Riding information crises: The performance of far-right Twitter users in Australia during the 2019-20 bushfires and the COVID-19 pandemic

Francesco Bailo
University of Sydney

Amelia Johns
University of Technology Sydney

Marian-Andrei Rizoiu
University of Technology Sydney

Version: Preprint (the published version is available here)

Date: 18 Apr 2023

Abstract

This paper focuses on the performance of the far-right community in the Australian Twittersphere during two information crises: the 2019-20 Australian bushfires and the early months of the 2020 COVID-19 pandemic. Using a mixed method approach to analysing the performance of far-right accounts active in both crises, and using an information disorder index to estimate the quality of information being shared on Twitter during the two events, we found that far-right accounts moved from the periphery of these disaster-driven conversations during the Australian bushfires to assume a more central location during the COVID-19 pandemic. We argue that an increase in information disorder and overperformance of far-right accounts during COVID-19 is suggestive of an association between the two, which warrants further investigation.

Keywords: Far-right; Twitter; Australian Bushfires; COVID-19; Network analysis; Information crisis; Epistemic crisis

Online appendix and replication materials: `https://doi.org/10.7910/DVN/QN1LUZ`**.**

Contents

1 Introduction

Between the end of 2019 and the beginning of 2020, the Australian public experienced two major crises: the 2019-20 bushfires and the onset of the COVID-19 pandemic. Both crises became primary public events capturing the attention of the media, monopolising the public conversation and producing conflicting information, narratives and interpretations. The public debate around the bushfire crisis was framed by two contrasting camps: supporters of the scientific consensus linking the fires to anthropogenic climate change and those with opposing views. The first debate was found by some to be fuelled by bot-like, coordinated networks (Graham & Keller, 2020) while others identified a higher prevalence of peer-to-peer, human activity (Weber et al., 2022). Similarly, the debate at the onset of the COVID-19 pandemic saw opposition between those who trusted the science around the virus, as communicated by national and international health authorities, and those who mistrusted their expertise and challenged their interpretation by, for example, sharing content or opinions which minimised the seriousness of the disease, or conspiracy theories which framed the pandemic as a global plot orchestrated by elites to control people (Graham et al., 2020; Kong et al., 2022).

Raising alarm about the prevalence of COVID-19 misinformation early in the pandemic, The World Health Organisation (WHO) likened its spread to an 'infodemic' (Zarocostas, 2020), claiming that misinformation had contributed to the spread of the virus. Studies have since adopted the term ('infodemic') to characterise the diffusion of low-credibility and manipulated sources of information that competed with and undermined public health information during the crisis, contributing to 'information disorder' (Hameleers et al., 2021; Zola et al., 2022). Evidence of coordinated behaviour to promote false or partially false claims have been identified as part of this dynamic (Gallotti et al., 2020; Graham et al., 2020).

Social networking platforms (SNPs) such as Twitter have been found in influential studies to be an effective "tool for public opinion manipulation" through computational operations (Woolley & Howard, 2019, p. 241) as well as becoming - in some contexts - an "echo chamber that radicalizes its inhabitants, destabilizes their ability to tell truth from fiction, and undermines their confidence in institutions" (Benkler et al., 2018, p. 383). And yet the pace of events during a crisis also makes SNPs a critical resource for an anxious public seeking up-to-date information. This makes research comparing information disorders on SNPs, across different crisis events, an important contribution to scholarly and policy discussions focused on maintaining the health and integrity of information systems.

We focused on far-right accounts active in the Australian Twittersphere owing to evidence that accounts connected to QAnon networks and other partisan far-right and conservative accounts, actively promoted false and harmful content across one or both crises (Graham et al., 2020; Kong et al., 2022; Zola et al., 2022). Further evidence that these accounts acted in violation of the platform's terms of use is Twitter's decision to suspend more than 38% of the accounts we analysed. Nonetheless, following Wardle & Derakhsan's conceptualisation of 'information disorder' as the spread and amplification of false, misleading or harmful content *either* with the intention to harm or deceive (i.e. disinformation) or unintentionally (i.e. misinformation), our first research question seeks to determine the performance of the far-right accounts we studied and their impact on the broader conversation space, without raising questions of motivation. Following Hameleers et al. (2021) we use the umbrella term of "misinformation" to describe the sharing of false or misleading content, regardless of intention.

To measure and compare the performance of the far-right community across the two events, we use a statistical technique known as 'matching' combined with network analysis. This helped us to determine the relative position of the community in the two online conversations, addressing our first research question:

> RQ1: Did far-right accounts overperform on Twitter across the two crises, becoming more central in the conversation?

Secondly, in seeking to quantify the degree of potentially false or misleading content that diffused across the entire public conversation (of which far-right accounts are a subset) we operationalise a metric of information disorder capturing the dispersion of attention as well as the authoritativeness of information sources shared by all accounts across the two periods. Despite some disagreement among scholars as to whether source-based classification is able to sufficiently measure the prevalence and degree of infodemic or information disorder (Gallotti et al., 2020; Shahi et al., 2021) – with some scholars preferring approaches where misinformation content is verified by fact-checking experts (see Gallotti et al., 2020) – it is accepted as an approach which allows automatic measurement of the credibility of information across large datasets (Lazer et al., 2018; Shahi et al., 2021; Yang et al., 2021). The development of this metric is a major contribution of this paper. We deploy network analysis combined with our information disorder index to address our second research question:

> RQ2: Is the public Twitter conversation across the two periods associated with a higher degree of information disorder?

Finally, studies examining information disorder across the Australian bushfires and COVID-19 have been concerned to detect evidence of coordinated behaviour between accounts. Coordinated, inauthentic behaviour on Twitter and other social media platforms has become a cause of alarm since 2016, when evidence emerged of Russian interference in the US election, using message coordination and amplification strategies where automated accounts, or "social bots", are utilised, as well as centrally organised, human-managed accounts, sometimes called cyborg accounts (Keller et al., 2020). These studies found that the presence of coordinated, inauthentic manipulation in information environments can produce confusion, distracting citizens from credible news and information and giving legitimacy to unverified or false narratives (Wardle & Derakhshan, 2017; Weber et al, 2022; Graham et al, 2020; Yang et al, 2021). In this way coordination behaviours contribute toward information disorder (Wardle & Derakhshan, 2017). In many studies, the agents responsible for coordinated manipulation of public discourse are associated with far-right networks Keller et al., 2020; Wardle & Derakhshan, 2017).

While studies examining the two crisis events investigated in this paper generally found that misinformation was spread by genuine human accounts (Gruzd & Mai, 2020; Shahi et al., 2021; Weber et al., 2022; Yang et al., 2021; Zola et al., 2022) some studies found evidence of suspected coordinated, inauthentic behaviour (Ferrara, 2020; Graham et al., 2020, 2021; Graham & Keller, 2020; Keller et al., 2020), with far-right accounts (QAnon and MAGA accounts) being identified as central agents. The removal of a number of accounts in our corpus indicates suspicious behaviour, but it also places limits on the ability to detect this behaviour in the wider conversation. This led us to use a series of independent measurements, including a bot detection algorithm, network analysis and manual coding, to address our last research question:

RQ3: Can we detect coordinated, inauthentic behaviour among the far-right accounts active in both crises?

2 Theoretical and conceptual framework

2.1 Epistemic gaps and information disorder during a crisis

Information crises are periods of high uncertainty concerning the selection of information and 'interpretations of contradictory and ambiguous messages from central or other sources' (Nohrstedt, 1991, p. 484). Information crises present a challenge to the *epistemic system*, the 'social system that houses a variety of procedures, institutions, and patterns of interpersonal influence that affect the *epistemic outcomes* [emphasis added] of its members' (Goldman, 2011, p. 13). These are characterised by higher levels of *epistemic uncertainty* as emerging information often lacks a clear connection to existing knowledge. The speed of the crisis might generate facts that the epistemic system is not able to explain yet and also outpace the capacity of the epistemic system in communicating to the public relevant epistemic outcomes to interpret unfolding events. The decoupling of information from interpretation generates *epistemic gaps* with the risk of further exacerbating an *epistemic crisis*, which ‘occurs when a group or community [...] finds itself with reasons to question the correctness of the rules and structures it has been using for fixing beliefs’ (Laudan, 2001, p. 273). Douglas et al. identify along with *existential* and *social* motives, also *epistemic* motives: “[s]pecific epistemic motives that causal explanations may serve include slaking curiosity when information is unavailable, reducing uncertainty and bewilderment when available information is conflicting, finding meaning when events seem random, and defending beliefs from disconfirmation” (2017, p. 538). Epistemic gaps foster the production of misinformation and misinterpretations as this content satisfies an urgent demand by people and Internet platforms (Golebiewski & Boyd, 2019) which is not met by content curated by the epistemic system (e.g. governments and intergovernmental institutions).

If information and epistemic crises have been studied well before the massification of Internet and mobile technologies (see Nohrstedt, 1991), two factors have radically changed them in recent years. First, SNPs can create unprecedented *volumes* of user-generated content - distributed alongside content generated by authoritative sources - and circulate it at unprecedented *speed.* This makes efforts to validate content by institutional actors and everyday users a more difficult task. Second, the sophistication of platforms– providing powerful communicative and organisational affordances to everyday users– facilitates the coordination of online behaviours (both authentic and inauthentic) with the aim of increasing reach and influencing public perceptions (Kim et al., 2019). This has created an environment in which the severity of an information crisis is associated with degrees of *information disorder*. Focusing on the agents involved in spreading low credibility information, the source and content of messages and how content is interpreted by its audience, Wardle and Derakhshan (2017) define information disorder as the production and diffusion of *misinformation* and *disinformation*[1] contributing to ‘pollution’ of information systems ‘at a global scale’ (2017, p. 4). *Misinformation* refers to the sharing of false content, unintentionally, while *disinformation* is identified as a range of techniques to intentionally create and share false or fabricated content to deceive or harm.

As mentioned, we use the term misinformation as an umbrella term to capture disorders relating to mis and disinformation. We also acknowledge that understanding the motivations of users sharing misinformation can be a difficult task. Fact-checking organisations have been tackling this problem by ‘authenticating’ sources, yet these efforts are not able to cover all sources, given the prevalence of unofficial, user generated content (Wardle & Darakhshan, 2017, p. 18). While Wardle & Derakhshan argue that research needs to account for the original sources creating and

[1] Malinformation is a third information disorder described by the author but it has less relevance to the current study.

sharing content, this makes assessing degrees of information disorder at scale a difficult task. In light of this, in this paper we operationalise an *information disorder* metric using two measures: the number of sources competing for attention across public conversations that form in crisis events, and the ratio between the number of authoritative and non-authoritative sources (RQ2). This provides insight into degrees of information overload *and* quality, which combine to 'contaminate public discourse' on a range of issues, from vaccination to climate change, overwhelming users' ability to agree on shared realities, let alone reach consensus on what the solutions should look like.

In identifying and defining low-credibility sources competing for attention, the literature has produced mixed results. Of relevance to our study, some scholars have singled out YouTube as a platform that is commonly used by fringe actors to amplify and spread misinformation and conspiracy theories across platforms, primarily Twitter (Li et al., 2020; Marwick & Lewis, 2017; Wilson & Starbird, 2020). This occurs due to the affordances YouTube and Twitter provide to users to game trending topic algorithms (Marwick & Lewis, 2017). Nonetheless, other studies found that YouTube was less likely to amplify unreliable sources of information compared to less moderated platforms (Cinelli et al., 2020).

2.2 Studying coordinated, inauthentic behaviour online

Coordinated, inauthentic behaviour refers to techniques employed by "malicious actors" while "hiding their real identities and intentions" (Giglietto et al., 2020, p. 872). Social media users can be deceived by creating and managing fake accounts giving the appearance of genuine engagement - a tactic employed by Russia's Internet Research Agency during the 2016 US election campaign (Jamieson, 2018). An audience and a recommender system can also be deceived by faking the number of people holding a particular view or "liking" a social media account or content.

Bot detection tools have been developed to determine whether the creation and activity of coordinating accounts are automated (Ferrara et al., 2016). Nonetheless, studies relying on these tools have become criticised (Graham et al., 2021; Keller et al., 2020), with evidence showing that 'machine learning algorithms in combination with human coding . . . likely misses many human-manned accounts or sophisticated bots' (Keller et al., 2020, p. 259). To address this limitation, studies have employed qualitative and quantitiative measures to detect coordinated inauthentic behaviour. We will employ both methods in this paper to address RQ3.

3 Data and Methods

3.1 Data collection

We purchased data capturing the conversation around the 2019–20 Australian bushfires and COVID-19 pandemic from Twitter in October 2020. It is therefore likely that some tweets and users were removed in the period interceding data creation and collection affecting the completeness of the data. Our queries filtered for tweets from users estimated to be in Australia, and for search terms such as 'bushfire', 'arson', 'climatecrisis', 'covid', 'plandemic', 'stayhome' (the full ruleset is available in the online supplemental material). This returned 1,864,011 tweets published by 119,057 unique users between 14 December 2019 and 25 January 2020 (bushfire related) and 5,038,308 tweets published by 253,028 users between 3 February 2020 and 13 March 2020 (COVID-19 related). For brevity, in the remainder of the paper, we refer to these two periods also as Period 1 (bushfires) and Period 2 (COVID-19) respectively.

3.2 Classification of Twitter accounts as far-right

We approached the research design with a broad definition of the far-right, which included ultra-nationalists but also accounts promoting QAnon themes (in 2020 QAnon enjoyed large support in the Australian far-right, see Ross, 2020). Instead of manually identifying far-right accounts we chose to validate accounts added by Twitter users to public lists linked to that phenomenon. First, a 'seed' Twitter account was manually selected from the data based on three criteria: the frequency of its activity in the Twitter conversation, its popularity and the degree of representativeness of the Australian far-right community based on our qualitative investigation of the conversation (Kong et al., 2022). Based on our data, this seed account (now suspended) increased the number of its followers from 10,866 to almost 50,000. Second, in October 2020, we queried the Twitter API requesting the Twitter lists that this account had been added to.[2] This returned 210 lists. Third, in a subsequent series of API calls, we requested all the members from these 210 lists.[3] These calls returned roughly 13,000 unique accounts, of which 4087 were found to be in our dataset. The 4087 accounts in this selection were manually checked using profile description, username and location to exclude 'false-positive' accounts - mostly journalists, political figures or parody accounts. Accounts were excluded if their profile information did not contain any explicit reference

[2] https://api.twitter.com/1.1/lists/memberships.json

[3] https://api.twitter.com/1.1/lists/members.json

to far-right topics, themes or ideology. This returned a final list of 208 Australian far-right accounts present in our data.

The use of Twitter lists identified through a seed account presents operational and methodological advantages over manual coding. First, the geography of far-right accounts changed constantly as accounts were suspended, particularly during the COVID-19 crisis, resulting in the creation of new accounts. Second, instead of relying on recommender systems or existing connections, by using public lists we leverage the ground-level knowledge of the broad Twitter community about constantly evolving social media communities.

We further validated these 208 accounts as far-right accounts using quantitative and qualitative methods. This included, first, identifying through *term frequency-inverse document frequency* (tf-idf) the most prevalent terms that appear in the descriptions of far-right accounts (a list is presented in the supplemental material). The analysis of the prevalent terms identified themes central to the Australian far-right community such as the defence of traditional values and roles - ‘western’, ‘christian’- against ‘socialism’ (Busbridge et al., 2020) but also ‘islam’ (Hutchinson, 2021). In addition to reference of Australian values - ‘aussie’, ‘australia’ - we also found a prominent connection with a transnational far-right and alt-right movement centred around the QAnon political movement (Ross, 2020) - ‘trump’, ‘maga’, ‘trump2020’, ‘q’, ‘wwg1wga’, ‘seeker’. Finally, we also found representation of an anti-establishment (and paradoxically anti-authoritarian) sentiment with terms like ‘free’ and ‘freedom’ denoting resentment towards state institutions (Davis, 2019).

3.3 Network analysis

We draw three networks from the Twitter data: first, a User-to-User network connecting users who have retweeted each other at least once with the edge’s weight capturing the number of mutual edge pairs (i.e. mutual retweets) and, second, a User-to-URL network connecting users to the URLs they have reshared, with edge’s weights capturing the frequency. We use the analyses on these two networks to answer RQ1 and RQ2. To better understand the relative position of accounts, we ran a “fast greedy” modularity optimization algorithm (Clauset et al., 2004), one of the most popular community detection algorithms for weighted social networks, on both the User-to-User and the User-to-URLs networks to create communities based on similarities in users’ tweeting patterns.

3.4 Classification of Twitter accounts on a left-right scale

We classified all acounts in our data sharing links to news resources on a traditional left-right scale to offer an ideological context to the analysis of networks of users and their network communities. We used responses to a survey conducted by Reuters and the University of Canberra (Park et al., 2020) on the media habits of 2131 Australians to estimate the position of 29 Australian publications as the average position of their audiences on the left-right spectrum. We estimated the left-right position of accounts sharing at least one link to the websites of these 29 publications as the average left-right position of the domains of the links.

3.5 Matching

To address RQ1 and estimate the performance of far-right accounts in terms of the frequency of rewteets, we used 'nearest neighbor matching' (Ho et al., 2011), a statistical method that identifies for any tweet of interest - the 'treated' unit - a comparable tweet to be used as 'untreated' control unit. The difference in Twitter performance between 'treated' and 'untreated' tweets is used to estimate the performance of the accounts publishing the 'treated' tweets.

Our 'treated' tweets either

1. were published by an account coded into one of these groups: journalists, politicians, far-right or climate action accounts; or
2. contained a link to the website of Facebook, YouTube and Twitter, to the website of a mainstream media organisation or to a website identified by *Media Bias / Fact Check*[4] as publishing low-credibility news stories.

A comparable, 'untreated' tweet was identified by the matching algorithm respectively among all the tweets or all the tweets containing a link published within 4 hours from the publication of the 'treated' tweet, using a predefined number of features, so to select the most similar control unit available. In our configuration, we used the following features to identify the matching tweet: the number of tweets the authoring account liked since its creation, the number of followers of the authoring account, the number of accounts followed by the authoring account, the number of Twitter lists the account was added to, the number of tweets the account published since its creation, whether the account was verified, whether the tweet contained any media, and whether the tweet contained any URL. Once a matching tweet was identified, we ran a Poisson regression for each 4-hour time window included in our two periods to estimate with cluster-robust standard

[4] https://mediabiasfactcheck.com/

errors (with the pair treated-untreated as the clustering variable) if treated tweets received more retweets than comparable and *non*-treated tweets (our control group). Because far-right accounts are found to systematically retweet each other, the number of retweets we use to estimate the relative performance of an account excludes retweets from all the accounts that reciprocated retweets with the authoring account of the 'treated' unit in the 7-days preceding the 4-hour time window of analysis (or as far back as the data allowed).

3.6 Information disorder index

Operationalising our definition, we measure the degree of information disorder using the tweets' embedded links. Each link was associated with an information source by extracting its domain. News media domains were identified using a list resulting from the combination of three collections of Australian and global English news media sites, curated by Media Cloud. Government domains were identified if the suffix of the domain was .gov, .gov.au or .int (reserved for intergovernmental organisations). The first measure of our index is the ratio between the number of unique sources of information (i.e. domains) found in the tweets and the number of unique users active in the conversation. The second measure is the ratio between the number of unique user-domain ties where the domain is an authoritative information source (i.e. news media or government domains) and the total number of unique user-domain ties found in the conversation. By counting unique user-domain ties instead of the number of tweets containing a link to a specific domain we focus on measuring the range of information sources accessed by users instead of the number of times users have been sharing content from any individual source.

The design of our index is based on two assumptions. First, independent of the quality of information sources, we expect a relatively larger pool of information sources to be associated with a higher degree of disorder as navigating the information becomes challenging. Second, we expect information disorder to be associated with users accessing a lower volume of information from news media and institutional sources relative to the volume of information from alternative sources. Our information disorder index responds to the necessity of classifying a large volume of multimedial content (i.e. text but also images, audios and videos) while returning a global measure for the quality of the pool of information being reshared during the conversation.

3.7 Measuring coordination and authenticity of accounts

To assess whether accounts demonstrated coordinated, inauthentic behaviour, we use a mixed-method approach and independent authenticity validating exercises.

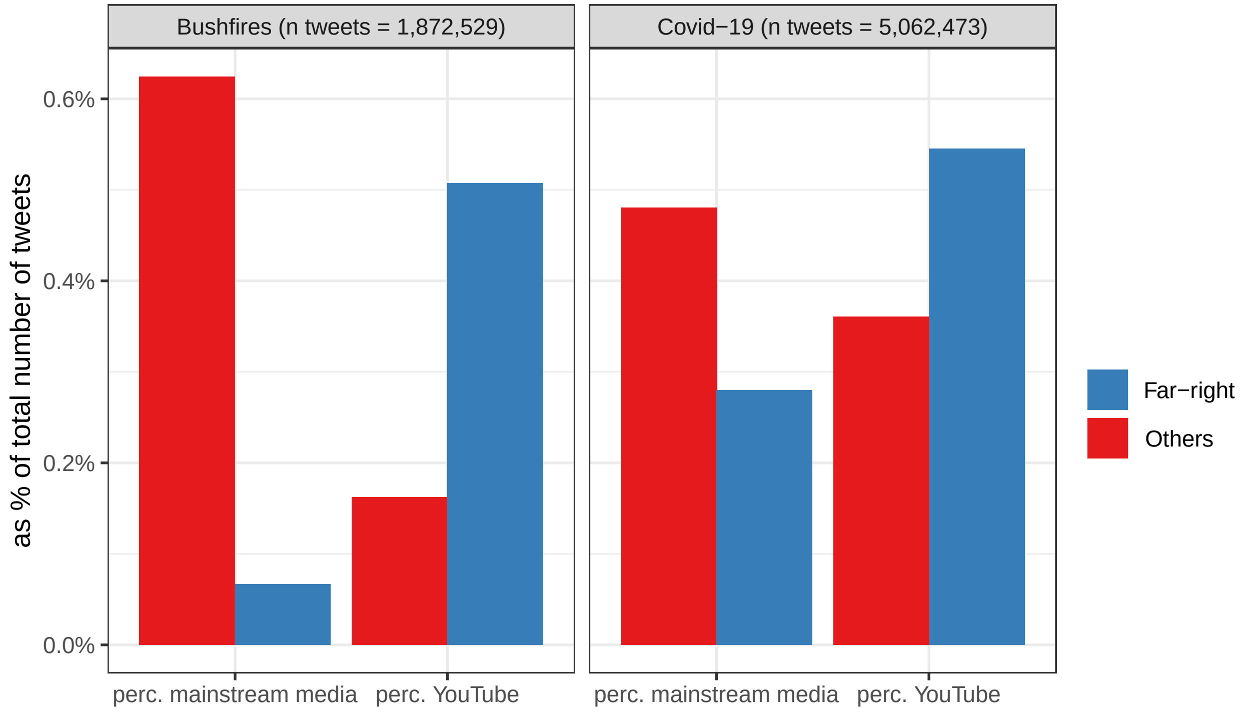
Firstly, drawing from Graham et al. (2021), we manually inspected 50 accounts randomly selected from the list of 208 far-right accounts. We used a codebook drawn from the previous authors, which checked profile information for mismatches between the "name" and "screen_name", between "name" and "url" or between the "description" and the "location"; checking if the "description" provided any biographical information; by reverse searching the profile's image and banner to check if these could be associated with any other person; finally inspecting up to 25 tweets from each account to determine whether the account has been tweeting about more than one topic. Second, we used a bot detection algorithm, using the tweetbotornot2 R package (Ferrara et al., 2016), to estimate the probability of each account being a bot based on a large number of features of the account and its posting patterns. Third, drawing from Kumar et al. (2017) and Keller et al. (2020), we analysed co-tweet and co-retweet networks to detect the presence of co-retweeting behaviours (i.e. retweeting within a few seconds from the retweet of another account) or co-tweeting behaviours (i.e. tweeting the exact same message).

4 Results

4.1 Tweeting behaviour of far-right versus other users

Figure 1 illustrates the proportion of tweets containing a URL linked to a mainstream media website and to a YouTube website (as discussed, a common source of misinformation) across both crises. The findings show far-right accounts shared relatively less links to the domains of news media organisations and relatively more links to YouTube videos. This is in line with the far-right anti-establishment stand and critique towards the mainstream media and the epistemic system that these media supposedly represent. This is relevant when we come to consider and measure the performance of far-right accounts relative to other users. Also relevant is the increase of popularity of YouTube links in Period 2 among other users along with a decrease of popularity of mainstream media links.

Figure 1: Tweets containing URLs to mainstream media websites and YouTube videos



4.2 RQ1: Did far-right accounts overperform on Twitter, becoming more central in the conversation?

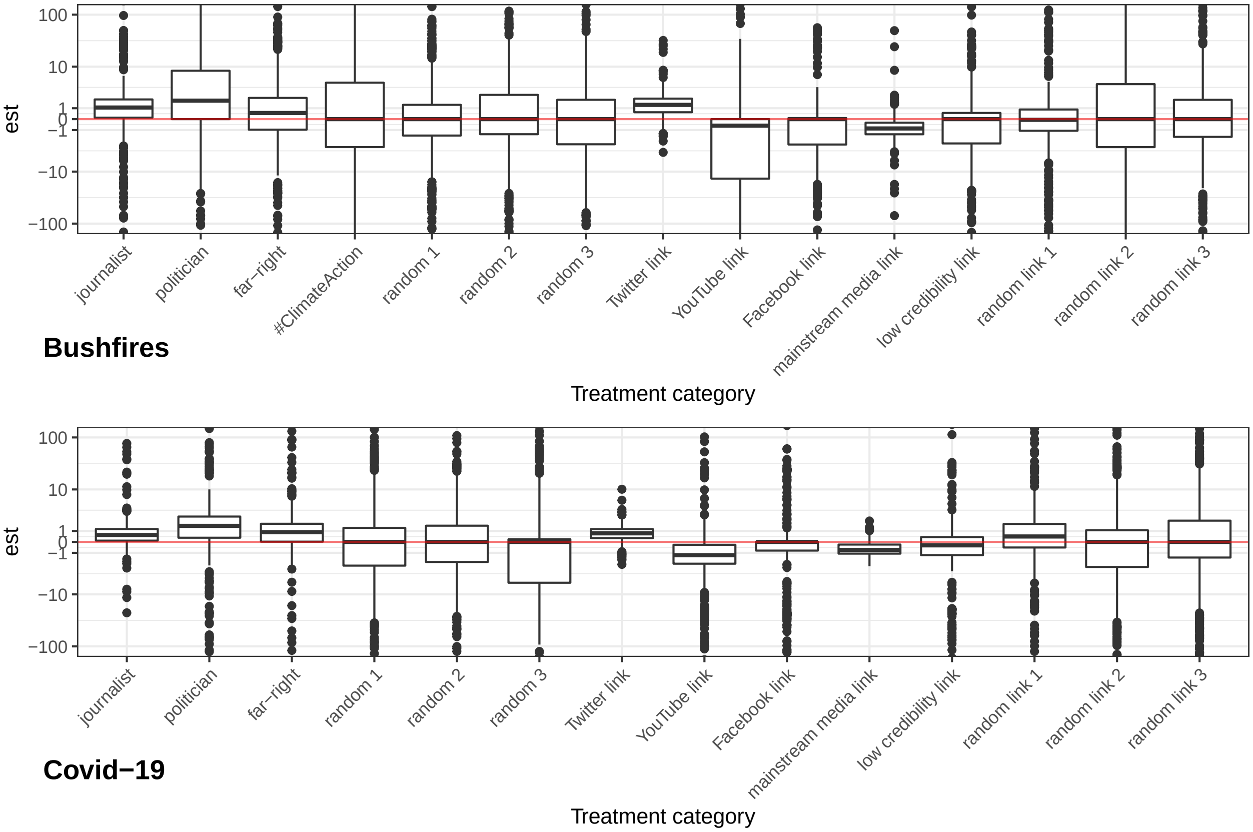
Figure 2 summarises the distribution of the 516 estimates for different categories of tweets (indicated on the horizontal axis) over the two periods. The vertical axis indicates performance in terms of retweets for that category relatively to that of other comparable (i.e. matched) tweets. In addition to far-right accounts, we ran the performance analysis on tweets published by journalists, politicians, and #ClimateAction activists as well as those containing specific typologies of URLs. This allows us to *longitudinally* compare performance but also to conduct a *cross-sectional* comparison for different typologies of tweets in the same period.

Tweets from the categories listed above were not found to perform differently from the control group when the median is zero. Instead we argue that a category of tweets can be considered to overperform when the lower hinge of the box (the 25th percentile of the distribution) is on or above zero, and it underperforms if the upper hinge (the 75th percentile) is on or below zero.

As a robustness test we ran the analysis on three sets of random tweets (in the Figure “random 1-3” and “random link 1-3”). As expected random tweets perform in line with the control group, with the median equalling zero.

When observed longitudinally, all the categories mantain a similar performance, with the exception of tweets published by far-right accounts. In Period 2, the far-right community joined journalists and politicians among the categories of accounts to overperform relative to the control group. In Period 2, tweets from far-right accounts significantly overperformed in 32% of the 516 regressions we ran while journalists overperformed in 35% (in Period 2, random tweets overperformed in only 19% of regressions). These results point to an overperformance of far-right accounts in Period 2 in a manner which is comparable to the overperformance of journalists.

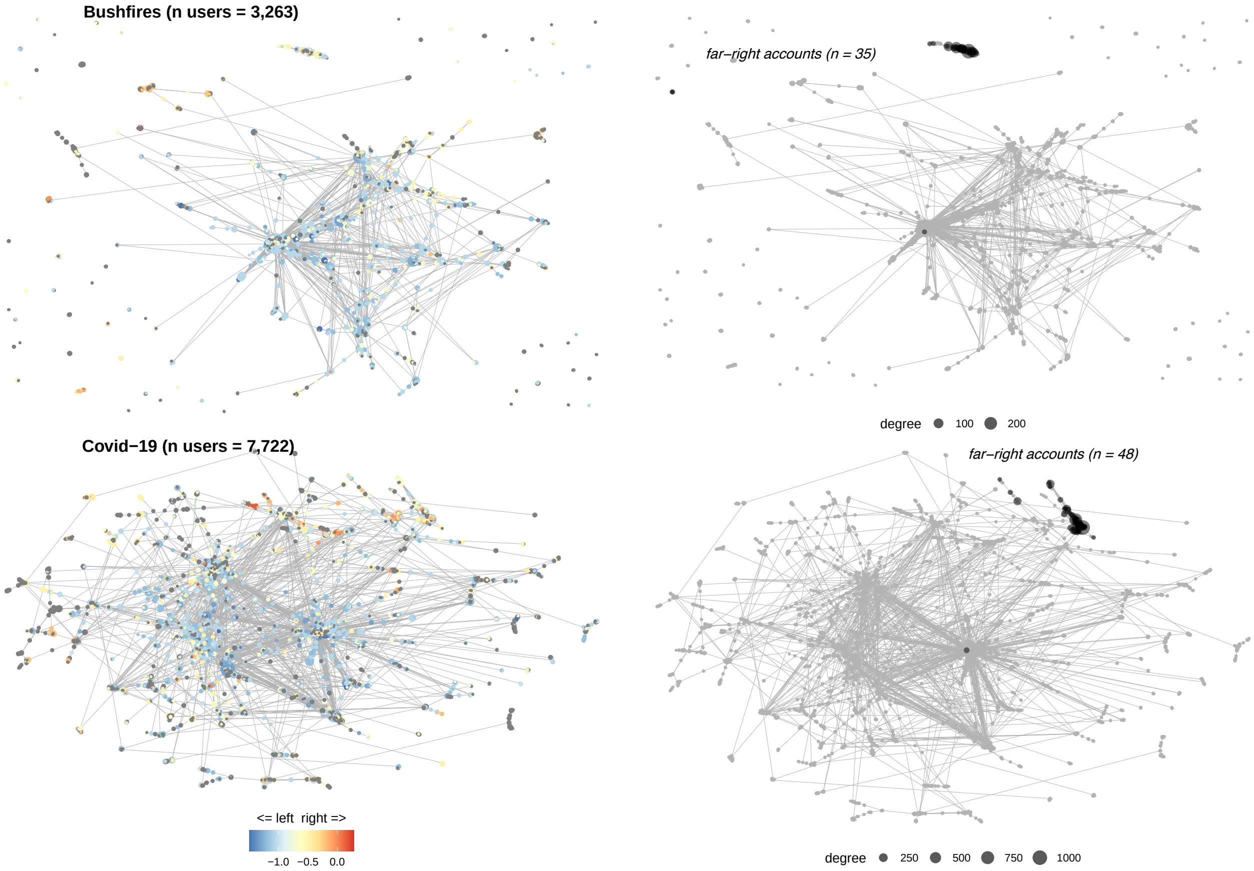
Figure 2: Distribution of regression estimates based on 'nearest neighbor matching' used for the performance analysis of different categories of tweets.



In line with the overperformance of far-right accounts in Period 2, we expect a more central position of their accounts in the conversation. This is found with an analysis of the mutual retweet networks. This is presented in Figure 3 where the position of 35 and 48 far-right accounts respectively is indicated.

In Period 1, far-right accounts are disconnected from the rest of the network and segregated in a peripheral community with no bridge to the rest of the network. In Period 2, the far-right is instead connected to the rest of the network, even if from a peripheral position. This is more evident when we group users into communities. In Figure 4, we observe the connections and relative position

Figure 3: Mutual retweet networks



of communities instead of single users. In Period 1, 27 far-right accounts (out of 35) are clustered in the third largest community of the network (with 160 members), which notably has no mutual retweets with accounts external to that community. In Period 2, 42 far-right accounts (out of 48) are still clustered in the third largest community (with 302 members), but in this case, they are connected to five other communities, including three of the four largest communities. The mere presence of a connection is in line with findings about an increase in the degree of influence of the far-right community in Period 2.

4.3 RQ2: Do we observe a change in the degree of information disorder in the two conversations?

To understand the degree of attention dispersal and the entry of a large number of competing information sources in the Twitter conversation during COVID-19, which is our first measure of information disorder, we rely again on network analysis but this time focus on the communities of users computed from User-to-URLs networks.

Figure 4: Communities of the mutual retweet network

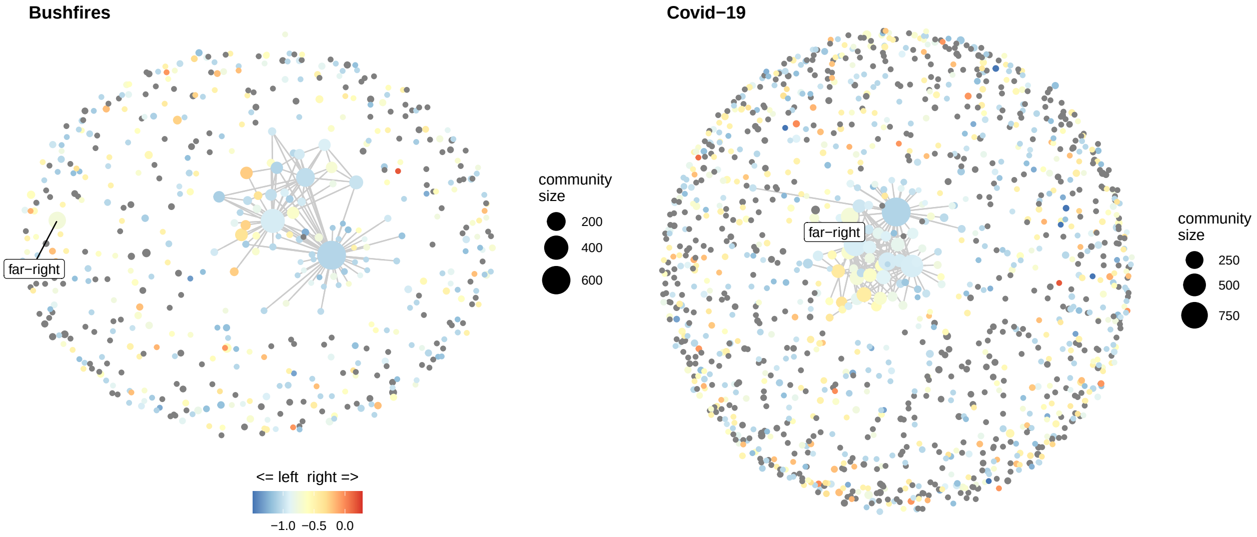


Figure 5 charts the largest communities in each network (respectively with more than 50 members in Period 1 and more than 500 members in Period 2) and the average political orientation of their members (on the horizontal axis) across the two crises. In Period 2 the more far-right actors are more central and the inclusion of more publications across the left-right spectrum has resulted in a greater dispersal of attention. This tendency to disperse attention over more diverse content is confirmed by network modularity, a global measure of the goodness of the clustering (Clauset et al., 2004), which is higher in Period 2 than in Period 1 (0.35 vs 0.17), indicating a stronger fragmentation of attention during the Covid pandemic. We consider this dispersal of attention across a larger number of competing sources to produce the type of confusion and disorientation that typify information disorder.

In Figure 6, we map our bidimensional information disorder index over the two periods. Based on our measurement, we clearly associate Period 2 with a higher degree of information disorder on both measures. Users tend to consistently refer to a wider number of domains while the proportion of tweets referring to institutional and news media sources is lower across all the daily observations.

4.4 RQ3: Can we detect coordinated, inauthentic behaviour among the far-right accounts active in both crises?

While it is difficult to measure evidence of coordinated, inauthentic behaviour in our sample given Twitter's removal of some of the accounts (which may be for this reason or others), our analysis of remaining accounts found little evidence of such behaviour. Nonetheless, because of

Figure 5: Community of users based on the URLs shared in their tweets with each red dot indicating a far-right account

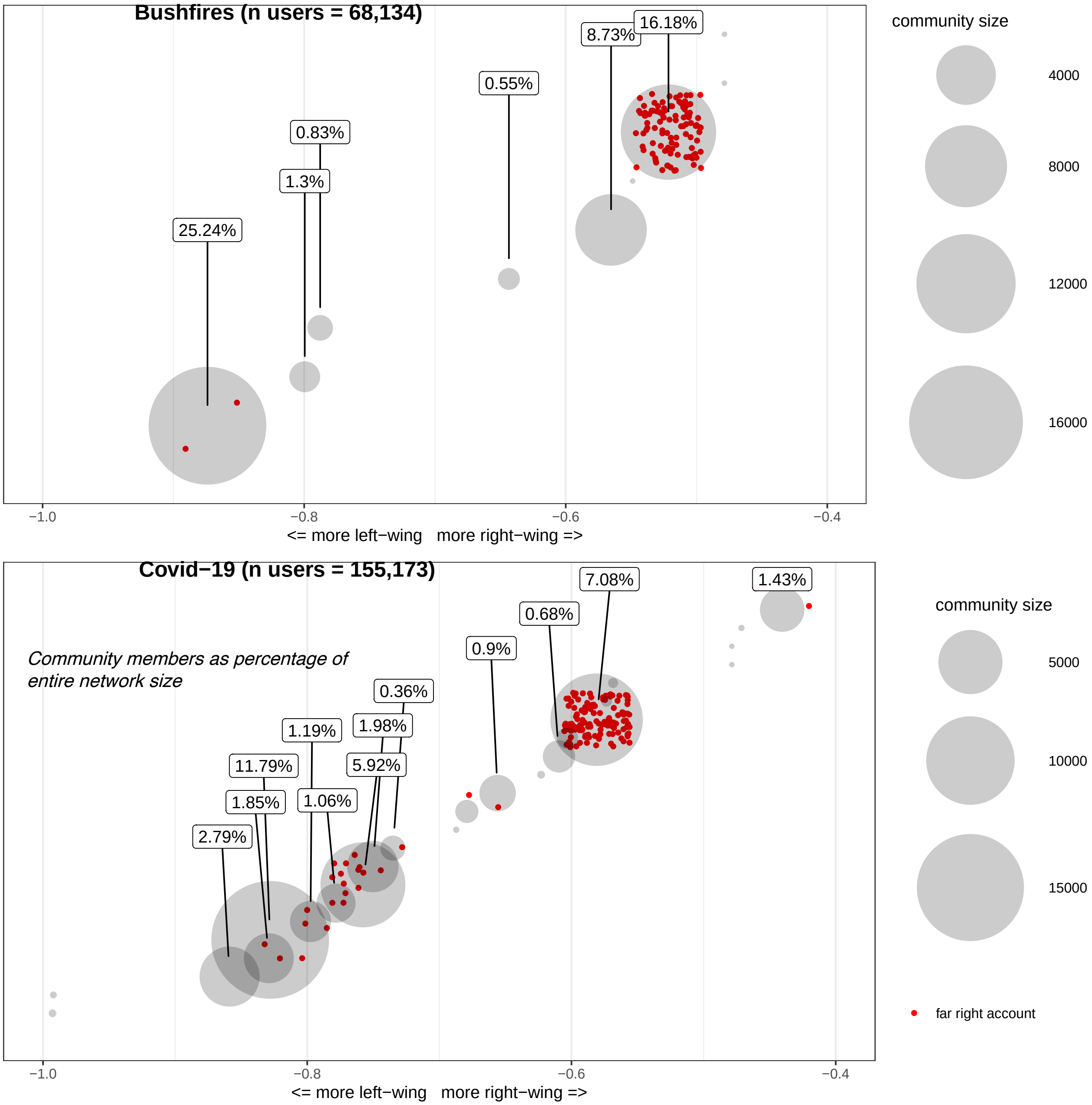
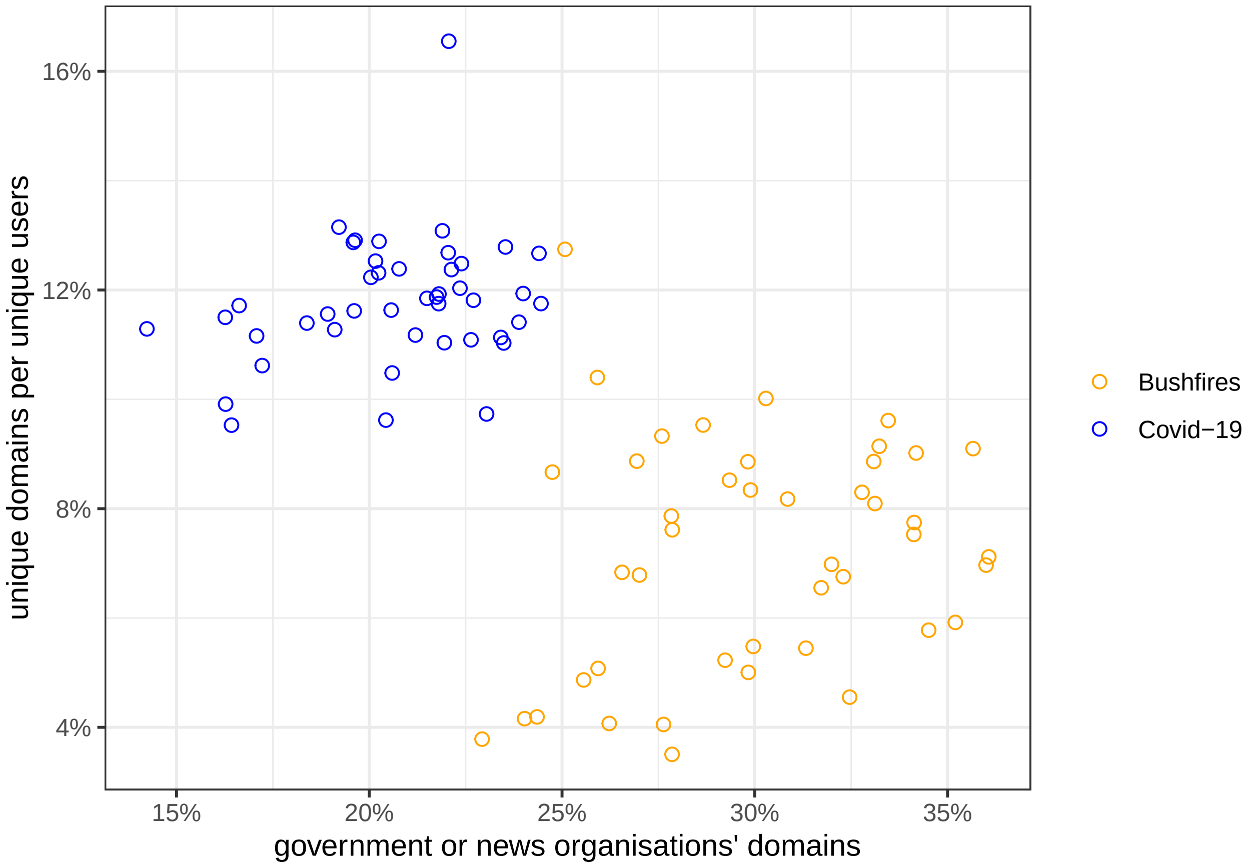


Figure 6: The two dimensions of information disorder measured daily in the two crisis



the characteristics of far-right online communities we are not able to unequivocally determine the authenticity of these accounts either.

Using results from the seven binary questions in our codebook, we constructed an index where 100% indicates that all answers that could be answered about an account point towards inauthenticity. The average for the 50 accounts is 17% across the seven questions and 23% if we only include questions for which information was available. Most of the accounts are found to present biographical information in their profile descriptions (88%) and to discuss different topics in their tweets (84%), which tends to be a marker of authenticity. Signs that might indicate inauthentic accounts were also present (23 out of 50), for example, more than 39% used images that can be found somewhere else on the web. Still, returning to these images, we found that no profile picture that appeared elsewhere on the web was planted with an intent to present a false identity as they did not represent people but instead were mostly images that served to characterise the political leaning of the account.

Second, using a bot detection algorithm (Ferrara et al., 2016) we found no clear bots detected among the 208 far-right accounts, with an estimated likelihood below 0.7% in Period 1 and 0.8% in Period 2. Third, our analysis of co-tweet (tweeting the same message) and co-retweet (retweeting the same tweet at about the same time) networks do not indicate any form of inauthentic coordination among far-right accounts or in support of their content. Far-right accounts are found to have engaged in near-simultaneous co-retweeting only eight times in Period 1 and only 14 times in Period 2 while publishing 7674 and 28,203 retweets in the two periods respectively. Only about 1% of all tweets published by far-right accounts in the two periods contain the text of another tweet but these tweets are mostly limited to single-hashtag tweets. Finally, the timestamp analysis of retweets of far-right content indicates that the time distribution of the retweeting cascade is almost identical to that of other tweets.

5 Discussion and Conclusion

In times of information crisis, such as the 2019-2020 summer of bushfires in Australia, and the first year of the global COVID-19 pandemic, concern regarding the credibility of information circulating among anxious publics is at the forefront of authorities responses, with the consequences of false or misleading content receiving greater public attention than authoritative sources driving or exacerbating a range of social, health and epistemic crises. The task of responding is a complex one, with the motivations and techniques of some communities and actors undermining efforts to inform the public and create consensus around goals to maintain public health and safety.

In this study, we aimed to understand some of the dynamics around information disorders (Wardle & , 2017) during crisis events by focusing primarily on agents (far-right actors) and diffusion of low-credibility sources. We provide a comparative analysis of the performance and behaviour of a small community of far-right Twitter accounts across two crisis periods (the Australian bushfires and COVID-19) – and related public Twitter conversations – where the circulation of misinformation created varying degrees of information disorder. The purpose of doing this was to: assess the tweeting behaviours of a community (far-right) found in the literature to share false and misleading information during crisis (Marwick & Lewis, 2018; Graham & Keller, 2020; Graham et al., 2020), to measure the centrality and performance of these accounts across the two public conversations, and to see if there are any connections between a high degree of information disorder and far right performance. Finally, we identify whether accounts were involved in coordinated, inauthentic behaviour, a major cause of concern which has contributed to a range of information disorders.

Two main findings emerged. First, we found that, while far-right accounts were disconnected from the rest of the Twitter user-network and their performance was similar to other accounts in Period 1 (bushfires), in Period 2 (during COVID-19) this changed considerably, with far-right accounts overperforming relative to other accounts, and moving to a more central location of influence. Worryingly, their relative overperformance was comparable to that of journalists. While journalists mainly tweeted credible news stories from trusted, mainstream news sources, we found that links shared by far right accounts pointed away from mainstream sources and toward less credible sources, i.e. YouTube. The latter finding is significant given that the literature has shown that fringe actors have benefited from YouTube and Twitter's recommendation algorithms to amplify and spread misinformation during COVID-19 and other related disinformation campaigns (Li et al, 2020; Wilson & Starbird, 2020; Marwick & Lewis, 2018). Second, by operationalising our information disorder index, we found that the Twitter conversation in Period 2 was more dispersed and had a higher degree of content being shared from non-authoritative sources, contributing to a higher degree of information disorder. The final question that we asked related to the presence of coordinated, inauthentic behaviour among the far-right accounts. The range of measurements we used here produced more ambivalent results. While we identified some suspicious accounts, trying to deduce the authenticity of accounts, and the motivations behind tweeting behaviour is a difficult task (Keller et al., 2020; Wardle & Derakhshan, 2017).

The paper does have some limitations. The results are based on analysis of a specific context - the Australian Twitter conversation - surrounding two crisis events and a relatively small community of far-right users. Moreover, if the first crisis is national, the second crisis is global with significantly more content (national and international) competing for local attention. We also acknowledge disagreement in the scholarship regarding source based classification and whether it is the most accurate instrument for assessing diffusion of misinformation and information disorder (Gallotti et al., 2020; Shahi et al., 2021). We recognise at least two limitations of this method. First, it necessarily collapses differences between content published on the same website by different users, that is, a YouTube video will always be considered low quality information even if published by the World Health Organisation and a government source will usually be considered high quality even if the leader is found to regularly publish false and misleading information. Second, we do not use existing lists of low-credibility domains such as that compiled by *Media Bias / Fact Check* (see again Yang et al., 2021), which as mentioned above we only use to measure the performance of different types of tweets. The main reason for not using the score from low-credibility lists is that these lists are limited to the most popular US or global sources and using them would have necessitated coding only a small fraction of the domains that appear in the Twitter conversation we analysed. Finally, we acknowledgre that future work could assess the structural importance and significance of the new connection emerging in Period 2 between the far-right community and the rest of the conversation. This is beyond the scope of this study and is noted as a limitation which will be addressed in future research, possibly by reducing the conversation network to its backbone (see for example Serrano et al., 2009).

Despite the limitations, we feel the paper offers unique insights and a measure of information disorder that may be used in other information crises where epistemic gaps emerge. For example, by focusing on the domain where the content was published instead of on the content we use a method that is applicable to very large datasets and has the best possibility of the findings being generalisable (Lazer et al., 2018; Shao et al., 2018; Yang et al., 2021). The novelty of being able to monitor a community that remained active across two proximate crises and to analyse performance and the quality of information they tweeted is also a unique contribution that this article makes. Based on the findings, we make the argument that information disorder is likely to be exacerbated by epistemic gaps that emerged early in the pandemic, when interpretations were not generally accessible. These gaps have been shown in the literature to encourage the production and adoption of misinformation and misinterpretations in other similar events that drive large amounts of attention (see Douglas et al., 2017). This is likely what happened in the first months of the COVID-19 pandemic as the epistemic system struggled to keep up with the pace of the crisis, the multiplication of sources of information as well as the reduction in the centrality of authoritative sources such as governmental agencies and news media.

In conclusion, if information crises are always an opportunity for actors offering interpretations that deviate from the scientific or authoritative consensus, and a space where the existence of coordinated behaviour to try to win the information war, or at least unsettle consensus opinion, is often prevalent, this research found that it is the degree of information disorder that is the most prominent factor that contributes to the severity of the information crisis and their margin of success. Connecting information crises to epistemic uncertainty and the emergence of epistemic gaps, we argue that information disorder can indicate that the epistemic system is failing to reach the public and that the public is turning to alternative sources of information during the global COVID-19 pandemic. These gaps represent demands for information that are not addressed by the epistemic system. The current information and communication infrastructure is not designed for slowing down the diffusion of content to allow the quality of authoritative information to catch up with public demand for information (see the concept of "data voids" in Golebiewski & Boyd, 2019). In fact, during crisis events, as people try to reduce the distress caused by uncertainty by searching for meaningful explanation (see Douglas et al., 2017), it is expected that this is when the infrastructure will experience the most duress. This is likely to result in even more information disorder as our findings indicate. In light of this, further research is needed to provide insights into the causal mechanisms that result in more disinformation being distributed on the Internet during an information crisis. Mainly we need to understand what or who is principally responsible for driving information disorder. Do fringe, coordinating actors significantly increase the degree of information disorder or are they mere beneficiaries? How much disorder is created by Internet platforms through their recommender systems, which respond to demands for more information, despite gaps in knowledge availability? Finally, how should epistemic institutions behave in an information ecosystem that does not tolerate epistemic uncertainty?

Acknowledgments

Previous versions of this article were presented at the ADM+S Misinformation Workshop (2022) and at the International Network for Social Network Analysis Sunbelt Conference (2022). We are grateful to Dr Emily Booth for support with research assistance and to Dr Michael Vaughan for his feedback on an early version of this paper. We have benefited from funds provided with the Facebook Content Policy Research Award.

The approval number of the ethics application for this study is UTS HREC REF NO. ETH19-3877.

References

- Benkler, Y., Faris, R., & Roberts, H. (2018). *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press.
- Busbridge, R., Moffitt, B., & Thorburn, J. (2020). Cultural Marxism: Far-right conspiracy theory in Australia's culture wars. *Social Identities*, *26*(6), 722–738. https://doi.org/10.1080/13504630.2020.1787822
- Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., & Scala, A. (2020). The COVID-19 social media infodemic. *Scientific Reports*, *10*(1), Article 1. https://doi.org/10.1038/s41598-020-73510-5
- Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very

large networks. *Physical Review E*, *70*(6), 066111. https://doi.org/10.1103/PhysRevE.70.066111

- Davis, M. (2019). Transnationalising the anti-public sphere: Australian anti-publics and reactionary online media. In M. Peucker & D. Smith (Eds.), *The Far-Right in contemporary Australia* (pp. 127–149). Palgrave Macmillan. https://doi.org/10.1007/978-981-13-8351-9
- Douglas, K. M., Sutton, R. M., & Cichocka, A. (2017). The psychology of conspiracy theories. *Current Directions in Psychological Science*, *26*(6), 538–542. https://doi.org/10.1177/0963721417718261
- Ferrara, E. (2020). What types of COVID-19 conspiracies are populated by Twitter bots? *First Monday*. https://doi.org/10.5210/fm.v25i6.10633
- Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, *59*(7), 96–104. https://doi.org/10.1145/2818717
- Gallotti, R., Valle, F., Castaldo, N., Sacco, P., & De Domenico, M. (2020). Assessing the risks of 'infodemics' in response to COVID-19 epidemics. *Nature Human Behaviour*, *4*(12), Article 12. https://doi.org/10.1038/s41562-020-00994-6
- Giglietto, F., Righetti, N., Rossi, L., & Marino, G. (2020). It takes a village to manipulate the media: Coordinated link sharing behavior during 2018 and 2019 Italian elections. *Information, Communication & Society*, *23*(6), 867–891. https://doi.org/10.1080/1369118X.2020.1739732
- Goldman, A. I. (2011). A guide to social epistemology. In A. I. Goldman & D. Whitcomb (Eds.), *Social epistemology: Essential readings* (pp. 11–37). Oxford University Press.
- Golebiewski, M., & Boyd, D. (2019). *Data voids: Where missing data can easily be exploited*.
- Graham, T., Bruns, A., Angus, D., Hurcombe, E., & Hames, S. (2021). #IStandWithDan versus #DictatorDan: The polarised dynamics of Twitter discussions about Victoria's COVID-19 restrictions. *Media International Australia*, *179*(1), 127–148. https://doi.org/10.1177/1329878X20981780
- Graham, T., Bruns, A., Zhu, G., & Campbell, R. (2020). *Like a virus: The coordinated spread of coronavirus disinformation*. Centre for Responsible Technology, The Australia Institute. https://apo.org.au/node/305864
- Graham, T., & Keller, T. R. (2020, January 10). *Bushfires, bots and arson claims: Australia flung in the global disinformation spotlight*. The Conversation. http://theconversatio

n.com/bushfires-bots-and-arson-claims-australia-flung-in-the-global-disinformation-spotlight-129556

- Gruzd, A., & Mai, P. (2020). Going viral: How a single tweet spawned a COVID-19 conspiracy theory on Twitter. *Big Data & Society*, *7*(2), 205395172093840. https://doi.org/10.1177/2053951720938405
- Hameleers, M., Humprecht, E., Möller, J., & Lühring, J. (2021). Degrees of deception: The effects of different types of COVID-19 misinformation and the effectiveness of corrective information in crisis times. *Information, Communication & Society*, *0*(0), 1–17. https://doi.org/10.1080/1369118X.2021.2021270
- Ho, D., Imai, K., King, G., & Stuart, E. A. (2011). Matchit: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*, 1–28. https://doi.org/10.18637/jss.v042.i08
- Hutchinson, J. (2021). The new-far-right movement in Australia. *Terrorism and Political Violence*, *33*(7), 1424–1446. https://doi.org/10.1080/09546553.2019.1629909
- Jamieson, K. H. (2018). *Cyberwar: How Russian hackers and trolls helped elect a president: What we don't, can't, and do know*. Oxford University Press.
- Keller, F. B., Schoch, D., Stier, S., & Yang, J. (2020). Political astroturfing on Twitter: How to coordinate a disinformation campaign. *Political Communication*, *37*(2), 256–280. https://doi.org/10.1080/10584609.2019.1661888
- Kim, D., Graham, T., Wan, Z., & Rizoiu, M.-A. (2019). Analysing user identity via time-sensitive semantic edit distance (t-SED): A case study of Russian trolls on Twitter. *Journal of Computational Social Science*, *2*(2), 331–351. https://doi.org/10.1007/s42001-019-00051-x
- Kong, Q., Booth, E., Bailo, F., Johns, A., & Rizoiu, M.-A. (2022). Slipping to the extreme: A mixed method to explain how extreme opinions infiltrate online discussions. *Proceedings of the International AAAI Conference on Web and Social Media*, *16*(1), 524–535.
- Kumar, S., Cheng, J., Leskovec, J., & Subrahmanian, V. S. (2017). An army of me: Sockpuppets in online discussion communities. *Proceedings of the 26th International Conference on World Wide Web*, 857–866. https://doi.org/10.1145/3038912.3052677
- Laudan, L. (2001). Epistemic crises and justification rules. *Philosophical Topics*, *29*(1/2), 271–317.
- Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A.,

Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

- Li, H. O.-Y., Bailey, A., Huynh, D., & Chan, J. (2020). YouTube as a source of information on COVID-19: A pandemic of misinformation? *BMJ Global Health*, *5*(5), e002604. https://doi.org/10.1136/bmjgh-2020-002604
- Marwick, A. E., & Lewis, B. (2017). *Media manipulation and disinformation online*. Data & Society Research Institute. https://datasociety.net/library/media-manipulation-and-disinfo-online/
- Nohrstedt, S. A. (1991). The information crisis in Sweden after Chernobyl. *Media, Culture & Society*, *13*(4), 477–497. https://doi.org/10.1177/016344391013004004
- Park, S., Fisher, C., Lee, J. Y., McGuinness, K., Sang, Y., O'Neil, M., Jensen, M., McCallum, K., & Fuller, G. (2020). *Digital news report: Australia 2020* (Australia) [Report]. News and Media Research Centre. https://apo.org.au/node/305057
- Rizoiu, M.-A., Graham, T., Zhang, R., Zhang, Y., Ackland, R., & Xie, L. (2018). #DebateNight: The Role and Influence of Socialbots on Twitter During the 1st 2016 U.S. Presidential Debate. *ArXiv:1802.09808 [Cs]*. http://arxiv.org/abs/1802.09808
- Ross, K. (2020, August 25). Why QAnon is attracting so many followers in Australia—And how it can be countered. *The Conversation*. http://theconversation.com/why-qanon-is-attracting-so-many-followers-in-australia-and-how-it-can-be-countered-144865
- Serrano, M. Á., Boguñá, M., & Vespignani, A. (2009). Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences*, *106*(16), 6483–6488. https://doi.org/10.1073/pnas.0808904106
- Shahi, G. K., Dirkson, A., & Majchrzak, T. A. (2021). An exploratory study of COVID-19 misinformation on Twitter. *Online Social Networks and Media*, *22*, 100104. https://doi.org/10.1016/j.osnem.2020.100104
- Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, *9*(1), Article 1. https://doi.org/10.1038/s41467-018-06930-7
- Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making* (DGI(2017)09). Council of Europe. http://tverezo.info/wp-content/uploads/2017/11/PREMS-162317-GBR-2018-Report-desinformation-A4-BAT.pdf

- Weber, D., Falzon, L., Mitchell, L., & Nasim, M. (2022). Promoting and countering misinformation during Australia's 2019–2020 bushfires: A case study of polarisation. *Social Network Analysis and Mining*, *12*(1), 64. https://doi.org/10.1007/s13278-022-00892-x
- Wilson, T., & Starbird, K. (2020). Cross-platform disinformation campaigns: Lessons learned and next steps. *Harvard Kennedy School Misinformation Review*, *1*(1). https://doi.org/10.37016/mr-2020-002
- Woolley, S., & Howard, P. N. (Eds.). (2019). *Computational propaganda: Political parties, politicians, and political manipulation on social media*. Oxford University Press.
- Yang, K.-C., Pierri, F., Hui, P.-M., Axelrod, D., Torres-Lugo, C., Bryden, J., & Menczer, F. (2021). The COVID-19 Infodemic: Twitter versus Facebook. *Big Data & Society*, *8*(1), 20539517211013860. https://doi.org/10.1177/20539517211013861
- Yang, K.-C., Torres-Lugo, C., & Menczer, F. (2020). *Prevalence of low-credibility information on Twitter during the Covid-19 outbreak*. *2020*, 16. https://doi.org/10.36190/2020.16
- Zarocostas, J. (2020). How to fight an infodemic. *The Lancet*, *395*(10225), 676. https://doi.org/10.1016/S0140-6736(20)30461-X
- Zola, P., Cola, G., Martella, A., & Tesconi, M. (2022). Italian top actors during the COVID-19 infodemic on Twitter. *International Journal of Web Based Communities*, *18*(2), 150–172. https://doi.org/10.1504/IJWBC.2022.124783