	0.77 (sd $7.45E-03$)
Random Forest	0.67 (sd $4.40E-02$)	0.68 (sd $4.87E-02$)	0.67 (sd $4.54E-02$)	0.74 (sd $3.93E-02$)
K-NN (N=5)	0.69 (sd $9.10E-03$)	0.71 (sd $1.33E-02$)	0.70 (sd $8.93E-03$)	0.75 (sd $1.51E-02$)
K-NN (N=10)	0.70 (sd $1.33E-02$)	0.74 (sd $1.50E-02$)	0.70 (sd $1.41E-02$)	0.77 (sd $1.64E-02$)
K-NN (N=20)	0.71 (sd $8.43E-03$)	0.75 (sd $1.33E-02$)	0.72 (sd $8.93E-03$)	0.79 (sd $1.30E-02$)
K-NN (N=50)	0.72 (sd $9.63E-03$)	0.75 (sd $1.43E-02$)	0.73 (sd $1.02E-02$)	0.80 (sd $1.34E-02$)
Gradient Boosting (exponential)	0.69 (sd $3.64E-02$)	0.72 (sd $4.46E-02$)	0.69 (sd $3.81E-02$)	0.77 (sd $3.60E-02$)
Gradient Boosting (deviance)	0.68 (sd $4.48E-02$)	0.69 (sd $5.36E-02$)	0.68 (sd $4.75E-02$)	0.75 (sd $5.50E-02$)
Balanced RF	0.67 (sd $3.61E-02$)	0.68 (sd $2.95E-02$)	0.67 (sd $3.44E-02$)	0.75 (sd $3.64E-02$)
Balanced ADABOOST	0.69 (sd $4.36E-02$)	0.73 (sd $1.46E-02$)	0.69 (sd $4.76E-02$)	0.76 (sd $4.25E-02$)

Table A.28: *Classification metrics for all classifiers evaluated using global network properties in D_{all}. Training and test data include all sources except Politicususa.com.*

	Evaluation Metrics for Classifiers in $D_{[0,100]}$			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.59 (sd $1.81E-02$)	0.77 (sd $3.07E-02$)	0.58 (sd $2.67E-02$)	0.70 (sd $4.67E-02$)
SVC RBF	0.68 (sd $1.77E-02$)	0.72 (sd $2.21E-02$)	0.69 (sd $1.92E-02$)	0.74 (sd $3.15E-02$)
Logistic Regression	0.64 (sd $2.02E-02$)	0.72 (sd $2.80E-02$)	0.65 (sd $2.38E-02$)	0.75 (sd $2.52E-02$)
Random Forest	0.67 (sd $2.54E-02$)	0.69 (sd $2.81E-02$)	0.68 (sd $2.66E-02$)	0.77 (sd $3.07E-02$)
K-NN (N=5)	0.66 (sd $1.98E-02$)	0.68 (sd $2.32E-02$)	0.67 (sd $2.09E-02$)	0.74 (sd $2.40E-02$)
K-NN (N=10)	0.66 (sd $1.98E-02$)	0.71 (sd $2.58E-02$)	0.67 (sd $2.20E-02$)	0.76 (sd $2.14E-02$)
K-NN (N=20)	0.67 (sd $2.07E-02$)	0.72 (sd $2.73E-02$)	0.68 (sd $2.29E-02$)	0.77 (sd $2.73E-02$)
K-NN (N=50)	0.67 (sd $1.63E-02$)	0.73 (sd $1.99E-02$)	0.68 (sd $1.79E-02$)	0.78 (sd $2.53E-02$)
Gradient Boosting (exponential)	0.68 (sd $1.45E-02$)	0.73 (sd $1.74E-02$)	0.69 (sd $1.56E-02$)	0.79 (sd $2.36E-02$)
Gradient Boosting (deviance)	0.68 (sd $1.60E-02$)	0.73 (sd $2.02E-02$)	0.69 (sd $1.74E-02$)	0.79 (sd $2.51E-02$)
Balanced RF	0.70 (sd $2.36E-02$)	0.68 (sd $2.19E-02$)	0.68 (sd $2.31E-02$)	0.78 (sd $2.82E-02$)
Balanced ADABOOST	0.69 (sd $2.44E-02$)	0.68 (sd $2.28E-02$)	0.68 (sd $2.39E-02$)	0.78 (sd $2.46E-02$)

Table A.29: *Classification metrics for all classifiers evaluated using global network properties in $D_{[0,100]}$. Training and test data include all sources except Politicususa.com.*

	Evaluation Metrics for Classifiers in $D_{[100,1000]}$			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.76 (sd $2.07E-02$)	0.76 (sd $2.16E-02$)	0.76 (sd $2.08E-02$)	0.85 (sd $1.56E-02$)
SVC RBF	0.76 (sd $2.03E-02$)	0.76 (sd $1.99E-02$)	0.76 (sd $2.04E-02$)	0.81 (sd $2.21E-02$)
Logistic Regression	0.75 (sd $2.13E-02$)	0.77 (sd $2.27E-02$)	0.75 (sd $2.17E-02$)	0.85 (sd $1.71E-02$)
Random Forest	0.76 (sd $1.65E-02$)	0.77 (sd $1.72E-02$)	0.76 (sd $1.66E-02$)	0.85 (sd $1.59E-02$)
K-NN (N=5)	0.75 (sd $2.05E-02$)	0.75 (sd $2.06E-02$)	0.75 (sd $2.04E-02$)	0.82 (sd $2.22E-02$)
K-NN (N=10)	0.76 (sd $1.28E-02$)	0.77 (sd $1.29E-02$)	0.76 (sd $1.30E-02$)	0.84 (sd $1.69E-02$)
K-NN (N=20)	0.76 (sd $1.45E-02$)	0.76 (sd $1.40E-02$)	0.76 (sd $1.46E-02$)	0.85 (sd $1.89E-02$)
K-NN (N=50)	0.77 (sd $1.36E-02$)	0.77 (sd $1.36E-02$)	0.77 (sd $1.36E-02$)	0.85 (sd $1.67E-02$)
Gradient Boosting (exponential)	0.76 (sd $1.42E-02$)	0.77 (sd $1.60E-02$)	0.76 (sd $1.43E-02$)	0.86 (sd $1.70E-02$)
Gradient Boosting (deviance)	0.77 (sd $1.75E-02$)	0.77 (sd $1.92E-02$)	0.77 (sd $1.76E-02$)	0.86 (sd $1.86E-02$)
Balanced RF	0.77 (sd $1.32E-02$)	0.77 (sd $1.39E-02$)	0.77 (sd $1.33E-02$)	0.85 (sd $1.62E-02$)
Balanced ADABOOST	0.76 (sd $1.91E-02$)	0.76 (sd $1.91E-02$)	0.76 (sd $1.92E-02$)	0.85 (sd $1.94E-02$)

Table A.30: *Classification metrics for all classifiers evaluated using global network properties in $D_{[100,1000]}$. Training and test data include all sources except Politicususa.com.*

individuals in the same domain, using the vectors of global network properties;

2. *farthest* individual: the network with the highest Euclidean distance from all other individuals in the other domain, using the vectors of global network properties.

Plots were obtained using Gephi and the Force Atlas 2 visualization algorithm with parameters: Stronger Gravity = ON, Approximate Repulsion = ON, Prevent Overlap =

	Evaluation Metrics for Classifiers in $D_{[1000,+\infty]}$			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.81 (sd $3.88E-02$)	0.82 (sd $3.80E-02$)	0.81 (sd $3.93E-02$)	0.90 (sd $3.13E-02$)
SVC RBF	0.78 (sd $5.02E-02$)	0.79 (sd $5.05E-02$)	0.78 (sd $5.10E-02$)	0.88 (sd $5.17E-02$)
Logistic Regression	0.84 (sd $4.48E-02$)	0.85 (sd $4.61E-02$)	0.84 (sd $4.49E-02$)	0.93 (sd $2.75E-02$)
Random Forest	0.83 (sd $5.68E-02$)	0.84 (sd $5.27E-02$)	0.83 (sd $5.83E-02$)	0.92 (sd $3.31E-02$)
K-NN (N=5)	0.83 (sd $6.50E-02$)	0.83 (sd $6.53E-02$)	0.83 (sd $6.51E-02$)	0.89 (sd $6.36E-02$)
K-NN (N=10)	0.81 (sd $6.52E-02$)	0.81 (sd $6.61E-02$)	0.81 (sd $6.53E-02$)	0.89 (sd $5.04E-02$)
K-NN (N=20)	0.81 (sd $4.77E-02$)	0.81 (sd $4.69E-02$)	0.81 (sd $4.80E-02$)	0.87 (sd $4.27E-02$)
K-NN (N=50)	0.77 (sd $7.91E-02$)	0.77 (sd $7.96E-02$)	0.77 (sd $7.97E-02$)	0.86 (sd $4.53E-02$)
Gradient Boosting (exponential)	0.82 (sd $4.72E-02$)	0.83 (sd $4.63E-02$)	0.82 (sd $4.76E-02$)	0.91 (sd $3.21E-02$)
Gradient Boosting (deviance)	0.81 (sd $5.15E-02$)	0.82 (sd $5.26E-02$)	0.81 (sd $5.16E-02$)	0.89 (sd $4.01E-02$)
Balanced RF	0.84 (sd $5.61E-02$)	0.85 (sd $5.70E-02$)	0.84 (sd $5.64E-02$)	0.93 (sd $3.62E-02$)
Balanced ADABOOST	0.79 (sd $6.24E-02$)	0.80 (sd $5.85E-02$)	0.79 (sd $6.57E-02$)	0.88 (sd $4.92E-02$)

Table A.31: *Classification metrics for all classifiers evaluated using global network properties in $D_{[1000,+\infty]}$. Training and test data include all sources except Politicususa.com.*

	Evaluation Metrics for Classifiers in D_{all}			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.66 (sd $1.10E-02$)	0.79 (sd $1.68E-02$)	0.67 (sd $1.35E-02$)	0.75 (sd $1.84E-02$)
SVC RBF	0.69 (sd $1.18E-02$)	0.78 (sd $1.56E-02$)	0.71 (sd $1.33E-02$)	0.72 (sd $2.91E-02$)
Logistic Regression	0.68 (sd $8.65E-03$)	0.77 (sd $2.08E-02$)	0.70 (sd $9.99E-03$)	0.78 (sd $1.59E-02$)
Random Forest	0.66 (sd $2.79E-02$)	0.75 (sd $1.73E-02$)	0.67 (sd $3.19E-02$)	0.74 (sd $3.82E-02$)
K-NN (N=5)	0.69 (sd $1.41E-02$)	0.73 (sd $1.05E-02$)	0.70 (sd $1.42E-02$)	0.74 (sd $1.66E-02$)
K-NN (N=10)	0.69 (sd $1.09E-02$)	0.78 (sd $7.69E-03$)	0.71 (sd $1.18E-02$)	0.77 (sd $1.74E-02$)
K-NN (N=20)	0.70 (sd $1.24E-02$)	0.78 (sd $1.46E-02$)	0.71 (sd $1.39E-02$)	0.78 (sd $1.94E-02$)
K-NN (N=50)	0.70 (sd $1.50E-02$)	0.78 (sd $1.35E-02$)	0.71 (sd $1.62E-02$)	0.79 (sd $1.54E-02$)
Gradient Boosting (exponential)	0.66 (sd $2.37E-02$)	0.76 (sd $3.15E-02$)	0.67 (sd $2.82E-02$)	0.75 (sd $4.30E-02$)
Gradient Boosting (deviance)	0.64 (sd $3.63E-02$)	0.68 (sd $7.26E-02$)	0.65 (sd $4.17E-02$)	0.73 (sd $5.15E-02$)
Balanced RF	0.68 (sd $2.95E-02$)	0.72 (sd $2.48E-02$)	0.69 (sd $2.52E-02$)	0.75 (sd $3.95E-02$)
Balanced ADABOOST	0.67 (sd $4.00E-02$)	0.72 (sd $1.59E-02$)	0.67 (sd $3.81E-02$)	0.74 (sd $4.37E-02$)

Table A.32: *Classification metrics for all classifiers evaluated using global network properties in D_{all}. Training and test data include all sources except Breitbart.com.*

	Evaluation Metrics for Classifiers in $D_{[0,100]}$			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.59 (sd $1.54E-02$)	0.80 (sd $3.72E-02$)	0.59 (sd $2.26E-02$)	0.66 (sd $6.02E-02$)
SVC RBF	0.61 (sd $2.55E-02$)	0.71 (sd $2.98E-02$)	0.62 (sd $3.35E-02$)	0.72 (sd $2.35E-02$)
Logistic Regression	0.61 (sd $1.81E-02$)	0.74 (sd $3.39E-02$)	0.62 (sd $2.34E-02$)	0.75 (sd $1.72E-02$)
Random Forest	0.64 (sd $2.10E-02$)	0.69 (sd $2.89E-02$)	0.66 (sd $2.38E-02$)	0.75 (sd $2.12E-02$)
K-NN (N=5)	0.64 (sd $2.43E-02$)	0.68 (sd $2.67E-02$)	0.65 (sd $2.66E-02$)	0.72 (sd $2.24E-02$)
K-NN (N=10)	0.63 (sd $1.49E-02$)	0.73 (sd $2.33E-02$)	0.65 (sd $1.82E-02$)	0.75 (sd $2.28E-02$)
K-NN (N=20)	0.64 (sd $1.47E-02$)	0.74 (sd $2.50E-02$)	0.66 (sd $1.80E-02$)	0.77 (sd $2.06E-02$)
K-NN (N=50)	0.63 (sd $1.31E-02$)	0.74 (sd $2.96E-02$)	0.64 (sd $1.66E-02$)	0.77 (sd $1.63E-02$)
Gradient Boosting (exponential)	0.63 (sd $2.38E-02$)	0.75 (sd $3.59E-02$)	0.64 (sd $2.95E-02$)	0.78 (sd $1.24E-02$)
Gradient Boosting (deviance)	0.63 (sd $2.62E-02$)	0.74 (sd $3.55E-02$)	0.65 (sd $3.21E-02$)	0.79 (sd $1.43E-02$)
Balanced RF	0.69 (sd $2.21E-02$)	0.65 (sd $1.72E-02$)	0.65 (sd $1.76E-02$)	0.77 (sd $1.96E-02$)
Balanced ADABOOST	0.68 (sd $1.36E-02$)	0.65 (sd $9.66E-03$)	0.66 (sd $9.60E-03$)	0.77 (sd $1.31E-02$)

Table A.33: *Classification metrics for all classifiers evaluated using global network properties in $D_{[0,100]}$. Training and test data include all sources except Breitbart.com.*

ON, Scaling = 100. We also adjust node sizes according to their degree.

Appendix A. Supplementary Information for "A network-based approach to detect online disinformation on Twitter"

	Evaluation Metrics for Classifiers in $D_{[100,1000]}$			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.71 (sd $1.87E-02$)	0.80 (sd $2.12E-02$)	0.72 (sd $2.13E-02$)	0.84 (sd $1.65E-02$)
SVC RBF	0.75 (sd $1.48E-02$)	0.78 (sd $1.24E-02$)	0.76 (sd $1.43E-02$)	0.80 (sd $2.05E-02$)
Logistic Regression	0.75 (sd $1.49E-02$)	0.78 (sd $1.97E-02$)	0.75 (sd $1.59E-02$)	0.85 (sd $1.60E-02$)
Random Forest	0.76 (sd $1.81E-02$)	0.77 (sd $1.98E-02$)	0.76 (sd $1.84E-02$)	0.85 (sd $1.30E-02$)
K-NN (N=5)	0.75 (sd $2.15E-02$)	0.76 (sd $2.18E-02$)	0.75 (sd $2.15E-02$)	0.82 (sd $1.60E-02$)
K-NN (N=10)	0.75 (sd $1.93E-02$)	0.78 (sd $1.78E-02$)	0.76 (sd $1.95E-02$)	0.84 (sd $1.47E-02$)
K-NN (N=20)	0.75 (sd $1.30E-02$)	0.78 (sd $1.08E-02$)	0.76 (sd $1.29E-02$)	0.85 (sd $1.59E-02$)
K-NN (N=50)	0.76 (sd $1.29E-02$)	0.79 (sd $1.38E-02$)	0.77 (sd $1.31E-02$)	0.85 (sd $1.42E-02$)
Gradient Boosting (exponential)	0.76 (sd $1.90E-02$)	0.80 (sd $1.19E-02$)	0.76 (sd $1.90E-02$)	0.86 (sd $1.26E-02$)
Gradient Boosting (deviance)	0.76 (sd $1.69E-02$)	0.80 (sd $1.36E-02$)	0.77 (sd $1.71E-02$)	0.86 (sd $1.21E-02$)
Balanced RF	0.76 (sd $1.74E-02$)	0.76 (sd $1.65E-02$)	0.76 (sd $1.69E-02$)	0.85 (sd $1.38E-02$)
Balanced ADABOOST	0.76 (sd $1.71E-02$)	0.75 (sd $1.88E-02$)	0.76 (sd $1.80E-02$)	0.85 (sd $1.23E-02$)

Table A.34: *Classification metrics for all classifiers evaluated using global network properties in $D_{[100,1000]}$. Training and test data include all sources except Breitbart.com.*

	Evaluation Metrics for Classifiers in $D_{[1000,+\infty]}$			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.75 (sd $5.88E-02$)	0.82 (sd $4.93E-02$)	0.76 (sd $6.45E-02$)	0.92 (sd $4.12E-02$)
SVC RBF	0.78 (sd $3.95E-02$)	0.83 (sd $5.38E-02$)	0.79 (sd $3.99E-02$)	0.89 (sd $5.37E-02$)
Logistic Regression	0.86 (sd $5.21E-02$)	0.85 (sd $5.70E-02$)	0.85 (sd $5.50E-02$)	0.93 (sd $3.47E-02$)
Random Forest	0.82 (sd $4.27E-02$)	0.83 (sd $3.78E-02$)	0.82 (sd $4.11E-02$)	0.90 (sd $3.15E-02$)
K-NN (N=5)	0.84 (sd $6.35E-02$)	0.84 (sd $5.90E-02$)	0.84 (sd $6.27E-02$)	0.91 (sd $4.30E-02$)
K-NN (N=10)	0.82 (sd $6.96E-02$)	0.83 (sd $6.69E-02$)	0.82 (sd $7.01E-02$)	0.90 (sd $4.34E-02$)
K-NN (N=20)	0.81 (sd $7.36E-02$)	0.82 (sd $7.33E-02$)	0.82 (sd $7.24E-02$)	0.89 (sd $4.66E-02$)
K-NN (N=50)	0.76 (sd $8.04E-02$)	0.79 (sd $7.58E-02$)	0.77 (sd $8.38E-02$)	0.88 (sd $5.60E-02$)
Gradient Boosting (exponential)	0.84 (sd $5.35E-02$)	0.84 (sd $5.29E-02$)	0.84 (sd $5.33E-02$)	0.89 (sd $4.31E-02$)
Gradient Boosting (deviance)	0.84 (sd $5.65E-02$)	0.83 (sd $5.63E-02$)	0.83 (sd $5.64E-02$)	0.89 (sd $4.59E-02$)
Balanced RF	0.83 (sd $5.38E-02$)	0.83 (sd $4.89E-02$)	0.83 (sd $5.03E-02$)	0.90 (sd $4.04E-02$)
Balanced ADABOOST	0.80 (sd $6.18E-02$)	0.80 (sd $6.23E-02$)	0.79 (sd $6.34E-02$)	0.90 (sd $3.41E-02$)

Table A.35: *Classification metrics for all classifiers evaluated using global network properties in $D_{[1000,+\infty]}$.Training and test data include all sources except Breitbart.com.*

	Evaluation Metrics for Classifiers in D_{all}			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.71 (sd $1.23E-02$)	0.71 (sd $1.15E-02$)	0.69 (sd $1.41E-02$)	0.78 (sd $1.30E-02$)
SVC RBF	0.72 (sd $1.32E-02$)	0.72 (sd $1.37E-02$)	0.72 (sd $1.37E-02$)	0.77 (sd $1.45E-02$)
Logistic Regression	0.71 (sd $3.08E-02$)	0.71 (sd $3.06E-02$)	0.70 (sd $2.72E-02$)	0.78 (sd $1.54E-02$)
Random Forest	0.67 (sd $2.38E-02$)	0.68 (sd $2.48E-02$)	0.67 (sd $2.50E-02$)	0.77 (sd $2.63E-02$)
K-NN (N=5)	0.69 (sd $2.22E-02$)	0.69 (sd $1.82E-02$)	0.69 (sd $2.05E-02$)	0.76 (sd $1.85E-02$)
K-NN (N=10)	0.70 (sd $2.16E-02$)	0.70 (sd $2.04E-02$)	0.70 (sd $1.83E-02$)	0.78 (sd $1.49E-02$)
K-NN (N=20)	0.72 (sd $1.57E-02$)	0.72 (sd $1.39E-02$)	0.71 (sd $1.39E-02$)	0.80 (sd $1.14E-02$)
K-NN (N=50)	0.72 (sd $1.63E-02$)	0.72 (sd $1.29E-02$)	0.72 (sd $1.47E-02$)	0.81 (sd $1.12E-02$)
Gradient Boosting (exponential)	0.71 (sd $2.08E-02$)	0.71 (sd $2.10E-02$)	0.71 (sd $2.20E-02$)	0.79 (sd $2.45E-02$)
Gradient Boosting (deviance)	0.70 (sd $2.67E-02$)	0.70 (sd $2.74E-02$)	0.70 (sd $2.69E-02$)	0.78 (sd $3.27E-02$)
Balanced RF	0.69 (sd $1.73E-02$)	0.69 (sd $1.55E-02$)	0.68 (sd $2.17E-02$)	0.78 (sd $2.21E-02$)
Balanced ADABOOST	0.71 (sd $2.08E-02$)	0.72 (sd $1.34E-02$)	0.70 (sd $3.02E-02$)	0.78 (sd $3.44E-02$)

Table A.36: *Classification metrics for all classifiers evaluated using global network properties in D_{all}. Mainstream training and test data include only Left sources.*

	Evaluation Metrics for Classifiers in $D_{[0,100]}$			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.68 (sd $1.55E-02$)	0.69 (sd $1.58E-02$)	0.67 (sd $1.64E-02$)	0.74 (sd $1.98E-02$)
SVC RBF	0.69 (sd $1.71E-02$)	0.69 (sd $1.66E-02$)	0.69 (sd $1.74E-02$)	0.75 (sd $2.01E-02$)
Logistic Regression	0.67 (sd $2.14E-02$)	0.68 (sd $2.27E-02$)	0.67 (sd $2.15E-02$)	0.75 (sd $1.86E-02$)
Random Forest	0.68 (sd $1.87E-02$)	0.68 (sd $1.87E-02$)	0.68 (sd $1.87E-02$)	0.77 (sd $2.14E-02$)
K-NN (N=5)	0.68 (sd $2.15E-02$)	0.68 (sd $2.13E-02$)	0.68 (sd $2.15E-02$)	0.75 (sd $2.13E-02$)
K-NN (N=10)	0.68 (sd $1.70E-02$)	0.69 (sd $1.66E-02$)	0.68 (sd $1.75E-02$)	0.77 (sd $1.91E-02$)
K-NN (N=20)	0.69 (sd $1.97E-02$)	0.69 (sd $1.99E-02$)	0.69 (sd $1.99E-02$)	0.78 (sd $1.90E-02$)
K-NN (N=50)	0.69 (sd $1.24E-02$)	0.69 (sd $1.24E-02$)	0.69 (sd $1.25E-02$)	0.78 (sd $1.44E-02$)
Gradient Boosting (exponential)	0.69 (sd $1.78E-02$)	0.70 (sd $1.73E-02$)	0.69 (sd $1.81E-02$)	0.79 (sd $1.66E-02$)
Gradient Boosting (deviance)	0.69 (sd $1.83E-02$)	0.70 (sd $1.79E-02$)	0.69 (sd $1.86E-02$)	0.79 (sd $1.71E-02$)
Balanced RF	0.68 (sd $1.83E-02$)	0.68 (sd $1.82E-02$)	0.68 (sd $1.83E-02$)	0.78 (sd $1.85E-02$)
Balanced ADABOOST	0.68 (sd $2.45E-02$)	0.69 (sd $2.42E-02$)	0.68 (sd $2.46E-02$)	0.77 (sd $1.90E-02$)

Table A.37: *Classification metrics for all classifiers evaluated using global network properties in $D_{[0,100]}$. Mainstream training and test data include only Left sources.*

	Evaluation Metrics for Classifiers in $D_{[100,1000]}$			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.76 (sd $1.99E-02$)	0.77 (sd $1.82E-02$)	0.76 (sd $1.88E-02$)	0.87 (sd $8.59E-03$)
SVC RBF	0.76 (sd $1.70E-02$)	0.78 (sd $1.70E-02$)	0.77 (sd $1.66E-02$)	0.82 (sd $1.04E-02$)
Logistic Regression	0.78 (sd $2.01E-02$)	0.77 (sd $1.92E-02$)	0.77 (sd $1.97E-02$)	0.87 (sd $1.14E-02$)
Random Forest	0.76 (sd $1.66E-02$)	0.77 (sd $1.68E-02$)	0.77 (sd $1.59E-02$)	0.87 (sd $9.46E-03$)
K-NN (N=5)	0.76 (sd $1.38E-02$)	0.76 (sd $1.58E-02$)	0.76 (sd $1.37E-02$)	0.84 (sd $1.29E-02$)
K-NN (N=10)	0.78 (sd $1.94E-02$)	0.78 (sd $1.84E-02$)	0.78 (sd $1.88E-02$)	0.86 (sd $1.29E-02$)
K-NN (N=20)	0.77 (sd $1.57E-02$)	0.78 (sd $1.67E-02$)	0.77 (sd $1.56E-02$)	0.87 (sd $8.59E-03$)
K-NN (N=50)	0.76 (sd $1.57E-02$)	0.78 (sd $1.54E-02$)	0.77 (sd $1.53E-02$)	0.87 (sd $1.06E-02$)
Gradient Boosting (exponential)	0.76 (sd $1.17E-02$)	0.78 (sd $1.11E-02$)	0.77 (sd $1.06E-02$)	0.88 (sd $9.11E-03$)
Gradient Boosting (deviance)	0.77 (sd $1.46E-02$)	0.78 (sd $1.43E-02$)	0.77 (sd $1.37E-02$)	0.88 (sd $9.62E-03$)
Balanced RF	0.79 (sd $1.63E-02$)	0.77 (sd $1.58E-02$)	0.77 (sd $1.68E-02$)	0.88 (sd $9.83E-03$)
Balanced ADABOOST	0.78 (sd $1.31E-02$)	0.77 (sd $1.26E-02$)	0.77 (sd $1.29E-02$)	0.87 (sd $7.09E-03$)

Table A.38: *Classification metrics for all classifiers evaluated using global network properties in $D_{[100,1000]}$. Mainstream training and test data include only Left sources.*

	Evaluation Metrics for Classifiers in $D_{[1000,+\infty]}$			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.66 (sd $3.75E-02$)	0.80 (sd $5.22E-02$)	0.67 (sd $4.56E-02$)	0.87 (sd $6.16E-02$)
SVC RBF	0.73 (sd $5.56E-02$)	0.80 (sd $4.68E-02$)	0.74 (sd $6.18E-02$)	0.83 (sd $7.35E-02$)
Logistic Regression	0.82 (sd $5.44E-02$)	0.84 (sd $4.29E-02$)	0.82 (sd $4.72E-02$)	0.90 (sd $4.17E-02$)
Random Forest	0.78 (sd $5.66E-02$)	0.79 (sd $5.76E-02$)	0.78 (sd $4.90E-02$)	0.88 (sd $4.88E-02$)
K-NN (N=5)	0.80 (sd $5.61E-02$)	0.80 (sd $5.72E-02$)	0.80 (sd $5.31E-02$)	0.86 (sd $4.04E-02$)
K-NN (N=10)	0.83 (sd $4.08E-02$)	0.82 (sd $4.66E-02$)	0.82 (sd $4.29E-02$)	0.89 (sd $3.85E-02$)
K-NN (N=20)	0.79 (sd $5.37E-02$)	0.82 (sd $5.19E-02$)	0.80 (sd $5.31E-02$)	0.87 (sd $4.07E-02$)
K-NN (N=50)	0.74 (sd $6.77E-02$)	0.81 (sd $4.64E-02$)	0.75 (sd $7.43E-02$)	0.86 (sd $5.24E-02$)
Gradient Boosting (exponential)	0.82 (sd $7.23E-02$)	0.83 (sd $6.45E-02$)	0.82 (sd $6.87E-02$)	0.88 (sd $5.75E-02$)
Gradient Boosting (deviance)	0.81 (sd $7.32E-02$)	0.82 (sd $7.01E-02$)	0.81 (sd $6.92E-02$)	0.88 (sd $5.38E-02$)
Balanced RF	0.79 (sd $4.60E-02$)	0.78 (sd $2.84E-02$)	0.77 (sd $3.57E-02$)	0.89 (sd $4.78E-02$)
Balanced ADABOOST	0.82 (sd $5.75E-02$)	0.81 (sd $4.58E-02$)	0.81 (sd $5.06E-02$)	0.88 (sd $4.34E-02$)

Table A.39: *Classification metrics for all classifiers evaluated using global network properties in $D_{[1000,+\infty]}$. Mainstream training and test data include only Left sources.*

Appendix A. Supplementary Information for "A network-based approach to detect online disinformation on Twitter"

	Evaluation Metrics for Classifiers in D_{all}			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.50 (sd $0.00E+00$)	0.41 (sd $5.55E-17$)	0.45 (sd $0.00E+00$)	0.73 (sd $2.78E-02$)
SVC RBF	0.58 (sd $1.49E-02$)	0.75 (sd $6.44E-02$)	0.59 (sd $2.20E-02$)	0.67 (sd $3.00E-02$)
Logistic Regression	0.55 (sd $2.20E-02$)	0.78 (sd $7.45E-02$)	0.54 (sd $3.69E-02$)	0.78 (sd $1.67E-02$)
Random Forest	0.60 (sd $2.96E-02$)	0.64 (sd $3.78E-02$)	0.61 (sd $2.84E-02$)	0.75 (sd $2.61E-02$)
K-NN (N=5)	0.63 (sd $4.41E-02$)	0.65 (sd $3.29E-02$)	0.63 (sd $4.28E-02$)	0.73 (sd $2.98E-02$)
K-NN (N=10)	0.63 (sd $4.41E-02$)	0.65 (sd $2.71E-02$)	0.63 (sd $4.20E-02$)	0.76 (sd $2.07E-02$)
K-NN (N=20)	0.62 (sd $3.78E-02$)	0.69 (sd $4.63E-02$)	0.63 (sd $4.17E-02$)	0.77 (sd $1.80E-02$)
K-NN (N=50)	0.59 (sd $3.52E-02$)	0.72 (sd $4.89E-02$)	0.61 (sd $4.79E-02$)	0.79 (sd $1.56E-02$)
Gradient Boosting (exponential)	0.59 (sd $3.34E-02$)	0.69 (sd $5.25E-02$)	0.60 (sd $3.75E-02$)	0.78 (sd $2.94E-02$)
Gradient Boosting (deviance)	0.58 (sd $2.47E-02$)	0.69 (sd $5.39E-02$)	0.59 (sd $3.21E-02$)	0.77 (sd $2.94E-02$)
Balanced RF	0.71 (sd $1.72E-02$)	0.63 (sd $1.09E-02$)	0.59 (sd $4.05E-02$)	0.77 (sd $2.55E-02$)
Balanced ADABOOST	0.69 (sd $2.55E-02$)	0.62 (sd $1.60E-02$)	0.57 (sd $6.00E-02$)	0.74 (sd $4.06E-02$)

Table A.40: *Classification metrics for all classifiers evaluated using global network properties in D_{all}. Mainstream training and test data include only Right sources.*

	Evaluation Metrics for Classifiers in $D_{[0,100]}$			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.50 (sd $0.00E+00$)	0.38 (sd $5.55E-17$)	0.43 (sd $0.00E+00$)	0.61 (sd $1.16E-01$)
SVC RBF	0.62 (sd $1.15E-02$)	0.82 (sd $3.82E-02$)	0.64 (sd $1.59E-02$)	0.61 (sd $3.27E-02$)
Logistic Regression	0.60 (sd $1.38E-02$)	0.82 (sd $2.23E-02$)	0.61 (sd $2.07E-02$)	0.74 (sd $2.04E-02$)
Random Forest	0.64 (sd $1.81E-02$)	0.71 (sd $1.59E-02$)	0.65 (sd $1.93E-02$)	0.75 (sd $1.61E-02$)
K-NN (N=5)	0.63 (sd $2.40E-02$)	0.69 (sd $2.99E-02$)	0.64 (sd $2.81E-02$)	0.71 (sd $1.72E-02$)
K-NN (N=10)	0.64 (sd $2.62E-02$)	0.72 (sd $3.40E-02$)	0.66 (sd $3.05E-02$)	0.74 (sd $2.45E-02$)
K-NN (N=20)	0.63 (sd $1.84E-02$)	0.76 (sd $2.96E-02$)	0.65 (sd $2.34E-02$)	0.75 (sd $2.30E-02$)
K-NN (N=50)	0.62 (sd $1.66E-02$)	0.80 (sd $4.21E-02$)	0.64 (sd $2.25E-02$)	0.76 (sd $1.63E-02$)
Gradient Boosting (exponential)	0.62 (sd $1.32E-02$)	0.80 (sd $2.89E-02$)	0.64 (sd $1.79E-02$)	0.77 (sd $2.40E-02$)
Gradient Boosting (deviance)	0.62 (sd $1.53E-02$)	0.78 (sd $4.21E-02$)	0.64 (sd $1.99E-02$)	0.77 (sd $2.38E-02$)
Balanced RF	0.68 (sd $1.73E-02$)	0.64 (sd $1.43E-02$)	0.64 (sd $1.87E-02$)	0.76 (sd $1.73E-02$)
Balanced ADABOOST	0.69 (sd $1.55E-02$)	0.64 (sd $1.19E-02$)	0.63 (sd $1.77E-02$)	0.76 (sd $1.78E-02$)

Table A.41: *Classification metrics for all classifiers evaluated using global network properties in $D_{[0,100]}$. Mainstream training and test data include only Right sources.*

	Evaluation Metrics for Classifiers in $D_{[100,1000]}$			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.50 (sd $0.00E+00$)	0.43 (sd $5.55E-17$)	0.46 (sd $0.00E+00$)	0.84 (sd $2.69E-02$)
SVC RBF	0.58 (sd $2.10E-02$)	0.83 (sd $6.85E-02$)	0.60 (sd $3.29E-02$)	0.71 (sd $4.60E-02$)
Logistic Regression	0.62 (sd $4.71E-02$)	0.79 (sd $9.21E-02$)	0.64 (sd $4.76E-02$)	0.85 (sd $2.99E-02$)
Random Forest	0.64 (sd $3.01E-02$)	0.75 (sd $4.58E-02$)	0.67 (sd $3.62E-02$)	0.84 (sd $2.37E-02$)
K-NN (N=5)	0.62 (sd $3.66E-02$)	0.70 (sd $4.72E-02$)	0.64 (sd $4.15E-02$)	0.79 (sd $3.20E-02$)
K-NN (N=10)	0.63 (sd $3.16E-02$)	0.72 (sd $2.93E-02$)	0.66 (sd $3.41E-02$)	0.82 (sd $2.42E-02$)
K-NN (N=20)	0.61 (sd $3.03E-02$)	0.76 (sd $6.50E-02$)	0.64 (sd $3.85E-02$)	0.83 (sd $2.30E-02$)
K-NN (N=50)	0.61 (sd $3.06E-02$)	0.80 (sd $7.98E-02$)	0.64 (sd $4.25E-02$)	0.83 (sd $2.26E-02$)
Gradient Boosting (exponential)	0.61 (sd $2.42E-02$)	0.79 (sd $6.35E-02$)	0.64 (sd $3.34E-02$)	0.84 (sd $2.53E-02$)
Gradient Boosting (deviance)	0.61 (sd $2.72E-02$)	0.76 (sd $6.48E-02$)	0.64 (sd $3.69E-02$)	0.84 (sd $2.27E-02$)
Balanced RF	0.76 (sd $1.78E-02$)	0.64 (sd $8.26E-03$)	0.64 (sd $1.22E-02$)	0.85 (sd $2.22E-02$)
Balanced ADABOOST	0.75 (sd $2.83E-02$)	0.64 (sd $1.37E-02$)	0.65 (sd $1.92E-02$)	0.84 (sd $2.78E-02$)

Table A.42: *Classification metrics for all classifiers evaluated using global network properties in $D_{[100,1000]}$. Mainstream training and test data include only Right sources.*

	Evaluation Metrics for Classifiers in $D_{[1000,+\infty]}$			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.50 (sd $0.00E+00$)	0.44 (sd $5.55E-17$)	0.47 (sd $0.00E+00$)	0.87 (sd $9.27E-02$)
SVC RBF	0.53 (sd $1.01E-01$)	0.50 (sd $1.61E-01$)	0.51 (sd $1.26E-01$)	0.70 (sd $2.03E-01$)
Logistic Regression	0.55 (sd $1.08E-01$)	0.58 (sd $2.11E-01$)	0.55 (sd $1.26E-01$)	0.86 (sd $1.03E-01$)
Random Forest	0.60 (sd $1.38E-01$)	0.66 (sd $2.32E-01$)	0.61 (sd $1.70E-01$)	0.81 (sd $7.79E-02$)
K-NN (N=5)	0.58 (sd $1.07E-01$)	0.59 (sd $1.77E-01$)	0.58 (sd $1.25E-01$)	0.68 (sd $1.07E-01$)
K-NN (N=10)	0.56 (sd $1.18E-01$)	0.60 (sd $2.41E-01$)	0.56 (sd $1.53E-01$)	0.71 (sd $1.03E-01$)
K-NN (N=20)	0.53 (sd $6.67E-02$)	0.55 (sd $2.07E-01$)	0.52 (sd $1.04E-01$)	0.76 (sd $1.10E-01$)
K-NN (N=50)	0.50 (sd $0.00E+00$)	0.44 (sd $5.55E-17$)	0.47 (sd $0.00E+00$)	0.75 (sd $1.43E-01$)
Gradient Boosting (exponential)	0.62 (sd $1.34E-01$)	0.61 (sd $1.48E-01$)	0.61 (sd $1.37E-01$)	0.86 (sd $8.87E-02$)
Gradient Boosting (deviance)	0.59 (sd $1.07E-01$)	0.58 (sd $1.30E-01$)	0.58 (sd $1.16E-01$)	0.83 (sd $1.09E-01$)
Balanced RF	0.81 (sd $1.07E-01$)	0.65 (sd $6.97E-02$)	0.64 (sd $1.08E-01$)	0.86 (sd $1.10E-01$)
Balanced ADABOOST	0.75 (sd $9.63E-02$)	0.69 (sd $9.33E-02$)	0.70 (sd $9.43E-02$)	0.86 (sd $6.77E-02$)

Table A.43: *Classification metrics for all classifiers evaluated using global network properties in $D_{[1000,+\infty]}$. Mainstream training and test data include only Right sources.*

	Evaluation Metrics for Classifiers in D_{all}			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.50 (sd $0.00E+00$)	0.42 (sd $0.00E+00$)	0.46 (sd $5.55E-17$)	0.71 (sd $4.59E-02$)
SVC RBF	0.58 (sd $2.62E-02$)	0.80 (sd $4.37E-02$)	0.60 (sd $3.74E-02$)	0.66 (sd $4.89E-02$)
Logistic Regression	0.53 (sd $1.90E-02$)	0.74 (sd $9.15E-02$)	0.52 (sd $3.43E-02$)	0.78 (sd $2.60E-02$)
Random Forest	0.60 (sd $3.90E-02$)	0.68 (sd $3.90E-02$)	0.61 (sd $3.96E-02$)	0.77 (sd $1.98E-02$)
K-NN (N=5)	0.61 (sd $3.73E-02$)	0.67 (sd $3.06E-02$)	0.62 (sd $3.65E-02$)	0.72 (sd $2.43E-02$)
K-NN (N=10)	0.62 (sd $4.01E-02$)	0.70 (sd $2.72E-02$)	0.63 (sd $3.89E-02$)	0.76 (sd $2.69E-02$)
K-NN (N=20)	0.60 (sd $4.21E-02$)	0.76 (sd $5.20E-02$)	0.62 (sd $4.93E-02$)	0.78 (sd $2.38E-02$)
K-NN (N=50)	0.58 (sd $3.08E-02$)	0.79 (sd $5.20E-02$)	0.60 (sd $4.34E-02$)	0.80 (sd $2.04E-02$)
Gradient Boosting (exponential)	0.59 (sd $4.79E-02$)	0.73 (sd $9.22E-02$)	0.60 (sd $5.34E-02$)	0.80 (sd $1.81E-02$)
Gradient Boosting (deviance)	0.58 (sd $4.22E-02$)	0.73 (sd $7.75E-02$)	0.59 (sd $5.47E-02$)	0.80 (sd $2.48E-02$)
Balanced RF	0.71 (sd $2.02E-02$)	0.62 (sd $1.22E-02$)	0.60 (sd $3.48E-02$)	0.79 (sd $1.67E-02$)
Balanced ADABOOST	0.71 (sd $3.00E-02$)	0.62 (sd $1.94E-02$)	0.59 (sd $5.79E-02$)	0.78 (sd $4.22E-02$)

Table A.44: *Classification metrics for all classifiers evaluated using global network properties in D_{all}. Mainstream training and test data include only Centre sources.*

	Evaluation Metrics for Classifiers in $D_{[0,100]}$			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.50 (sd $0.00E+00$)	0.40 (sd $5.55E-17$)	0.44 (sd $0.00E+00$)	0.61 (sd $8.87E-02$)
SVC RBF	0.61 (sd $9.73E-03$)	0.83 (sd $3.56E-02$)	0.63 (sd $1.40E-02$)	0.62 (sd $2.48E-02$)
Logistic Regression	0.56 (sd $1.62E-02$)	0.83 (sd $4.11E-02$)	0.55 (sd $2.72E-02$)	0.75 (sd $1.80E-02$)
Random Forest	0.62 (sd $2.06E-02$)	0.69 (sd $3.65E-02$)	0.64 (sd $2.47E-02$)	0.75 (sd $2.44E-02$)
K-NN (N=5)	0.62 (sd $1.82E-02$)	0.69 (sd $2.49E-02$)	0.64 (sd $2.10E-02$)	0.71 (sd $2.62E-02$)
K-NN (N=10)	0.63 (sd $2.29E-02$)	0.72 (sd $4.22E-02$)	0.65 (sd $2.77E-02$)	0.73 (sd $2.74E-02$)
K-NN (N=20)	0.61 (sd $1.48E-02$)	0.75 (sd $4.43E-02$)	0.63 (sd $1.99E-02$)	0.75 (sd $1.87E-02$)
K-NN (N=50)	0.60 (sd $1.40E-02$)	0.80 (sd $5.58E-02$)	0.63 (sd $2.00E-02$)	0.78 (sd $2.44E-02$)
Gradient Boosting (exponential)	0.62 (sd $1.25E-02$)	0.79 (sd $3.41E-02$)	0.64 (sd $1.69E-02$)	0.77 (sd $2.46E-02$)
Gradient Boosting (deviance)	0.62 (sd $1.09E-02$)	0.78 (sd $2.54E-02$)	0.64 (sd $1.42E-02$)	0.77 (sd $2.35E-02$)
Balanced RF	0.69 (sd $2.54E-02$)	0.63 (sd $1.64E-02$)	0.63 (sd $1.79E-02$)	0.77 (sd $2.44E-02$)
Balanced ADABOOST	0.69 (sd $2.81E-02$)	0.62 (sd $1.91E-02$)	0.61 (sd $2.42E-02$)	0.75 (sd $2.67E-02$)

Table A.45: *Classification metrics for all classifiers evaluated using global network properties in $D_{[0,100]}$. Mainstream training and test data include only Centre sources.*

Appendix A. Supplementary Information for "A network-based approach to detect online disinformation on Twitter"

	Evaluation Metrics for Classifiers in $D_{[100,1000]}$			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.50 (sd $0.00E+00$)	0.44 (sd $0.00E+00$)	0.47 (sd $5.55E-17$)	0.84 (sd $2.80E-02$)
SVC RBF	0.58 (sd $1.52E-02$)	0.90 (sd $5.15E-02$)	0.60 (sd $2.26E-02$)	0.70 (sd $4.20E-02$)
Logistic Regression	0.66 (sd $5.94E-02$)	0.77 (sd $9.60E-02$)	0.68 (sd $5.18E-02$)	0.85 (sd $2.92E-02$)
Random Forest	0.64 (sd $2.94E-02$)	0.77 (sd $5.69E-02$)	0.67 (sd $3.01E-02$)	0.83 (sd $2.46E-02$)
K-NN (N=5)	0.62 (sd $2.99E-02$)	0.72 (sd $3.22E-02$)	0.64 (sd $3.16E-02$)	0.79 (sd $2.52E-02$)
K-NN (N=10)	0.62 (sd $2.55E-02$)	0.74 (sd $4.37E-02$)	0.65 (sd $3.11E-02$)	0.81 (sd $2.87E-02$)
K-NN (N=20)	0.62 (sd $2.12E-02$)	0.78 (sd $5.86E-02$)	0.65 (sd $2.86E-02$)	0.83 (sd $2.22E-02$)
K-NN (N=50)	0.61 (sd $1.55E-02$)	0.84 (sd $4.65E-02$)	0.65 (sd $2.02E-02$)	0.84 (sd $2.43E-02$)
Gradient Boosting (exponential)	0.62 (sd $1.99E-02$)	0.79 (sd $5.86E-02$)	0.65 (sd $2.37E-02$)	0.85 (sd $2.18E-02$)
Gradient Boosting (deviance)	0.63 (sd $1.96E-02$)	0.78 (sd $6.38E-02$)	0.66 (sd $2.44E-02$)	0.84 (sd $2.06E-02$)
Balanced RF	0.76 (sd $2.75E-02$)	0.63 (sd $1.21E-02$)	0.62 (sd $1.61E-02$)	0.85 (sd $1.90E-02$)
Balanced ADABOOST	0.75 (sd $3.54E-02$)	0.63 (sd $1.77E-02$)	0.64 (sd $2.28E-02$)	0.84 (sd $2.54E-02$)

Table A.46: *Classification metrics for all classifiers evaluated using global network properties in $D_{[100,1000]}$. Mainstream training and test data include only Centre sources.*

	Evaluation Metrics for Classifiers in $D_{[1000,+\infty]}$			
Classifier	Recall	Precision	F1-Score	AUROC
SVC Linear	0.50 (sd $0.00E+00$)	0.46 (sd $0.00E+00$)	0.48 (sd $5.55E-17$)	0.87 (sd $1.64E-01$)
SVC RBF	0.50 (sd $6.25E-03$)	0.46 (sd $4.62E-04$)	0.48 (sd $3.18E-03$)	0.81 (sd $1.74E-01$)
Logistic Regression	0.61 (sd $1.71E-01$)	0.62 (sd $2.12E-01$)	0.61 (sd $1.84E-01$)	0.85 (sd $1.90E-01$)
Random Forest	0.69 (sd $1.48E-01$)	0.73 (sd $2.03E-01$)	0.70 (sd $1.64E-01$)	0.84 (sd $1.61E-01$)
K-NN (N=5)	0.59 (sd $1.20E-01$)	0.64 (sd $2.32E-01$)	0.60 (sd $1.56E-01$)	0.77 (sd $1.87E-01$)
K-NN (N=10)	0.50 (sd $6.25E-03$)	0.46 (sd $4.62E-04$)	0.48 (sd $3.18E-03$)	0.82 (sd $1.49E-01$)
K-NN (N=20)	0.50 (sd $0.00E+00$)	0.46 (sd $0.00E+00$)	0.48 (sd $5.55E-17$)	0.80 (sd $1.73E-01$)
K-NN (N=50)	0.50 (sd $0.00E+00$)	0.46 (sd $0.00E+00$)	0.48 (sd $5.55E-17$)	0.80 (sd $2.09E-01$)
Gradient Boosting (exponential)	0.66 (sd $1.51E-01$)	0.65 (sd $1.74E-01$)	0.65 (sd $1.49E-01$)	0.85 (sd $1.34E-01$)
Gradient Boosting (deviance)	0.65 (sd $1.44E-01$)	0.60 (sd $1.22E-01$)	0.62 (sd $1.23E-01$)	0.82 (sd $1.66E-01$)
Balanced RF	0.78 (sd $1.07E-01$)	0.64 (sd $1.25E-01$)	0.61 (sd $1.17E-01$)	0.85 (sd $1.68E-01$)
Balanced ADABOOST	0.72 (sd $1.66E-01$)	0.64 (sd $1.43E-01$)	0.65 (sd $1.21E-01$)	0.83 (sd $1.42E-01$)

Table A.47: *Classification metrics for all classifiers evaluated using global network properties in $D_{[1000,+\infty]}$. Mainstream training and test data include only Centre sources.*

	Evaluation Metrics for Classifiers in D_{all}			
Classifier	Recall	Precision	F1-Score	AUROC
Balanced RF	0.69 (sd $3.02E-02$)	0.72 (sd $3.02E-02$)	0.69 (sd $3.35E-02$)	0.76 (sd $3.04E-02$)
Balanced ADABOOST	0.67 (sd $3.95E-02$)	0.70 (sd $2.46E-02$)	0.67 (sd $4.66E-02$)	0.75 (sd $3.66E-02$)
SVC Linear	0.59 (sd $4.74E-02$)	0.77 (sd $1.12E-02$)	0.52 (sd $7.95E-02$)	0.74 (sd $2.61E-02$)
SVC RBF	0.58 (sd $2.94E-02$)	0.77 (sd $1.41E-02$)	0.52 (sd $4.98E-02$)	0.70 (sd $2.88E-02$)
Logistic Regression	0.58 (sd $2.02E-02$)	0.77 (sd $1.13E-02$)	0.51 (sd $3.52E-02$)	0.77 (sd $2.14E-02$)
Random Forest	0.61 (sd $4.14E-02$)	0.76 (sd $2.16E-02$)	0.56 (sd $6.44E-02$)	0.74 (sd $2.98E-02$)
K-NN (N=5)	0.62 (sd $3.51E-02$)	0.74 (sd $2.46E-02$)	0.58 (sd $5.33E-02$)	0.71 (sd $3.06E-02$)
K-NN (N=10)	0.61 (sd $3.38E-02$)	0.77 (sd $1.44E-02$)	0.55 (sd $5.55E-02$)	0.75 (sd $2.36E-02$)
K-NN (N=20)	0.61 (sd $3.66E-02$)	0.77 (sd $1.25E-02$)	0.56 (sd $5.93E-02$)	0.77 (sd $2.03E-02$)
K-NN (N=50)	0.60 (sd $3.06E-02$)	0.76 (sd $1.24E-02$)	0.55 (sd $5.03E-02$)	0.79 (sd $1.24E-02$)
Gradient Boosting (exponential)	0.60 (sd $4.33E-02$)	0.77 (sd $9.61E-03$)	0.54 (sd $7.15E-02$)	0.75 (sd $3.90E-02$)
Gradient Boosting (deviance)	0.60 (sd $4.18E-02$)	0.76 (sd $1.56E-02$)	0.55 (sd $6.69E-02$)	0.75 (sd $4.17E-02$)

Table A.48: *Classification metrics for classifiers evaluated using global network properties in D_{all}. Training on left-biased Mainstream and Misleading sources only and testing on all sources regardless of bias.*

	Evaluation Metrics for Classifiers in $D_{[0,100]}$			
Classifier	Recall	Precision	F1-Score	AUROC
Balanced RF	0.70 (sd $1.79E-02$)	0.69 (sd $1.71E-02$)	0.69 (sd $1.74E-02$)	0.78 (sd $1.38E-02$)
Balanced ADABOOST	0.67 (sd $5.34E-03$)	0.67 (sd $6.12E-03$)	0.67 (sd $5.69E-03$)	0.76 (sd $1.11E-02$)
SVC Linear	0.59 (sd $6.94E-03$)	0.78 (sd $9.89E-03$)	0.56 (sd $1.06E-02$)	0.64 (sd $4.32E-02$)
SVC RBF	0.52 (sd $9.81E-03$)	0.72 (sd $4.44E-02$)	0.42 (sd $1.99E-02$)	0.68 (sd $2.30E-02$)
Logistic Regression	0.56 (sd $1.07E-02$)	0.76 (sd $2.60E-02$)	0.50 (sd $1.78E-02$)	0.74 (sd $1.80E-02$)
Random Forest	0.61 (sd $1.30E-02$)	0.76 (sd $2.39E-02$)	0.59 (sd $1.85E-02$)	0.77 (sd $1.49E-02$)
K-NN (N=5)	0.61 (sd $1.32E-02$)	0.74 (sd $1.21E-02$)	0.59 (sd $2.00E-02$)	0.73 (sd $1.16E-02$)
K-NN (N=10)	0.59 (sd $1.10E-02$)	0.77 (sd $1.72E-02$)	0.56 (sd $1.75E-02$)	0.74 (sd $1.30E-02$)
K-NN (N=20)	0.60 (sd $9.89E-03$)	0.78 (sd $1.59E-02$)	0.56 (sd $1.51E-02$)	0.76 (sd $1.58E-02$)
K-NN (N=50)	0.59 (sd $8.52E-03$)	0.79 (sd $1.08E-02$)	0.55 (sd $1.38E-02$)	0.76 (sd $1.46E-02$)
Gradient Boosting (exponential)	0.59 (sd $7.79E-03$)	0.78 (sd $1.82E-02$)	0.56 (sd $1.15E-02$)	0.78 (sd $1.58E-02$)
Gradient Boosting (deviance)	0.60 (sd $1.04E-02$)	0.78 (sd $1.70E-02$)	0.57 (sd $1.49E-02$)	0.78 (sd $1.40E-02$)

Table A.49: *Classification metrics for classifiers evaluated using global network properties in $D_{[0,100]}$. Training on left-biased Mainstream and Misleading sources only and testing on all sources regardless of bias.*

	Evaluation Metrics for Classifiers in $D_{[100,1000]}$			
Classifier	Recall	Precision	F1-Score	AUROC
Balanced RF	0.75 (sd $3.89E-02$)	0.76 (sd $3.24E-02$)	0.75 (sd $4.33E-02$)	0.84 (sd $3.10E-02$)
Balanced ADABOOST	0.72 (sd $2.36E-02$)	0.73 (sd $1.86E-02$)	0.71 (sd $2.72E-02$)	0.81 (sd $3.13E-02$)
SVC Linear	0.68 (sd $5.79E-02$)	0.75 (sd $1.64E-02$)	0.64 (sd $9.06E-02$)	0.84 (sd $1.14E-02$)
SVC RBF	0.65 (sd $3.81E-02$)	0.76 (sd $1.05E-02$)	0.59 (sd $5.97E-02$)	0.75 (sd $1.82E-02$)
Logistic Regression	0.66 (sd $3.19E-02$)	0.77 (sd $1.33E-02$)	0.61 (sd $4.67E-02$)	0.85 (sd $1.86E-02$)
Random Forest	0.69 (sd $5.14E-02$)	0.77 (sd $2.76E-02$)	0.65 (sd $7.42E-02$)	0.81 (sd $4.31E-02$)
K-NN (N=5)	0.69 (sd $4.46E-02$)	0.76 (sd $2.54E-02$)	0.66 (sd $6.11E-02$)	0.80 (sd $2.68E-02$)
K-NN (N=10)	0.67 (sd $4.12E-02$)	0.77 (sd $1.43E-02$)	0.62 (sd $6.31E-02$)	0.82 (sd $2.89E-02$)
K-NN (N=20)	0.66 (sd $4.33E-02$)	0.76 (sd $1.70E-02$)	0.60 (sd $6.65E-02$)	0.83 (sd $2.14E-02$)
K-NN (N=50)	0.64 (sd $4.16E-02$)	0.76 (sd $1.11E-02$)	0.57 (sd $6.82E-02$)	0.84 (sd $1.74E-02$)
Gradient Boosting (exponential)	0.68 (sd $5.11E-02$)	0.77 (sd $2.24E-02$)	0.64 (sd $7.52E-02$)	0.82 (sd $4.09E-02$)
Gradient Boosting (deviance)	0.68 (sd $5.00E-02$)	0.77 (sd $2.01E-02$)	0.64 (sd $7.39E-02$)	0.82 (sd $4.17E-02$)

Table A.50: *Classification metrics for classifiers evaluated using global network properties in $D_{[100,1000]}$. Training on left-biased Mainstream and Misleading sources only and testing on all sources regardless of bias.*

	Evaluation Metrics for Classifiers in $D_{[1000,+\infty]}$			
Classifier	Recall	Precision	F1-Score	AUROC
Balanced RF	0.79 (sd $5.84E-02$)	0.79 (sd $6.02E-02$)	0.78 (sd $5.80E-02$)	0.90 (sd $3.88E-02$)
Balanced ADABOOST	0.79 (sd $4.17E-02$)	0.79 (sd $4.33E-02$)	0.78 (sd $4.21E-02$)	0.87 (sd $3.49E-02$)
SVC Linear	0.54 (sd $1.46E-02$)	0.74 (sd $3.96E-03$)	0.39 (sd $2.86E-02$)	0.93 (sd $3.38E-02$)
SVC RBF	0.55 (sd $2.78E-02$)	0.68 (sd $1.56E-01$)	0.42 (sd $4.62E-02$)	0.75 (sd $5.65E-02$)
Logistic Regression	0.68 (sd $5.04E-02$)	0.77 (sd $3.92E-02$)	0.63 (sd $6.61E-02$)	0.91 (sd $3.81E-02$)
Random Forest	0.72 (sd $4.84E-02$)	0.78 (sd $3.59E-02$)	0.69 (sd $5.81E-02$)	0.90 (sd $4.62E-02$)
K-NN (N=5)	0.67 (sd $5.12E-02$)	0.74 (sd $5.01E-02$)	0.63 (sd $6.20E-02$)	0.81 (sd $4.45E-02$)
K-NN (N=10)	0.59 (sd $4.47E-02$)	0.73 (sd $5.33E-02$)	0.49 (sd $6.80E-02$)	0.81 (sd $5.01E-02$)
K-NN (N=20)	0.55 (sd $1.91E-02$)	0.74 (sd $5.29E-03$)	0.41 (sd $3.66E-02$)	0.79 (sd $5.97E-02$)
K-NN (N=50)	0.51 (sd $9.55E-03$)	0.38 (sd $2.32E-01$)	0.33 (sd $2.06E-02$)	0.76 (sd $7.36E-02$)
Gradient Boosting (exponential)	0.75 (sd $5.63E-02$)	0.78 (sd $4.43E-02$)	0.74 (sd $6.75E-02$)	0.87 (sd $5.12E-02$)
Gradient Boosting (deviance)	0.75 (sd $5.20E-02$)	0.77 (sd $4.73E-02$)	0.74 (sd $5.88E-02$)	0.87 (sd $5.27E-02$)

Table A.51: *Classification metrics for classifiers evaluated using global network properties in $D_{[1000,+\infty]}$. Training on left-biased Mainstream and Misleading sources only and testing on all sources regardless of bias.*

Appendix A. Supplementary Information for "A network-based approach to detect online disinformation on Twitter"

	Evaluation Metrics for Classifiers in D_{all}			
Classifier	Recall	Precision	F1-Score	AUROC
Balanced RF	0.70 (sd $2.05E-02$)	0.70 (sd $2.26E-02$)	0.70 (sd $2.21E-02$)	0.78 (sd $1.79E-02$)
Balanced ADABOOST	0.70 (sd $2.52E-02$)	0.70 (sd $2.32E-02$)	0.69 (sd $2.69E-02$)	0.77 (sd $2.30E-02$)
SVC Linear	0.50 (sd $0.00E+00$)	0.23 (sd $2.78E-17$)	0.31 (sd $5.55E-17$)	0.76 (sd $2.64E-02$)
SVC RBF	0.53 (sd $2.41E-02$)	0.72 (sd $2.27E-02$)	0.38 (sd $4.93E-02$)	0.60 (sd $6.95E-02$)
Logistic Regression	0.52 (sd $1.24E-02$)	0.71 (sd $1.75E-02$)	0.36 (sd $2.75E-02$)	0.77 (sd $1.18E-02$)
Random Forest	0.58 (sd $4.06E-02$)	0.66 (sd $2.37E-02$)	0.50 (sd $7.62E-02$)	0.76 (sd $1.73E-02$)
K-NN (N=5)	0.59 (sd $3.85E-02$)	0.65 (sd $3.13E-02$)	0.52 (sd $5.95E-02$)	0.71 (sd $2.91E-02$)
K-NN (N=10)	0.59 (sd $4.19E-02$)	0.66 (sd $3.56E-02$)	0.51 (sd $6.80E-02$)	0.74 (sd $2.56E-02$)
K-NN (N=20)	0.56 (sd $4.30E-02$)	0.69 (sd $2.29E-02$)	0.45 (sd $8.33E-02$)	0.77 (sd $2.01E-02$)
K-NN (N=50)	0.55 (sd $3.86E-02$)	0.70 (sd $1.88E-02$)	0.42 (sd $7.87E-02$)	0.78 (sd $1.44E-02$)
Gradient Boosting (exponential)	0.55 (sd $5.00E-02$)	0.70 (sd $2.47E-02$)	0.43 (sd $1.05E-01$)	0.78 (sd $1.29E-02$)
Gradient Boosting (deviance)	0.56 (sd $5.15E-02$)	0.71 (sd $2.28E-02$)	0.43 (sd $1.06E-01$)	0.78 (sd $1.31E-02$)

Table A.52: *Classification metrics for classifiers evaluated using global network properties in D_{all}. Training on right-biased Mainstream and Misleading sources only and testing on all sources regardless of bias.*

	Evaluation Metrics for Classifiers in $D_{[0,100]}$			
Classifier	Recall	Precision	F1-Score	AUROC
Balanced RF	0.69 (sd $1.29E-02$)	0.68 (sd $1.22E-02$)	0.68 (sd $1.21E-02$)	0.78 (sd $1.52E-02$)
Balanced ADABOOST	0.68 (sd $1.98E-02$)	0.67 (sd $1.89E-02$)	0.67 (sd $1.92E-02$)	0.75 (sd $1.99E-02$)
SVC Linear	0.50 (sd $0.00E+00$)	0.19 (sd $0.00E+00$)	0.28 (sd $5.55E-17$)	0.70 (sd $2.08E-02$)
SVC RBF	0.60 (sd $1.01E-02$)	0.69 (sd $1.36E-02$)	0.49 (sd $1.46E-02$)	0.67 (sd $1.39E-02$)
Logistic Regression	0.59 (sd $8.37E-03$)	0.68 (sd $8.85E-03$)	0.46 (sd $1.63E-02$)	0.74 (sd $1.75E-02$)
Random Forest	0.64 (sd $1.59E-02$)	0.67 (sd $1.76E-02$)	0.57 (sd $1.89E-02$)	0.76 (sd $1.69E-02$)
K-NN (N=5)	0.64 (sd $1.30E-02$)	0.67 (sd $1.47E-02$)	0.57 (sd $1.40E-02$)	0.74 (sd $1.70E-02$)
K-NN (N=10)	0.65 (sd $1.11E-02$)	0.67 (sd $1.19E-02$)	0.59 (sd $1.20E-02$)	0.76 (sd $1.12E-02$)
K-NN (N=20)	0.63 (sd $1.13E-02$)	0.67 (sd $1.50E-02$)	0.56 (sd $1.14E-02$)	0.76 (sd $1.51E-02$)
K-NN (N=50)	0.61 (sd $8.07E-03$)	0.69 (sd $1.07E-02$)	0.51 (sd $1.12E-02$)	0.76 (sd $1.53E-02$)
Gradient Boosting (exponential)	0.61 (sd $1.12E-02$)	0.69 (sd $1.30E-02$)	0.50 (sd $1.58E-02$)	0.76 (sd $1.82E-02$)
Gradient Boosting (deviance)	0.61 (sd $1.27E-02$)	0.69 (sd $1.44E-02$)	0.50 (sd $1.79E-02$)	0.76 (sd $1.66E-02$)

Table A.53: *Classification metrics for classifiers evaluated using global network properties in $D_{[0,100]}$. Training on right-biased Mainstream and Misleading sources only and testing on all sources regardless of bias.*

	Evaluation Metrics for Classifiers in $D_{[100,1000]}$			
Classifier	Recall	Precision	F1-Score	AUROC
Balanced RF	0.76 (sd $1.17E-02$)	0.76 (sd $1.19E-02$)	0.76 (sd $1.18E-02$)	0.85 (sd $1.16E-02$)
Balanced ADABOOST	0.73 (sd $2.44E-02$)	0.74 (sd $1.96E-02$)	0.73 (sd $2.56E-02$)	0.83 (sd $1.34E-02$)
SVC Linear	0.52 (sd $6.15E-03$)	0.74 (sd $4.43E-02$)	0.38 (sd $1.19E-02$)	0.85 (sd $1.72E-02$)
SVC RBF	0.57 (sd $8.55E-03$)	0.77 (sd $1.31E-02$)	0.49 (sd $1.50E-02$)	0.71 (sd $3.12E-02$)
Logistic Regression	0.59 (sd $4.77E-02$)	0.76 (sd $2.35E-02$)	0.52 (sd $8.04E-02$)	0.84 (sd $1.07E-02$)
Random Forest	0.63 (sd $1.94E-02$)	0.76 (sd $1.61E-02$)	0.59 (sd $2.96E-02$)	0.84 (sd $1.65E-02$)
K-NN (N=5)	0.64 (sd $1.80E-02$)	0.74 (sd $2.00E-02$)	0.61 (sd $2.46E-02$)	0.79 (sd $1.30E-02$)
K-NN (N=10)	0.64 (sd $9.29E-03$)	0.76 (sd $1.59E-02$)	0.61 (sd $1.18E-02$)	0.82 (sd $1.58E-02$)
K-NN (N=20)	0.62 (sd $6.56E-03$)	0.76 (sd $1.01E-02$)	0.57 (sd $1.02E-02$)	0.84 (sd $1.09E-02$)
K-NN (N=50)	0.61 (sd $1.03E-02$)	0.76 (sd $8.99E-03$)	0.55 (sd $1.74E-02$)	0.84 (sd $1.17E-02$)
Gradient Boosting (exponential)	0.61 (sd $1.80E-02$)	0.76 (sd $1.34E-02$)	0.55 (sd $2.86E-02$)	0.84 (sd $1.00E-02$)
Gradient Boosting (deviance)	0.61 (sd $1.55E-02$)	0.75 (sd $1.71E-02$)	0.55 (sd $2.44E-02$)	0.84 (sd $1.04E-02$)

Table A.54: *Classification metrics for classifiers evaluated using global network properties in $D_{[100,1000]}$. Training on right-biased Mainstream and Misleading sources only and testing on all sources regardless of bias.*

	Evaluation Metrics for Classifiers in $D_{[1000,+\infty]}$			
Classifier	Recall	Precision	F1-Score	AUROC
Balanced RF	0.77 (sd $1.15E-01$)	0.80 (sd $1.03E-01$)	0.76 (sd $1.34E-01$)	0.89 (sd $6.53E-02$)
Balanced ADABOOST	0.72 (sd $9.75E-02$)	0.82 (sd $5.17E-02$)	0.71 (sd $1.26E-01$)	0.87 (sd $8.51E-02$)
SVC Linear	0.50 (sd $0.00E+00$)	0.27 (sd $5.55E-17$)	0.35 (sd $0.00E+00$)	0.92 (sd $3.29E-02$)
SVC RBF	0.53 (sd $3.17E-02$)	0.58 (sd $2.52E-01$)	0.42 (sd $6.14E-02$)	0.80 (sd $7.39E-02$)
Logistic Regression	0.55 (sd $2.45E-02$)	0.73 (sd $1.54E-01$)	0.44 (sd $4.72E-02$)	0.91 (sd $3.65E-02$)
Random Forest	0.62 (sd $8.86E-02$)	0.76 (sd $1.64E-01$)	0.56 (sd $1.40E-01$)	0.85 (sd $9.94E-02$)
K-NN (N=5)	0.59 (sd $4.64E-02$)	0.80 (sd $1.40E-02$)	0.52 (sd $7.61E-02$)	0.78 (sd $5.31E-02$)
K-NN (N=10)	0.56 (sd $2.97E-02$)	0.79 (sd $8.32E-03$)	0.47 (sd $5.34E-02$)	0.81 (sd $4.54E-02$)
K-NN (N=20)	0.52 (sd $1.50E-02$)	0.63 (sd $2.32E-01$)	0.39 (sd $3.09E-02$)	0.82 (sd $6.35E-02$)
K-NN (N=50)	0.50 (sd $0.00E+00$)	0.27 (sd $5.55E-17$)	0.35 (sd $0.00E+00$)	0.85 (sd $5.01E-02$)
Gradient Boosting (exponential)	0.62 (sd $8.88E-02$)	0.70 (sd $2.14E-01$)	0.55 (sd $1.41E-01$)	0.84 (sd $1.14E-01$)
Gradient Boosting (deviance)	0.63 (sd $1.01E-01$)	0.70 (sd $2.16E-01$)	0.58 (sd $1.57E-01$)	0.83 (sd $1.20E-01$)

Table A.55: *Classification metrics for classifiers evaluated using global network properties in $D_{[1000,+\infty]}$. Training on right-biased Mainstream and Misleading sources only and testing on all sources regardless of bias.*

Appendix A. Supplementary Information for "A network-based approach to detect online disinformation on Twitter"

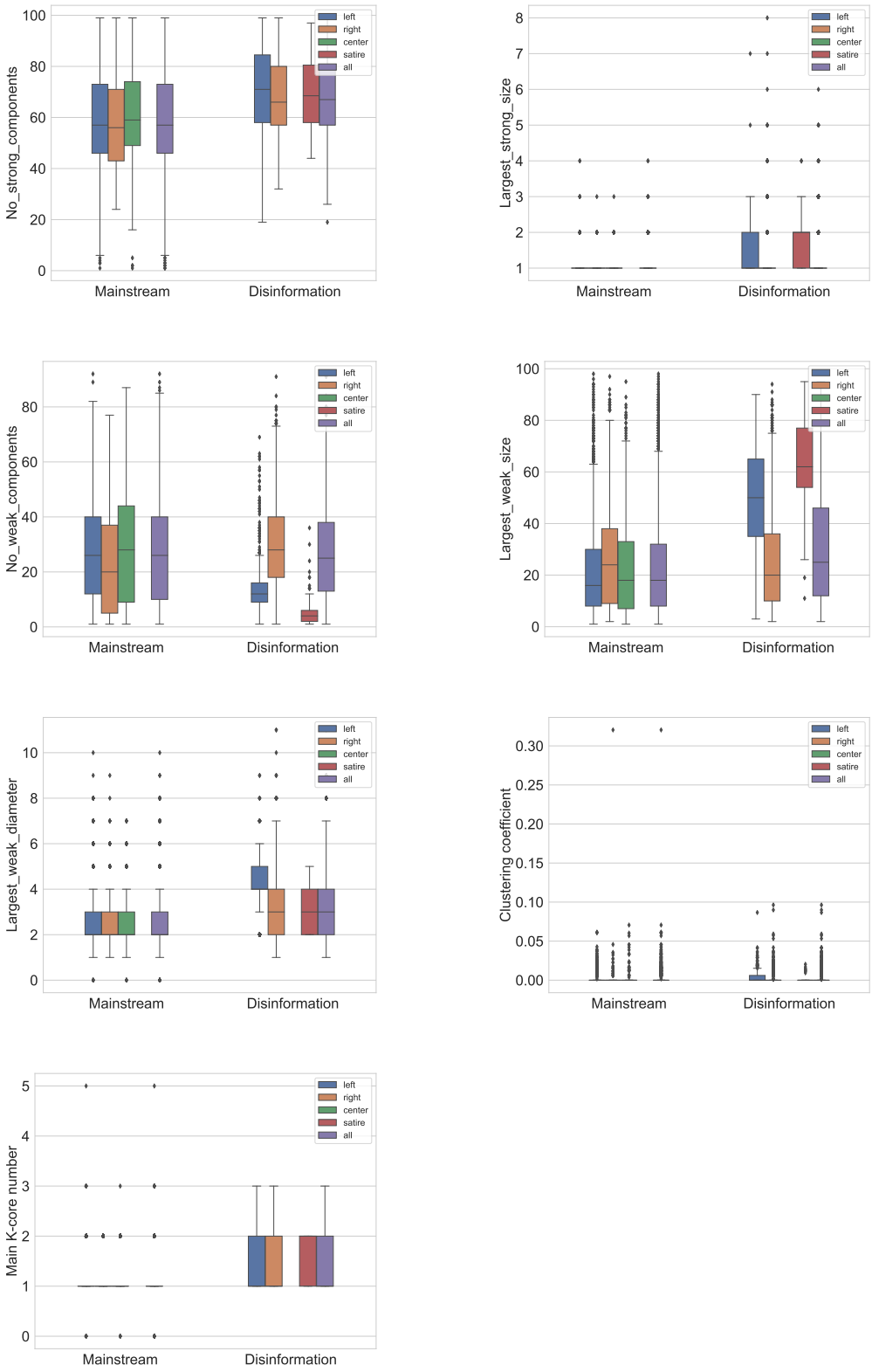


Figure A.11: *Box plots for all global network properties in* $D_{[0,100)}$

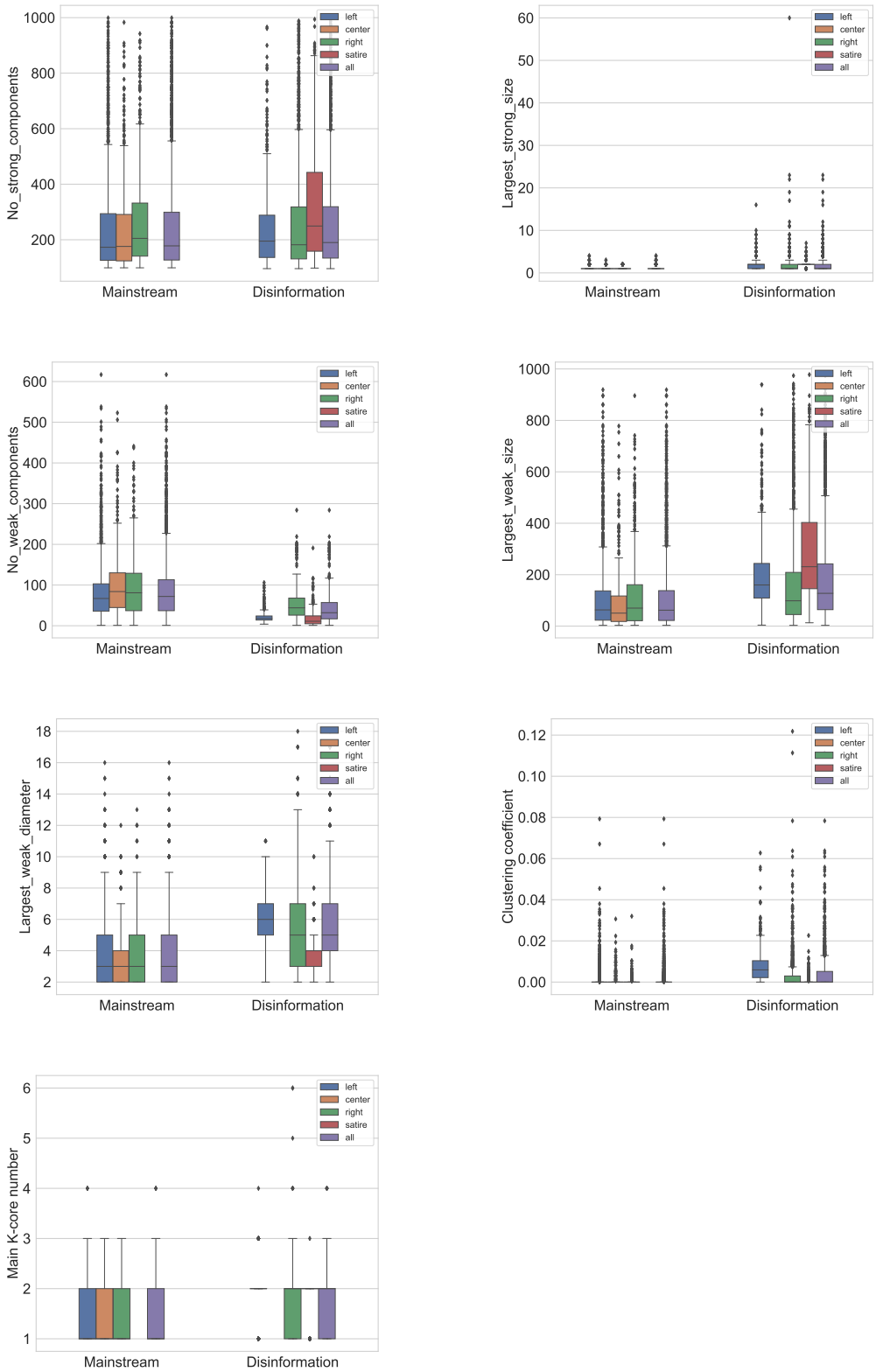


Figure A.12: *Box plots for all global network properties in* $D_{[100,1000)}$

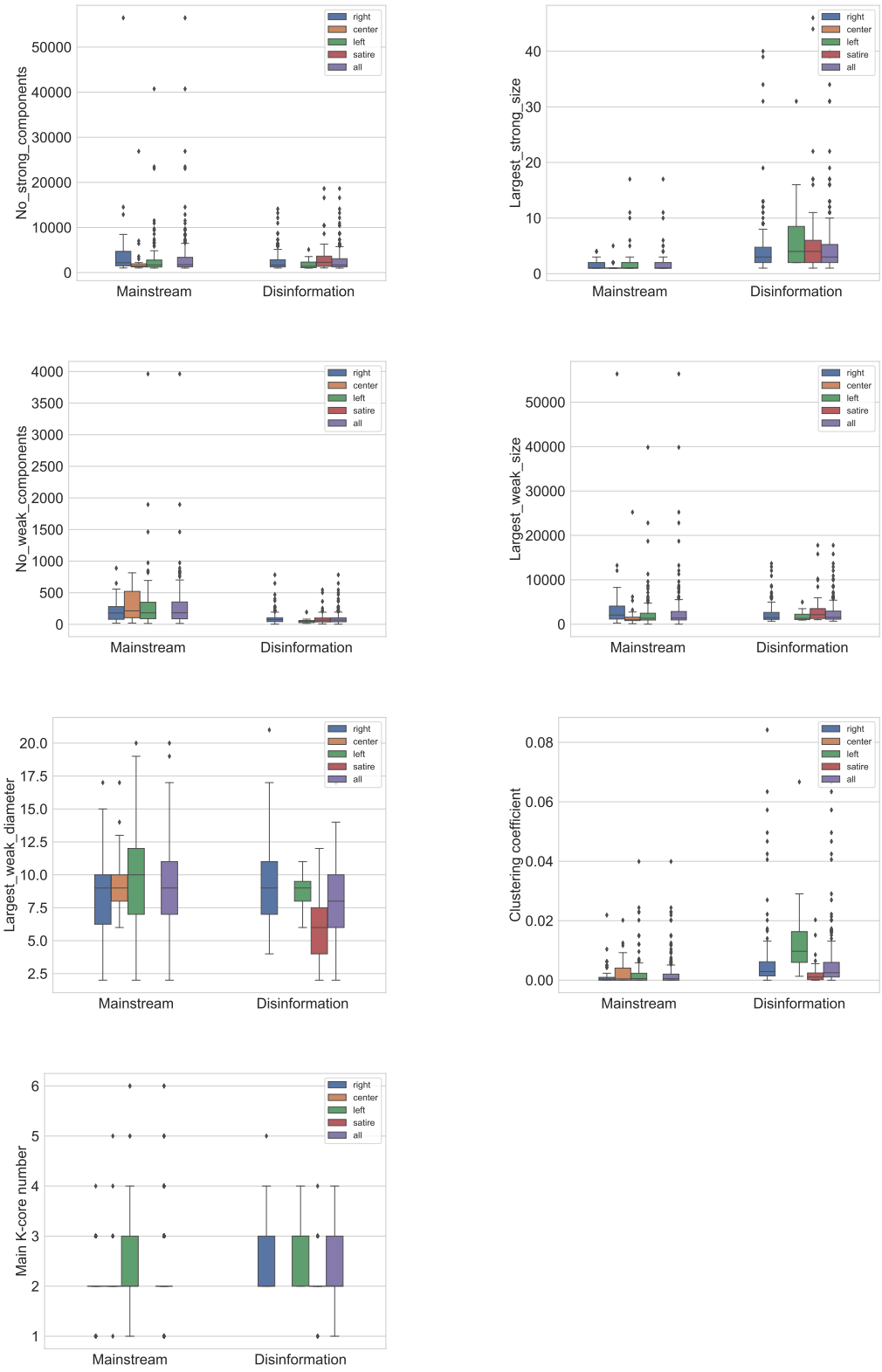


Figure A.13: *Box plots for all global network properties in* $D_{[1000,+\infty)}$

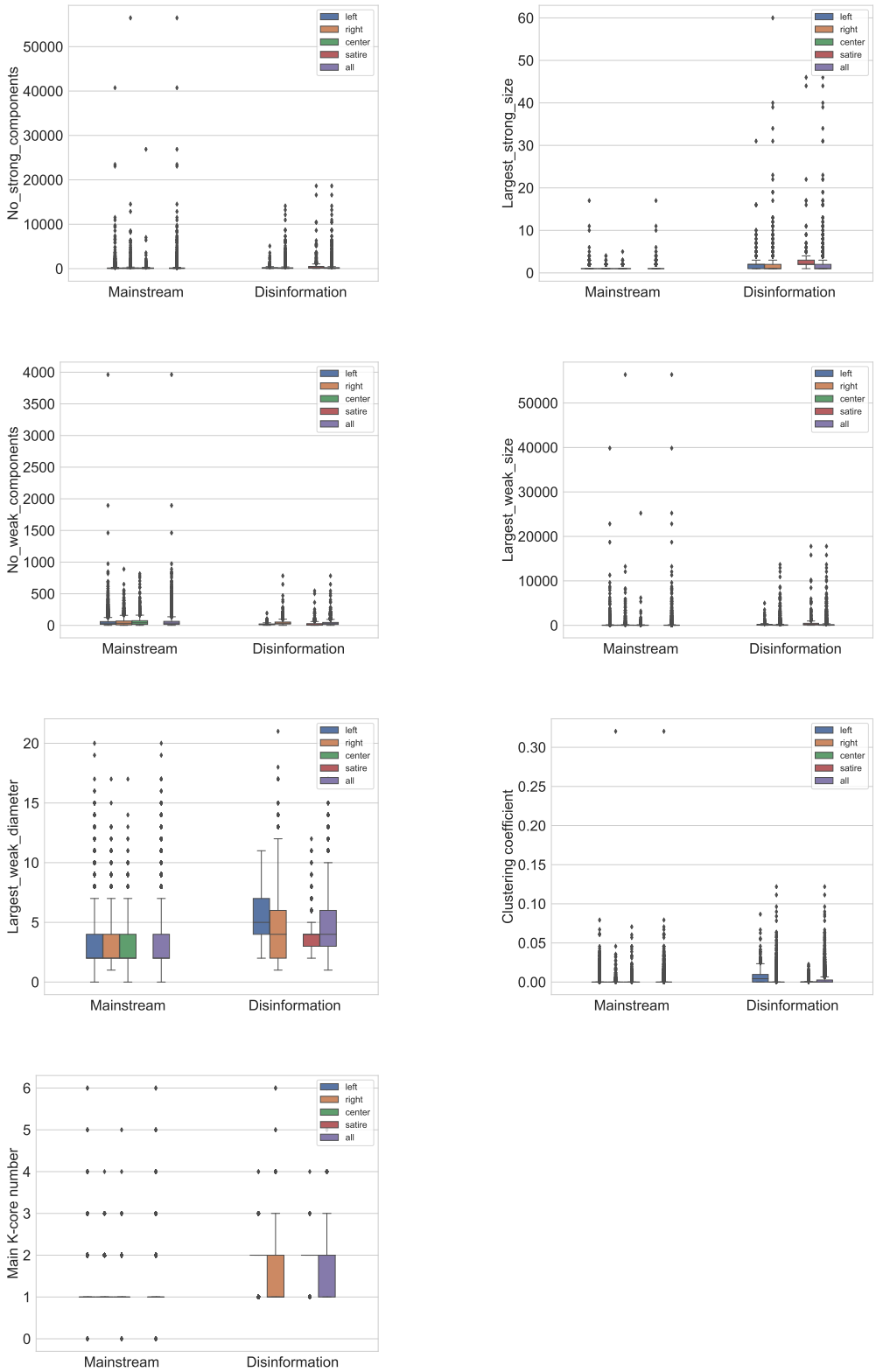
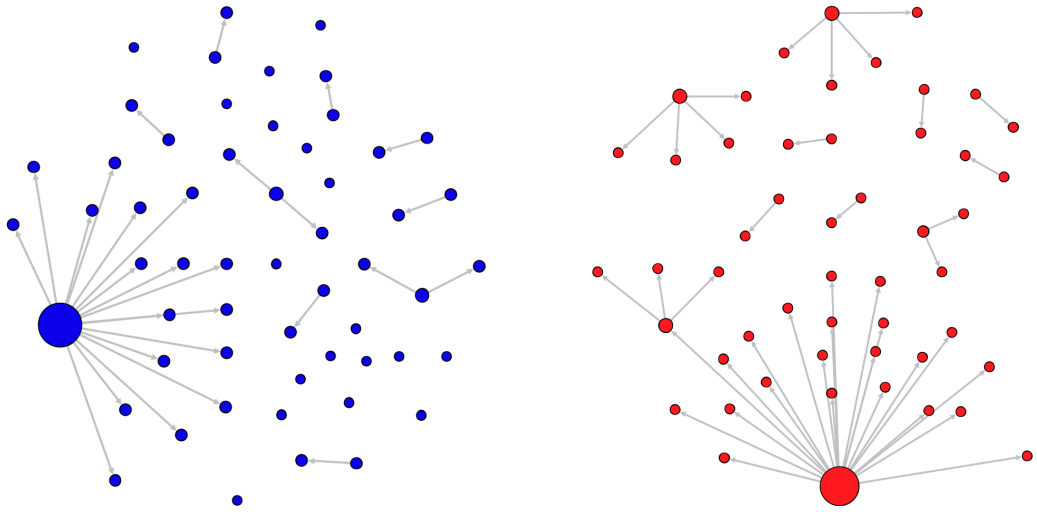
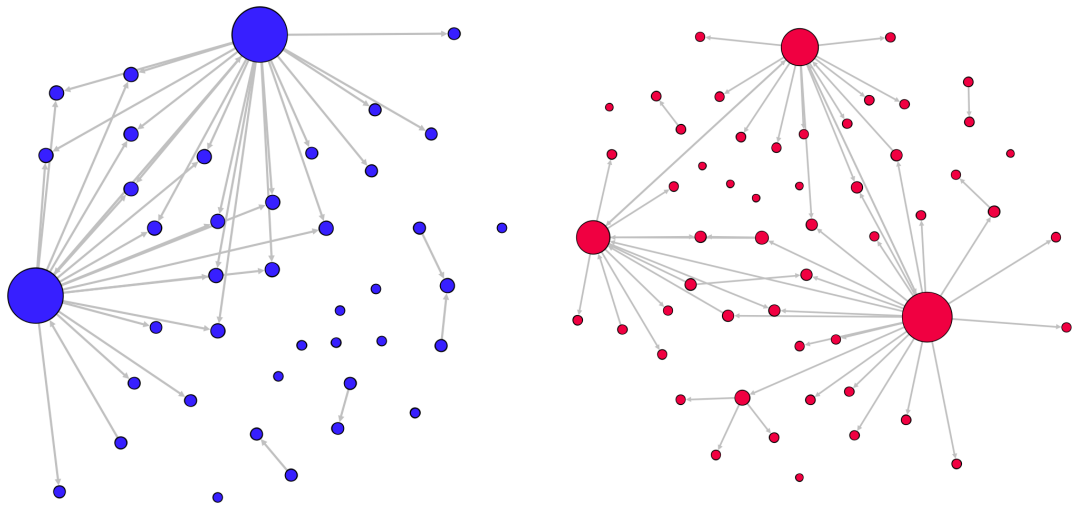


Figure A.14: *Box plots for all global network properties in* D_{all}



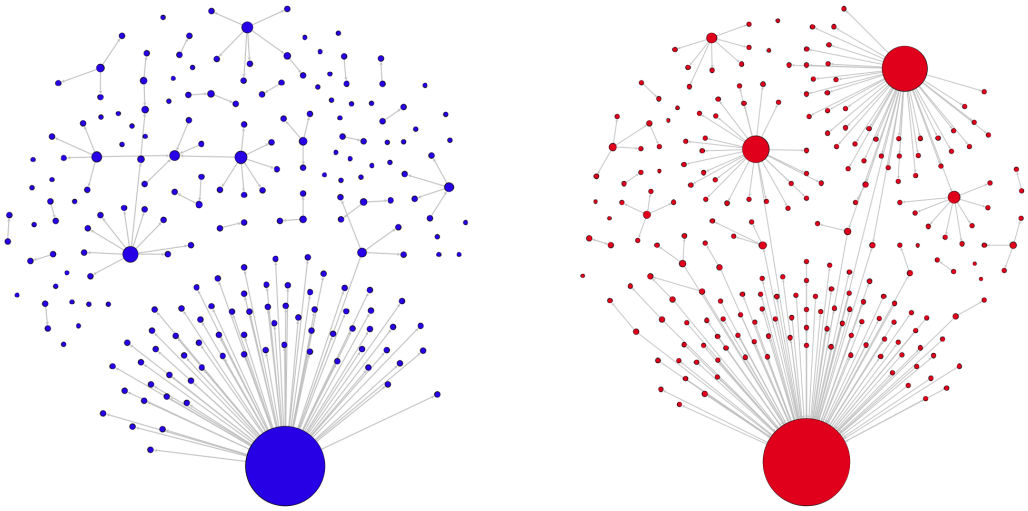
a) Mainstream diffusion network b) Misleading diffusion network

Figure A.15: *(bottom) The nearest diffusion networks in both news domains belonging to* $D_{[0,100)}$*. The misleading network has a larger size and diameter of the largest weakly connected component.*



a) Mainstream diffusion network b) Misleading diffusion network

Figure A.16: *(bottom) The farthest diffusion networks in both news domains belonging to* $D_{[0,100)}$*. The misleading network has a larger size and diameter of the largest weakly connected component.*



a) Mainstream diffusion network

b) Misleading diffusion network

Figure A.17: *(bottom) The farthest diffusion networks in both news domains belonging to* $D_{[100,1000)}$*. The misleading network has a larger diameter and size of the largest weakly connected component, and a smaller number of weakly connected components.*

APPENDIX B

Supplementary Information for "The impact of vaccine-related disinformation"

In this chapter we provide supplementary information for Chapter 5.

B.1 Data collection and sources

B.1.1 Twitter data

In our CoVaxxy [**?**] project, we collected around 55 M English-language posts about vaccines on Twitter by means of the Twitter POST statuses/filter v1.1 API, in the period from January 4th, 2021 to March 25th, 2021. Data collection and analysis was done using the Extreme Science and Engineering Discovery Environment (XSEDE) [245].

To define as complete a set as possible of English language keywords related to vaccines, we employed a snowball sampling methodology in December 20201 (see reference for full details on the data collection pipeline). The final list contains almost 80 keywords, and it is accessible in the online repository associated with the reference [191]. As a robustness test, we further perform sensitivity analyses using a restricted set of keywords ("vaccine", "vaccinate", "vaccination", "vax") which covers almost 95% of the total number of geolocated tweets. Results are equivalent to those presented in the main text and are described in the section "Sensitivity Analyses".

To match Twitter posts with US states and counties, we first identified a collection

of Twitter accounts that disclosed a location in their Twitter profile. We then employed the carmen Python library [68] to match each location to US states and counties. We were able to match around 1.67 M users to 50 US states, and a subset of 1.15 M users to over 1,300 US counties; the larger set accounts for a total number of almost 11 M shared tweets.

To analyze the spread of low-credibility information, we identified all URLs shared in Twitter posts that originated from a list of low-credibility sources, following a large corpus of literature [**?**, **?**, **?**, **?**, 270]. We employ the Iffy+ Misinfo/Disinfo list of low-credibility sources [93], which is based on information provided by the Media Bias/Fact Check website (MBFC, `https://mediabiasfactcheck.com`), an independent organization that reviews and rates the reliability of news sources. As defined in the related methodology, political leaning is not a factor for inclusion. The list includes sites labeled by MBFC as having a "Very Low" or "Low" factual-reporting level as well as those classified as "Questionable" or "Conspiracy-Pseudoscience". The list also includes fake-news websites flagged by BuzzFeed, FactCheck.org, PolitiFact, and Wikipedia, for a total number of 674 low-credibility sources.

Based on this list, we measure the prevalence of low-credibility information about vaccines in each region by (1) calculating the proportion of vaccine-related tweets containing URLs pointing to a low-credibility news website, for each geo-located account; and (2) taking the average of this proportion across all accounts within a specific region. We refer to this average as the state-wide (county-wide) prevalence of misinformation.

At the county level, we omit observations without vaccine hesitancy data (see next section), and we used different thresholds for the minimum number of geolocated accounts, respectively 10, 50, and 100. In the main paper, we present results when using 100 as a threshold. We provide sensitivity analyses using versions including counties with at least 10 and 50 Twitter accounts (see "Sensitivity Analyses" section). The larger threshold is likely to contain less error but also omits more counties.

B.1.2 Election data

We use data provided by the MIT Election Lab to extract state-level returns for the 2020 US presidential election [158]. For counties, we use data provided by Fox News, Politico, and the New York Times. They are publicly available at `https://github.com/tonmcg/US_County_Level_Election_Results_08-20`.

B.1.3 Vaccine hesitancy data

To compute vaccine hesitancy rates in each state (county), we leverage daily COVID-19 Symptom Surveys produced by the Delphi Group at Carnegie Mellon University [76]. These surveys are voluntarily answered by a random sample of users on Facebook

(total reported sample size $N = 22,128,855$). Within the Vaccination Indicators of the survey, we extract the estimated percentage of respondents (for each state/county) "who either have already received a COVID vaccine or would definitely or probably choose to get vaccinated, if a vaccine were offered to them today." Results are available daily, for all 50 US states and for 764 US counties. We compute state-wide (county-wide) vaccine hesitancy rates by taking the proportion of negative responses in the period from January 4th to March 25th.

B.1.4 Vaccine uptake data

Vaccination uptake statistics are derived from the Centers for Disease Control and Prevention (CDC) dataset (`https://covid.cdc.gov/covid-data-tracker/#vaccinations`). Doses monitored for each state include those administered in jurisdictional partner clinics, retail pharmacies, long-term care facilities, Federal Emergency Management Agency partner sites, Health Resources and Services Administration partner sites, and federal facilities. The data have been compiled on a daily basis by ourworldindata.org, and we have downloaded them for the period from January 12 to March 25, 2021. The data are available at `https://github.com/owid/covid-19-data/tree/master/public/data/vaccinations`.

B.1.5 COVID-19 data

We extracted the number of COVID-19 cases and fatalities at the state and county level based on reports made by USAFacts (`https://usafacts.org`). In particular, we summed the number of daily confirmed COVID-19 cases and fatalities, referring to these as "recent", in the period from January 4 to March 25, 2021. We then computed the cumulative number of cases and fatalities on March 25th, referring to these as "total".

B.1.6 Socioeconomic data

To include socioeconomic covariates in our regression model, we use data from the Atlas of Rural and Small-Town America (available at `https://www.ers.usda.gov/data-products/atlas-of-rural-and-small-town-america/`), which includes data at the state and county level from the American Community Survey (ACS), the Bureau of Labor Statistics, and other sources. We employ data last updated on July 2, 2020, which include county population estimates and annual unemployment/employment data for 2019. County-level measurements about religion are derived from surveys by the Association of Religion Data Archives (accessible at `https://www.thearda.com/Archive/ChCounty.asp`).

Appendix B. Supplementary Information for "The impact of vaccine-related disinformation"

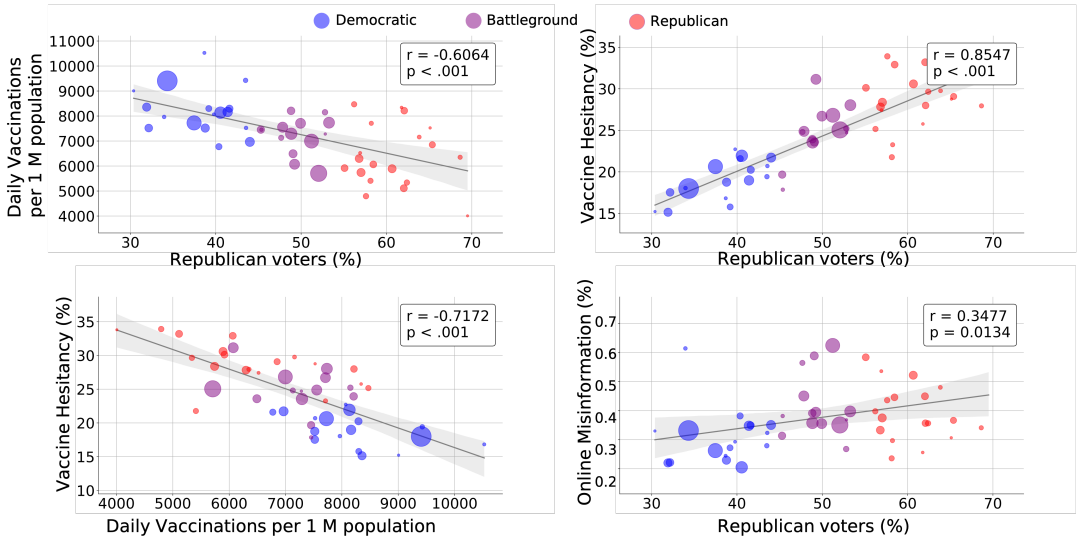


Figure B.1: ***Correlations between vaccine demand, vaccine hesitancy, political partisanship, and online misinformation at the state level.*** *Vaccine demand is computed as the mean number of daily vaccinations per million population in the period 19-25 March 2021. Vaccine hesitancy corresponds to the proportion of individuals who would not get vaccinated according to Facebook daily surveys administered in the period from January 4th to March 25th, 2021. Partisanship is measured as the percentage of Republican voters in the 2020 US Presidential elections. Online misinformation about vaccines shared on Twitter is measured during the period from Jan 4th to March 25th, 2021. Each dot represents a U.S. state, sized according to population and colored according to Republican vote share (battleground states have a share between 45% and 55%).*

B.2 Additional correlation results

Figures B.1 and B.2 present additional results about correlations between vaccine demand, vaccine hesitancy, political partisanship, and online misinformation at state and county levels.

B.3 Main findings from regression analysis

Table S1 presents results from the weighted (Models 1 and 2) and ordinary (Models 3 and 4) least-squares regression of state-level vaccine hesitancy and vaccination rate, respectively, on covariates. As shown in Model 1, the misinformation variable and the percent of GOP voters explain nearly 80% of the variation in vaccine hesitancy at the state level. These predictors remain significant after the addition of multiple control variables (see Model 2). Misinformation and republican vote percentage explain nearly half of the variation in vaccination rate (see Model 3), and are also significantly associated with vaccination rate at the state level net of controls (see Model 4).

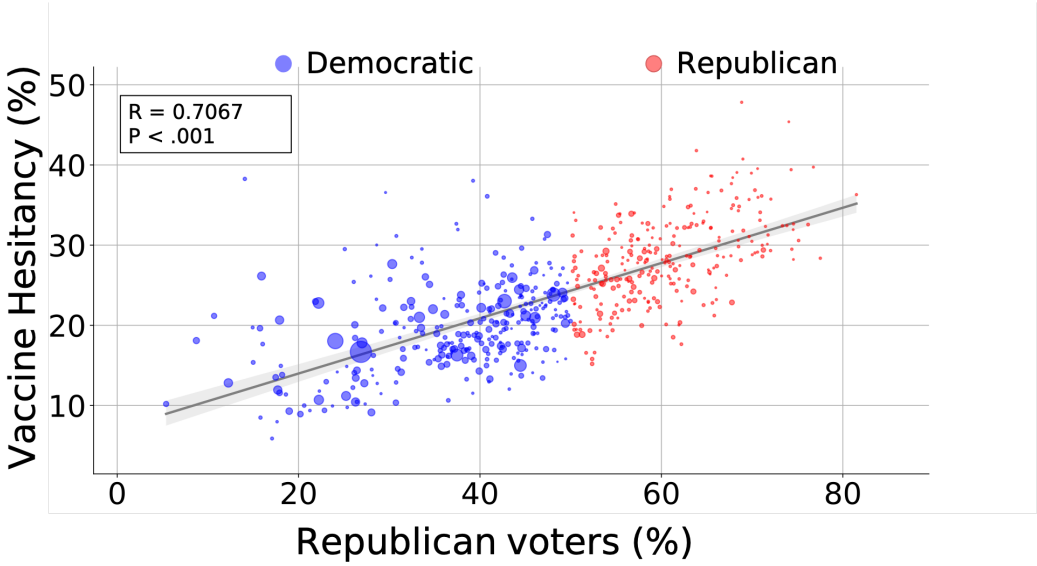


Figure B.2: ***Political partisanship is correlated with vaccine hesitancy at the U.S. county level.*** *Vaccine hesitancy corresponds to the proportion of individuals who would not get vaccinated according to Facebook daily surveys administered in the period from January 4th to March 25th, 2021. Partisanship is measured as the percentage of Republican voters in the 2020 US Presidential elections. Each dot represents a U.S. county, sized according to population and colored according to Republican vote share.*

B.4 Sensitivity analyses

We conduct a set of sensitivity analyses to ensure that our findings are robust to alternative variable and model specifications. First, we run standard diagnostics for non-linearity, skewness, multicollinearity, and heteroskedasticity, correcting any problems we discover. Second, because the misinformation measure at the state level is slightly positively skewed, we conduct a model using a natural logarithmic transformation of mean percent misinformation. Results from these models are consistent with the main findings (Table S2). The untransformed variable has a better model fit (lower BIC). Third, because the effect of misinformation may depend on political partisanship, we test for an interaction between misinformation and the percent of GOP voters. There is no evidence of such interaction at the state level. Fourth, we rerun the above models using versions of the mean percentage of vaccine-related misinformation shared by Twitter users by considering a restricted set of keywords to gather tweets (see previous "Twitter Data" section). As shown in Table S3, findings are consistent and robust to this alternate definition of misinformation sharing.

We also conduct a similar set of sensitivity analyses at the county level. First, we test multiple versions of the misinformation variable, which is highly skewed and zero-inflated at the county level. We use the log-transformed version for the main findings due to the best model fit, but obtain significant results with the untransformed variable

and very similar findings with a polynomial model that also captures the nonlinear relationship between misinformation and vaccine hesitancy. Second, we test for an interaction between misinformation and percent of GOP voters, finding that being in a majority Republican versus Democratic state moderates the association between misinformation and vaccine hesitancy (Table S4). A scatterplot of republican and democratic-leaning counties confirms the moderation finding (Fig.2 in the main manuscript). Third, we run models adding the number of tweets per county as a control variable to address variation in the volume of Twitter activity across counties. Adding this covariate did not affect results. Fourth, as at the state level, we generate versions of the vaccine misinformation variable using a restricted set of keywords. Again, these results are consistent with our main findings (Table S5). Fifth, we examine the robustness of the threshold of 100 Twitter accounts per county for inclusion in the analysis, setting thresholds of 50 and 10. These results are similar to the main findings (Tables S6 and S7), demonstrating that results are robust to different variable specifications.

To confirm the relationship between misinformation and GOP vote share, we compute a negative binomial regression model predicting mean percent information (untransformed) at the county level using percent GOP vote and a set of control variables. This multivariate analysis confirms the bivariate correlation, indicating a strong relationship between these factors net of potential confounding variables (Table S8).

Table S9 describes all the covariates considered in the regression analyses. Table S10 and S11 provide results of the OLS regression for the Granger causality analysis respectively at county and state level.

Table S1. Weighted/ordinary least squares regression of state-level percent vaccine hesitancy and daily vaccination rate per million on misinformation and covariates (N=50 states). In this and the following tables, columns correspond to different models.

	(1) Vaccine hesitancy b (SE)	(2) Vaccine hesitancy b (SE)	(3) Vaccination rate b (SE)	(4) Vaccination rate b (SE)
Mean % low credibility tweets	8.093*	6.877**	-3444.858**	-3518.002**
	(3.04)	(2.43)	(1240.20)	(1277.08)
% GOP vote (10% change)	3.996***	2.960***	-606.567***	-640.319**
	(0.38)	(0.42)	(140.32)	(208.11)
% below poverty line		0.530**		18.173
		(0.15)		(81.84)
% aged 65+		-0.197		171.533
		(0.15)		(100.14)
% Asian		0.011		13.213
		(0.07)		(27.74)
% Black		0.124**		-40.491
		(0.04)		(22.54)
% Hispanic		-0.066*		4.564
		(0.03)		(19.71)
% Indigenous		-0.138		71.890
		(0.12)		(51.00)
COVID deaths/thousand		-0.221		217.490
		(0.42)		(262.06)
Constant	1.858	3.024	11586.785***	9126.137***
	(1.65)	(2.72)	(708.20)	(1537.38)
R^2	0.797***	0.937***	0.457***	0.641***
BIC	225.217	194.454	836.580	843.252

Notes: Vaccine hesitancy is based on state-level means from Facebook survey data. The vaccination rate is vaccines administered per million (CDC data). For models predicting vaccine hesitancy (i.e., state means), analytic weights based on sample size are applied. Unstandardized betas and standard errors are provided. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table S2. Weighted/ordinary least squares regression of state-level percent vaccine hesitancy and daily vaccination rate per million on misinformation (logged) and covariates (N=50 states).

	(1) Vaccine hesitancy b (SE)	(2) Vaccine hesitancy b (SE)	(3) Vaccination rate b (SE)	(4) Vaccination rate b (SE)
Logged mean % low cred tweets	4.136**	3.257**	-1669.206*	-1593.010*
	(1.53)	(1.19)	(636.52)	(660.59)
% GOP vote (10% change)	3.945***	2.962***	-601.418***	-676.915**
	(0.38)	(0.42)	(143.03)	(210.70)
% below poverty line		0.515**		29.711
		(0.15)		(83.31)
% aged 65+		-0.158		158.518
		(0.14)		(101.53)
% Asian		0.009		8.878
		(0.07)		(28.09)
% Black		0.130**		-42.750
		(0.04)		(22.90)
% Hispanic		-0.062*		1.398
		(0.03)		(19.93)
% Indigenous		-0.129		70.503
		(0.12)		(51.98)
COVID deaths/thousand		-0.235		224.368
		(0.42)		(268.26)
Constant	8.318**	7.683	8981.085***	6852.773**
	(2.63)	(3.90)	(1015.40)	(2048.22)
R^2	0.798***	0.936***	0.448***	0.627***
BIC	225.049	194.982	837.352	845.150

Notes: Vaccine hesitancy is based on state-level means from Facebook survey data. The vaccination rate is actual vaccines administered per million (CDC data). For models predicting vaccine hesitancy (i.e., state means), analytic weights based on sample size are applied. Unstandardized betas and standard errors are provided. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table S3. Weighted/ordinary least squares regression of state-level percent vaccine hesitancy and daily vaccination rate per million on misinformation (restricted key words) and covariates (N=50 states).

	(1) Vaccine hesitancy b (SE)	(2) Vaccine hesitancy b (SE)	(3) Vaccination rate b (SE)	(4) Vaccination rate b (SE)
Mean % low credibility tweets	8.320**	7.108**	-3342.575**	-3517.510**
	(2.97)	(2.37)	(1200.22)	(1236.41)
% GOP vote (10% change)	3.982***	2.944***	-611.854***	-648.565**
	(0.37)	(0.41)	(139.58)	(204.44)
% below poverty line		0.517**		27.129
		(0.15)		(81.32)
% aged 65+		-0.206		170.945
		(0.15)		(99.35)
% Asian		0.003		16.019
		(0.07)		(27.87)
% Black		0.125**		-42.464
		(0.04)		(22.25)
% Hispanic		-0.065*		2.774
		(0.03)		(19.42)
% Indigenous		-0.132		68.678
		(0.12)		(50.75)
COVID deaths/thousand		-0.216		225.119
		(0.42)		(259.70)
Constant	1.841	3.313	11575.126***	9085.430***
	(1.64)	(2.71)	(706.47)	(1530.36)
R^2	0.800***	0.938***	0.457***	0.645***
BIC	224.530	193.465	836.543	842.724

Notes: Vaccine hesitancy is based on state-level means from Facebook survey data. The vaccination rate is actual vaccines administered per million (CDC data). For models predicting vaccine hesitancy (i.e., state means), analytic weights based on sample size are applied. Unstandardized betas and standard errors are provided. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table S4. Weighted least squares regression of county-level percent vaccine hesitancy on misinformation (logged) and covariates (N=548 counties, minimum 100 accounts/county).

	(1) b (SE)	(2) b (SE)	(3) b (SE)	(4) b (SE)
Logged mean % low credibility tweets	1.411**	4.304***	1.018***	4.278***
	(0.47)	(0.78)	(0.28)	(0.59)
% GOP vote (10% change)	2.926***		3.663***	
	(0.29)		(0.16)	
Majority GOP state (1=GOP; 0=Dem)		3.892***		3.340***
		(1.02)		(0.66)
GOP state * Logged low credibility		-3.585***		-3.414***
		(0.99)		(0.76)
% below poverty line			0.376***	0.398***
			(0.07)	(0.08)
% aged 65+			-0.056	-0.091
			(0.05)	(0.05)
% Asian			0.028	-0.173**
			(0.03)	(0.05)
% Black			0.202***	0.090***
			(0.02)	(0.03)
% Hispanic			0.002	-0.030
			(0.02)	(0.02)
% Indigenous			0.033	-0.108
			(0.19)	(0.14)
Rural-urban continuum code			0.447	0.617
			(0.26)	(0.34)
COVID deaths/thousand			0.547*	0.925**
			(0.27)	(0.29)
Constant	10.227***	23.668***	-1.535	17.834***
	(1.63)	(1.03)	(1.12)	(1.45)
R^2	0.500***	0.419***	0.805***	0.662***
BIC	3151.490	3240.010	2686.806	2993.820

Notes: Vaccine hesitancy is based on county-level means from Facebook survey data. Misinformation is measured using mean percent of low credibility tweets for counties with at least 100 Twitter accounts. Analytic weights based on Facebook survey sample size are applied, and models use cluster robust standard errors to account for counties being nested in states. Unstandardized betas and standard errors are provided. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table S5. Weighted least squares regression of county-level percent vaccine hesitancy on misinformation (logged, restricted key words) and covariates (N=548 counties, minimum 100 accounts/county).

	(1) b (SE)	(2) b (SE)	(3) b (SE)	(4) b (SE)
Logged mean % low credibility tweets	1.510**	4.382***	1.074***	4.319***
	(0.46)	(0.73)	(0.27)	(0.53)
% GOP vote (10% change)	2.905***		3.641***	
	(0.29)		(0.15)	
Majority GOP state (1=GOP; 0=Dem)		12.010***		11.132***
		(1.49)		(1.16)
GOP state * Logged low credibility		-3.530***		-3.392***
		(0.94)		(0.70)
% below poverty line			0.375***	0.394***
			(0.07)	(0.08)
% aged 65+			-0.058	-0.095
			(0.05)	(0.05)
% Asian			0.028	-0.171**
			(0.03)	(0.05)
% Black			0.202***	0.091***
			(0.02)	(0.03)
% Hispanic			0.002	-0.030
			(0.02)	(0.02)
% Indigenous			0.038	-0.101
			(0.19)	(0.13)
Rural-urban continuum code			0.451	0.648
			(0.26)	(0.33)
COVID deaths/thousand			0.546*	0.916**
			(0.26)	(0.28)
Constant	6.937***	13.673***	-3.849***	7.981***
	(1.14)	(0.95)	(0.93)	(1.29)
R^2	0.501***	0.423***	0.805***	0.665***
BIC	3136.899	3222.391	2673.021	2975.819

Notes: Vaccine hesitancy is based on county-level means from Facebook survey data. Misinformation is measured using mean percent of low credibility tweets for counties with at least 100 Twitter accounts. Analytic weights based on Facebook survey sample size are applied, and models use cluster robust standard errors to account for counties being nested in states. Unstandardized betas and standard errors are provided. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table S6. Weighted least squares regression of county-level percent vaccine hesitancy on misinformation (logged) and covariates (N=658 counties, minimum 10 accounts/county).

	(1) b (SE)	(2) b (SE)	(3) b (SE)	(4) b (SE)
Logged mean % low credibility tweets	1.078*	3.252**	0.941***	3.673***
	(0.47)	(1.11)	(0.22)	(0.75)
% GOP vote (10% change)	3.140***		3.748***	
	(0.29)		(0.15)	
Majority GOP state (1=GOP; 0=Dem)		5.627***		4.247***
		(1.55)		(0.85)
GOP state * Logged low credibility		-2.467*		-2.746**
		(1.16)		(0.84)
% below poverty line			0.369***	0.378***
			(0.07)	(0.07)
% aged 65+			-0.059	-0.114*
			(0.06)	(0.05)
% Asian			0.023	-0.223***
			(0.02)	(0.05)
% Black			0.204***	0.089***
			(0.02)	(0.02)
% Hispanic			0.002	-0.030
			(0.02)	(0.02)
% Indigenous			-0.002	-0.065
			(0.12)	(0.11)
Rural-urban continuum code			0.600**	0.749*
			(0.22)	(0.32)
COVID deaths/thousand			0.549*	1.054***
			(0.27)	(0.29)
Constant	9.047***	22.464***	-2.034	17.582***
	(1.65)	(1.58)	(1.07)	(1.56)
R^2	0.534***	0.421***	0.812***	0.664***
BIC	3796.413	3945.657	3251.830	3639.761

Notes: Vaccine hesitancy is based on county-level means from Facebook survey data. Misinformation is measured using mean percent of low credibility tweets for counties with at least 10 Twitter accounts. Analytic weights based on Facebook survey sample size are applied, and models use cluster robust standard errors to account for counties being nested in states. Unstandardized betas and standard errors are provided. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table S7. Weighted least squares regression of county-level percent vaccine hesitancy on misinformation (logged) and covariates (N=628 counties, minimum 50 accounts/county).

	(1) b (SE)	(2) b (SE)	(3) b (SE)	(4) b (SE)
Logged mean % low credibility tweets	1.347**	4.241***	1.028***	4.233***
	(0.42)	(0.78)	(0.24)	(0.59)
% GOP vote (10% change)	3.039***		3.718***	
	(0.27)		(0.15)	
Majority GOP state (1=GOP; 0=Dem)		4.480***		3.731***
		(0.99)		(0.65)
GOP state * Logged low credibility		-3.350***		-3.236***
		(0.90)		(0.69)
% below poverty line			0.378***	0.407***
			(0.07)	(0.08)
% aged 65+			-0.059	-0.102
			(0.06)	(0.05)
% Asian			0.030	-0.173**
			(0.03)	(0.05)
% Black			0.202***	0.087***
			(0.02)	(0.02)
% Hispanic			0.001	-0.034
			(0.02)	(0.02)
% Indigenous			-0.008	-0.083
			(0.12)	(0.10)
Rural-urban continuum code			0.559*	0.716*
			(0.23)	(0.31)
COVID deaths/thousand			0.538	0.972**
			(0.27)	(0.28)
Constant	9.757***	23.600***	-1.842	17.708***
	(1.48)	(1.03)	(1.09)	(1.49)
R^2	0.524***	0.439***	0.809***	0.667***
BIC	3619.976	3729.469	3099.337	3453.070

Notes: Vaccine hesitancy is based on county-level means from Facebook survey data. Misinformation is measured using mean percent of low credibility tweets for counties with at least 50 Twitter accounts. Analytic weights based on Facebook survey sample size are applied, and models use cluster robust standard errors to account for counties being nested in states. Unstandardized betas and standard errors are provided. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table S8. Negative binomial regression of county-level misinformation on percent GOP vote and covariates (N=548 counties).

	b (SE)
% GOP vote (10% change)	0.263***
	(0.04)
% below poverty line	-0.019*
	(0.01)
% aged 65+	0.043***
	(0.01)
% Asian	0.017
	(0.01)
% Black	0.013***
	(0.00)
% Hispanic	0.006*
	(0.00)
% Indigenous	0.031*
	(0.02)
Rural-urban continuum code	-0.068
	(0.04)
COVID deaths/thousand	-0.098
	(0.06)
Constant	-2.647***
	(0.23)
Wald chi-squared	232.330***
BIC	774.836

Notes: Misinformation is measured using mean percent of low credibility tweets for counties with at least 100 Twitter accounts. Models use cluster robust standard errors to account for counties being nested in states. Negative binomial regression is employed due to zero-inflated Poisson distribution. Unstandardized betas and standard errors are provided. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table S9. Description of covariates used during analyses.

Stata variable	Description	Year	Source
vaxrate	Daily number of people vaccinated per million	2021	Centers for Disease Control and Prevention
lowcred	Mean percentage of low credibility shared (per user)	2021	Twitter API
loglowcred	Natural logarithm of the mean percentage of low credibility shared (per user)	2021	Twitter API
propgop	Proportion of votes for Republican candidate	2020	Fox News, Politico, New York Times
covidmortality	Total COVID 19 deaths	2021	Centers for Disease Control and Prevention
population	Census Population	2010	United States Census
vMedHHInc	Median Household Income	2010	United States Department of Agriculture (Atlas of Rural and Small-Town America)
ppoverty	Percentage of people of all ages in poverty	2019	United States Department of Agriculture (County-Level Datasets)
vPercBachelors	Percent of adults with a bachelor's degree or higher	2015-2019	United States Department of Agriculture (County-Level Datasets)
vUnemployment_rate_2019	Unemployment rate	2019	United States Department of Agriculture (County-Level Datasets)
vTOTRATE	Rates of religious adherence per 1,000 population (200+ religions)	2010	Association of Religious Data Archives
vUnder18Pct2010	Percentage of population age 18 years or younger	2010	United States Department of Agriculture (Atlas of Rural and Small-Town America)
vAge65AndOlderPct2010	Percentage of population age 65 years or older	2010	United States Department of Agriculture (Atlas of Rural and Small-Town America)
vAsianNonHispPct2010	Percentage of population Asians (Non-Hispanic)	2010	United States Department of Agriculture (Atlas of Rural and Small-Town America)
vBlackNonHispPct2010	Percentage of population Black (Non-Hispanic)	2010	United States Department of Agriculture (Atlas of Rural and Small-Town America)
vHispanicPct2010	Percentage of population Hispanic	2010	United States Department of Agriculture (Atlas of Rural and Small-Town America)

vNatAmNonHispPct2010	Percentage of population Native American (Non-Hispanic)	2010	United States Department of Agriculture (Atlas of Rural and Small-Town America)

Table S10. Ordinary Least Squares regression of lagged variates for Granger Causality analysis. (N = 610 counties).

	(1)	(2)	(3)	(4)	(5)	(6)
	coef	std err	t	P>|t|	[0.025	0.975]
hesitancy t-1	0.8852	0.005	174.943	0	0.875	0.895
hesitancy t-2	0.0039	0.007	0.571	0.568	-0.009	0.017
hesitancy t-3	-0.0044	0.007	-0.645	0.519	-0.018	0.009
hesitancy t-4	-0.0004	0.007	-0.061	0.951	-0.014	0.013
hesitancy t-5	0.0074	0.007	1.088	0.277	-0.006	0.021
hesitancy t-6	-0.124	0.005	-24.543	0	-0.134	-0.114
misinfo t-1	0.006	0.004	1.362	0.173	-0.003	0.015
misinfo t-2	0.0087	0.004	1.972	0.049	5.36E-05	0.017
misinfo t-3	0.0156	0.004	3.598	0	0.007	0.024
misinfo t-4	0.0027	0.004	0.625	0.532	-0.006	0.011
misinfo t-5	-0.0014	0.004	-0.337	0.736	-0.01	0.007
misinfo t-6	0.0179	0.004	4.396	0	0.01	0.026
AIC:	56910					
R-squared (uncentered):	0.743					

Null model

	(1)	(2)	(3)	(4)	(5)	(6)
	coef	std err	t	P>|t|	[0.025	0.975]
hesitancy t-1	0.8854	0.005	174.954	0	0.875	0.895
hesitancy t-2	0.0037	0.007	0.549	0.583	-0.01	0.017
hesitancy t-3	-0.0041	0.007	-0.605	0.545	-0.017	0.009
hesitancy t-4	-0.0005	0.007	-0.079	0.937	-0.014	0.013
hesitancy t-5	0.0076	0.007	1.128	0.26	-0.006	0.021
hesitancy t-6	-0.1239	0.005	-24.526	0	-0.134	-0.114
R-squared (uncentered):	0.743					
AIC:	56940					

Table S11. Ordinary Least Squares regression of lagged variates for Granger Causality analysis. (N = 50 states).

	(1)	(2)	(3)	(4)	(5)	(6)
	coef	std err	t	P>|t|	[0.025	0.975]
hesitancy t-1	0.9599	0.016	58.889	0	0.928	0.992
hesitancy t-2	0.024	0.023	1.062	0.288	-0.02	0.068
hesitancy t-3	-0.0748	0.022	-3.325	0.001	-0.119	-0.031
hesitancy t-4	0.1014	0.023	4.501	0	0.057	0.146
hesitancy t-5	-0.0904	0.023	-3.988	0	-0.135	-0.046
hesitancy t-6	-0.0533	0.016	-3.268	0.001	-0.085	-0.021
misinfo t-1	0.0016	0.006	0.262	0.793	-0.011	0.014
misinfo t-2	0.021	0.006	3.351	0.001	0.009	0.033
misinfo t-3	0.0018	0.006	0.295	0.768	-0.01	0.014
misinfo t-4	-0.0161	0.006	-2.603	0.009	-0.028	-0.004
misinfo t-5	0.0133	0.006	2.153	0.031	0.001	0.025
misinfo t-6	0.0003	0.006	0.044	0.965	-0.012	0.012
R-squared (uncentered):	0.842					
AIC:	3133					

Null model

	(1)	(2)	(3)	(4)	(5)	(6)
	coef	std err	t	P>|t|	[0.025	0.975]
hesitancy t-1	0.9593	0.016	58.935	0	0.927	0.991
hesitancy t-2	0.0254	0.023	1.127	0.26	-0.019	0.07
hesitancy t-3	-0.0725	0.023	-3.22	0.001	-0.117	-0.028
hesitancy t-4	0.0982	0.023	4.353	0	0.054	0.142
hesitancy t-5	-0.0879	0.023	-3.873	0	-0.132	-0.043
hesitancy t-6	-0.0548	0.016	-3.358	0.001	-0.087	-0.023
R-squared (uncentered):	0.841					
AIC:	3143					

Bibliography

[1] KFF COVID-19 Vaccine Monitor Dashboard, February 2021.

[2] WHO Coronavirus (COVID-19) Dashboard. https://covid19.who.int, 2021.

[3] Odd Aalen, Ornulf Borgan, and Hakon Gjessing. *Survival and event history analysis: a process point of view*. Springer Science & Business Media, 2008.

[4] U.S. Food & Drug Administration. Pfizer-BioNTech COVID-19 Vaccine. *FDA*, Fri, 04/09/2021 - 13:28.

[5] AGCOM. News vs fake nel sistema dell'informazione. *Report available at: https://www.agcom.it/documents/10179/12791486/Pubblicazione+23-11-2018/93869b4f-0a8d-4380-aad2-c10a0e426d83?version=1.0*, 2018.

[6] Ricardo Aguas, Rodrigo M. Corder, Jessica G. King, Guilherme Gonçalves, Marcelo U. Ferreira, and M. Gabriela M. Gomes. Herd immunity thresholds for SARS-CoV-2 estimated from unfolding epidemics. *medRxiv*, page 2020.07.23.20160762, November 2020.

[7] Selim Aksoy and Robert M Haralick. Feature normalization and likelihood-based similarity measures for image retrieval. *Pattern recognition letters*, 22(5):563–582, 2001.

[8] M. S. Al-Rakhami and A. M. Al-Amri. Lies Kill, Facts Save: Detecting COVID-19 Misinformation in Twitter. *IEEE Access*, 8:155961–155970, 2020.

[9] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *nature*, 406(6794):378, 2000.

[10] Righi Alessandra, Mauro M Gentile, and Domenico M Bianco. Who tweets in italian? demographic characteristics of twitter users. In *Convegno della Società Italiana di Statistica*, pages 329–344. Springer, 2017.

[11] Max Aliapoulios, Emmi Bevensee, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, and Savvas Zannettou. An early look at the parler online social network. *arXiv preprint arXiv:2101.03820*, 2021.

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

[13] Francesco Aquino, Gabriele Donzelli, Emanuela De Franco, Gaetano Privitera, Pier Luigi Lopalco, and Annalaura Carducci. The web and public confidence in mmr vaccination in italy. *Vaccine*, 35(35):4494–4498, 2017.

[14] Solomon E Asch and H Guetzkow. Effects of group pressure upon the modification and distortion of judgments. *Groups, leadership, and men*, pages 222–236, 1951.

[15] Blake E Ashforth and Fred Mael. Social identity theory and the organization. *Academy of management review*, 14(1):20–39, 1989.

[16] Avaaz. Far right networks of deception. *Available at: https://avaazimages.avaaz.org/Avaaz%20Report%20Network%20Deception%2020190522.pdf*, 2019.

[17] Adam Badawy, Emilio Ferrara, and Kristina Lerman. Analyzing the digital traces of political manipulation: the 2016 russian interference Twitter campaign. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 258–265. IEEE, 2018.

[18] J. P. Bagrow, E. M. Bollt, J. D. Skufca, and D. ben Avraham. Portraits of complex networks. *EPL (Europhysics Letters)*, 81(6):68004, feb 2008.

[19] James P Bagrow and Erik M Bollt. An information-theoretic, all-scales approach to comparing networks. *Applied Network Science*, 4(1):1–15, 2019.

[20] Chi Y Bahk, Melissa Cumming, Louisa Paushter, Lawrence C Madoff, Angus Thomson, and John S Brownstein. Publicly available online tool facilitates real-time monitoring of vaccine conversations and sentiments. *Health affairs*, 35(2):341–347, 2016.

[21] Sean Baird, Doug Sibley, and Yuxi Pan. Talos targets disinformation with fake news challenge victory. *Fake News Challenge*, 2017.

[22] Albert-László Barabási. *Network science*. Cambridge university press, 2016.

[23] Pablo Barberá, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10):1531–1542, 2015.

[24] Marc Barthélemy, Alain Barrat, Romualdo Pastor-Satorras, and Alessandro Vespignani. Characterization and modeling of weighted networks. *Physica a: Statistical mechanics and its applications*, 346(1-2):34–43, 2005.

[25] Marco T Bastos and Dan Mercea. The brexit botnet and user-generated hyperpartisan news. *Social Science Computer Review*, 37(1):38–54, 2019.

[26] Vladimir Batagelj and Matjaz Zaversnik. An o(m) algorithm for cores decomposition of networks. *arXiv preprint cs/0310049*, 2003.

[27] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[28] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[29] 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, 2008.

[30] Stefano Boccaletti, Vito Latora, Yamir Moreno, Martin Chavez, and D-U Hwang. Complex networks: Structure and dynamics. *Physics reports*, 424(4-5):175–308, 2006.

[31] Kaustubh Bora, Dulmoni Das, Bhupen Barman, and Probodh Borah. Are internet videos useful sources of information during global public health emergencies? A case study of YouTube videos during the 2015–16 Zika virus pandemic. *Pathogens and Global Health*, 112(6):320–328, 2018.

[32] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

[33] Alexandre Bovet and Hernán A Makse. Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications*, 10(1):7, 2019.

[34] Alexandre Bovet, Flaviano Morone, and Hernán A Makse. Validation of Twitter opinion trends with national polling aggregates: Hillary Clinton vs Donald Trump. *Scientific Reports*, 8(1):8673, 2018.

[35] Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. In *2010 43rd Hawaii International Conference on System Sciences*, pages 1–10. IEEE, 2010.

[36] Lia Bozarth and Ceren Budak. Toward a better performance evaluation framework for fake news classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 60–71, 2020.

[37] David A Broniatowski, Amelia M Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C Quinn, and Mark Dredze. Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate. *American journal of public health*, 108(10):1378–1384, 2018.

[38] David A. Broniatowski, Daniel Kerchner, Fouzia Farooq, Xiaolei Huang, Amelia M. Jamison, Mark Dredze, and Sandra Crouse Quinn. The COVID-19 social media Infodemic reflects uncertainty and state-sponsored propaganda. *arXiv:2007.09682*, July 2020.

[39] E. K. Brunson. The Impact of Social Networks on Parents' Vaccination Decisions. *Pediatrics*, 131(5):e1397–e1404, May 2013.

[40] Talha Burki. Vaccine misinformation and social media. *The Lancet Digital Health*, 1(6):e258–e259, October 2019.

[41] Timothy Callaghan, Ali Moghtaderi, Jennifer A. Lueck, Peter Hotez, Ulrich Strych, Avi Dor, Erika Franklin Fowler, and Matthew Motta. Correlates and disparities of intention to vaccinate against COVID-19. *Social Science & Medicine (1982)*, 272:113638, March 2021.

[42] Michele Cantarella, Nicolò Fraccaroli, and Roberto Volpe. Does fake news affect voting behaviour? *Available at SSRN: https://ssrn.com/abstract=3402913*, 2019.

[43] Centers for Disease Control and Prevention. COVID Data Tracker. https://covid.cdc.gov/covid-data-tracker, March 2020.

[44] Cynthia Chew and Gunther Eysenbach. Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak. *PloS one*, 5(11):e14118, 2010.

[45] Wen-Ying Sylvia Chou, April Oh, and William MP Klein. Addressing health-related misinformation on social media. *JAMA*, 320(23):2417–2418, 2018.

[46] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[47] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. The COVID-19 social media Infodemic. *Scientific Reports*, 10(1):16598, 2020.

[48] European Commission. Tackling online disinformation, 2019.

[49] Michael D Conover, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. Partisan asymmetries in online political activity. *EPJ Data Science*, 1(6), 2012.

[50] Niall J Conroy, Victoria L Rubin, and Yimin Chen. Automatic deception detection: Methods for finding fake news. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, page 82. American Society for Information Science, 2015.

[51] Nicolo Conti. Elezioni europee, ma poca europa. *La Repubblica*, 2019.

[52] Alessandro Cossard, Gianmarco De Francisci Morales, Kyriaki Kalimeri, Yelena Mejova, Daniela Paolotti, and Michele Starnini. Falling into the echo chamber: the italian vaccination debate on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 130–140, 2020.

[53] Carolyn Crist. States Begin Opening COVID-19 Vaccines to All Adults. https://www.webmd.com/vaccines/covid-19-vaccine/news/20210324/states-begin-opening-covid-19-vaccines-to-all-adults, 2021.

[54] CrowdTangle Team. CrowdTangle. Menlo Park, CA: Facebook., 2020. Accessed December 2020.

[55] Clayton A Davis, Giovanni Luca Ciampaglia, Luca Maria Aiello, Keychul Chung, Michael D Conover, Emilio Ferrara, Alessandro Flammini, et al. OSoMe: the IUNI observatory on social media. *PeerJ Computer Science*, 2:e87, 2016.

[56] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 273–274. International World Wide Web Conferences Steering Committee, 2016.

[57] Juan Manuel Ortiz de Zarate, Marco Di Giovanni, Esteban Zindel Feuerstein, and Marco Brambilla. Measuring controversy in social networks through nlp. In Christina Boucher and Sharma V. Thankachan, editors, *String Processing and Information Retrieval*, pages 194–209, Cham, 2020. Springer International Publishing.

[58] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559, 2016.

[59] Michela Del Vicario, Sabrina Gaito, Walter Quattrociocchi, Matteo Zignani, and Fabiana Zollo. News consumption during the italian referendum: A cross-platform analysis on facebook and twitter. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 648–657. IEEE, 2017.

[60] Susi Dennison and Pawel Zerka. The 2019 european election: How anti-europeans plan to wreck europe and what can be done to stop it. *European council on foreign relations*, 2019.

[61] Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, 2017.

[62] Matthew DeVerna, Francesco Pierri, Bao Truong, John Bollenbacher, David Axelrod, Niklas Loynes, Cristopher Torres-Lugo, Kai-Cheng Yang, Fil Menczer, and John Bryden. Covaxxy: A global collection of english twitter posts about covid-19 vaccines. *Proceedings of the International AAAI Conference on Web and Social Media*, 2021.

[63] Matthew DeVerna, Francesco Pierri, Bao Tran Truong, John Bollenbacher, David Axelrod, Nikals Loynes, Christopher Torres-Lugo, Kai-Cheng Yang, Filippo Menczer, and John Bryden. Covaxxy tweet ids dataset. Zenodo, February 2021.

[64] Matthew R. DeVerna, Francesco Pierri, Bao Truong, John Bollenbacher, David Axelrod, Niklas Loynes, Cristopher Torres-Lugo, Kai-Cheng Yang, Fil Menczer, and John Bryden. Data for CoVaxxy: A collection of English-language Twitter posts about COVID-19 vaccines. https://github.com/osome-iu/CoVaxxy, February 2021.

[65] Andrey A Dobrynin, Roger Entringer, and Ivan Gutman. Wiener index of trees: theory and applications. *Acta Applicandae Mathematica*, 66(3):211–249, 2001.

[66] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet Infectious Diseases*, 20(5):533–534, May 2020.

[67] Gabriele Donzelli, Giacomo Palomba, Ileana Federigi, Francesco Aquino, Lorenzo Cioni, Marco Verani, Annalaura Carducci, and Pierluigi Lopalco. Misinformation on vaccination: a quantitative analysis of youtube videos. *Human vaccines & immunotherapeutics*, 14(7):1654–1659, 2018.

[68] Mark Dredze, Michael J. Paul, Shane Bergsma, and Hieu Tran. *Carmen: A Twitter Geolocation System with Applications to Public Health*.

[69] A. Dutta, N. Beriwal, L. M. Van Breugel, et al. YouTube as a source of medical and epidemiological information during COVID-19 pandemic: A cross-sectional study of content across six languages around the globe. *Cureus*, 12(6):e8622, 2020.

[70] Arianna D'Ulizia, Maria Chiara Caschera, Fernando Ferri, and Patrizia Grifoni. Fake news detection: a survey of evaluation datasets. *PeerJ Computer Science*, 7:e518, 2021.

[71] Bradley Efron and Trevor Hastie. *Computer age statistical inference*, volume 5. Cambridge University Press, 2016.

[72] Gunther Eysenbach. Infodemiology: The epidemiology of (mis) information. *The American Journal of Medicine*, 113(9):763–765, 2002.

[73] Gunther Eysenbach, John Powell, Oliver Kuss, and Eun-Ryoung Sa. Empirical studies assessing the quality of health information for consumers on the world wide web: a systematic review. *JAMA*, 287(20):2691–2700, 2002.

[74] FactCheckEU. Good news and bad news after election week-end. 2019.

[75] J. Fairbanks et al. Credibility assessment in the news: Do we need to read? 2018.

[76] David C. Farrow, Logan C. Brooks, Aaron Rumack, Ryan J. Tibshirani, and Roni Rosenfeld. Delphi Epidata API. https://github.com/cmu-delphi/delphi-epidata, 2015.

[77] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[78] Miriam Fernandez and Harith Alani. Online misinformation: Challenges and future directions. In *Companion of the The Web Conference 2018 on The Web Conference 2018*, pages 595–602. International World Wide Web Conferences Steering Committee, 2018.

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

[80] Emilio Ferrara, Stefano Cresci, and Luca Luceri. Misinformation, manipulation, and abuse on social media in the era of COVID-19. *Journal of Computational Social Science*, 3(2):271–277, November 2020.

[81] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *Communications of the ACM*, 59(7):96–104, 2016.

[82] William Ferreira and Andreas Vlachos. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1163–1168, 2016.

[83] Santo Fortunato and Darko Hric. Community detection in networks: A user guide. *Physics reports*, 659:1–44, 2016.

[84] Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

[85] Isaac Chun-Hai Fung, King-Wa Fu, Chung-Hong Chan, Benedict Shing Bun Chan, Chi-Ngai Cheung, Thomas Abraham, and Zion Tsz Ho Tse. Social media's initial reaction to information and misinformation on Ebola, August 2014: facts and rumors. *Public Health Reports*, 131(3):461–473, 2016.

[86] Cary Funk and Alec Tyson. Growing Share of Americans Say They Plan To Get a COVID-19 Vaccine – or Already Have, March 2021.

[87] Sebastian Funk, Marcel Salathé, and Vincent A. A. Jansen. Modelling the influence of human behaviour on the spread of infectious diseases: A review. *Journal of The Royal Society Interface*, 7(50):1247–1256, September 2010.

[88] Riccardo Gallotti, Francesco Valle, Nicola Castaldo, Pierluigi Sacco, and Manlio De Domenico. Assessing the risks of 'infodemics' in response to COVID-19 epidemics. *Nature Human Behaviour*, 4:1285–1293, 2020.

[89] Michael T. Gastner, Vivien Seguy, and Pratyush More. Fast flow-based algorithm for creating density-equalizing map projections. *Proceedings of the National Academy of Sciences*, 115(10):E2156–E2164, March 2018.

[90] Fabio Giglietto, Laura Iannelli, Luca Rossi, Augusto Valeriani, Nicola Righetti, Francesca Carabini, Giada Marino, Stefano Usai, and Elisabetta Zurovac. Mapping italian news media political coverage in the lead-up to 2018 general election. *Available at SSRN: https://ssrn.com/abstract=3179930*, 2018.

[91] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[92] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J Watts. The structural virality of online diffusion. *Management Science*, 62(1):180–196, 2015.

[93] Barrett Golding. Iffy+ Mis/Disinfo Sites. https://iffy.news/iffy-plus/, December 2020.

[94] M. Gabriela M. Gomes, Rodrigo M. Corder, Jessica G. King, Kate E. Langwig, Caetano Souto-Maior, Jorge Carneiro, Guilherme Gonçalves, Carlos Penha-Gonçalves, Marcelo U. Ferreira, and Ricardo Aguas. Individual variation in susceptibility or exposure to SARS-CoV-2 lowers the herd immunity threshold. *medRxiv*, page 2020.04.27.20081893, May 2020.

[95] Gillian C. Goobie, Sabina A. Gulera, Kerri A. Johannson, Jolene H. Fisher, and Christopher J. Ryerson. YouTube videos as a source of misinformation on idiopathic pulmonary fibrosis. *Annals of the American Thoracic Society*, 16(5):572—-579, 2019.

[96] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[97] Jeffrey Gottfried and Elisa Shearer. *News Use Across Social Medial Platforms 2016*. Pew Research Center, 2016.

[98] Przemyslaw A Grabowicz, José J Ramasco, Esteban Moro, Josep M Pujol, and Victor M Eguiluz. Social features of online networks: The strength of intermediary ties in online social media. *PLOS ONE*, 7(1), 2012.

[99] C. W. J. Granger. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37(3):424–438, 1969.

[100] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. Fake news on Twitter during the 2016 U.S. presidential election. *Science*, 363(6425):374–378, 2019.

[101] Anatoliy Gruzd, Manlio De Domenico, Pier Luigi Sacco, and Sylvie Briand. Studying the covid-19 infodemic at scale, 2021.

[102] Stefano Guarino, Francesco Pierri, Marco Di Giovanni, and Alessandro Celestini. Information disorders during the covid-19 infodemic: The case of italian facebook. *Online Social Networks and Media*, 22:100124, 2021.

[103] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[104] James Douglas Hamilton. *Time Series Analysis*. Princeton, N.J, 1st edition edition, January 1994.

[105] Andreas Hanselowski, PVS Avinesh, Benjamin Schiller, and Felix Caspelherr. Description of the system developed by team athene in the fnc-1. *Fake News Challenge*, 2017.

[106] Andreas Hanselowski, PVS Avinesh, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M Meyer, and Iryna Gurevych. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874, 2018.

[107] Freja Hedman, Fabian Sivnert, and PN Howard. News and political information consumption in sweden: Mapping the 2018 swedish general election on twitter, 2018.

[108] Jon Henley. How populism emerged as an electoral force in europe. *The Guardian*, 2018.

[109] Peter Hernon. Disinformation and misinformation through the internet: Findings of an exploratory study. *Government Information Quarterly*, 12(2):133–139, 1995.

[110] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[111] Benjamin D Horne and Sibel Adali. This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *arXiv preprint arXiv:1703.09398*, 2017.

[112] Seyedmehdi Hosseinimotlagh and Evangelos E Papalexakis. Unsupervised content-based identification of fake news articles with tensor decomposition ensembles. 2018.

[113] Peter Hotez, Carolina Batista, Onder Ergonul, J. Peter Figueroa, Sarah Gilbert, Mayda Gursel, Mazen Hassanain, Gagandeep Kang, Jerome H. Kim, Bhavna Lall, Heidi Larson, Denise Naniche, Timothy Sheahan, Shmuel Shoham, Annelies Wilder-Smith, Nathalie Strub-Wourgaft, Prashant Yadav, and Maria Elena Bottazzi. Correcting COVID-19 vaccine misinformation: Lancet Commission on COVID-19 Vaccines and Therapeutics Task Force Members. *EClinicalMedicine*, 33, March 2021.

[114] Philip N Howard, Samantha Bradshaw, Bence Kollanyi, and Gillian Bolsolver. Junk news and bots during the french presidential election: What are french voters sharing over twitter in round two?

[115] Philip N Howard and Bence Kollanyi. Bots,# strongerin, and# brexit: computational propaganda during the uk-eu referendum. 2016.

[116] Pik-Mai Hui, Chengcheng Shao, Alessandro Flammini, Filippo Menczer, and Giovanni Luca Ciampaglia. The hoaxy misinformation and fact-checking diffusion network. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[117] Azhar Hussain, Syed Ali, Madiha Ahmed, and Sheharyar Hussain. The Anti-vaccination Movement: A Regression in Modern Medicine. *Cureus*, 10(7), 2018.

[118] Amnesty International. Il barometro dell'odio - elezioni europee 2019. *Available at: https://www.amnesty.it/cosa-facciamo/elezioni-europee/*, 2019.

[119] Shalev Itzkovitz, Ron Milo, Nadav Kashtan, Guy Ziv, and Uri Alon. Subgraphs in random networks. *Physical review E*, 68(2):026127, 2003.

[120] S Mo Jang, Tieming Geng, Jo-Yun Queenie Li, Ruofan Xia, Chin-Tser Huang, Hwalbin Kim, and Jijun Tang. A computational approach for examining the roots and spreading patterns of fake news: Evolution tree analysis. *Computers in Human Behavior*, 84:103–113, 2018.

[121] Fang Jin, Wei Wang, Liang Zhao, Edward Dougherty, Yang Cao, C Lu, and Naren Ramakrishnan. Misinformation propagation in the age of Twitter. *IEEE Annals of the History of Computing*, 47(12):90–94, 2014.

[122] Neil F. Johnson, Nicolas Velásquez, Nicholas Johnson Restrepo, Rhys Leahy, Nicholas Gabriel, Sara El Oud, Minzhang Zheng, Pedro Manrique, Stefan Wuchty, and Yonatan Lupu. The online competition between pro- and anti-vaccination views. *Nature*, 582(7811):230–233, 2020.

[123] David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 51–57, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[124] Maurice George Kendall. Rank correlation methods. *Griffin*, 1948.

[125] Jagdish Khubchandani, Sushil Sharma, James H. Price, Michael J. Wiblishauser, Manoj Sharma, and Fern J. Webb. COVID-19 Vaccination Hesitancy in the United States: A Rapid National Assessment. *Journal of Community Health*, 46(2):270–277, April 2021.

[126] Jerome H. Kim, Florian Marks, and John D. Clemens. Looking beyond COVID-19 vaccine phase 3 trials. *Nature Medicine*, 27(2):205–211, February 2021.

[127] Jooyeon Kim, Behzad Tabibian, Alice Oh, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 324–332. ACM, 2018.

[128] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, 2014.

[129] Aleksi Knuutila, Aliaksandr Herasimenka, Hubert Au, Jonathan Bright, Rasmus Nielsen, and Philip N Howard. COVID-related misinformation on YouTube: The spread of misinformation videos on social media and the effectiveness of platform policies. Oxford, UK: Project on Computational Propaganda, 2020. `https://comprop.oii.ox.ac.uk/research/posts/youtube-platform-policies/`.

[130] Bence Kollanyi and Philip N Howard. Junk news and bots during the german parliamentary election: What are german voters sharing over twitter, 2017.

[131] Moshe Koppel, Jonathan Schler, and Elisheva Bonchek-Dokow. Measuring differentiability: Unmasking pseudonymous authors. *Journal of Machine Learning Research*, 8(Jun):1261–1276, 2007.

[132] 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.

[133] Srijan Kumar and Neil Shah. False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559, To appear in the book titled Social Media Analytics: Advances and Applications, by CRC press, 2018*, 2018.

[134] Srijan Kumar, Robert West, and Jure Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*, pages 591–602. International World Wide Web Conferences Steering Committee, 2016.

[135] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600, 2010.

[136] Heidi J. Larson and David A. Broniatowski. Volatility of vaccine confidence. *Science*, 371(6536):1289–1289, March 2021.

[137] David Lazer, Katherine Ognyanova, Matthew Baum, James Druckman, Jon Green, Adina Gitomer, Matthew Simonson, Roy H. Perlis, Mauricio Santillana, Alexi Quintana, Jennifer Lin, and Ata Uslu. The COVID States Project #43: COVID-19 vaccine rates and attitudes among Americans, March 2021.

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

[139] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

[140] Heidi Oi-Yee Li, Adrian Bailey, David Huynh, and James Chan. YouTube as a source of information on COVID-19: a pandemic of misinformation? *BMJ Global Health*, 5(5):e002604, 2020.

[141] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM Comput. Surv.*, 50(6):94:1–94:45, December 2017.

[142] Dimitra Liotsiou, Bence Kollanyi, and Philip N Howard. The junk news aggregator: examining junk news posted on facebook, starting with the 2018 us midterm elections. *arXiv preprint arXiv:1901.07920*, 2019.

[143] Yang Liu and Yi-Fang Brook Wu. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[144] Sahil Loomba, Alexandre de Figueiredo, Simon 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*, 2021.

[145] Dan Lu, Alberto Aleta, Marco Ajelli, Romualdo Pastor-Satorras, Alessandro Vespignani, and Yamir Moreno. Data-driven estimate of SARS-CoV-2 herd immunity threshold in populations with individual contact pattern variations. *medRxiv*, page 2021.03.19.21253974, March 2021.

[146] Wei Lyu and George L. Wehby. Community use of face masks and COVID-19: Evidence from a natural experiment of state mandates in the US. *Health Affairs*, 39(8):1419–1425, June 2020.

[147] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 3818–3824. AAAI Press, 2016.

[148] Noni E. MacDonald. Vaccine hesitancy: Definition, scope and determinants. *Vaccine*, 33(34):4161–4164, August 2015.

[149] L Meghan Mahoney, Tang Tang, Kai Ji, and Jessica Ulrich-Schad. The digital distribution of public health news surrounding the human papillomavirus vaccination: a longitudinal infodemiology study. *JMIR Public Health and Surveillance*, 1(1):e2, 2015.

[150] Henry B Mann. Nonparametric tests against trend. *Econometrica: Journal of the Econometric Society*, pages 245–259, 1945.

[151] Nahema Marchal, Bence Kollanyi, Lisa-Maria Neudert, and Philip N Howard. Junk news during the eu parliamentary elections: Lessons from a seven-language study of twitter and facebook. 2019.

[152] Sergei Maslov and Kim Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913, 2002.

[153] Maxwell McCombs. *Setting the agenda: Mass media and public opinion*. John Wiley & Sons, 2018.

[154] Maxwell E McCombs, Donald L Shaw, and David H Weaver. New directions in agenda-setting theory and research. *Mass communication and society*, 17(6):781–802, 2014.

[155] Yelena Mejova and Nicolas Kourtellis. Youtubing at home: Media sharing behavior change as proxy for mobilityaround covid-19 lockdowns. *arXiv preprint arXiv:2103.14601*, 2021.

[156] Shahan Ali Memon and Kathleen M. Carley. Characterizing COVID-19 misinformation communities using a novel Twitter dataset. *arXiv:2008.00791*, September 2020.

[157] J Millman. The inevitable rise of ebola conspiracy theories. *The Washington Post*, 2014.

[158] MIT Election Data and Science Lab. U.S. President 1976–2020, January 2021.

[159] Amy Mitchell, Jeffrey Gottfried, Jocelyn Kiley, and Katerina Eva Matsa. Political polarization & media habits. Pew Research Center, 2014. `http://pewrsr.ch/1vZ9MnM` (Accessed November 2020).

[160] Amy Mitchell and J. Baxter Oliphant. Americans immersed in coronavirus news; most think media are doing fairly well covering it. Pew Research Center, March 2020. `https://pewrsr.ch/3dbTpxs` (Accessed November 2020).

[161] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[162] Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M Bronstein. Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*, 2019.

[163] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *Seventh international AAAI conference on weblogs and social media*, 2013.

[164] Subhabrata Mukherjee and Gerhard Weikum. Leveraging joint interactions for credibility analysis in news communities. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 353–362. ACM, 2015.

[165] Kevin P Murphy, Yair Weiss, and Michael I Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 467–475. Morgan Kaufmann Publishers Inc., 1999.

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

[167] Rasmus Kleis Nielsen, Nic Newman, Richard Fletcher, and Antonis Kalogeropoulos. Reuters institute digital news report 2019. *Report of the Reuters Institute for the Study of Journalism*, 2019.

[168] Sophie Nightingale, Marc Faddoul, and Hany Farid. Quantifying the reach and belief in COVID-19 misinformation. *arXiv:2006.08830*, June 2020.

[169] Dimitar Nikolov, Alessandro Flammini, and Filippo Menczer. Right and left, partisanship predicts (asymmetric) vulnerability to misinformation. *Harvard Kennedy School Misinformation Review*, 1(7), February 2021.

[170] Dimitar Nikolov, Mounia Lalmas, Alessandro Flammini, and Filippo Menczer. Quantifying biases in online information exposure. *Journal of the Association for Information Science and Technology*, 70(3):218–229, 2019.

[171] Leonardo Nizzoli, Serena Tardelli, Marco Avvenuti, Stefano Cresci, and Maurizio Tesconi. Coordinated behavior on social media in 2019 uk general election. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 443–454, 2021.

[172] Brendan Nyhan and Jason Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330, 2010.

[173] Daniel J O'Keefe. Elaboration likelihood model. *The international encyclopedia of communication*, 2008.

[174] Walter A. Orenstein and Rafi Ahmed. Simply put: Vaccination saves lives. *Proceedings of the National Academy of Sciences*, 114(16):4031–4033, April 2017.

[175] Diogo Pacheco, Pik-Mai Hui, Christopher Torres-Lugo, Bao Tran Truong, Alessandro Flammini, and Filippo Menczer. Uncovering coordinated networks on social media. *arXiv preprint arXiv:2001.05658*, 2020.

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

[177] Donatella Panatto, Daniela Amicizia, Lucia Arata, Piero Luigi Lai, and Roberto Gasparini. A comprehensive analysis of Italian web pages mentioning squalene-based influenza vaccine adjuvants reveals a high prevalence of misinformation. *Human Vaccines & Immunotherapeutics*, 14(4):969–977, 2018.

[178] Eli Pariser. *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011.

[179] Irene V. Pasquetto, Briony Swire-Thompson, et al. Tackling misinformation: What researchers could do with social media data. *Harvard Kennedy School Misinformation Review*, 1(8), 2020.

[180] Ranjan Pathak, Dilli Ram Poudel, Paras Karmacharya, Amrit Pathak, Madan Raj Aryal, Maryam Mahmood, and Anthony A Donato. YouTube as a source of information on Ebola virus disease. *North American Journal of Medical Sciences*, 7(7):306, 2015.

[181] V Paul Pauca, Farial Shahnaz, Michael W Berry, and Robert J Plemmons. Text mining using non-negative matrix factorizations. In *Proceedings of the 2004 SIAM International Conference on Data Mining*, pages 452–456. SIAM, 2004.

[182] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[183] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of liwc2015. Technical report, 2015.

[184] Gordon Pennycook and David G Rand. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 116(7):2521–2526, 2019.

[185] Gordon Pennycook and David G Rand. Who falls for fake news? the roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of personality*, 88(2):185–200, 2020.

[186] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401. Association for Computational Linguistics, 2018.

[187] Andrew Perrin and Monica Anderson. Share of US adults using social media, including Facebook, is mostly unchanged since 2018. Pew Research Center, 2019. `https://pewrsr.ch/2VxJuJ3` (Accessed February 2021).

[188] Francesco Pierri. The diffusion of mainstream and disinformation news on twitter: the case of italy and france. *Companion Proceedings of the Web Conference 2020 (WWW '20 Companion)*, 2020.

[189] Francesco Pierri, Alessandro Artoni, and Stefano Ceri. Investigating italian disinformation spreading on twitter in the context of 2019 european elections. *PloS one*, 15(1):e0227821, 2020.

[190] Francesco Pierri and Stefano Ceri. False news on social media: a data-driven survey. *ACM Sigmod Record*, 48(2), 2019.

[191] Francesco Pierri, Brea Perry, Matthew R. DeVerna, Alessandro Flammini, Kai-Cheng Yang, Filippo Menczer, and John Bryden. Reproducibility code for "The impact of online misinformation on U.S. COVID-19 vaccinations". https://github.com/osome-iu/CoVaxxy-Misinfo, April 2021.

[192] Francesco Pierri, Brea Perry, Matthew R DeVerna, Kai-Cheng Yang, Alessandro Flammini, Filippo Menczer, and John Bryden. The impact of online misinformation on us covid-19 vaccinations. *arXiv preprint arXiv:2104.10635*, 2021.

[193] Francesco Pierri, Carlo Piccardi, and Stefano Ceri. A multi-layer approach to disinformation detection in us and italian news spreading on twitter. *EPJ Data Science*, 9(35), 2020.

[194] Francesco Pierri, Carlo Piccardi, and Stefano Ceri. Topology comparison of Twitter diffusion networks effectively reveals misleading news. *Scientific Reports*, 10:1372, 2020.

[195] Francesco Pierri, Andrea Tocchetti, Lorenzo Corti, Marco Di Giovanni, Silvio Pavanetto, Marco Brambilla, and Stefano Ceri. Vaccinitaly: monitoring italian conversations around vaccines on twitter and facebook. 2021.

[196] Dean Pomerleau and Delip Rao. Fake news challenge. http://www.fakenewschallenge.org, 2017.

[197] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. In *International symposium on computer and information sciences*, pages 284–293. Springer, 2005.

[198] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. Declare: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, 2018.

[199] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240. Association for Computational Linguistics, 2018.

[200] Cristina M Pulido, Beatriz Villarejo-Carballido, Gisela Redondo-Sama, and Aitor Gómez. COVID-19 infodemic: More retweets for science-based information on coronavirus than for false information. *International Sociology*, 35(4):377–392, July 2020.

[201] Kennet Rapoza. Can 'fake news' impact the stock market? *Forbes*, 2017.

[202] Jacob Ratkiewicz, Michael D Conover, Mark Meiss, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer Menczer. Detecting and tracking political abuse in social media. In *Fifth international AAAI conference on weblogs and social media*, 2011.

[203] Edward S Reed, Elliot Turiel, and Terrance Brown. Naive realism in everyday life: Implications for social conflict and misunderstanding. In *Values and knowledge*, pages 113–146. Psychology Press, 2013.

[204] Benjamin Riedel, Isabelle Augenstein, Georgios P Spithourakis, and Sebastian Riedel. A simple but tough-to-beat baseline for the fake news challenge stance detection task. *arXiv preprint arXiv:1707.03264*, 2017.

[205] Bernhard Rieder. Studying facebook via data extraction: the netvizz application. In *Proceedings of the 5th annual ACM web science conference*, pages 346–355. ACM, 2013.

[206] Nicola Righetti. Health politicization and misinformation on twitter. a study of the italian twittersphere from before, during and after the law on mandatory vaccinations, Apr 2020.

[207] Megan Risdal. Fake news dataset. https://www.kaggle.com/mrisdal/fake-news. 2017.

[208] Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

[209] Jon Roozenbeek, Claudia R. Schneider, Sarah Dryhurst, John Kerr, Alexandra L. J. Freeman, Gabriel Recchia, Anne Marthe van der Bles, and Sander van der Linden. Susceptibility to misinformation about COVID-19 around the world. *Royal Society Open Science*, 7(10):201199, 2020. Publisher: Royal Society.

[210] Jon Roozenbeek, Claudia R. Schneider, Sarah Dryhurst, John Kerr, Alexandra L. J. Freeman, Gabriel Recchia, Anne Marthe van der Bles, and Sander van der Linden. Susceptibility to misinformation about COVID-19 around the world. *Royal Society Open Science*, 7(10):201199, October 2020.

[211] Nir Rosenfeld, Aron Szanto, and David C Parkes. A kernel of truth: Determining rumor veracity on twitter by diffusion pattern alone. In *Proceedings of The Web Conference 2020*, pages 1018–1028, 2020.

[212] Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17, 2016.

[213] Victoria L Rubin, Yimin Chen, and Niall J Conroy. Deception detection for news: three types of fakes. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, page 83. American Society for Information Science, 2015.

[214] Natali Ruchansky, Sungyong Seo, and Yan Liu. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 797–806. ACM, 2017.

[215] Britta Ruhnau. Eigenvector-centrality—a node-centrality? *Social networks*, 22(4):357–365, 2000.

[216] Marcel Salathé and Sebastian Bonhoeffer. The effect of opinion clustering on disease outbreaks. *Journal of the Royal Society Interface*, 5(29):1505–1508, December 2008.

[217] Giovanni Santia and Jake Williams. Buzzface: A news veracity dataset with facebook user commentary and egos, 2018.

[218] Anida Sarajlić, Noël Malod-Dognin, Ömer Nebil Yaveroğlu, and Nataša Pržulj. Graphlet-based characterization of directed networks. *Scientific reports*, 6:35098, 2016.

[219] Jari Saramäki, Mikko Kivelä, Jukka-Pekka Onnela, Kimmo Kaski, and Janos Kertesz. Generalizations of the clustering coefficient to weighted complex networks. *Physical Review E*, 75(2):027105, 2007.

[220] Katherine Schaeffer. Nearly three-in-ten Americans believe COVID-19 was made in a lab. Pew Research Center, April 2020. `https://pewrsr.ch/2XlJqAa` (Accessed November 2020).

[221] Ana Lucía Schmidt, Fabiana Zollo, Antonio Scala, Cornelia Betsch, and Walter Quattrociocchi. Polarization of the vaccination debate on facebook. *Vaccine*, 36(25):3606–3612, 2018.

[222] Tara Kirk Sell, Divya Hosangadi, and Marc Trotochaud. Misinformation and the us ebola communication crisis: analyzing the veracity and content of social media messages related to a fear-inducing infectious disease outbreak. *BMC Public Health*, 20:1–10, 2020.

[223] EK Seltzer, E Horst-Martz, M Lu, and RM Merchant. Public sentiment and discourse about Zika virus on Instagram. *Public Health*, 150:170–175, 2017.

[224] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, pages 745–750, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

[225] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. The spread of low-credibility content by social bots. *Nature Communications*, 9:4787, 2018.

[226] 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):1–23, 04 2018.

[227] Karishma Sharma, Emilio Ferrara, and Yan Liu. Identifying coordinated accounts in disinformation campaigns. *arXiv:2008.11308*, 2020.

[228] Megha Sharma, Kapil Yadav, Nitika Yadav, and Keith C Ferdinand. Zika virus pandemic—analysis of Facebook as a social media health information platform. *American Journal of Infection Control*, 45(3):301–302, 2017.

[229] 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.

[230] Kai Shu, Suhang Wang, and Huan Liu. Beyond news contents: The role of social context for fake news detection. *arXiv preprint arXiv:1712.07709, to appear in Proceedings of 12th ACM International Conference on Web Search and Data Mining (WSDM 2019)*, 2017.

[231] C. Silverman. This analysis shows how fake election news stories outperformed real news on facebook. buzzfeed, `https://zenodo.org/record/1239675`, 2016.

[232] C Silverman and J Singer-Vine. Most americans who see fake news believe it, new survey says, buzzfeed, 2016.

[233] Lisa Singh, Shweta Bansal, Leticia Bode, Ceren Budak, Guangqing Chi, Kornraphop Kawintiranon, Colton Padden, Rebecca Vanarsdall, Emily Vraga, and Yanchen Wang. A first look at COVID-19 information and misinformation sharing on Twitter. *arXiv:2003.13907*, March 2020.

[234] Lisa Singh, Leticia Bode, Ceren Budak, Kornraphop Kawintiranon, Colton Padden, and Emily Vraga. Understanding high-and low-quality URL Sharing on COVID-19 Twitter streams. *Journal of Computational Social Science*, 3:343–366, 2020.

[235] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

[236] Massimo Stella, Emilio Ferrara, and Manlio De Domenico. Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences*, 115(49):12435–12440, 2018.

[237] Craig A. Stewart, Timothy M. Cockerill, Ian Foster, David Hancock, Nirav Merchant, Edwin Skidmore, Daniel Stanzione, James Taylor, Steven Tuecke, George Turner, Matthew Vaughn, and Niall I. Gaffney. Jetstream: A self-provisioned, scalable science and engineering cloud environment. In *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*, XSEDE '15, pages 1–8. Association for Computing Machinery, 2015.

[238] Leo G Stewart, Ahmer Arif, and Kate Starbird. Examining trolls and polarization with a retweet network. In *Proceedings ACM WSDM, Workshop on Misinformation and Misbehavior Mining on the Web*, 2018.

[239] Cass Sunstein. On rumors: How falsehoods spread, why we believe them, what can be done. new york: Farrar, straus and giroux; solove, daniel j. the future of reputation, 2007.

[240] Cass R Sunstein. *Echo chambers: Bush v. Gore, impeachment, and beyond*. Princeton University Press, 2001.

[241] Eugenio Tacchini, Gabriele Ballarin, Marco L Della Vedova, Stefano Moret, and Luca de Alfaro. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506*, 2017.

[242] Mattia Tantardini, Francesca Ieva, Lucia Tajoli, and Carlo Piccardi. Comparing methods for comparing networks. *Scientific reports*, 9(1):1–19, 2019.

[243] James Thorne and Andreas Vlachos. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, 2018.

[244] J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. R. Scott, and N. Wilkins-Diehr. Xsede: Accelerating scientific discovery. *Computing in Science & Engineering*, 16(5):62–74, Sept.-Oct. 2014.

[245] John Towns, Timothy Cockerill, Maytal Dahan, Ian Foster, Kelly Gaither, Andrew Grimshaw, Victor Hazlewood, Scott Lathrop, Dave Lifka, Gregory D. Peterson, Ralph Roskies, J. Ray Scott, and Nancy Wilkins-Diehr. XSEDE: Accelerating Scientific Discovery. *Computing in Science Engineering*, 16(5):62–74, September 2014.

[246] Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. Fake news detection in social networks via crowd signals. In *Companion of the The Web Conference 2018 on The Web Conference 2018*, pages 517–524. International World Wide Web Conferences Steering Committee, 2018.

[247] Chris J Vargo, Lei Guo, and Michelle A Amazeen. The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *New Media & Society*, 20(5):2028–2049, 2018.

[248] Chris J Vargo, Lei Guo, Maxwell McCombs, and Donald L Shaw. Network issue agendas on twitter during the 2012 us presidential election. *Journal of Communication*, 64(2):296–316, 2014.

[249] Michela Del Vicario, Walter Quattrociocchi, Antonio Scala, and Fabiana Zollo. Polarization and fake news: Early warning of potential misinformation targets. *ACM Transactions on the Web (TWEB)*, 13(2):10, 2019.

[250] Nguyen Vo and Kyumin Lee. The rise of guardians: Fact-checking url recommendation to combat fake news. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 275–284, New York, NY, USA, 2018. ACM.

[251] Svitlana Volkova and Jin Yea Jang. Misleading or falsification: Inferring deceptive strategies and types in online news and social media. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 575–583, Republic and Canton of Geneva, Switzerland, 2018. International World Wide Web Conferences Steering Committee.

[252] Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 647–653, 2017.

[253] Luis Von Ahn, Manuel Blum, Nicholas J Hopper, and John Langford. Captcha: Using hard ai problems for security. In *International Conference on the Theory and Applications of Cryptographic Techniques*, pages 294–311. Springer, 2003.

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

[255] Patrick Wang, Rafael Angarita, and Ilaria Renna. Is this the era of misinformation yet: combining social bots and fake news to deceive the masses. In *Companion Proceedings of the The Web Conference 2018*, pages 1557–1561. International World Wide Web Conferences Steering Committee, 2018.

[256] William Yang Wang. " liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 422–426, 2017.

[257] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 849–857. ACM, 2018.

[258] Yuxi Wang, Martin McKee, Aleksandra Torbica, and David Stuckler. Systematic literature review on the spread of health-related misinformation on social media. *Social Science & Medicine*, 240:112552, 2019.

[259] Claire Wardle and Hossein Derakhshan. Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe*, 27, 2017.

[260] Stanley Wasserman and Katherine Faust. Social network analysis: Methods and applications. 8, 1994.

[261] Lillian Weng, Alessandro Flammini, Alessandro Vespignani, and Fillipo Menczer. Competition among memes in a world with limited attention. *Scientific reports*, 2:335, 2012.

[262] Steven Lloyd Wilson and Charles Wiysonge. Social media and vaccine hesitancy. *BMJ Global Health*, 5(10):e004206, October 2020.

[263] Tom Wilson and Kate Starbird. Cross-platform disinformation campaigns: Lessons learned and next steps. *Harvard Kennedy School Misinformation Review*, 1(1), 2020.

[264] Stefan Wojick and Adam Hughes. Sizing up twitter users. Pew Research Center, `https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/` (accessed January, 2021), 2020.

[265] Michael J Wood. Propagating and debunking conspiracy theories on Twitter during the 2015–2016 Zika virus outbreak. *Cyberpsychology, Behavior, and Social Networking*, 21(8):485–490, 2018.

[266] Samuel C Woolley and Philip N Howard. *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media*. Oxford University Press, 2018.

[267] Liang Wu and Huan Liu. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 637–645. ACM, 2018.

[268] Kai-Cheng Yang, Pik-Mai Hui, and Filippo Menczer. Bot electioneering volume: Visualizing social bot activity during elections. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 214–217, 2019.

[269] Kai-Cheng Yang, Francesco Pierri, Pik-Mai Hui, David Axelrod, Christopher Torres-Lugo, John Bryden, and Filippo Menczer. Dataset for paper: The COVID-19 Infodemic: Twitter versus Facebook. Zenodo, 2020. `https://doi.org/10.5281/zenodo.4313903`.

[270] Kai-Cheng Yang, Francesco Pierri, Pik-Mai Hui, David Axelrod, Christopher Torres-Lugo, John Bryden, and Filippo Menczer. The COVID-19 Infodemic: Twitter versus Facebook. *Big Data & Society*, 8(1):20539517211013861, January 2021.

[271] Kai-Cheng Yang, Christopher Torres-Lugo, and Filippo Menczer. Prevalence of low-credibility information on twitter during the COVID-19 outbreak. In *Proceedings of the ICWSM International Workshop on Cyber Social Threats*, 2020.

[272] Kai-Cheng Yang, Onur Varol, Pik-Mai Hui, and Filippo Menczer. Scalable and generalizable social bot detection through data selection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(1):1096–1103, 2020.

[273] Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, and Huan Liu. Unsupervised fake news detection on social media: A generative approach. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5644–5651, 2019.

[274] Ömer Nebil Yaveroğlu, Tijana Milenković, and Nataša Pržulj. Proper evaluation of alignment-free network comparison methods. *Bioinformatics*, 31(16):2697–2704, 2015.

[275] Robert B Zajonc. Mere exposure: A gateway to the subliminal. *Current directions in psychological science*, 10(6):224–228, 2001.

[276] John Zarocostas. How to fight an infodemic. *The Lancet*, 395(10225):676, 2020.

[277] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, pages 9051–9062, 2019.

[278] Zilong Zhao, Jichang Zhao, Yukie Sano, Orr Levy, Hideki Takayasu, Misako Takayasu, Daqing Li, Junjie Wu, and Shlomo Havlin. Fake news propagates differently from real news even at early stages of spreading. *EPJ Data Science*, 9(1):1–14, 2020.

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

[280] Fabiana Zollo, Alessandro Bessi, Michela Del Vicario, Antonio Scala, Guido Caldarelli, Louis Shekhtman, Shlomo Havlin, and Walter Quattrociocchi. Debunking in a world of tribes. *PloS one*, 12(7):e0181821, 2017.

[281] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. Detection and resolution of rumours in social media: A survey. *ACM Comput. Surv.*, 51(2):32:1–32:36, February 2018.