Social Network Analysis and Mining (2021) 11:111
https://doi.org/10.1007/s13278-021-00815-2

ORIGINAL ARTICLE

Amplifying influence through coordinated behaviour in social networks

Derek Weber[1,2] · **Frank Neumann**[1]

Received: 13 February 2021 / Revised: 4 August 2021 / Accepted: 6 August 2021
© Crown 2021

Abstract

Political misinformation, astroturfing and organised trolling are online malicious behaviours with significant real-world effects that rely on making the voices of the few sounds like the roar of the many. These are especially dangerous when they influence democratic systems and government policy. Many previous approaches examining these phenomena have focused on identifying campaigns rather than the small groups responsible for instigating or sustaining them. To reveal latent (i.e. hidden) networks of cooperating accounts, we propose a novel temporal window approach that can rely on account interactions and metadata alone. It detects groups of accounts engaging in various behaviours that, in concert, come to execute different goal-based amplification strategies, a number of which we describe, alongside other inauthentic strategies from the literature. The approach relies upon a pipeline that extracts relevant elements from social media posts common to the major platforms, infers connections between accounts based on criteria matching the coordination strategies to build an undirected weighted network of accounts, which is then mined for communities exhibiting high levels of evidence of coordination using a novel community extraction method. We address the temporal aspect of the data by using a windowing mechanism, which may be suitable for near real-time application. We further highlight consistent coordination with a sliding frame across multiple windows and application of a decay factor. Our approach is compared with other recent similar processing approaches and community detection methods and is validated against two politically relevant Twitter datasets with ground truth data, using content, temporal, and network analyses, as well as with the design, training and application of three one-class classifiers built using the ground truth; its utility is furthermore demonstrated in two case studies of contentious online discussions.

Keywords Coordinated amplification · Coordinated behaviour · Online social networks · Information campaigns

1 Introduction

Online social networks (OSNs) have established themselves as flexible and accessible systems for activity coordination and information dissemination. This benefit was illustrated during the Arab Spring (Carvin 2012) but inherent dangers are increasingly apparent in ongoing political interference and disinformation (Howard and Kollanyi 2016; Ferrara 2017; Keller et al. 2017; Neudert 2018; Singer and Brooking 2019; Nimmo et al. 2020). Modern Strategic Information Operations (SIOs) are participatory activities, which aim to use their audiences to amplify their desired narratives, not just receive it (Starbird et al. 2019). The widespread use of social media for political communication and its identity-obscuring nature have made it a prime target for politically-driven influence, both legitimate and illegitimate. Through cyclical reporting (i.e. social media feeding stories and narratives to traditional news media, which then sparks more social media activity), social media users can unknowingly become "unwitting agents" as "sincere activists" of concerted operations (Benkler et al. 2018; Starbird and Wilson 2020). The use of *political* bots and trolls to influence the framing and discussion of issues in the mainstream media (MSM) remains prevalent (Bessi and Ferrara 2016; Woolley 2016; Woolley and Guilbeault 2018; Rizoiu et al. 2018; Cresci 2020). The use of bots and sockpuppet accounts

✉ Derek Weber
derek.weber@adelaide.edu.au;
derek.weber@dst.defence.gov.au

Frank Neumann
frank.neumann@adelaide.edu.au

[1] School of Computer Science, University of Adelaide, Adelaide, SA, Australia

[2] Defence Science and Technology Group, Adelaide, SA, Australia

Published online: 31 October 2021

to amplify individual voices above the crowd, sometimes referred to as the *megaphone effect*, requires coordinated action and a degree of regularity that may leave traces in the digital record.

Relevant research has focused on high-level analyses of campaign detection and classification (Lee et al. 2013; Varol et al. 2017; Alizadeh et al. 2020), the identification of botnets and other dissemination groups (Vo et al. 2017; Woolley and Guilbeault 2018), and coordination at the community level (Kumar et al. 2018; Hine et al. 2017; Cresci 2020). Some have considered generalised approaches to social media analytics (e.g. Weber 2019; Graham et al. 2020; Nizzoli et al. 2021; Pacheco et al. 2021), but unanswered questions regarding the clarification of coordination strategies and their detection remain. Forensic studies of SIOs and other influence campaigns using these strategies (e.g. Benkler et al. 2018; Jamieson 2020; Nimmo et al. 2020) currently require significant human input to reveal the covert ties underpinning them, and could benefit greatly from enhanced automation.

In this work, we expand upon the novel approach to detect groups engaging in potentially coordinated amplification activities, revealed through anomalously high levels of coincidental behaviour, which we presented at ASONAM'20 (Weber and Neumann 2020). Links between the group members are inferred from behaviours that, when used intentionally, are used to execute a number of identifiable coordination strategies. We use a range of techniques to validate our new approach on two relevant datasets, as well as comparison with ground truth and a synthesised dataset, and show it successfully identifies coordinating communities.

Our approach infers ties between accounts to construct *latent coordination networks* (LCNs) of accounts, using criteria specific to different coordination strategies, which are based on features common to major OSNs. The accounts may not be directly connected, thus we use the term 'latent' to mean 'hidden' when describing these connections. The inference of connections is performed solely on the accounts' interactions, i.e. not their content or friending/following behaviour, only metadata and temporal information, though it could incorporate them.

Highly coordinating communities (HCCs) are then detected and extracted from the LCN. We propose a variant of *focal structures analysis* (FSA, Şen et al. 2016) to do this, in order to take advantage of FSA's focus on finding influential sets of nodes in a network while also reducing the computational complexity of the algorithm. A window-based approach is used to enforce temporal constraints.

The following research questions guided our evaluation:

- RQ1: How can HCCs be found in an LCN?
- RQ2:How do the discovered communities differ?
- RQ3: Are the HCCs internally or externally focused?
- RQ4: How consistent is the HCC messaging?
- RQ5: What evidence is there of consistent coordination?
- RQ6: How well can HCCs in one dataset inform the discovery of HCCs in another?

This paper expands upon Weber and Neumann (2020) by providing further methodological detail and experimental validation, and case studies in which the technique is applied to new real-world Twitter datasets relating to contentious political issues, as well as consideration of algorithmic complexity and comparison with several similar techniques. Prominent among the extra validation provided is the use of machine learning classifiers to show that our datasets contain similar coordination to our ground truth, and the application of a sliding frame across the time windows as a way to search for consistent coordination.

This paper provides an overview of relevant literature, followed by a discussion of online coordination strategies and their execution. Our approach is then explained, and its experimental validation is presented. Following the validation, the algorithmic complexity and performance of the technique are presented, and two case studies are explored, demonstrating the utility of the approach with real-world politically relevant datasets, and we compare our technique to those of Pacheco et al. (2021), Graham et al. (2020), Nizzoli et al. (2021) and Giglietto et al. (2020b).

1.1 A motivating example

In the aftermath of the 2020 US Presidential election, a data scientist noticed a pattern emerging on Twitter.[1] Figure 1a shows a tweet by someone who was so upset with their partner voting for Joe Biden in the election that they decided to divorce them immediately and move to Pakistan (in the midst of the COVID-19 pandemic). This might seem an extreme reaction, but the interesting thing was that the person was not alone. The researcher had identified dozens of similar, but not always identical, tweets by people leaving for other cities but for the same reason (Fig. 1b). Analysis of these accounts also revealed they were not automated accounts. This kind of pattern of tweeting identical text is sometimes referred to as "copypasta" and can be used to give the appearance of a genuine grassroots movement on a particular issue. It had been previously used by ISIS terrorists as they approached the city of Mosul, Iraq, in 2014, which they occupied for several years after the local forces believed a giant army was invading based on the level of relevant online activity (Brooking and Singer 2016).

[1] Tweeted 2020-11-17: https://twitter.com/conspirator0/status/1328479128908132358.

(a) One upset partner. (b) Many upset partners.

Fig. 1 Copypasta tweets noticed in the aftermath of the 2020 US Presidential election, which may be a coordinated campaign to undermine confidence in American society's ability to accept electoral outcomes, or may just be a prank similar to a flashmob

It is unclear whether this "copypasta" campaign is part of a deliberate SIO, designed to damage trust in the electoral system and ability of Americans to accept the loss of a preferred political party in elections, or simply a group of like-minded jokers starting a viral gag or engaging in a kind of flashmob. At the very least, it is important to be able to identify which accounts are participating in the event, and how they are coordinating their actions.

1.2 Online information campaigns and related work

Social media has been increasingly used for communication in recent years (particularly political communication), and so the market has followed, with media organisations using it for cheap, wide dissemination of their products and consumers increasingly looking to it for news (Shearer and Grieco 2019). Over that same time period, people have begun to explore how to exploit the features of the internet and social media that bring us benefits: the ability to target marketing to specific audiences that connects businesses with the most receptive customers (e.g. Kosinski et al. 2013) also enables highly targeted non-transparent political advertising (Woolley and Guilbeault 2018; Singer and Brooking 2019) and the ability to expose people to propaganda and recruit them to extremist organisations (Badawy and Ferrara 2018; Benkler et al. 2018; The Soufan Center 2021); the anonymity that supports the voiceless in society to express themselves also enables trolls to attack others without repercussions (Hine et al. 2017; Burgess and Matamoros-Fernández 2016); and the automation that enables news aggregators also facilitates social and political bots (Ferrara et al. 2016; Woolley 2016; Cresci 2020). In summary, targeted marketing and automation coupled with anonymity provide the tools required for potentially significant influence in the online sphere, perhaps enough to swing an election, but certainly enough to be associated with real-world violence (The Soufan Center 2021; Karell et al. 2021).

Effective influence campaigns relying on these capabilities will somehow coordinate the actions of their participants. Early work on the concept of coordination by Malone and Crowston (1994) described it as the dependencies between the tasks and resources required to achieve a goal. One task may require the output of another task to complete. Two tasks may share, and require exclusive access to, a resource or they may both need to use the resource simultaneously.

At the other end of the spectrum, sociological studies of influence campaigns can reveal their intent and how they are conducted, but they consider coordination at a much higher level. Starbird et al. (2019) highlight three kinds of campaigns: *orchestrated*, centrally controlled campaigns that are run from the top down (e.g. paid teams, Chen 2015; King et al. 2017); *cultivated* campaigns that infiltrate existing issue-based movements to drive them to particular extreme positions (e.g. encouraging political violence during elections, Nimmo et al. 2020; Jamieson 2020; The Soufan Center 2021); and *emergent* campaigns arising from enthusiastic communities centred around a shared ideology (e.g. conspiracy groups and other fringe movements). Though their strategies differ, they use the same online interactions as normal users (e.g. posts, shares, mentions, hashtags, URLs), but their patterns differ. Fundamentally, however, they rely on influencing others by spreading an agenda-driven message or narrative.

At the scale of nation states, multiple disinformation campaigns may be run as part of an operation, each with different targets and different intended outcomes. The 2016 US Presidential election has received significant academic (as well as political and diplomatic) attention, and deep analysis of the interference by Russia has revealed a variety of

Table 1 Detecting inauthentic behaviour in the computer science literature

Automation	Ferrara et al. (2016), Davis et al. (2016), Grimme et al. (2017), Cresci (2020)
Campaigns	
—By content	Lee et al. (2013), Assenmacher et al. (2020), Alizadeh et al. (2020), Graham et al. (2020)
—By URL	Ratkiewicz et al. (2011), Cao et al. (2015), Giglietto et al. (2020b), Broniatowski (2021), Yu (2021)
—By hashtag	Ratkiewicz et al. (2011), Burgess and Matamoros-Fernández (2016), Varol et al. (2017), Weber et al. (2020)
Synchronicity	Chavoshi et al. (2017), Hine et al. (2017), Nasim et al. (2018), Mazza et al. (2019), Pacheco et al. (2020), Magelinski et al. (2021)
Communities	Vo et al. (2017), Morstatter et al. (2018), Gupta et al. (2019)
Political bots	Bessi and Ferrara (2016), Woolley (2016), Rizoiu et al. (2018), Woolley and Guilbeault (2018) *(particularly embeddedness and organisation)*

such campaigns were employed to promote Donald Trump, detract from Hilary Clinton, sow doubt in the country's democratic system and generally exacerbate divisions in society (Benkler et al. 2018; Mueller 2018; Jamieson 2020). Furthermore, much of the social media activity in particular was conducted by accounts made to look like average Americans, including "personable swing-voters" (p. 134, Jamieson 2020) and comparatively simple analyses of individual accounts over long periods has revealed how they were used to build audiences susceptible to their narratives (Dawson and Innes 2019). America is clearly not the only target—campaigns have been directed across any national border as well as within (Woolley and Howard 2018; Singer and Brooking 2019; Nimmo et al. 2020). Many of the analyses mentioned in these works rely on direct connections between entities (e.g. Benkler et al.'s mentions of articles and YouTube videos and Nimmo et al's follower networks, and studies of retweet and mention networks in chapters of Woolley and Howard's book), but Jamieson makes it clear that covert or at least indirect behaviour-related connections were a key part of the Russian operation during the 2016 US presidential election.

Disinformation campaigns effectively trigger human cognitive heuristics, such as individual and social biases to believe what we hear first (*anchoring*) and what we hear frequently and can remember easily (*availability* cascades) (Tversky and Kahneman 1973; Kuran and Sunstein 1999); thus the damage is already done by the time lies are exposed. This is especially true if they are promoted under the guise of authority, such as from accounts purporting to be media outlets, like `@TodayPittsburgh` or `@KansasDailyNews` (p. 188, Miller 2018). Persuasive messaging also relies on emotion, especially fear, and appeals to religion (Jamieson 2020), and have been effective even when such claims border on the ridiculous and conspiratorial (The Soufan Center 2021). Recent experiences of false information moving beyond social media during Australia's 2019–2020 bushfires highlight that identifying these campaigns as they occur can aid OSN monitors and the media to better inform the public (Graham and Keller 2020; Weber et al. 2020).

In between task level coordination and entire SIOs, at the level of social media interactions, as demonstrated by Graham and Keller (2020), we can directly observe the online actions and effects of such activities, and infer links between accounts based on pre-determined criteria. Relevant efforts in computer science have focused on a variety of methods and domains (see Table 1). These efforts have uncovered a new field of research: the computer science study of the "orchestrated activities" of accounts in general, as Grimme et al. (2018) put it, regardless of their degree of automation (Cresci et al. 2017; Alizadeh et al. 2020; Nizzoli et al. 2021; Vargas et al. 2020). It must be noted that bot activity, even coordinated activity, may be entirely benign and even useful (Ferrara et al. 2016; Graham and Ackland 2017).

Though some studies have observed the existence of strategic behaviour in and between online groups (e.g. Keller et al. 2017; Kumar et al. 2018; Hine et al. 2017; Keller et al. 2019; Giglietto et al. 2020b; Broniatowski 2021), the challenge of identifying a broad range of their interaction strategies and their underpinning execution methods remains to be fully explored, especially as new strategies are constantly be devised (Nimmo et al. 2020).

Inferring social networks from OSN data requires attendance to the temporal aspect to understand information (and influence) flow and degrees of activity (Holme and Saramäki 2012). Real-time processing of OSN posts can enable tracking narratives via text clusters (Assenmacher et al. 2020), but to process networks requires graph streams (McGregor 2014) or window-based pipelines (e.g. Weber 2019), otherwise processing is limited to post-collection activities (Graham et al. 2020; Alizadeh et al. 2020; Vargas et al. 2020; Pacheco et al. 2021).

This work contributes to the identification of interaction-based strategic coordination behaviours observable over relatively short time frames, along with a general technique to enable detection of groups using them. As such, this enhances the toolbox of techniques available to higher level explorations of information campaigns and operations (e.g. Benkler et al. 2018; Jamieson 2020; Nimmo et al. 2020; The Soufan Center 2021).

Table 2 Coordinated amplification strategies

Pollution	Flooding a community with repeated or objectionable content, causing the OSN to shut it down
Observed by	(Ratkiewicz et al. 2011; Woolley 2016; Hegelich and Janetzko 2016; Hine et al. 2017; Nasim et al. 2018; Fisher 2018; Mariconti et al. 2019)
Boost	Heavily reposting or duplicating content to make it appear popular
Observed by	(Ratkiewicz et al. 2011; Cao et al. 2015; Varol et al. 2017; Vo et al. 2017; Gupta et al. 2019; Keller et al. 2019; Graham et al. 2020; Assenmacher et al. 2020)
Bully	Groups engaging in organised harassment of an individual or community.
Observed by	(Ratkiewicz et al. 2011; Burgess and Matamoros-Fernández 2016; Hine et al. 2017; Kumar et al. 2018; Datta and Adar 2019; Mariconti et al. 2019)

2 Coordinated amplification strategies

Influencing others online, especially on political and social issues, relies on two primary mechanisms to maximise the reach of a given narrative thus amplifying its effect: *mass dissemination* and *engagement*. For example, an investigation of social media activity following UK terrorist attacks in 2017[2] identified accounts promulgating contradictory narratives, inflaming racial tensions and simultaneously promoting tolerance to sow division. By engaging aggressively, the accounts drew in participants who then spread the message.

Mass dissemination aims to maximise audience, to convince through repeated exposure and, in the case of malicious use, to cause outrage, polarisation and confusion, or at least attract attention to distract from other content.

Engagement is a form of dissemination that solicits a response. It relies on targeting individuals or communities through mentions, replies and the use of hashtags as well as rhetorical approaches that invite responses (e.g. inflammatory comments or, as present in the UK terrorist example above and observed by Nimmo et al. (2020), pleas to highly popular accounts).

A number of online coordination strategies have been observed in the literature making use of both dissemination and engagement to amplify their effect, including specifically those identified in Table 2. These in particular are all potentially observable in short periods of online activity, e.g. a political debate (Rizoiu et al. 2018). Other coordinated behaviour observed in the literature require some ability to identify accounts of interest and track them over extended periods of time. *Metadata shuffling* involves groups of accounts hiding through changing and even swapping their names and other metadata (Mariconti et al. 2017; Ferrara 2017). Related to this is *narrative switching*, in which an account suddenly deletes all their posts and then, potentially after a significant period of time, starts posting about different themes and issues (perhaps also having changed their account's appearance) (Dawson and Innes 2019). Dawson and Innes (2019) also observed changes in accounts' follower counts to identify the purchase of *fake followers* and *follower fishing* (used to boost reputation metrics), both of which require records of potentially lengthy periods of activity. Dawson and Innes (2019) also use *synchronicity* to identify groups temporally correlated through activity, but neglect to describe their specific method.

Different behaviour primitives, such as those in Table 3, can be used to execute the amplification strategies mentioned. Many of these behaviour primitives have analogies on multiple OSNs, so techniques devised to detect them on one could be employed effectively on others. Dissemination can be carried out by reposting, using hashtags, or mentioning highly connected individuals in the hope they spread a message further. Accounts doing this covertly will avoid direct connections, and thus inference is required for identification. Giglietto et al. (2020b) propose detecting anomalous levels of coincidental URL use as a way to do this; we expand this approach to other interactions.

Some strategies require more sophisticated detection: detecting bullying through *dogpiling* (e.g. as happened during the `#GamerGate` incident, studied by Burgess and Matamoros-Fernández (2016), or to those posing questions to public figures at political campaign rallies[3]) requires collection of (mostly) entire conversation trees, which, while trivial to obtain on forum-based sites (e.g. Facebook and Reddit), are difficult on stream-of-post sites (e.g. Twitter,[4] Parler and Gab). As mentioned, detecting metadata shuffling requires long term collection on broad issues to detect the same accounts being active in different contexts, and other follower and narrative analyses can also require extended collection periods.

Figure 2 shows representations of the strategies highlighted above, offering clues about how they might be

[2] https://crestresearch.ac.uk/resources/russian-influence-uk-terrorist-attacks/.

[3] https://www.bbc.co.uk/bbcthree/article/72686b6d-abd2-471b-ae1d-8426522b1a97.

[4] Changes introduced with Twitter's Application Programming Interface (API) version 2.0 aim to make this easier: https://developer.twitter.com/en/docs/twitter-api/conversation-id.

Table 3 Social media interaction equivalents

OSN	POST	REPOST	REPLY	MENTION	TAG	LIKE
Twitter	Tweet	Retweet	Reply tweet	@Mention	#Hashtags	Favourite
Facebook	Post	Share	Comment	Mention	#Hashtag	Reactions
Tumblr	Post	Repost	Comment	@Mention	#Tag	Heart
Reddit	Post	Crosspost	Comment	/u/Mention	Subreddit	Up/down vote

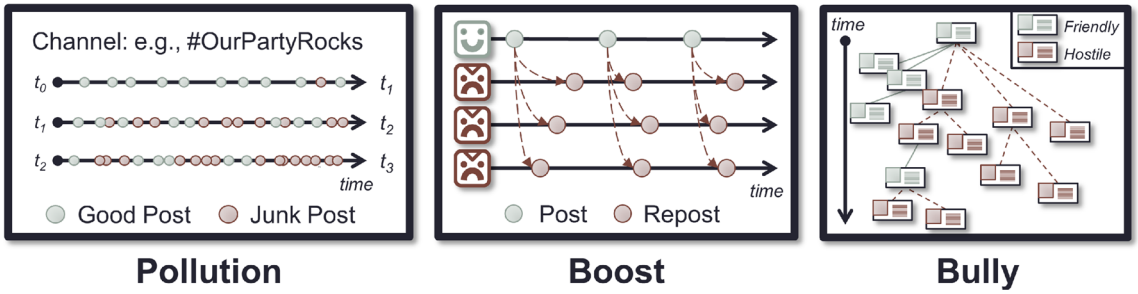


Fig. 2 Patterns matching the mentioned coordinated amplification strategies. Green posts and avatars are benign, whereas red or maroon ones are malign

identified. To detect *Pollution*, we match the authors of posts mentioning the same (hash)tag. This way we can reveal not just those who are using the same hashtags with significantly greater frequency than the average but also those who use more hashtags than is typical. To detect a variant of *Boost*, we match authors reposting the same original post, and can explore which sets of users not only repost more often than the average, but those who repost content from a relatively small pool of accounts. Alternatively, we can match authors who post identical, or near-identical text, as seen in our motivating example (Sect. 1.1); Graham et al. (2020) have recently developed open-sourced methods for this kind of matching, which have previously been used for campaign analysis (Lee et al. 2013). Considering reposts like retweets, however, it is unclear whether platforms deprioritise them when responding to stream filtering and search requests, so special consideration may be required when designing data collection plans. Finally, to detect *Bully*, we match authors whose replies are transitively rooted in the same original post, thus they are in the *same conversation*. This requires collection strategies that result in complete conversation trees, and also stipulates a somewhat strict definition of 'conversation'. On forum-based OSNs, the edges of a 'conversation' may be relatively clear: by commenting on a post, one is 'joining' the 'conversation'. Delineating smaller sets of interactions within all the comments on a post to find smaller conversations may be achieved by regarding each top-level comment and its replies as a conversation, but this may not be sufficient. Similarly, on stream-based OSNs, a conversation may be engaged in by a set of users if they all mention each other in their posts, as it is not possible to *reply* to more than one post at a time.

2.1 Problem statement

A clarification of our challenge at this point is:

> *To identify groups of accounts whose behaviour, though typical in nature, is anomalous in degree.*

There are two elements to this. The first is *discovery*. How can we identify not just behaviour that appears more than coincidental, but also the accounts responsible for it? That is the topic of the next section. The second element is *validation*. Once we identify a group of accounts via our method, what guarantee do we have that the group is a real, coordinating set of users? This is especially difficult given inauthentic behaviour is hard for humans to judge by eye (Cresci et al. 2017; Benkler et al. 2018; Jamieson 2020).

3 Methodology

The major OSNs share a number of features, primarily in how they permit users to interact with each other, digital media and the platforms (e.g. Table 3); hashtags, URLs, and mentions work much the same way across many OSNs. By focusing on these commonalities, we can develop approaches that generalise across OSNs.

Traditional social network analysis relies on long-standing relationships between actors (Wasserman and Faust 1994; Borgatti et al. 2009). On OSNs this requirement is typically fulfilled by friend/follower relations. These are expensive to collect and quickly degrade in meaning if not followed with frequent activity. By focusing on active interactions, however, it is possible to understand not just

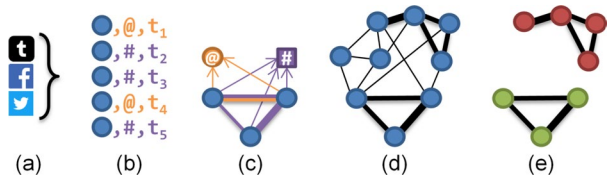


Fig. 3 Conceptual LCN construction and HCC discovery process

who is interacting with whom, but to what degree. This provides a basis for constructing (or inferring) social networks, acknowledging they may be transitory.

LCNs are built from inferred links between accounts. Supporting criteria relying on interactions alone, as observed in the literature (Ratkiewicz et al. 2011; Keller et al. 2019), include retweeting the same tweet (*co-retweet*), using the same hashtags (*co-hashtag*) or URLs (*co-URL*), or mentioning the same accounts (*co-mention*). To these we add joining the same 'conversation' (a tree of *reply* chains with a common root tweet) (*co-conv*). As mentioned earlier, other ways to link accounts rely on similar or identical content, metadata and temporal patterns (see Sect. 2). The criteria underpinning LCN links may be a combination of these and other interaction types.

3.1 The LCN/HCC pipeline

The key steps to extract HCCs from raw social media data are shown in Fig. 3 and documented in Algorithm 1. The example in Fig. 3 is explained after the algorithm has been explained, in Sect. 3.1.2.

Step 4. Construct an LCN, L, from the pairings in M. This network $L = (V, E)$ is a set of vertices V representing accounts connected by undirected weighted edges E of inferred links. These edges represent evidence of different criteria linking the adjacent vertices. The weight of each edge $e \in E$ between vertices representing accounts u and v for each criterion c is $w^c(e)$, and is equal to $\beta^c_{\{u,v\}}$.

Most community detection algorithms will require the multi-edges be collapsed to single edges. The edge weights are incomparable (e.g. retweeting the same tweet is not equivalent to using the same hashtag), however, for practical purposes, the inferred links can be collapsed and the weights combined for cluster detection using a simple summation, e.g. Eq. (1), or a more complex process like varied criteria weighting.

$$w(e) = \sum_{c=1}^{q} w^c(e) \tag{1}$$

Some criteria may result in highly connected LCNs, even if its members never interact directly. Not all types of coordination will be meaningful—people will co-use the same hashtag repeatedly if that hashtag defines the topic of the discussion (e.g. `#auspol` for Australian politics), in which case it is those accounts who co-use it significantly more often than others which are of interest. If required, the final step filters out these coincidental connections.

Step 5. Identify the highest coordinating communities H in L (Fig. 3e) using a suitable community detection algorithm, such as Blondel et al. (2008)'s Louvain algorithm (used by Morstatter et al. 2018; Nasim et al. 2018; Vosoughi et al. 2018; Nizzoli et al. 2021), *k nearest neighbour* (*kNN*) (used by Cao et al. 2015), edge weight thresholding (used by Lee et al. 2013; Pacheco et al. 2021), or FSA (Şen et al. 2016), an algorithm from the Social Network Analysis community that focuses on extracting sets of highly influential nodes from a network. Depending on the size of the dataset under consideration, algorithms suitable for very large networks may need to be considered (Fang et al. 2019). Some algorithms may not require the LCN's multi-edges to be merged (e.g. Bacco et al. 2017). We present a variant of FSA (Şen et al. 2016), FSA_V (Algorithm 2), designed to take advantage of FSA's benefits while addressing some of its costs. FSA does not just divide a network

Algorithm 1 FindHCCs

Input: P: Social media posts, C: Coordination criteria, θ: Extraction parameter
Output: H: A list of HCCs

1: $I_{all} \leftarrow$ ParseInteractionsFrom(P)
2: $I_C \leftarrow$ FilterInteractions(I_{all}, C)
3: $M \leftarrow$ FindCoordination(I_C, C)
4: $L \leftarrow$ ConstructLCN(M)
5: $H \leftarrow$ ExtractHCCs(L, θ)

Step 1. Convert social media posts P to common interaction primitives, I_{all}. This step removes extraneous data and provides an opportunity for the fusion of sources by standardising all interactions (thus including only the elements required for the coordination being sought).

Step 2. From I_{all}, filter the interactions, I_C, relevant to the set $C = \{c_1, c_2, \ldots, c_q\}$ of criteria (e.g. co-mentions and co-hashtags).

Step 3. Infer links between accounts given C, ensuring links are typed by criterion. The result, M, is a collection of inferred pairings. The count of inferred links between accounts u and v due to criterion $c \in C$ is $\beta^c_{\{u,v\}}$.

into communities (so that every node belongs to a community), but extracts only subsets of adjacent nodes that form influential communities within the overall network. FSA_V reduces the computational complexity introduced by FSA, which recursively applies Louvain to divide the network into smaller components and then, under certain circumstances, stitches them back together. The reason for this is to make FSA_V more suitable for application to a streaming scenario, in which execution speed is a priority.

Similar to FSA, FSA_V initially divides L into communities using the Louvain algorithm but then builds candidate HCCs within each, starting with the 'heaviest' (i.e. highest weight) edge (representing the most evidence of coordination). It then attaches the next heaviest edge until the candidate's mean edge weight (MEW) is no less than θ ($0 < \theta \leq 1$) of the previous candidate's MEW, or is less than L's overall MEW. In testing, edge weights appeared to follow a power law, so θ was introduced to identify the point at which the edge weight drops significantly; θ requires tuning. A final filter ensures no HCC with a MEW less than L's is returned. Unlike in FSA, recursion is not used, nor stitching of candidates, resulting in a simpler algorithm.

Algorithm 2 ExtractHCCs (FSA_V)

Input: L=(V, E): An LCN, θ: HCC threshold
Output: H: Highly coordinating communities

```
E' ← MergeMultiEdges(E)
g_mean ← MeanWeight(E')
louvain_communities ← ApplyLouvain(L)
Create new list, H
for l ∈ louvain_communities do
    Create new community candidate, h = (V_h, E_h)
    Add heaviest edge e ∈ l to h
    growing ← true
    while growing do
        Find heaviest edge e ∈ l connected to h not in h
        old_mean ← MeanWeight(E_h)
        new_mean ← MeanWeight(Concatenate(E_h, e))
        if new_mean < g_mean or
        new_mean < (old_mean × θ) then
            growing ← false
        else
            Add e to h
    if MeanWeight(E_h) > g_mean then
        Add h to H
```

This algorithm prioritises edge weights while maintaining an awareness of the network topology by examining adjacent edges, something ignored by simple edge weight filtering. Our goal is to find sets of strongly coordinating users, so we prioritise strongly tied communities while still acknowledging coordination can also be achieved with weak ties (e.g. 100 accounts paid to retweet one tweet).

The complexity of the entire pipeline is low order polynomial due primarily to the pairwise comparison of accounts to infer links in Step 3, which can be constrained by window size when addressing the temporal aspect. For large networks (i.e. networks with many accounts), that may be too costly to be of practical use; the solution for this relies on the application domain inasmuch as it either requires a tighter temporal constraint (i.e. a smaller time window) or tighter stream filter criteria, causing a reduction in the number of accounts, potentially along with a reduction in posts. Algorithmic complexity is discussed in Sect. 3.3.

3.1.1 Addressing the temporal aspect

Temporal information is a key element of coordination, and thus is critical for effective coordination detection. Frequent posts within a short period may represent genuine discussion or deliberate attempts to game trend algorithms (Grimme et al. 2018; Varol et al. 2017; Assenmacher et al. 2020). We treat the post stream as a series of discrete windows to constrain detection periods. An LCN is constructed from each window (Step 4), and these are aggregated and mined for HCCs (Step 5). We assume posts arrive in order, and assign them to windows by timestamp.

3.1.2 A brief example

Figure 3 gives an example of searching for co-hashtag and co-mention coordination across Facebook, Twitter, and Tumblr posts. The posts are converted to their interaction primitives in Step 1, shown in Fig. 3a. The information required from each post is the identity of the post's author,[5] the timestamp of the post for addressing the temporal aspect, and the hashtag or account mentioned (there may be many, resulting in separate records for each). This is done in Fig. 3b, which shows the filtered mentions (in orange) and hashtag uses (in purple), ordered according to timestamp.

[5] Linking identities across social media platforms is beyond the scope of this work, but the interested reader is referred to Adjali et al. (2020) for a recent contribution to the subject.

Step 3 in Fig. 3c involves searching for evidence of coordination through searching for our target coordination strategies through pairwise examination of accounts and their interactions. Here, three accounts co-use a hashtag while only two of them co-mention another account.

By Step 4 in Fig. 3d, the entire LCN has been constructed, and then Fig. 3e shows its most highly coordinating communities.

As mentioned above, to account for the temporal aspect, the LCNs produced for each time window in Fig. 3d can be aggregated and then mined for HCCs, or HCCs could be extracted from each window's LCN and then they can be aggregated, or analysed in near real-time, as dictated by the application domain.

3.1.3 Opportunities for fusion

As mentioned above, many of the interaction we consider have analogies on multiple OSNs, so a technique applied to Twitter, for example, may also be effective on Reddit or Tumblr. Misinformation was widely disseminated over Facebook, Tiktok, Twitter, and WhatsApp during the 2021 Israeli/Palestinian conflict as links to misattributed videos, images of blocks of text, and audio files.[6] Our technique could be used to study coordinated link (i.e. URL) sharing across these platforms in an appropriate time period, similar to the work of Giglietto et al. (2020b) and Broniatowski (2021)—all that is required from each platform's posts are the identity of the posting account, the link posted[7] and the post's timestamp. The identities of accounts posting the URLs will differ between platforms, of course, but this technique may also provide a mechanism for cross-platform identity matching, associating accounts that frequently post the same or similar content. Nimmo et al. (2020) essentially performed this task manually by searching for the same article content across different platforms, and then confirming similarity between the account names found. Our technique could be incorporated into the researcher's workflow to make this task easier by searching for duplication of text, and automatically linking instances where it is found, and then highlighting those connections.

3.2 Validation methods

As mentioned in Sect. 2.1, the second element of addressing our research challenge is that of validation. Once HCCs have been discovered, it is necessary to confirm that what has been found are examples of genuine coordinating groups. This step is required before addressing the further question of whether the coordination is authentic (e.g. grassroots activism) or inauthentic (e.g. astroturfing).

3.2.1 Datasets

In addition to relevant datasets, we make use of a ground truth (GT), in which we expect to find coordination (cf., Keller et al. 2017; Vargas et al. 2020). By comparing the evidence of coordination (i.e. HCCs) we find within the ground truth with the coordination we find in the other datasets, we can develop confidence that: (a) our method finds coordination where we expect to find it (in the ground truth); and (b) our method also finds coordination of a similar type where it was not certain to exist. Furthermore, to represent the broader population (which is not expected to exhibit coordination), similar to Cao et al. (2015), we create a randomised HCC network from the non-HCC accounts in a given dataset, and then compare its HCCs with the HCCs that had been discovered by our method.

3.2.2 Membership comparison

While our primary factors include the HCC extraction method (using FSA_V, *kNN*, or thresholds), the temporal window size, γ, and the strategy being targeted (*Boost*, *Pollution* or *Bully*), our interest prioritises the grouping of accounts over how they are individually connected, and so for each pair of results we compare the number, edge count and membership of the HCCs discovered. These figures provide context for the degree of overlap between the HCC members identified under different conditions (i.e. factor values). We use Jaccard and overlap similarity measures (Verma and Aggarwal 2020) to compare the accounts appearing in each (ignoring their groupings) and render them as heatmaps. The Jaccard similarity coefficient of two sets of items, X and Y, is:

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} = \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|}. \tag{2}$$

If there is significant imbalance in the sizes of X and Y, then their similarity may be low, even if one is a subset of the other. An alternative measure, the Overlap coefficient (Verma and Aggarwal 2020), takes this imbalance into account by using the size of the smaller of the two sets as the denominator:

$$\text{overlap}(X, Y) = \frac{|X \cap Y|}{\min(|X|, |Y|)}. \tag{3}$$

[6] https://www.nytimes.com/2021/05/14/technology/israel-palestine-misinformation-lies-social-media.html.

[7] More sophisticated content matching can also be used in Step 3, comparing what media the links refer to, rather than just the link itself (cf. Yu 2021).

In a circumstance such as ours, it is unclear whether a longer time window will garner more results after HCC extraction is applied. The Jaccard and overlap coefficients can be used to quickly understand two facts about the sets of accounts identified as HCC members with different values of γ:

- *Is one set a subset of the other?* If so, the overlap coefficient will reach 1.0, while the Jaccard coefficient will not if the two sets differ in size. If they are disjoint, the overlap coefficient will be 0.0 along with the Jaccard coefficient.
- *Do the sets differ in size?* If the sets are different sizes, but one is a subset of the other, the overlap coefficient will hide this fact, while the Jaccard coefficient will expose it. If both coefficients have values close to 0.0, then the sets are clearly different in membership and potentially also in size. If the coefficient values are very close, then the sets are close in size, because the denominators are similar in size, meaning $|X \cup Y| \approx \min(|X|, |Y|)$, but this will only occur if they share many members (i.e. $|X \cap Y|$ is high).

Alongside the heatmaps, we provide exact numbers of the accounts which are common to the discovered HCCs, to better inform the reader of the overall influence of the particular factor(s) being varied. For example, by being able to compare the results for each value for γ in one visualisation, it is possible to see the progression of the coefficient values as the window size increases (in both the provided raw numbers and by the colour scale in the heatmaps).

3.2.3 Network visualisation

A second subjective method of analysis for networks is to visualise them. We use two visualisation tools, *visone* (https://visone.info) and Gephi (https://gephi.org), both of which make use of force directed layouts, which help to clarify clusters in the network structure. Node colour is used to represent cluster membership detected with the Louvain method (Blondel et al. 2008). Each connected component is an HCC, and node colour can be used to represent the number of posts, and edge weight can be represented by thickness and, depending on the density of the network, darkness of colour. For analyses that involve multiple criteria (e.g. co-conv and co-mention), we use node shape to represent which combination of criteria an HCC is bound by (e.g. just co-mention or a combination of co-mention and co-conv or just co-conv).

By extending the HCC account networks with nodes to represent the 'reasons' or instances of evidence that link each pair of nodes, e.g. the tweets they retweet in common, or accounts they both mention or reply to, thereby creating a two-level *account-reason* network, we can investigate how HCCs relate to one another. In this case, the account-reason network has two types of nodes and two types of edges ('coordinates with' links between accounts and 'caused by' or 'associated because' links between 'reasons' and accounts). Visualising the two-level network by colouring nodes by their HCC and using a force-directed layout highlights how closely HCCs associate with each other, not only revealing what reasons draw HCCs together (i.e. HCCs may be bound by a single reason, or an HCC may be entirely isolated from others in the broader community), but also how many reasons bind them (i.e. many reasons may bind an HCC together or just one). Deeper insights can be revealed from this point using multi-layer network analyses.

3.2.4 Consistency of content

To help answer RQ2, it is helpful to look beyond network structures and consider how consistent the content produced by an HCC is relative to other HCCs and the population in general. This will be most applicable when the type of strategy the HCC is suspected to have engaged in relies on repetition, e.g. co-retweeting or copypasta. If an HCC is boosting a message, it is reasonable to assume the content posted by the members of the HCC will be more similar internally than when compared externally (i.e. to the content of non-members). To analyse this internal consistency of content, we treat each HCC member's tweets as a single document and create a doc-term matrix using 5 character n-grams for terms to maintain phrase ordering (which is lost with bag-of-word approaches). Comparing the members' document vectors using cosine similarity in a pairwise fashion creates a $n \cdot n$ matrix where n is the number of accounts in the HCC network. This approach was chosen for its performance with non-English corpora (Damashek 1995), and because using individual tweets as documents produced too sparse a matrix in a number of tests we conducted. The pairwise account similarity matrix can be visualised, using a spectrum of colours to represent similarity. By ordering the accounts on both the x and y axes to ensure they are grouped within their HCCs, if our hypothesis is correct that similarity within HCCs is higher than outside, then we should observe clear bright squares representing entire HCCs along the diagonal of the resulting similarity matrix. The diagonal itself will be the brightest because it represents each account's similarity with itself.

If HCCs contribute few posts, which are similar or identical to other HCCs, then bright squares may appear off the diagonal, and this would be evidence similar to clusters of account nodes around a small number of reason nodes in the two-level account-reason networks mentioned above.

This method offers no indication of how active each HCC or HCC member is, so displays of high similarity may imply low levels of coincidental activity as well as high content similarity, just because of the lower likelihood that highly active accounts are highly similar in content (by contributing more posts, there are simply more opportunities for accounts' content to diverge). The use of the 5-character

n-gram approach is designed to offset this because each tweet in common between two accounts will yield a large number of points of similarity, as will the case when the same two tweets are posted in the same order (i.e. two accounts both post tweet t_1 and then t_2), because the overlap between the tweets will yield at least four points of similarity.

3.2.5 Variation of content

Converse to the consistency of content within HCC is the question of content variation, and how does the variation observed in detected HCCs differ from that of RANDOM groupings. Highly coordinated behaviour such as co-retweeting involves reusing the same content frequently, resulting in low feature variation (e.g. hashtags, URLs, mentioned accounts), which can be measured as entropy (Cao et al. 2015). A frequency distribution of each HCC's use of each feature type is used to calculate each entropy score. Low feature variation corresponds to low entropy values. As per Cao et al. (2015), we compare the entropy of features used by detected HCCs to RANDOM ones and visualise their cumulative frequency. Entries for HCCs which did not use a particular feature are omitted, as their scores would inflate the number of groups with 0 entropy.

3.2.6 Hashtag analysis

Hashtags can be used to define discussion groups or chat channels (Woolley 2016), so hashtag analysis can be used to study those communities. It is another aspect to content analysis that relies upon social media users declaring the topic of their post through the hashtags they include. At the minimum, we can plot the frequency of the most frequently used hashtags as used by the most active HCCs. In doing so, we can quickly see which hashtags different HCCs make use of, and how they relate by how they overlap. Some hashtags will be unique to HCCs, while others will be used by many. This exposes the nature of HCC behaviour: they may focus their effort on a single hashtag, perhaps to get it trending, or they may use many hashtags together, perhaps to spread their message to different communities.

To further explore how hashtags are used together, we perform *hashtag co-occurrence analysis*, creating networks of hashtags linked when they are mentioned in the same tweet (as distinct from the co-hashtag linking introduced above). These hashtag co-occurrence networks are sometimes referred to as *semantic networks* (Radicioni et al. 2020). When visualised with force-directed layouts it is possible to see themes in the groupings of hashtags, and to gain insights from how the theme clusters are connected (including when they are isolated from one another). Colouring hashtags by their clusters detected using the Louvain method (Blondel et al. 2008) can provide a statistical measure of hashtag relations.

3.2.7 Temporal patterns

Campaign types can exhibit different temporal patterns (Lee et al. 2013), so we use the same temporal averaging technique as Lee et al. (2013) (dynamic time warping barycenter averaging) to compare the daily activities of the HCCs in the GT and RANDOM datasets with those in the test datasets. The temporal averaging technique produces a single time series made by averaging together each account's activity time series. Using this technique avoids averaging out of time series that are off-phase from one another by aligning them before averaging them.

Another aspect of temporal analysis is the comparison of HCCs detected in different time windows, including specifically observing whether such HCCs share members and what the implications are for the behaviour of those members. This is non-trivial for any moderately large dataset, but examination of the ground truth can provide insight into the behaviours exhibited by known collaborators.

3.2.8 Focus of connectivity

Groups that retweet or mention themselves create direct connections between their members, meaning if one is discovered, it may be trivial to find its collaborators. To be covert, therefore, it would be sensible to have a low *internal retweet* and *mention ratios* (IRR and IMR, respectively). Formally, if RT_{int} and M_{int} are the the sets of retweets and mentions of accounts within an HCC, respectively, and RT_{ext} and M_{ext} are the corresponding sets of retweets and mentions of accounts outside the HCC, then, for a single HCC

$$\text{IRR} = \frac{|RT_{\text{int}}|}{|RT_{\text{int}}| + |RT_{\text{ext}}|} \tag{4}$$

$$\text{IMR} = \frac{|M_{\text{int}}|}{|M_{\text{int}}| + |M_{\text{ext}}|} \tag{5}$$

3.2.9 Consistency of coordination

The method presented Sect. 3.1 highlights HCCs that coordinate their activity at a high level over an entire collection period. Further steps can be taken to determine which HCCs are coordinating their behaviour repeatedly and consistently across adjacent time windows. In this case, for each time window, we consider not just the nodes and edges from the current LCN, but additionally from previous windows, applying a degradation factor the contribution of their edge weights. To build an LCN from a sliding frame of T time windows, the new LCN includes the union of the nodes and edges of the individual LCNs from the current and previous windows, but to calculate the edge weights, we apply a decay

factor, α, to the weights of edges appearing in windows before the current one. In this way, we apply a multiplier of α^x to the edge weights, where x is the number of windows into the past: the current window is 0 windows into the past, so its edges are multiplied by $\alpha^0 = 1$; the immediate previous window is 1 window back, so its edge multiplier is α^1; the one before that uses α^2, and so on until the farthest window back uses α^{T-1}. Generalising from Step 4, the weight $w^{c,t}(e)$ for an edge $e \in E$ between accounts u and v for criterion c at window t and a sliding window T windows wide is given by

$$w^{c,t}(e) = \sum_{x=0}^{T-1} w^{c,(t-x)}(e) \cdot \alpha^x. \tag{6}$$

In this way, to create a baseline in which the sliding frame is only one window wide, one only need choose $T = 1$, regardless of the value of α. As $\alpha \to 1$, the contributions of previous windows are given more consideration.

3.2.10 Supervised machine learning with one-class classifiers

An approach that aids in the management of data with many features is classification through machine learning. This is an approach that has been used extensively in campaign detection, in which tweets are classified, rather than accounts (e.g. Lee et al. 2013; Chu et al. 2012; Wu et al. 2018). Because of its 'black box' nature, its application should be considered carefully, however. Our intent is to use classification to validate that entire HCCs (not just individual tweets or accounts) detected in datasets are similar to those found in ground truth. Such classifiers will not be generally applicable, as they rely on ground truth (which is historical by nature) for training data. Tactics and strategies used in information operations will change over time, as shown by Alizadeh et al. (2020); this is not just to avoid detection but also because OSN features change over time. As our focus is only on a positive answer to whether one HCC is similar to others, it is acceptable to rely on one-class classification (i.e. an HCC detected in a dataset is recognised as COORDINATING/positive or is regarded as NON-COORDINATING/unknown). The more common binary classification approach was used by Vargas et al. (2020), however our approach has two distinguishing features:

1. We rely on one-class classification because we have positive examples of what we regard as COORDINATING from the ground truth, and everything else is regarded as unknown, rather than definitely 'not coordinating'. These are sometimes referred to as *positive and unlabeled*, or PU, classifiers. A one-class classifier can, for example, suggest a new book from a wide range (such as a library) based on a person's reading history. In such a circumstance, the classifier designer has access to positive examples (books the reader likes or has previously borrowed) but all other instances (books, in this case) are either positive or negative. When our one-class classifier recognises HCC accounts as positive instances, it provides confidence that the HCC members are coordinating their behaviour in the same manner as the accounts in the ground truth. We can therefore prioritise Precision over Recall (discussed below).
2. We rely on features from both the HCCs and the HCC members and use the HCC members as the instances for classification, given it is unclear how many members an HCC may have, and accounts that are members of HCCs may have traits in common that are distinct from 'normal' accounts. In contrast, Vargas et al. (2020) relied on features of "coordination networks" (i.e. HCCs) alone, as they were their classification instances. For this reason the feature vectors that our classifier is trained and tested on will comprise features drawn from the individual accounts and their behaviour as well as the behaviour of the HCC of which they are a member. Feature vectors for members of the same HCC will naturally share the feature values drawn from their grouping.

Regarding the construction of the feature vector, at a group level, we consider not just features from the HCC itself, which is a weighted undirected network of accounts, but of the activity network built from the interactions of the HCC members within the corpus. The activity network is a multi-network (i.e. supports multiple edges between nodes) with nodes and edges of different types. The three node types are *accounts*, *URLs*, and *hashtags*. Edges represent interactions and the following types are modelled: hashtag uses, URL uses, mentions, repost/retweets, quotes (cf. comments on a Facebook share or Tumblr repost), reply, and 'in conversation' (meaning that one account replied to a post that was transitively connected via replies to an original post by an account in the corpus). This activity network therefore represents not just the members of the HCC but also their degree of activity in context.

3.2.10.1 Classifier algorithms We use the GT to train three classifiers. A bagging PU classifier (BPU, Mordelet and Vert 2014) was used, the implementation[8] for which was based on a Random Forest (RF) classifier configured with 1000 trees (estimators). We also used a standard 1000 tree RF, as used by Vargas et al. (2020), to compare directly with BPU. A Support Vector Machine (SVM) classifier was also used, given the technique's known high performance with non-linear recognition problems even with small feature sets due

[8] Thanks to Roy Wright for his implementation: https://github.com/roywright/pu_learning/blob/master/baggingPU.py.

its use of the kernel trick. Furthermore, Mordelet and Vert (2014) employed a variety of SVMs as part of their experimentation, though our choice of implementation differed. Both SVM and RF implementations were drawn from the `scikit-learn` Python library (Pedregosa et al. 2011). Contrasting "unlabeled" training instances were created from the RANDOM dataset. Feature vector values were standardised prior to classification and upsampling was applied to create balanced training sets of approximately 400 positive and random elements each. 10-fold cross validation was used.

The classifiers predict whether instances provided to them are in the positive or unlabeled classes, which, to aid readability, we refer to as 'COORDINATING' and 'NON-COORDINATING', respectively.

3.2.10.2 Performance metrics The performance metrics used include the classifier's accuracy, F_1 scores for each class, and the Precision and Recall measures that the F_1 scores are based upon. High precision implies the classifier is good at recognising samples correctly, and high Recall implies that a classifier does not miss instances of the class they are trained on in any testing data. For example, a good apple classifier will successfully recognise an apple when it is presented with one, and when presented with a bowl of fruit, it will successfully find all the apples in it. The F_1 score combines these two measures:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

and provides insight into to the balance between the classifier's Precision and Recall. The accuracy of a classifier is the proportion of instances in a test data set that the classifier labeled correctly. In this way, the accuracy is the most coarse of these measures, because it offers little understanding of whether the classifier is missing instances it should find (false negatives) or labeling non-matching instances incorrectly (false positives). The F_1 score begins to address this failing, but direct examination of the Precision and Recall provides the most insight into each classifier's performance.

3.2.11 Bot analysis

Although coordinated behaviour in online campaigns is often conducted without automation (Starbird et al. 2019), automation is still commonly present in campaigns, especially in the form of social bots, which aim to present themselves as typical human users (Ferrara et al. 2016; Woolley and Guilbeault 2018; Cresci 2020). For this reason, the technique presented here is a valid tool for exposing teams of cooperating bot and social bot accounts. We use the Botometer (Davis et al. 2016) service to evaluate selected accounts for bot-like behaviour. The primary summary measure for bot classification is the Complete Automation Probability (CAP),[9] provided as a value in [0, 1] in two variants: one for predominantly English-speaking accounts and one language-agnostic. Other studies have relied on a CAP of 0.5 as a threshold for labelling an account as a bot, but there is a significant overlap between humans that act in a very bot-like manner and bots that are quite human-like, so we adopt the practice of Rizoiu et al. (2018) and regard scores below 0.2 to be human and above 0.6 to be bots.

3.3 Complexity analysis

The steps in processing timeline presented in Sect. 3.1 are reliant on two primary factors: the size of the corpus of posts, P, being processed, and the size of the set of accounts, A, that posted them. Therefore $|A| \leq |P|$ and the complexity of Step 1 is linear, $O(|P|)$, because it requires processing each post, one-by-one. The set of interactions, I_{all}, it produces may be larger than $|P|$, because a post may include many hashtags, mentions, or URLs, but given posts are not infinitely long (even long Facebook and Tumblr posts can only include several thousand words), the number of interactions will also be linear, i.e. $|I| = k|P|$, for some constant k. Step 2 filters these interactions down to only those of interest, I_C, based on the type of coordinated activity sought, C, so $|I_{\text{all}}| \geq |I_C|$, and again the complexity of this step is also linear, $O(|I_{\text{all}}|)$, as it requires each interaction to be considered. Step 3 seeks to find evidence of coordination between the accounts in the dataset, and so requires examining each filtered interaction and building up data structures to associate each account with their interactions ($O(|I_{\text{all}}|)$), then emitting pairs of accounts matching the coordination criteria, producing the set M, which requires the pairwise processing of all accounts, and so is $|A|^2$ steps with a subsequent complexity of $O(|A|^2)$. This, however, also depends on the pairwise comparison of each account's interactions, which is likely to be small, practically, but theoretically could be as large as $|I_C|$ if one user is responsible for every single interaction in the corpus (but then $|A|$ would be 1). On balance, as a result, we will regard the processing of each pair of users' interactions as linear with a constant factor k (i.e. $O(k|A|^2) = O(|A|^2)$). In Step 4, producing the LCN, L, from the criteria is a matter of considering each match one-by-one, so is again linear (though potentially large, depending on $|M|$). The final step (5) is to extract the HCCs from the LCN, and its performance and complexity very much depend upon the algorithm employed, but significant research has been applied in this field (Bedru et al. 2020, as considered in, e.g.] [). For FSA_V, which relies on the Louvain algorithm with complexity $O(|A| \log_2 |A|)$ (Blondel et al. 2008), it considers

[9] https://botometer.osome.iu.edu/faq#which-score.

Table 4 Experiment dataset statistics

	Tweets	Retweets (%)		Accounts	Tweet rate	Retweet rate
DS1	115,913	63,164	(54.5%)	20,563	0.31	0.17
(GT)	4193	2505	(59.7%)	134	1.74	1.04
DS2	1,571,245	729,937	(56.5%)	1381	3.12	1.45

Rates are per account per day

edges within each community to build its HCC candidates, so has a complexity of less than $O(|E|)$, where $|E|$ is the number of edges in the LCN, meaning its complexity is linear. FSA_V's complexity is therefore $O(|A| \log_2 |A| + |E|)$.

We regard the computation complexity of the entire pipeline as the highest complexity of its steps, which are:

1. Extract interactions from posts: $O(|P|)$
2. Filter interactions: $O(|I_{\text{all}}|)$
3. Find evidence of coordination: $O(|A|^2)$
4. Build LCN from the evidence: $O(|M|)$
5. Extract HCCs from LCN using, e.g. FSA_V: $O(|A| \log_2 |A| + |E|)$

The maximum of these is Step 3, the search for evidence of coordination, $O(|A|^2)$. Though in theoretical terms the method is potentially very costly, in practical terms we are bound by the number of accounts in the collection (which is determined by the manner in which the data was collected and the nature of the online discussion to which it pertains) and may be managed by constraining the time window, further reducing the number of posts (and therefore accounts) considered, as long as that suits the type of coordination being sought.

4 Evaluation

Our approach was evaluated in two phases:

- The first was conducted as an experiment using the validation methods mentioned above and two datasets known to include coordinated behaviour, as well as a ground truth dataset.
- The second phase involved two case studies in which we apply our approach against datasets relating to politically contentious topics expected to include polarised groups.

The first stage of the evaluation involved searching for *Boost* by co-retweet and other strategies while varying window sizes (γ). FSA_V was compared against two other community detection algorithms, when applied to the LCNs built in Step 4 (aggregated). We then validated the resulting HCCs through a variety of network, content, and temporal analyses and machine learning classification, guided by the research questions posed in Sect. 1. Discussion of further applications and performance metrics is also presented.

4.1 The experiment datasets

The two real-world datasets selected (shown in Table 4) represent two collection techniques: filtering a live stream of posts using keywords direct from the OSN (DS1) and collecting the posts of specific accounts (DS2):

- DS1: Tweets relating to a regional Australian election in March 2018, including a ground truth subset (GT); and
- DS2: A large subset of the Internet Research Agency (IRA, Chen 2015; Mueller 2018) dataset published by Twitter in October 2018.[10]

DS1 was collected using RAPID (Lim et al. 2019) over an 18 day period (the election was on day 15) in March 2018. The filter terms included nine hashtags and 134 political handles (candidate and party accounts). The dataset was expanded by retrieving all replied to, quoted and political account tweets posted during the collection period. The political account tweets formed our ground truth. It was our expectation that some of the coordinated political influence techniques observed on the international stage may have been adopted by political parties and issue-motivated groups at the regional level by 2018 (especially given the use of political bots had been reported in the Australian setting five years prior, as reported in Woolley 2016), and hence would be present in this dataset.

The IRA dataset released by Twitter covers 2009 to 2018, but DS2 is the subset of tweets posted in 2016, the year of the US Presidential election. Because DS2 consists entirely of IRA accounts which Twitter believed to be connected with an SIO, it was expected to include evidence of coordinated amplification. It was also much larger than DS1, and our intent was that our findings would complement forensic studies of the activity (e.g. Benkler et al. 2018; Jamieson 2020) and also contrast with techniques from more focused studies (e.g. Dawson and Innes 2019).

[10] https://about.twitter.com/en_us/values/elections-integrity.html.

Table 5 HCCs by coordination strategy

Strategy	γ	GT			DS1			DS2		
		Nodes	Edges	Comp.	Nodes	Edges	Comp.	Nodes	Edges	Comp.
LCN										
Boost	15	44	112	5	8855	80,702	419	855	23,022	14
Pollute	15	51	154	2	13,831	1,281,134	73	1203	65,949	5
Bully	60	70	482	1	16,519	1,925,487	222	1103	37,368	5
FSA_V										
Boost	15	9	6	3	633	753	167	113	758	19
Pollute	15	9	5	4	135	93	50	24	15	9
Bully	60	11	7	4	338	280	119	109	1123	16
kNN										
Boost	15	9	21	1	1041	33,621	1	675	22,494	1
Pollute	15	11	37	1	724	153,424	1	1040	65,280	1
Bully	60	18	135	1	1713	663,413	1	692	35,136	1
Threshold										
Boost	15	11	16	3	85	68	31	8	10	2
Pollute	15	24	26	3	44	37	10	6	13	1
Bully	60	15	19	3	25	23	8	10	10	3

4.2 Experimental set up

The size of the window γ was set at $\{15, 60, 360, 1440\}$ (in minutes) and the three community detection methods used on the aggregated LCNs were:

- FSA_V ($\theta = 0.3$);
- *kNN* with $k = ln(|V|)$ (cf., Cao et al. 2015); and
- a simple threshold retaining the edges with a normalised value above 0.1.

4.2.1 Parameter selection

Other than a value of $k = ln(|V|)$ for *kNN* (taken from Cao et al. 2015), the choice of values for parameters γ, θ and the threshold were determined as follows. Our intent was to search for human-driven coordination, i.e. teams of humans manipulating potentially several accounts each, meaning that the timeframes under examination would need to allow for the time required to switch between accounts. As discussed by Dawson and Innes (2019), the motivation for even paid coordinated behaviour may be based on numbers of posts made, rather than how tightly coordinated they are, so by examining a relatively wide 'short' window of 15 minutes allows for such people to react to each others' posts as they see them (rather than the sub-minute coordination sought by others, e.g. Giglietto et al. 2020a; Pacheco et al. 2021; Dawson and Innes 2019). The 60 minute window allows for people motivated by personal interest as well as paid trolls, who check their social media frequently throughout the day while attending to other duties (e.g. preparing new content, Nimmo et al. 2020). The six hour time frame is of medium length and allows for users who check social media over breakfast, at lunch, and then at dinner who also may be more motivated by personal reasons to coordinate their behaviour. Finally, the long term time frame of a whole day allows for accounts that only check social media in concentrated sessions once a day, but who coordinate their actions with others each day outside of the six hour window. Furthermore, automated coordinated accounts (i.e. bots) can react to posts very quickly (i.e. within seconds), and simple implementations can be revealed by their consistent short response times rather than relying on the more sophisticated co-activity methods presented here. More complex bot implementations vary their response times to avoid this (Cresci et al. 2017; Cresci 2020), however if they wish to game OSN trending algorithms to improve their reach, their posts must occur near to each other in time. Values for γ were also informed by the observation of Zhao et al. (2015) that 75% of retweets occur within six hours of posting. This implies that if attempts were made to boost a tweet, retweeting it in much shorter times would be required for it to stand out from typical traffic. Varol et al. (2017) checked Twitter's trending hashtags every 10 minutes, which is an indication of how quickly a concerted *Boost*ing effort may have an effect. Values chosen for γ therefore ranged from 15 minutes to a day, growing by a factor of approximately four at each increment. Deliberate coordinated retweeting (i.e. covert *Boost*ing masquerading as grassroots activity) was expected to occur in the smaller windows, but then be replaced by coincidental co-retweeting as the window size increases.

Values for θ and the threshold were based on experimenting with values in [0.1, 0.9], maximising the MEW to HCC

Table 6 HCCs by window size γ (Boost, FSA_V)

γ	Network attributes			HCC sizes				Nodes in common			
	Nodes	Edges	HCCs	Min.	Max.	Mean	SD	$\gamma = 15$	$\gamma = 60$	$\gamma = 360$	$\gamma = 1440$
GT											
15	9	6	3	3	3	3.00	0.00	9	9	8	8
60	14	9	5	2	3	2.80	0.40	–	14	10	12
360	13	9	5	2	3	2.60	0.49	–	–	13	12
1440	17	12	6	2	3	2.80	0.37	–	–	–	17
DS1											
15	633	753	167	2	18	3.79	2.21	633	218	93	100
60	619	1293	151	2	13	4.10	2.30	–	619	208	193
360	503	1119	127	2	19	3.96	2.58	–	–	503	350
1440	815	2019	141	2	110	5.78	12.60	–	–	–	815
DS2											
15	113	758	19	2	65	5.95	13.94	113	34	29	25
60	77	394	18	2	27	4.28	5.64	–	77	62	54
360	98	775	15	2	32	6.53	9.13	–	–	98	56
1440	69	380	15	2	27	4.60	6.15	–	–	–	69

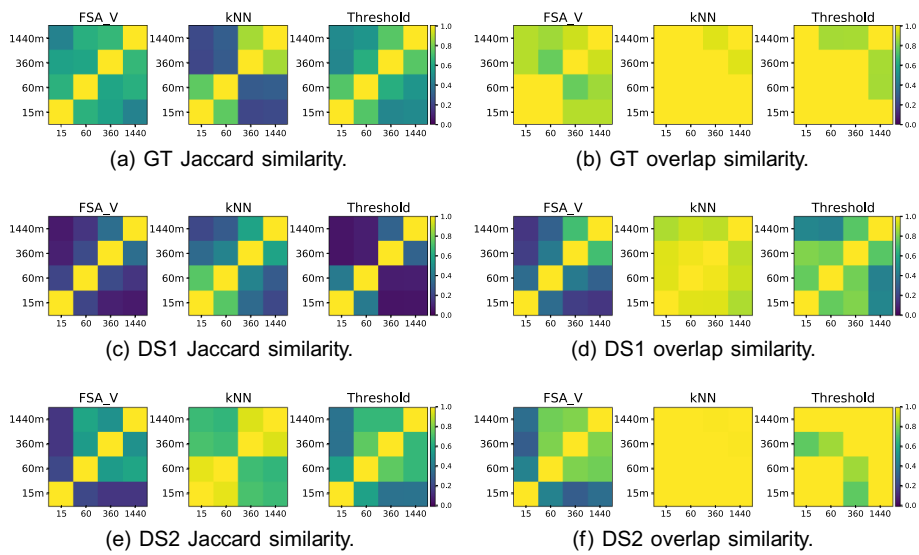


Fig. 4 Similarity matrices of HCC account sets found using different window sizes (FSA_V). The similarity measured here relates to the accounts found not to the similarity in groupings of accounts into HCCs. Yellow implies a high similarity (Jaccard: account sets are identical, Overlap: one set is a subset), while blue implies low similarity (i.e. account sets are disjoint)

size ratio, using the DS1 and DS2 aggregated LCNs when $\gamma = \{15, 1440\}$.

4.3 Experimental results

The research questions introduced in Sect. 1 guide our discussion, but we also present follow-up analyses.

4.3.1 HCC detection (RQ1)

4.3.1.1 Detecting different strategies The three detection methods all found HCCs when searching for *Boost* (via co-retweets), *Pollute* (via co-hashtags), and *Bully* (via co-mentions), details of which are shown in Table 5. Notably, *kNN* consistently builds a single large HCC, highlighting the need to filter the network prior to applying it (cf., Cao et al. 2015). The *kNN* HCC is also consistently nearly as large as the original LCN for DS2, perhaps due to the low number of accounts and the fact that *kNN* retains every edge adjacent to the retained vertices, regardless of weight. It is not clear, then, that *kNN* is producing meaningful results used in this way, even if it can extract a community.

4.3.1.2 Varying window size γ Different strategies may be executed over different time periods, based on their aims. *Boost*ing a message to game trending algorithms requires the messages to appear close in time, whereas some forms of *Bully*ing exhibit only consistency and low variation (men-

Table 7 HCCs by detection method (Boost, $\gamma = 15$)

	Network attributes			HCC sizes		Nodes in common		
	Nodes	Edges	HCCs	Min.	Max.	FSA_V	*kNN*	Threshold
DS1								
FSA_V	633	753	167	2	18	633	56	36
kNN	1041	33,621	1	1041	1041	–	1041	44
Threshold	85	68	31	2	14	–	–	85
DS2								
FSA_V	113	758	19	2	65	113	88	4
kNN	675	22,494	1	675	675	–	675	8
Threshold	8	10	2	2	6	–	–	8

tioning the same account repeatedly). Polluting a user's timeline on Twitter can also be achieved by frequently joining their conversations over a sustained period.

Varying γ searching for *Boost*, we found different accounts were prominent over different time frames (Table 6); the overlap in the accounts detected in each time frame differed considerably even though the number of HCCs stayed relatively similar. Figure 4 shows the Jaccard and overlap similarity between the sets of accounts appearing in each window size (agnostic of HCC membership). The overlap results for *kNN* shows very high levels of similarity, but lower levels of Jaccard similarity. For all datasets, as γ grows *kNN* finds more and more HCC members, including all the ones it found with smaller window sizes (overlap similarity values appear close to 1.0, shown as yellow). The highest Jaccard similarities for *kNN* seem to group the shorter periods ($\gamma = \{15, 60\}$) and the medium and long periods ($\gamma = \{360, 1440\}$). FSA_V finds different sets of members in each time window without significant overlap, though for DS2 it appears that the windows longer than 15 minutes have many members in common, but have very few in common with the $\gamma = 15$ HCCs. As might be expected, thresholding by LCN edge weight results in the identification of additional accounts as γ increases, and the Jaccard similarity of GT and DS1 (Fig. 4c) reveals that accounts identified in the shorter time windows ($\gamma = \{15, 60\}$) are very different to those from the longer time windows, but they still overlap somewhat (Fig. 4d). This suggests that although there are some accounts that coordinate in short periods, other accounts coordinate *more* over the medium and long time periods. These include media accounts that are consistently highly active over longer periods and differ from the active discussion participants who might log on to Twitter in the evening for a few hours whose behaviour is more bursty in nature.

Other than in GT, which revealed very few HCCs, the sizes of the HCCs found seemed to follow a rough power law; most were very small but one or a few were very large (see the HCC Sizes section in Table 6). The number of HCCs did not vary significantly nor consistently as γ increased. The number of edges retrieved tells us in DS1, as the window increased, more edges had weights high enough to be retained, whereas DS2 edge counts diminished, implying that the LCNs were progressively dominated by a smaller number of very *heavy* edges, while other remained relatively *light*.

4.3.1.3 HCC detection methods Similarly, HCCs discovered by the three community extraction methods (Table 7) exhibit large discrepancies, suggesting that whichever method is used, tuning is required to produce interpretable results. This is evident in the literature: Cao et al. conducted significant pre-processing when identifying URL sharing campaigns across two years of Twitter activity (Cao et al. 2015), and Pacheco et al. showed how specific strategies could identify groups in the online narrative surrounding the Syrian White Helmet organisation (Pacheco et al. 2020). Here we present the variation in results while controlling methods and other variables and keeping the coordination strategy constant, as our interest here is to validate the effectiveness of the method.

The networks were visualised using the FR layout in Fig. 5, revealing further structure within the *kNN* networks, each of which consisted of a single connected component. To examine the structure of the single *kNN* component more closely, we applied Louvain analysis (Blondel et al. 2008) and coloured the largest detected clusters. The clustering reveals distinct communities within both the lone *kNN* HCC found in each of the datasets. It is possible the DS2 ones are more easily discernible either due to the smaller number of accounts (675 compared with 1041) or because the accounts were, in fact, organised teams of malicious actors acting over a longer time frame. In either case, it makes clear that *kNN*, configured as it was, failed to distinguish communities clearly extractable via other means. This is less an indictment on *kNN* and more an indication that community extraction is likely to be a multi-step process embedded in particular domains and datasets, and in the particular types of networks to which they are applied. The networks in Fig. 5b, e bear a passing resemblance to many in, e.g. the

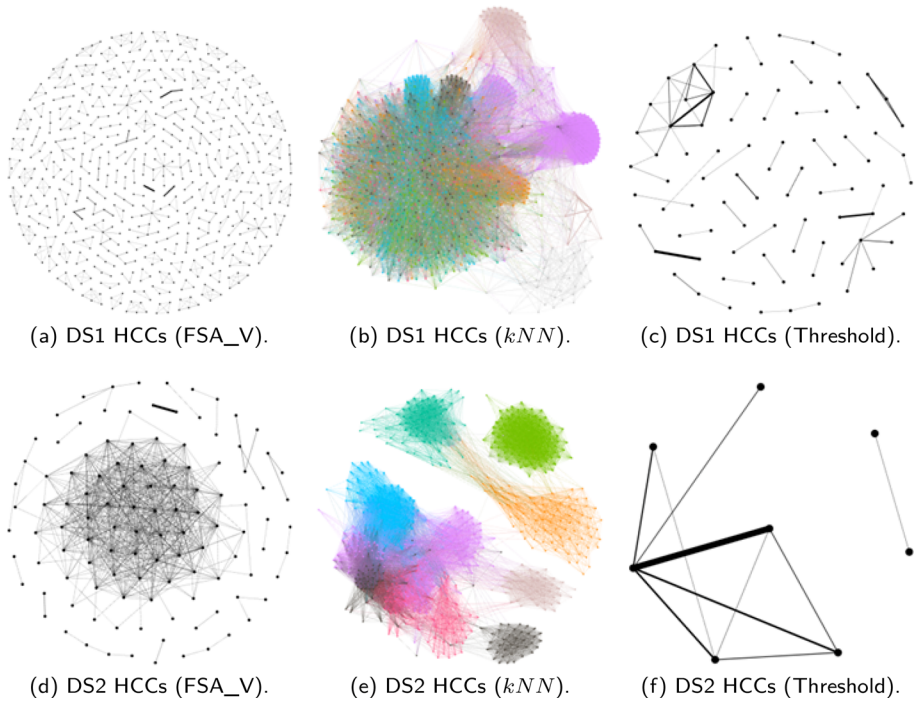


Fig. 5 HCCs discovered using different methods in DS1 and DS2 (Boost, $\gamma = 15$). Each *kNN* network consists of a single connected component, but detected clusters have been coloured to highlight internal structures

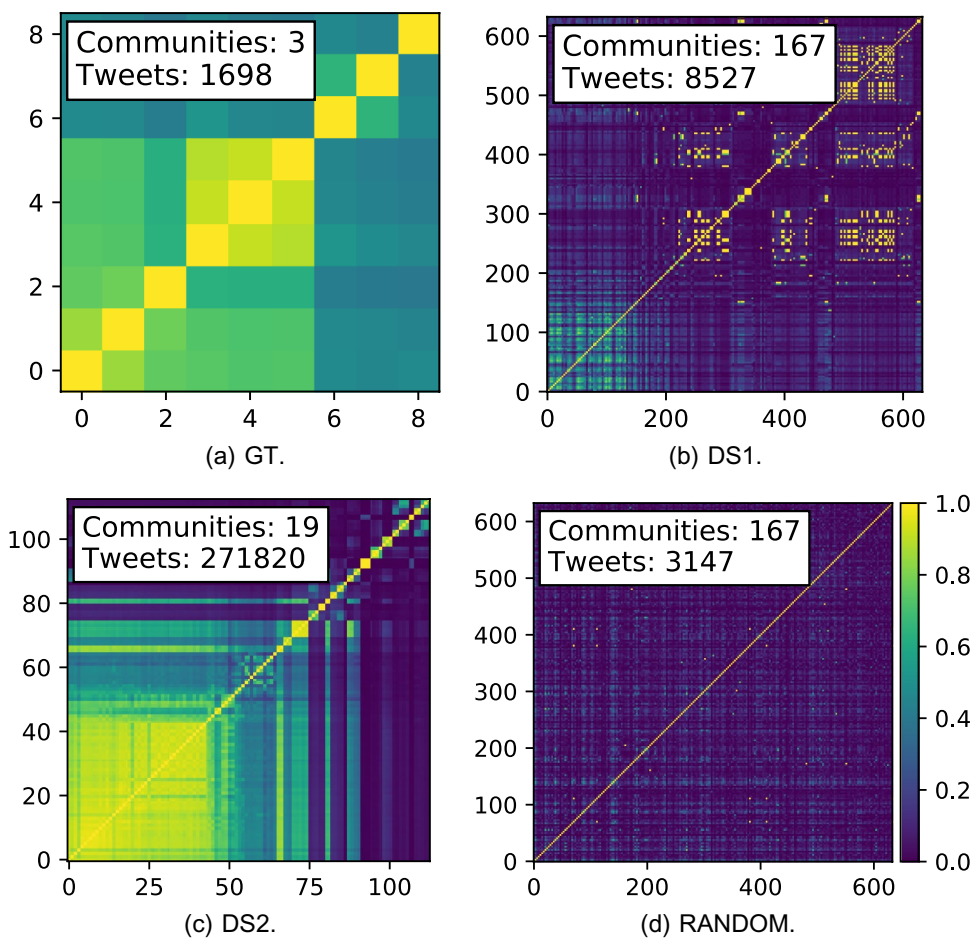
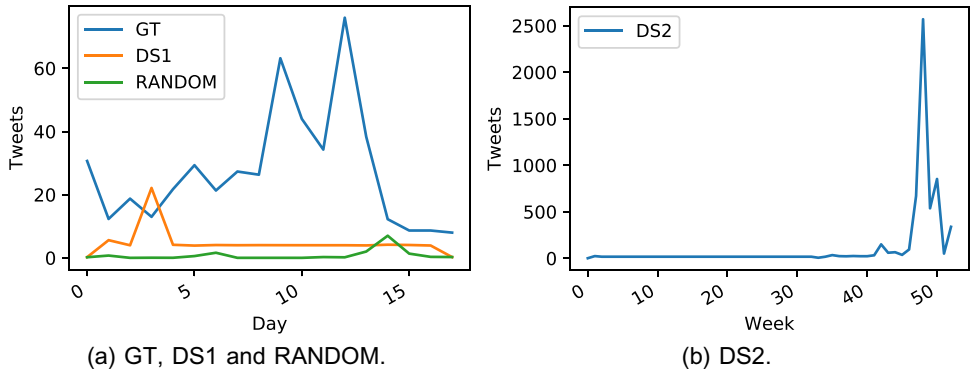


Fig. 6 Similarity matrices of content posted by HCC accounts (FSA_V, $\gamma = 15$). Each axis has an entry for each account, grouped by HCC. Each cell represents the similarity between the two corresponding accounts' content, calculated using cosine similarity (yellow = high similarity). Each account's content is modeled as a vector of 5 character n-grams of their combined tweets



(a) GT, DS1 and RANDOM. (b) DS2.

Fig. 7 Averaged temporal graphs of HCC activities (FSA_V, $\gamma = 15$)

deep analysis of the media landscape during the 2016 US election by Benkler et al. (2018) (which relied on simpler methods to build their networks), however these examples are networks of accounts rather than media organisations or sites, and, importantly, are not necessarily directly linked, offering the possibility of uncovering otherwise hidden connections between actors. This could be especially valuable when searching multiple OSNs.

4.3.2 HCC differentiation (RQ2)

How similar are the discovered HCCs to each other and to the rest of the corpus? The HCC detection methods used relied on network information; in contrast we examine content, metadata and temporal information to validate the results. We contrast DS1 and DS2 results with GT and a RANDOM dataset, constructed to match the HCC distributions in DS1 (FSA_V, $\gamma = 15$). As DS2 consisted entirely of bad actors, and GT consisted entirely of political accounts, it was felt non-HCC accounts from DS1 would offer more 'normal' non-coordinating accounts.

4.3.2.1 Internal consistency Visualising the similarities between accounts using the method in Sect. 3.2.4 (Fig. 6), the HCCs are discernible as being internally similar. The RANDOM groupings demonstrated little to no similarity, internal or external, as expected, while the DS2 HCCs demonstrated high internal similarity, as expected of organised accounts over an extended period. The internal consistency of the DS1 HCCs is not as clear as for DS2, possibly due to the greater number of HCCs. Where HCCs are highly similar to others (indicated by yellow cells off the diagonal), it is highly likely these are due to small HCCs (e.g. with two or three members) retweeting the same small set of tweets (fewer than ten). The use of filtering in conjunction with FSA_V may help remove potentially spurious HCCs, as could a final merge phase, joining HCC candidates whose evidence for coordination matches closely (e.g. two small HCCs retweeting 90% of the same tweets, kept separate by FSA_V but clearly similar).

4.3.2.2 Temporal patterns We applied the temporal averaging technique described in Sect. 3.2.7 to compare the daily activities of the HCCs found in GT, DS1 and RANDOM (all of which occur over the same time period) in Fig. 7a and weekly activities in DS2 in Fig. 7b. The GT accounts were clearly most active at two points prior to the election (around day 15), during the last leaders' debate and just prior to the mandatory electoral advertising blackout. DS1 and RANDOM HCCs were only consistently active at different times: around the day 3 leaders' debate and on election day, respectively. Inter-HCC variation may have dragged the mean activity value down, as many small HCCs were inactive each day. Reintroducing FSA's stitching element to FSA_V may avoid this. In DS2, HCC activity increased in the second half of 2016, culminating in a peak around the election, inflated by two very active HCCs, both of which had used many predominantly benign hashtags over the year.

4.3.2.3 Hashtag use The most frequent hashtags in the most active HCCs revealed the most in GT (Fig. 8a). It is possible to assign some HCCs to political parties via the examination of partisan hashtags (e.g. `#voteliberals` and `#orangelibs`), although the hashtags of contemporaneous cultural events are also prominent; for example, `#silentinvasion`, `#detours` and `#adlww` all relate to a contemporaneous international writers' festival. DS1 hashtags are all politically relevant, but are dominated by a single small HCC (rendered in pale green) which used many hashtags very often (Fig. 8b). These accounts clearly attempted to widely disseminate their tweets by using 1621 hashtags in 354 tweets. Furthermore, the hashtags they use relate to political discussions in many regions around the country (all listed hashtags that end in `pol` relate to the political discussion communities for each Australian state or the national community). Their prominence in hashtag use effectively hampers our ability to analyse the hashtag use of other HCCs, however, but seeing the results in context is important, as it helps to confirm that the pale green HCC is probably engaging in inauthentic behaviour. We can

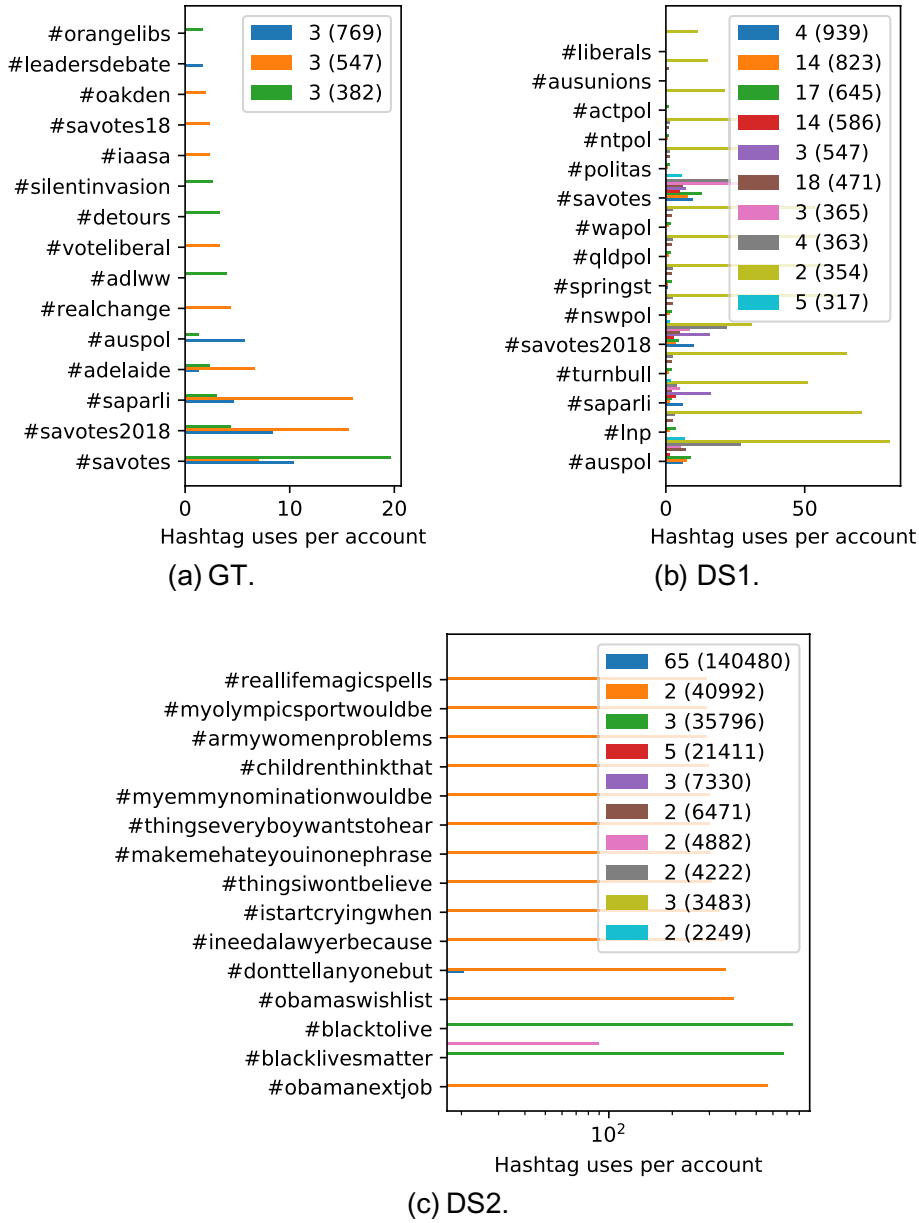


Fig. 8 Most used hashtags (per account) of the most active HCCs ($\gamma = 15$, FSA_V). The labels indicate HCC identifiers and post counts. Many HCCs are too inactive to be visible

Fig. 9 Clusters of hashtags relating to non-election events, including a writers festival, International Women's Day, and a multicultural festival, connected only when they appeared in the same tweet (GT). Wider edges represent a higher tweet count. Node colour implies the frequency of hashtag occurrences (darker means more)

still see that a large portion of hashtag use amongst the other listed HCCs relates to `#savotes`, `#savotes2018`, and `#saparli`, focussing on the South Australian election. If the hashtags had been irrelevant to the election, that could have provided evidence of accounts attempting to divert the discussion to other topics (because those tweets would still have needed to include the collection filter terms—i.e. ones relating to the election—to have been captured in the first

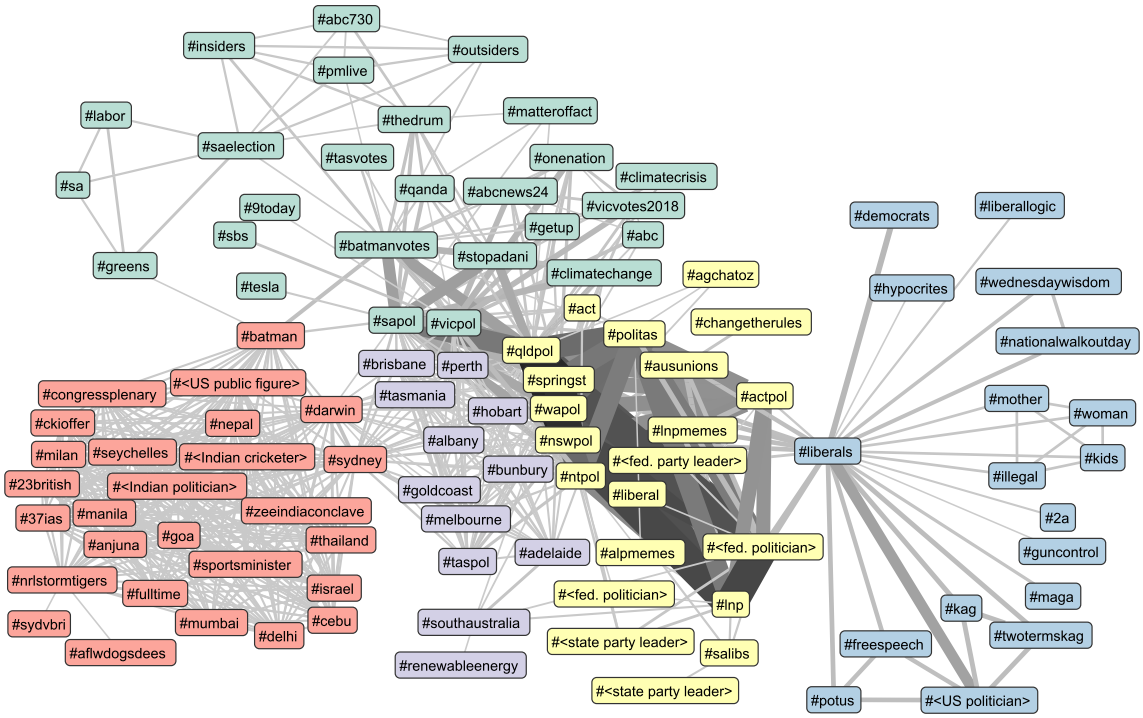


Fig. 10 Semantic network of hashtags used in DS1, connected only when they appeared in the same tweet. The minimum edge weight is 100 and the most highly co-occurring hashtags (`#savotes`, `#savotes2018`, `#saparli` and `#auspol`) have been excluded. Nodes are coloured according to Louvain clustering (Blondel et al. 2008), and some hashtags have been anonymised. Wider and darker edges represent a higher tweet count, and a darker background has been provided to improve contrast

place). Similarly, DS2 hashtags were dominated by a single HCC (using 41,317 relatively general hashtags in 40,992 tweets) and one issue-motivated HCC (Fig. 8c). Given DS2 covers an entire year, it is unsurprising the largest HCCs use such a variety of hashtags that their hashtags do not appear on the chart (little evidence of most of the HCCs listed in the legend appear visible in the barchart, despite the use of a log scale on the *x* axis), but it is revealing that at least a few of HCCs devoted much of their content to using hashtags, while the other most active HCCs did not, indicating that different HCCs detected by searching for one coordination strategy (co-retweet) are engaging (perhaps even more strongly) in other strategies. Perhaps these hashtag disseminator HCCs acted as distractors, supporters or even polluters, contributing messages sporadically but not consistently.

Analysing hashtag co-occurrences can help further explore the HCC discussions to determine if HCCs are truly single groups or merged ones. Applied to GT HCC activities (Fig. 8a), it was possible to delineate subsets of hashtags in use: e.g. one HCC promoted a political narrative in some tweets with `#orangelibs` (a partisan hashtag) and discussed cultural events such as the writers' festival in others with `#adlww` (Fig. 9), but was definitely one group.

Given the great number of hashtags used in even moderate sized datasets such as DS1, using hashtag co-occurrence analysis to examine the broader election discussion in DS1 requires filtering to reveal the core structure underlying the semantic network. We limited the minimum frequency of co-occurrences to 100 and also removed the most frequently occurring hashtags (`#savotes`, `#savotes2018`, `#saparli` and `#auspol`) to produce Fig. 10. Application of Louvain cluster detection (Blondel et al. 2008) exposes five clear clusters, though domain knowledge tells us that there is interesting conflation of topics within some of the clusters. The green cluster contains subclusters relating to current affairs television programmes (`#pmlive`, `#abc730`, `#insiders`, `#outsiders`, `#qanda` and `#thedrum`), political parties and advocacy groups (`#onenation`, `#labor`, `#greens`, and `#getup`) and relevant issues (`#climatechange`, `#climatecrisis`, and `#stopadani`). It also includes political hashtags (e.g. hashtags ending with `pol` and `votes`) that might fit better in the yellow cluster, which is dominated by them and forms the core of the semantic network by including the heaviest edges. The purple cluster consists primarily of location names, apart from `#renewableenergy` which hangs off `#southaustralia` (the focus of the election collection).

The other two clusters make apparent the fact that Twitter is an international network and hashtag clashes can draw in content irrelevant to local issues. The hashtag `#liberals`

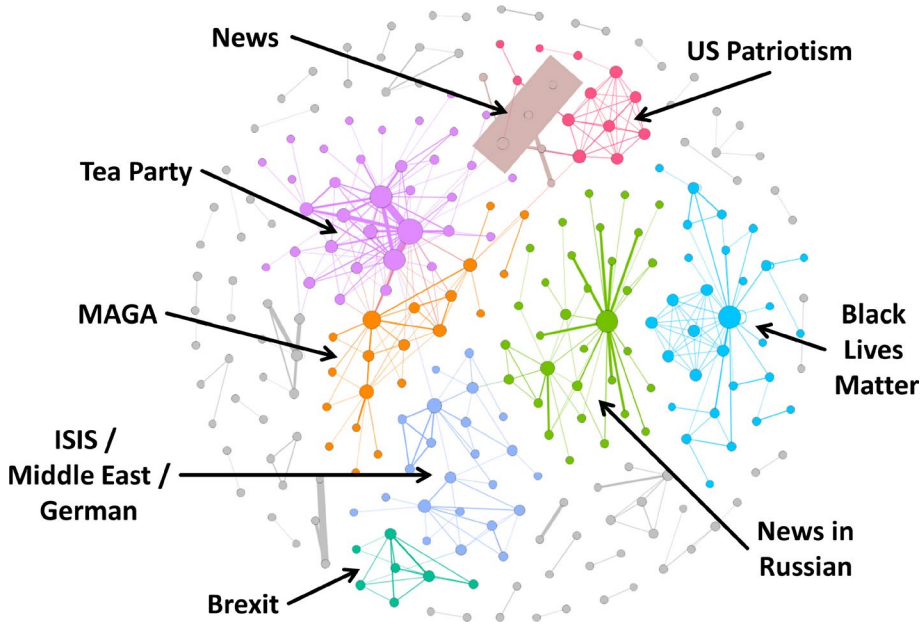


Fig. 11 Clusters of hashtags used in DS2, connected only when they appeared in the same tweet. The minimum edge weight is 100. Nodes are coloured according to Louvain clustering (Blondel et al. 2008), the most prominent of which have been annotated with their topic of discussion. Wider and darker edges represent a higher tweet count

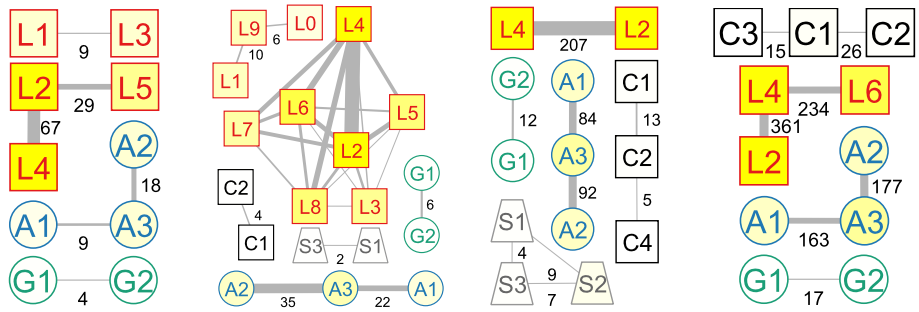
in the blue cluster can refer either to the Liberal party in South Australia (the major party that ultimately won the election) but is also used as a focus in American politics, especially right-wing politics, as reflected by the links to `#maga`, `#guncontrol` and `#2a` (i.e. the 2nd Amendment of the United States' Constitution, which refers to the right to bare arms), as well as `#nationalwalkoutday`. During the collection period, high school students in the United States staged a national day of protest against gun violence following a mass school shooting.[11] The red cluster also highlights content from outside the area of interest, with many terms relating to locations in other countries, possibly bound by sports, given the presence of `#fulltime`, `#nrlstormtigers`, `#aflwdogsdees`, and `#sydvbri`, the last three of which refer to Australian sporting matches between specific teams.

DS2 covers a longer period and seemed to consist of different teams of accounts driving different topics. As a consequence, its semantic network reveals clearly delineated (but often connected) discussion topics, as shown in Fig. 11. It is immediately notable that although the accounts in the dataset were flagged as trolls implicated in attempting to influence the US election, a lot of content is not in English and, in fact, appears to target other countries. This would be consistent with at least one other Russian campaign that targeted many Western audiences as well as Russians ("Secondary Infektion", Nimmo et al. 2020). Three non-English examples are apparent:

- The green cluster in the centre consists primarily of Russian news-related hashtags, perhaps aimed at a Russian audience to direct their attention to US election-related content.
- The pale blue central cluster has many hashtags related to the Middle East, including the ISIS terrorist group, but also German politicians and German names for nearby countries, such as Turkey. Germany's response to refugees from Syria escaping ISIS was politically contentious and may have been seen as an opportunity to foster divisions in the European Union and within Germany.
- The green cluster on the lower left is aimed at discussions of the United Kingdom's (UK) exit from the European Union (EU), otherwise referred to as Brexit. The UK held a referendum in 2016 on whether it should leave the EU and the campaigning caused significant division within the UK and Europe.

Other significant communities in the semantic network are the pink Tea Party/Conservatives Online (`#tcot` and `#ccot`) cluster, tightly connected to the emerging `#MAGA` cluster supporting Donald Trump, the red cluster focused on American patriotism and the highly active brown cluster including the terms `#news`, `#local`, `#business` and `#world`. The activity of HCCs shown previously in Fig. 8c presents a different and complementary view into hashtag use in the dataset, as very little of it apparent in the semantic network—it is the combination of not only which hashtags are associated together, but also which groups of accounts are using them that provides deeper insights. By finding groups that are using otherwise entirely disjoint sets of hashtags it may be possible to identify changes in narrative,

[11] https://www.nytimes.com/2018/03/14/us/school-walkout.html.



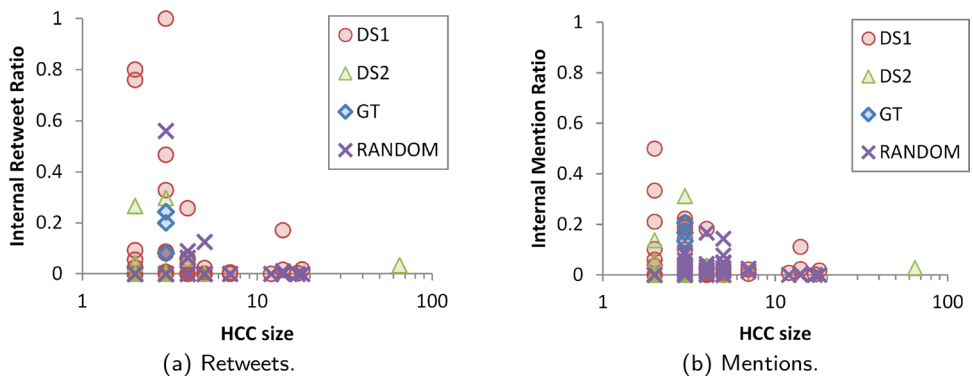
(a) γ= 15. (b) γ= 60. (c) γ= 360. (d) γ= 1440.

Fig. 12 Ground truth HCCs identified with FSA_V. Vertex shape = ideology (centre, left, right), colour = activity (brighter = higher), label and border colour = political party (L = red, A = blue, G = green, C = black), label = party and account identifier (e.g. 'G1' is Party G's account #1), link width = co-retweet count (some omitted for clarity)

Table 8 The most retweeted tweet in each GT HCC (FSA_V $\gamma = 15$)

RT `@alpsa`: A message from former `@AustralianLabor` Prime Minister, ⟨REDACTED⟩. https://t.co/⟨URL⟩*
RT @⟨`Reporter`⟩: Liberals promise $40m to tackle elective surgery waiting times in South Australian hospitals. `#SAVotes2018`... https://t.co/⟨URL⟩
RT `@SALibMedia`: Under Labor there aren't enough job opportunities for young South Australians. Here's what they are saying about @⟨`LaborPolitician`⟩ and `@alpsa` `#saparli` https://t.co/⟨URL⟩
RT @⟨`GreensPolitician`⟩: The results of this state election are clear—celebrity candidates and pop up parties come and go, but the Greens...https://t.co/⟨URL⟩

*N.B. URLs starting with 'https://t.co/' refer back to the original retweeted tweet's URL, and are obscured here for readability and anonymity



(a) Retweets. (b) Mentions.

Fig. 13 The proportions of each HCCs retweets and mentions referring to accounts within the HCC ($\gamma = 15$, FSA_V)

especially if HCCs can be tracked over time to see when they use which sets of hashtags.

4.3.2.4 Examining the ground truth The importance of having ground truth in context is demonstrated by Keller et al. (2017, 2019). By analysing the actions of known bad actors in a broad dataset, they could identify not just different subteams within the actors and their strategies, but their effect on the broader discussion. Many datasets comprising only bad actors (e.g. DS2) miss this context.

Considering GT alone, the HCCs identified consist only of members within the same political party, across all values of γ. Some accounts appeared in each window size. HCCs of six major parties were identified. Figure 12 shows the HCCs for each γ value. Some accounts and parties appeared at each window size, (e.g. parties L, A, G, and nodes L2, A1, G2), while some only appear in a few (e.g. parties C and S). This shows that different parties exhibited different approaches to retweeting and different members were involved over different time frames. Although party S members co-retweeted enough to appear in two time windows, they were not consistently active enough to re-appear in the largest time window, where their activity was overtaken by other accounts. It is particularly noticeable that the L party had two core cooperating accounts, L2 and L4, who were active enough to appear in each time window, and then a large team active in the hour-long window, implying that a deliberate strategy of team-based co-retweeting was employed (rather than a coincidental one). Rather than the posting times being highly coordinated (so that retweets could appear nearly simultaneously), it appears as if the L accounts were simply more attentive to their colleagues' tweets and retweets and retweeted them when they saw them

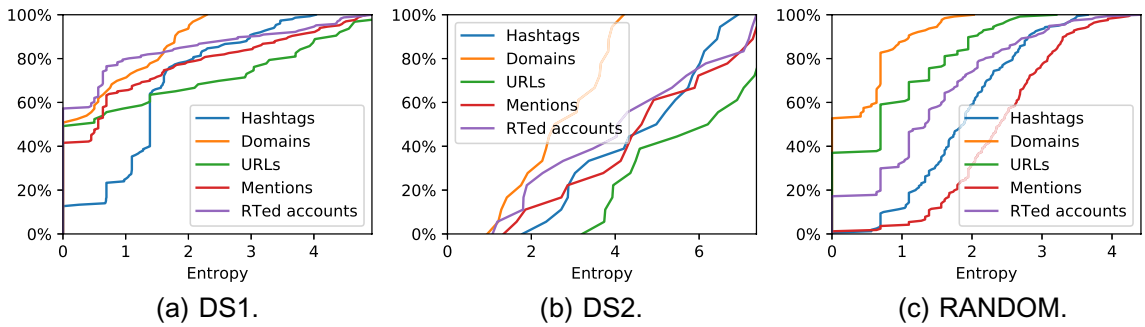


Fig. 14 Cumulative frequency of HCCs' entropy scores for five tweet features, comparing DS1 and DS2 with RANDOM (FSA_V, $\gamma = 15$). Feature variation increases along the *x* axis

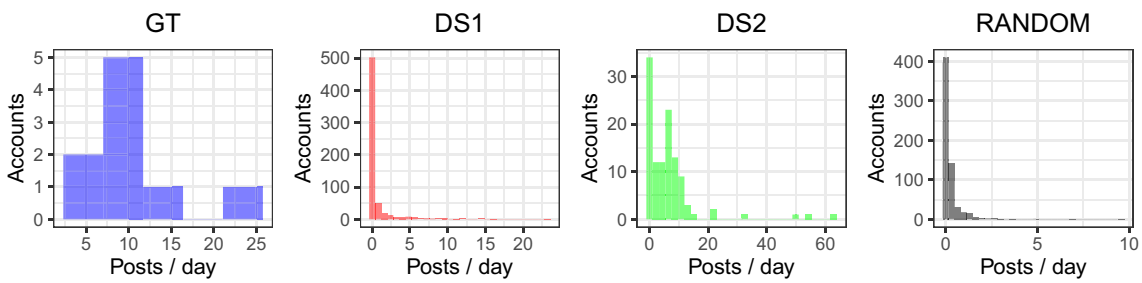


Fig. 15 Histograms of the daily posting rates of accounts in the GT, DS1, DS2, and RANDOM HCCs (FSA_V, $\gamma = 15$). Because the datasets cover different periods of time, the posting rate enables a fairer comparison. The distributions of DS1 and RANDOM posting rates are similar and notably different to those of DS2, while GT includes a higher proportion of more active accounts than the other datasets

(which often occurred within an hour), as could be expected of any social media-savvy group.

Examining the content of these HCCs confirmed that they were genuine communities engaging in co-retweeting (though not necessarily deliberately). The top retweeted tweets of each HCC (FSA_V, $\gamma = 15$) are shown in Table 8. Using the tweets each HCC posted, it is possible to attribute each to a political affiliation, if not a party, without resorting to inspecting each member's identity.

4.3.3 Focus of connectivity (RQ3)

The IRRs and IMRs for the HCCs in the DS1, DS2, GT and RANDOM datasets are shown in Fig. 13. The larger the HCC size, the greater the likelihood of retweeting or mentioning internally, so it is notable that DS2's largest HCC has IRR and IMR's of around 0, though even the smaller HCCs have low ratios. Ratios for the smallest HCCs seem largest, possibly due to low numbers of posts, many of which may be retweets or include a mention, inflating the ratios. The hypothesis that political accounts would retweet and mention themselves frequently is not confirmed by these results, possibly because they are retweeting and mentioning official or party accounts outside the HCCs.

4.3.4 Content variation (RQ4)

We compared the entropy of features used by DS1 and DS2 HCCs to RANDOM ones (Fig. 14). Many of DS1's small HCCs used only one of a particular feature, resulting in an entropy score of 0 (Fig. 14a). In contrast, DS2's fewer HCCs have higher entropy values (Fig. 14b), likely because more of their activity was collected (365, not 18, days' worth) and they therefore had more opportunity to use more feature values. The majority of HCCs used few hashtags and URL domains, which is to be expected as the dominating domain is *twitter.com*; this domain is embedded in all retweets as part of the link back to the original (retweeted) tweet. Compared to the RANDOM HCCs (Fig. 14c), DS1 HCCs had lower variation in all features, while the longer activity period of DS2 resulted in distinctly different entropy distributions. Because DS1 HCC activity appears to have been more deliberate, and perhaps coordinated, it may be that the HCCs were more focused on their topic of conversation (especially when contrasted with RANDOM HCCs).

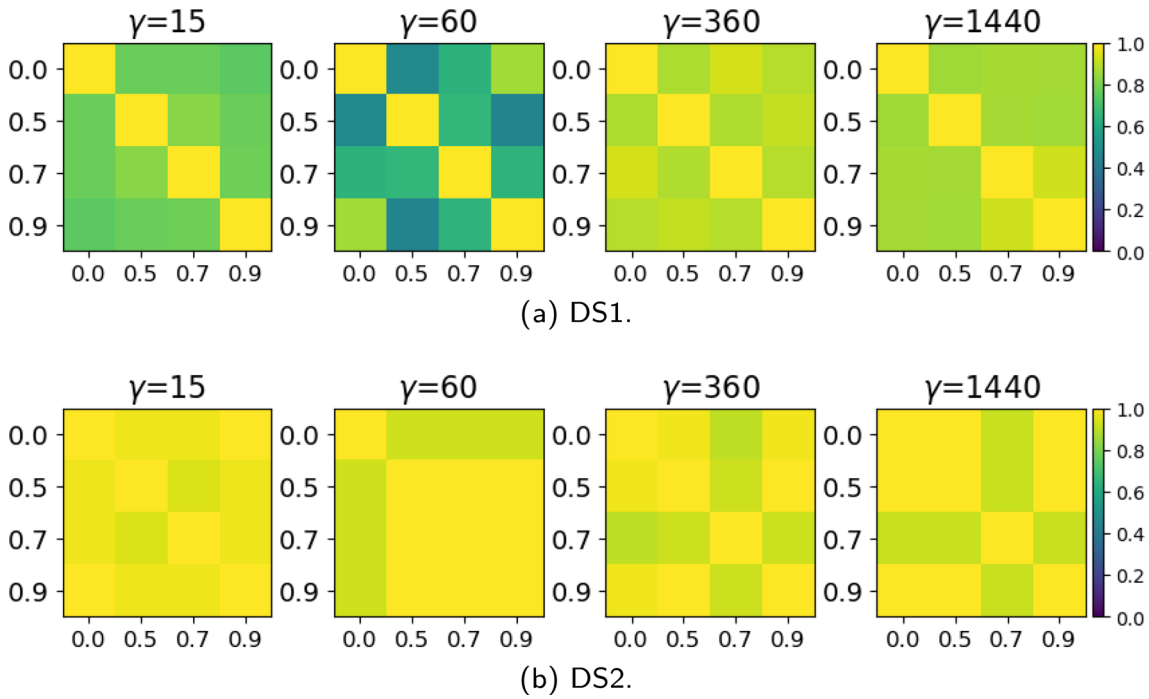


Fig. 16 Jaccard similarity of HCC membership when varying α (0.0 is the Baseline)

Table 9 Statistics of discovered HCCs while varying α (FSA_V, Boost)

		$\gamma = 15$			$\gamma = 60$			$\gamma = 360$			$\gamma = 1440$		
		N	E	C	N	E	C	N	E	C	N	E	C
DS1	Baseline	633	753	167	619	1293	151	503	1119	127	815	2019	141
	$\alpha = 0.5$	604	711	168	1178	2121	149	519	1183	129	800	2037	137
	$\alpha = 0.7$	578	697	160	847	1569	149	518	1155	130	792	1997	136
	$\alpha = 0.9$	596	706	165	585	1223	145	530	1188	134	796	1995	141
DS2	Baseline	113	758	19	77	394	18	98	775	15	69	380	15
	$\alpha = 0.5$	116	760	20	79	396	18	100	776	16	69	380	15
	$\alpha = 0.7$	110	756	18	79	395	18	102	777	17	69	381	15
	$\alpha = 0.9$	113	758	19	79	396	18	100	776	16	69	381	15

$T = 5$ except in the Baseline condition. N = node count, E = edge count, C = HCC count

Compared with RANDOM HCCs, DS1 HCCs retweeted fewer accounts, used fewer URLs (though they were from a similar distribution of domains), and many fewer mentions and hashtags. Many non-HCC accounts posted only a single retweet as their contribution to the discussion, and so it may be that a relatively high proportion the RANDOM HCC members only posted a single tweet, causing the distributions observed. The RANDOM HCC members posted 3147 tweets compared with DS1 HCCs' 8527 tweets, despite having the same number of members, so DS1 HCC members posted more than 2.7 times as often. Although DS1 accounts posted more tweets per individual than the RANDOM accounts, their distribution appears similar, and notably different to those of both DS2 and GT (Fig. 15).

4.3.5 Consistent coordination (RQ5)

The sliding frame technique from Sect. 3.2.9 was applied to DS1 and DS2 to reveal HCCs engaging in coordination consistently in adjacent time windows. The baseline used $T = 1$ (i.e. a sliding frame a single time window wide) and $\alpha = 0.0$. For the three other conditions, T was set to 5 (as γ increases approximately five times each time) and $\alpha = \{0.5, 0.7, 0.9\}$. In this way, the choice of $\alpha = 0.9$ would most strongly consider the contribution of LCNs from preceding time windows. Once applied for each time window, the aggregated LCNs were mined for HCCs and then the membership of these were compared in the same manner as in Sect. 4.3.1 using Jaccard similarity (Eq. 2). As noted earlier, Jaccard

Table 10 Features selected from both account activity and collective HCC activity based on their activity network (described in Sect. 3.2.10)

	Account-level	Group-level
Instances (Uses)	Posts, reposts, replies, mentions, hashtags, URLs	Posts, interactions, user nodes, hashtags, URLs, reposts, quotes, mentions, replies, in-conversations (see Sect. 3)
Unique	Mentions, hashtags, URLs	HCC members, Nodes in the network (including URLs and hashtags), hashtags, URLs
Rates	Posts/minute	Reposts of HCC members/all reposts (cf. IRR), mentions of HCC members/all mentions (cf. IMR), replies to HCC members/all replies
Profile	Default image (boolean) Characters in description Characters in URL	–

IRR and IMR are defined in Sect. 4.3.3

similarity is stricter about set matching than the Overlap method (Eq. 3). Even so, as can be seen in Fig. 16, changes introduced by using the decaying sliding frame with different α values were insignificant in all cases, except for DS1 and $\gamma = 60$. The implication, which is borne out when the exact network sizes (in nodes) are compared in Table 9, is that the previous windows did not add significant numbers of nodes, but instead increased the weight of existing edges, so the HCCs that were detected consisted of the same members working together over time, rather than splitting into subsets. To hide a team's coordination, one might expect that its members would associate separately in different time windows, but that does not appear to have happened significantly in these datasets, except in the shorter time windows in DS1, the majority of which may very well be coincidental.

The greatest variation in node and edge count occurs in the shorter windows in DS1 ($\gamma = \{15, 60\}$), probably because of the greater number of accounts active in DS1 (compared to DS2): accounts have more alters to form HCCs with in DS1, which has 20.5k accounts, whereas choice in DS2 is limited to 1.3k accounts. The near doubling of accounts in DS1's HCCs when $\gamma = 60$ implies accounts co-retweeted often just within a single hour, and then not again (at least not for $T = 5$ h). This effect is swamped by the much more active consistent co-retweeting of a smaller set of users when α is increased to 0.7 and above. Given the membership varies so little in the other conditions, an analysis of how these HCCs form and change over time is required. It is clear, however, that this approach would be best suited to filter-based collections, as they are likely to capture more accounts.

4.3.6 Validation via HCC classification (RQ6)

Our final validation method relies on the HCCs in GT as positive examples of coordinating sets of accounts, given it is reasonable to assume that they ought to be coordinating their activities during an election campaign (an intuition shared by Vargas et al. 2020). The purpose of this particular activity is not to build a classifier for coordinated behaviour in general, or coordinated amplification specifically, but to provide a degree of confidence that the HCCs detected in DS1 and DS2 are exhibiting similar behaviour to those in the GT.

4.3.6.1 Feature selection As mentioned in Sect. 3.2.10, features are drawn from individual accounts *and* their groupings, specifically based on their individual and collective behaviour and homophily. For this reason, we select account-level features as well as group-level features to make up each account's feature vector, meaning that some of the values for HCC co-members will be identical. The account-level features are all drawn from their activity within the dataset, while the group-level features are drawn from the HCC's activity network (see Sect. 3.2.10) and are included in the feature vector of each member of the HCC. The account- and group-level features used are shown in Table 10.

4.3.6.2 Classification results After being trained on the GT HCCs, the classifiers were then applied to the HCCs in DS1 and DS2. We use COORDINATING and NON-COORDINATING to represent the positive and unlabeled classes, respectively. A second disjoint subset of RANDOM HCCs were created for this testing by sampling accounts outside the ground truth and training sets. Upsampling was also used to ensure the classes were balanced with at least 400 instances each.

The accuracy of the best classifier for each dataset and time window ranged from 0.69 to 0.91 (shown in Table 11), with performance varying between classifiers and window sizes, but mostly recognising HCC members in DS1 slightly better than DS2. This difference may be because the training data was sourced from the same online discussion (though using the behaviour of completely different accounts). F_1 scores (outside $\gamma = 360$) for the COORDINATING (F_{1P}) and NON-COORDINATING (F_{1U}) instances ranged from 0.80 to 0.91 and 0.67 to 0.91, respectively. Each classifier performed best for DS1 in different time windows, except for $\gamma = 360$, but all classifiers performed well, with the worst

Table 11 Accuracy (Acc.), positive class (COORDINATING) F_1-scores (F_{1P}) and unlabeled class (NON-COORDINATING) F_1-scores (F_{1U}) from the HCC classifiers

Classifier	$\gamma = 15$			$\gamma = 60$			$\gamma = 360$			$\gamma = 1440$		
	Acc.	F_{1P}	F_{1U}	Acc.	F_{1P}	F_{1U}	Acc.	F_{1P}	F_{1U}	Acc.	F_{1P}	F_{1U}
DS1												
SVM	**0.91**	**0.91**	**0.90**	**0.80**	**0.82**	**0.77**	0.56	0.39	0.65	0.88	0.88	0.88
RF	0.72	0.76	0.67	0.63	0.63	0.64	0.59	0.40	0.68	**0.90**	**0.89**	**0.91**
BPU	0.70	0.74	0.64	0.66	0.65	0.67	**0.69**	**0.63**	**0.73**	0.88	0.86	0.89
DS2												
SVM	**0.84**	**0.86**	**0.81**	0.73	0.79	0.64	0.81	0.84	0.76	0.81	0.84	0.77
RF	0.81	0.84	0.77	**0.75**	**0.80**	**0.67**	**0.85**	**0.87**	**0.82**	**0.89**	**0.90**	**0.88**
BPU	0.81	0.84	0.76	**0.75**	**0.80**	**0.67**	0.84	0.86	0.80	0.88	0.89	0.86

Table 12 Precision and recall for the positive (COORDINATING) class

Classifier	$\gamma = 15$		$\gamma = 60$		$\gamma = 360$		$\gamma = 1440$	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
DS1								
SVM	**0.85**	**0.99**	**0.74**	**0.92**	0.62	0.28	0.90	**0.85**
RF	0.67	0.88	0.64	0.62	0.73	0.28	**1.00**	0.80
BPU	0.65	0.86	0.67	0.64	**0.78**	**0.54**	**1.00**	0.75
DS2								
SVM	**0.76**	**1.00**	0.65	**1.00**	0.72	**1.00**	0.73	**1.00**
RF	0.73	**1.00**	**0.67**	**1.00**	**0.77**	**1.00**	**0.82**	**1.00**
BPU	0.72	**1.00**	**0.67**	**1.00**	0.75	**1.00**	0.80	**1.00**

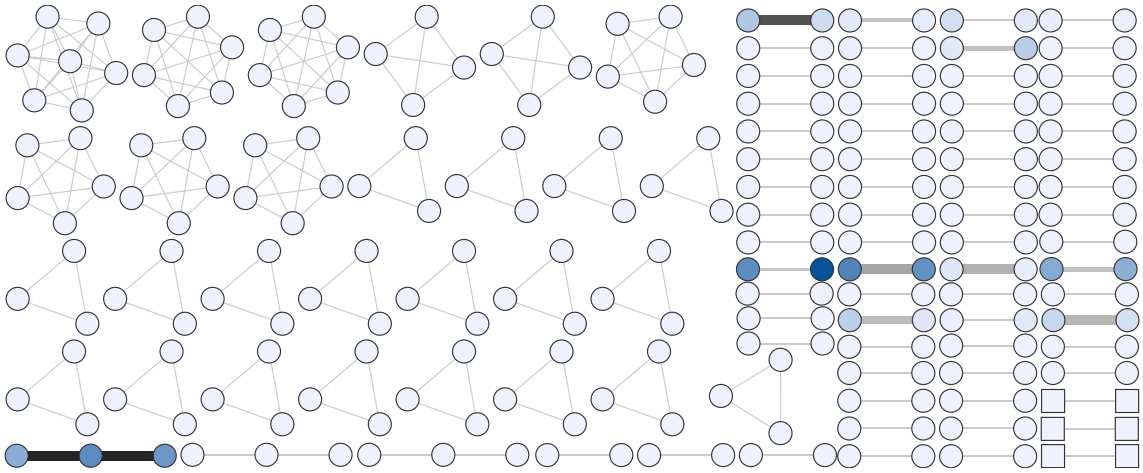


Fig. 17 While searching for *Bully*ing behaviour in DS1, these are HCCs of accounts found engaging in co-mentions (circles) and co-mentions plus co-convs, i.e. engaged in both (square vertices in bottom right) ($\gamma = 360$, FSA_V, $\theta = 0.01$). Edge thickness and darkness = inferred connections (darker = more). Vertex colour = tweets posted by that account (darker = more)

accuracy at 0.69. All classifiers also performed the least well in the six hour time window for DS1, possibly because the GT HCCs' activity coordination was most prominent over the short time frames of an hour or less, and otherwise at the day level. Even so, F_{1U} scores consistently hover around 0.7 when $\gamma = 360$, which is significantly better than random, though the F_{1P} scores around 0.40 for SVM and RF indicate difficulty identifying all COORDINATING HCC members, a detail which is discussed in more detail below. The accuracy and F_1 results show that the the classifiers could all be successfully trained to recognise GT HCCs in most time windows and that the GT HCCs represented most of the HCCs in DS1 and DS2, despite the different levels of activity (DS2 HCC members interacted more than DS1 or GT HCC members in their corpus, primarily because the collection period was longer).

Table 12 shows Precision and Recall across all classifiers and datasets for the COORDINATING class. (Given our emphasis on recognising COORDINATING instances, we do not present the corresponding results for the NON-COORDINATING class here.) For all time windows, Precision is high for the classifiers against DS1 (ranging from 0.62 to 1.00) and moderate against DS2 (ranging from 0.67 to 0.82), meaning that the HCCs are clearly discernible from the NON-COORDINATING instances (i.e. if an instance was classified as COORDINATING, then it was almost certainly a member of an HCC). Recall varies significantly for DS1 (between 0.28 and 0.99), but is perfect (i.e. 1.00) for DS2, meaning that some DS1 HCCs were rejected incorrectly, while all DS2 HCC members were identified. The Recall scores for $\gamma = 360$ explain why the F_{1P} scores were so low in Table 11, because the corresponding Precision scores are still relatively high. As mentioned above, there is something particular about the 6 h time window ($\gamma = 360$), as the GT HCC members (via their account features and group behavioural features) were less easily distinguishable from the randomised NON-COORDINATING accounts, resulting in poorer classifier performance. The reason for this is possibly the choice of window boundaries. The time window boundaries rested at 0000, 0600, 1200, and 1800 h, while boundaries defined more by work activity (e.g. 0200, 0800, 1400, 2000 h) may better match human activity patterns. For other, less geographically bound datasets (i.e. ones where the activity comes from around the world, rather than from a single or small group of adjacent timezones), other ground truth may be required.

SVM was the best performing classifier for COORDINATING accounts in DS1 in the shorter time windows ($\gamma = \{15, 60\}$) and had close to equal top performance in $\gamma = 1440$, but BPU clearly performed best in the challenging six hour window, including with moderately better Precision and markedly better Recall than SVM and RF. For DS2, all classifiers performed well, with RF most often performing best, but only marginally. SVM struggled to compete in the day long period, though still achieved moderate scores for Precision and Accuracy. For that reason, we can argue that RF performed best overall, but the margin was minimal. Importantly, classifiers found all DS2 HCC members, though they incorrectly included some false positives.

Consequently, by accepting the assumption that the ground truth HCCs exhibited at least one type of coordination, these classifiers provide confidence that the other HCCs appear similar to the GT ones and thus may have behaved in similar ways. The question of intent remains, however. Examining the content subjected to coordination will likely provide clues, but deeper examination of behaviour to identify, e.g. Principal-Actor patterns (Giglietto et al. 2020b), may also be enlightening. More examples of similar

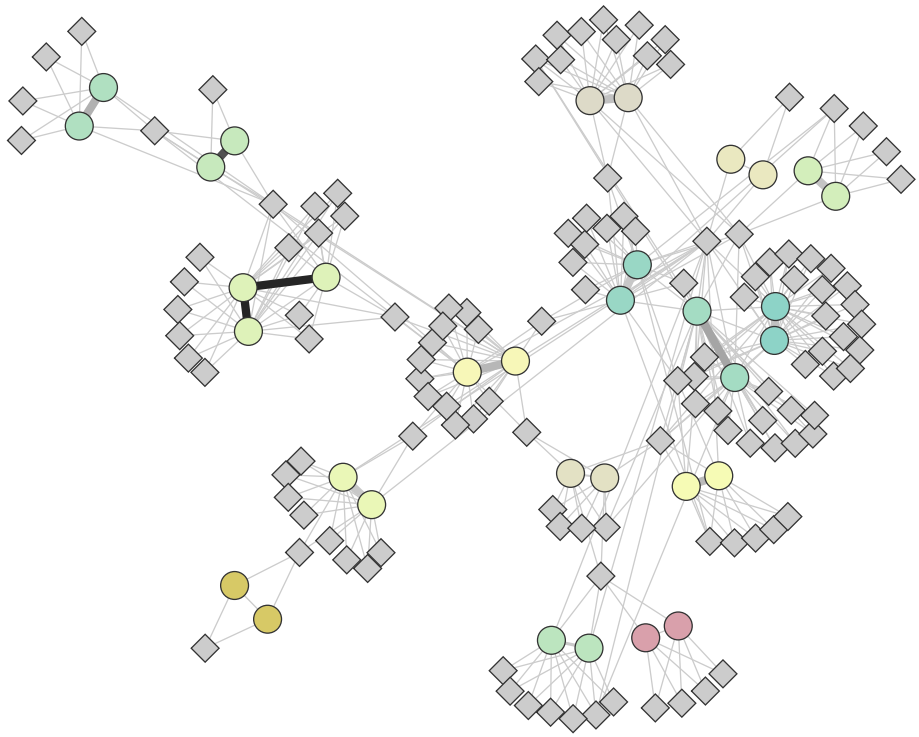


Fig. 18 A network of DS1 HCC accounts (circle vertices) connected to the accounts they mention or conversations they join (diamonds). Accounts in the same HCC share a colour. Clear communities surrounding HCCs indicate who they converse with, and which conversants are co-mentioned by multiple HCC accounts. The width and darkness of the edges between HCC accounts indicates the weight of evidence linking them (darker implies more)

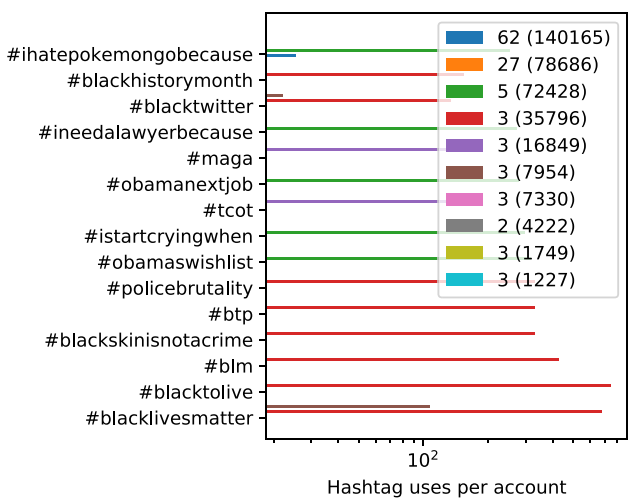


Fig. 19 Hashtag uses of the most active HCCs boosting accounts (FSA_V, $\gamma = 15$). The labels indicate HCC member and tweet counts. Many HCCs are too inactive to be visible

coordination activities as well as other coordination types would bolster the positive training and testing sets, as well as expand knowledge regarding coordination strategies in use online. Furthermore, Vargas et al. (2020) make the point in their work on detecting SIOs that "SIO coordination should be seen as a spectrum and not a binary state...[which could lead to] ...an overestimation of accounts that are part of disinformation campaigns" (p. 142, Vargas et al. 2020) potentially silencing those who need their voice to heard the most. For this reason, the application of binary classifiers for SIO detection ought to be part of a larger overall process with strong oversight.

4.3.7 Multiple criteria: *Bully*ing

Some strategies can involve a combination of actions. Behaviours that contribute to *Bully*ing by dogpiling, for example, include joining conversations started by the target's posts and mentioning the target repeatedly, within a confined time frame. As DS1 included all replied to tweets, we investigated it inferring links via co-mentions and co-convs (FSA_V, $\theta = 0.001$, $\gamma = 10$ min), having maximised the ratio of MEW to HCC size. Of 142 HCCs discovered, the largest had five accounts and most only had two. Only 32 had more than ten inferred connections, but five had more than 1000. These heavily connected accounts, after deep analysis, were simply very active Twitter users who engaged others in conversation via mentions, which outweighed the more strict co-conv criterion of participants *reply*ing into the same conversation reply tree.

A larger window size was considered ($\gamma = 360$) in case co-conv interactions were more prevalent. FSA_V ($\theta = 0.01$) exposed little further evidence of co-conv (Fig. 17), finding 98 small HCCs again dominated by co-mentions, not many of which had more than one inferred connection, implying most links were incidental; FSA_V did not filter these out.

This provides an argument for a more sophisticated approach to combining LCN edge weights for analysis than

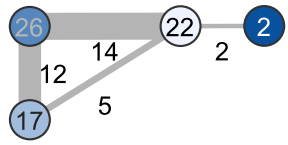


Fig. 20 The most active DS1 co-retweet HCC (γ = 10 s). Node label = post count, node colour = Botometer scores (higher = darker), link thickness and label = co-retweet occurrences

Table 13 Execution times (in seconds)

	DS1				DS2			
Tweets	115,913				1,571,245			
Parse raw (Step 1)	19.0 (from JSON)				74.0 (from CSV)			
Window size γ (min)	15	60	360	1440	15	60	360	1440
Find evidence and build LCNs	15.0	28.0	123.0	427.0	121.0	106.0	246.0	567.0
Aggregate LCNs	27.0	65.6	168.5	170.7	70.4	55.2	35.6	22.7
HCCs: FSA_V	28.3	58.2	126.1	209.3	6.3	4.2	5.8	5.0
HCCs: kNN	9.0	22.7	97.5	206.4	4.3	4.3	4.7	4.6
HCCs: Threshold	5.2	11.9	34.6	64.0	2.2	2.3	2.7	2.7

Eq. (1), and that FSA_V could be modified to better balance HCC size and edge weight. Furthermore, it is likely that bullying accounts will not just co-mention accounts frequently, but have low diversity in the accounts they co-mention, i.e. they will repeatedly co-mention a small set of accounts, and spend a disproportional number of their tweets doing so. A further consideration is that participants in long discussions (reply trees) often include the author of the original tweet that sparked the discussion, and it would be misleading to include their account in results, implying that they *bull*ied themselves. Finally, patterns of behaviour that would clearly qualify as conversations were observed in the datasets that did not fit the strict 'conversation tree' model: accounts would mention several collocutors at the start of every tweet, but only reply to a tweet of one of them while continuing the conversation. Importantly, sometimes the mentioned accounts included in tweets were prominent individuals whose names were included not because they were active participants in the conversation, but because the tweeter wanted to draw their attention to the conversation (regardless of the likelihood that the attempt would succeed; e.g. some tweets included references to prominent and busy politicians who would be unlikely to wade into arbitrary online discussions).

4.3.8 HCC inter-relationships

To study the relationships between HCCs, we create two-level networks starting with the HCC network and then adding nodes representing the elements of evidence linking them, known as *reason* nodes (e.g. the tweets they co-retweet or the hashtags they use in common). Figure 18 shows the largest component after such expansion was conducted on the HCCs in Fig. 17. HCC accounts (circles) share colours and the distribution of the reasons for their connection (diamonds) show which other accounts are uniquely mentioned by an HCC and which are mentioned by more than one HCC. Heavy links between HCC accounts with few adjacent reason vertices imply these accounts are mentioning a small set of other accounts on many occasions.

4.3.9 *Boost*ing accounts, not just posts

It is possible to *Boost* an account rather than just a post. Returning to DS2, we sought HCCs from accounts retweeting the same account (FSA_V, γ = 15), and found that the hashtag use revealed further insights (Fig. 19). No longer does one HCC dominate the hashtags. Instead clear themes are exhibited by different HCCs, but again, they are not the largest HCCs. The red HCC uses `#blacklivesmatter` and other Black rights-related hashtags (including `#blm`, `#blacktolive`, `#blackskinisnotacrime`, `#policebrutality` and `#btp`[12]), while the purple HCC uses pro-Republican ones (`#maga` and `#tcot`), and the green HCC is more general. Given the number of tweets these HCCs posted over 2016 (at least 16, 849), it is clear

[12] BTP refers to the British Transport Police, the conduct of which was discussed in accounts of the arrest of a Black man at a London train station in mid-2016, e.g. https://www.theguardian.com/uk-news/2016/jul/28/man-complains-after-police-place-spit-hood-over-head-during-arrest-london-bridge.

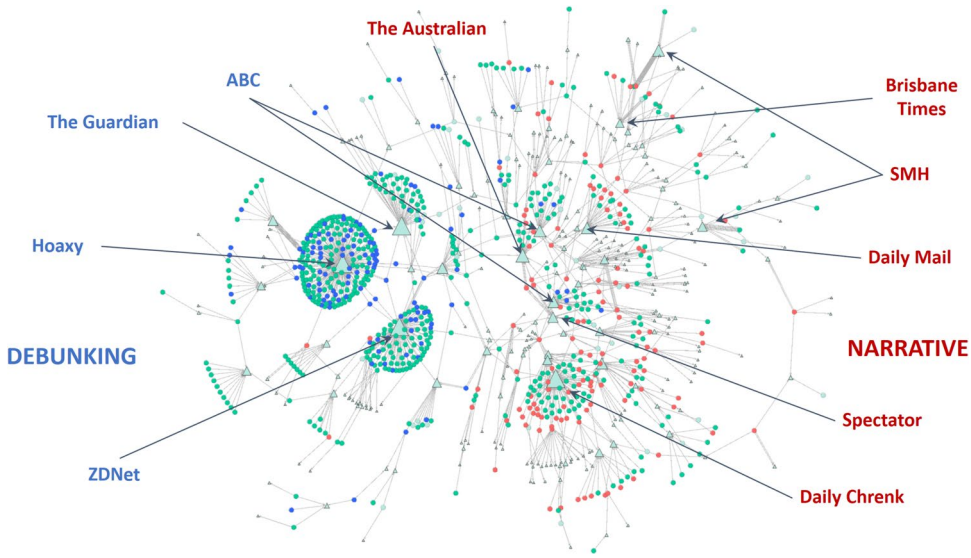


Fig. 21 Annotated account/URL bipartite network constructed from co-URL analysis of the ArsonEmergency dataset. Circles represent HCC accounts (Threshold, $t = 0.1$, $\gamma = 10$ s) and triangles represent URLs mostly referring to news articles. Accounts are linked to the URLs they shared, with multi-edges representing each use of a particular URL. URL nodes are sized by in-degree, and all coloured pale green. *Supporter* nodes are coloured red, *Opposer* nodes are blue, while *Unaffiliated* ones are green. The most widely shared articles are annotated with the website on which they are hosted (**N.B.** ABC = Australian Broadcasting Corporation, SMH = Sydney Morning Herald). Blue annotated articles are categorised as DEBUNKING, while red ones are categorised as supporting or prominently discussing the 'arson' NARRATIVE

they concentrated their messaging on particular topics, some politically charged. It is arguable that their contributions helped inflame tensions and stoke divisions in socially sensitive topics, not just in the United States, but in the UK as well, and at the very least sought to draw the attention of others.

The green HCC may be acting in distractor or polluter roles, as previously suggested, given their contribution of 72,428 tweets over the year (an average of nearly 40 tweets per account every day).

4.3.10 Validation of inauthentic behaviour detection

The approach presented can be used to perform analytics similar to the Rapid Retweet Network used by Pacheco et al. (2020), who used it to expose tight clusters of bot-like accounts, which retweeted the same tweet within 10 s of it appearing. We varied this for the DS1 dataset (due to its small nature) and searched for accounts which retweeted the same tweet within 10 s, regardless of the age of the original tweet. We discovered a tight cluster of accounts, most with relatively high Botometer CAP scores[13] (Davis et al. 2016), shown in Fig. 20. The scores were as follows: node 26: 0.787; node 22: 0.381; node 2: 0.949; and node 17: 0.464. All were high relative to the other accounts in the corpus, most of which had scores well below 0.2; all four were had scores well above 0.2, but the scores of two were also well above the 'bot' threshold of 0.6. On further inspection, they appeared to support vocational training and left-wing issues and posted retweets almost exclusively, but the content all related to the election. This finding enhances the bot ratings by making it clear which bots (or bot-like accounts) appear to work together. It also raises further questions regarding bot detection systems, however, as some of the accounts appeared to be genuinely human, though unusually active. These accounts appeared to work together to actively disseminate messages aligned with their preferred narrative, though with a very low IRR (just shy of 10%) despite most of their activity being retweets (97.7%), so to a certain degree it matters not whether they are automated or genuinely human-driven, but whether they are engaging in astroturfing or other inauthentic behaviour. In this circumstance, they may be genuine agenda-driven users, but they were definitely all highly attentive to the same sources. Alternatively, when we consider their bot ratings more closely, it is possible that there is a mixture of account types, with node 26, in particular, acting as an automated 'cheerleader' for nodes 22 and 17. Examining relative timings of their posts (to answer whether node 26 consistently was the second co-retweeter when paired with nodes 22 and 17) could reveal support for this hypothesis.

[13] The English score variant was used as both the datasets were either primarily in English or aimed at English speaking audiences.

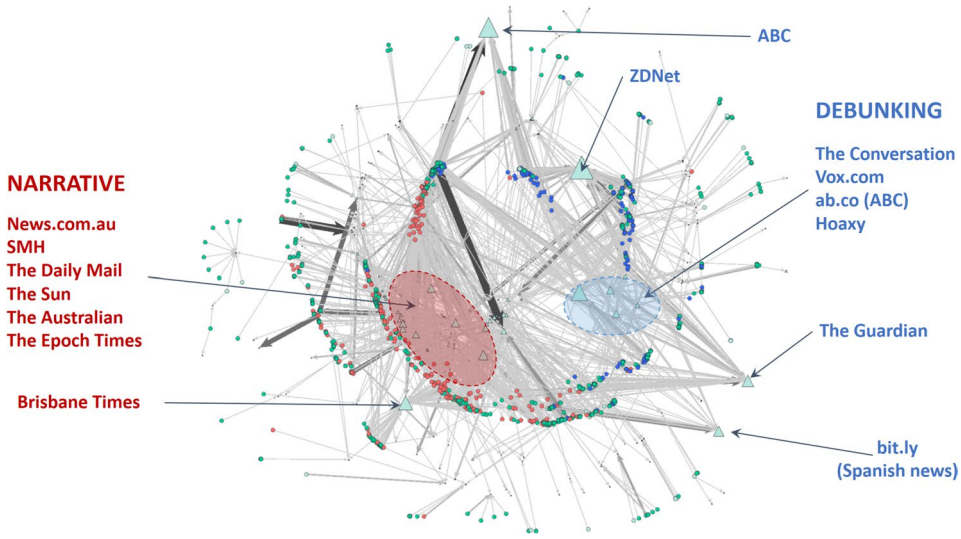


Fig. 22 Annotated account/URL domain bipartite network constructed from co-domain analysis of the ArsonEmergency dataset. Circles are HCC accounts (Threshold, $t = 0.1$, $\gamma = 10$ s) and triangles represent the domains of URLs used in tweets. Accounts are linked to the domains of URLs they shared, with thicker, darker edges representing frequent use of a particular domain. Domain nodes are sized by in-degree, and all coloured pale green. *Supporter* nodes are coloured red, *Opposer* nodes are blue, while *Unaffiliated* ones are green. The most frequently referred to domains are annotated with the organisation to which they belong (**N.B.** ABC and ab.co = Australian Broadcasting Corporation, SMH = Sydney Morning Herald, News.com.au = News Corporation). Blue annotated domains are categorised as DEBUNKING, while red ones are categorised as supporting or prominently discussing the 'arson' NARRATIVE. The red zone includes a number of DEBUNKING domains and is mostly referred to by *Supporters* while the blue zone includes academic and centre and left wing domains categorised as DEBUNKING domains, which are referred to predominantly by *Opposers*

4.3.11 Performance

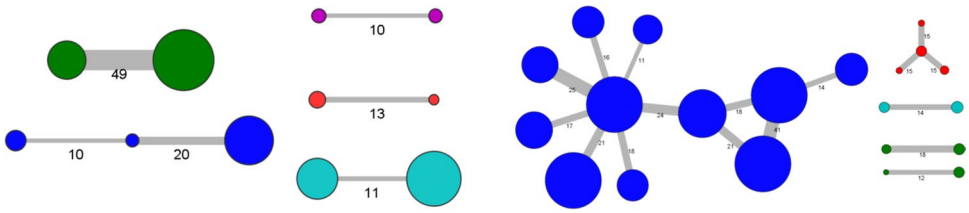
In Table 13 we present the timings observed for the stages of processing for DS1 and DS2 conducted on a Dell Precision 5520 laptop equipped with an Intel Core i7-7820HQ CPU (2.9 GHz), 32Gb RAM, and an NVMe PC300 480Gb SSD, running Windows 10. Parsing raw data is relatively cheap, with DS2's 1.5m tweets processed in just over a minute, and LCN construction is dependent on the degree of activity and the number of accounts. DS1's larger account pool increased the size of the networks generated, and all associated post-processing. The size of DS1 LCNs were an order of magnitude greater than DS2's (in nodes and edges), resulting in increasing execution times for aggregation and HCC extraction.

4.4 Applications

Complementing the detailed validation presented above, in this section, we offer two case studies in which our method has been used to demonstrate its utility. Extending a study of polarised online communities in a discussion of bushfires in Australia (Weber et al. 2020), the co-URL and co-URL domain analysis we conducted revealed how sources of information were used by discussion participants, and how that use differed between the polarised communities. A second study of Twitter activity during the Democratic and Republican Conventions in the United States in August, 2020, makes use of co-retweeting analysis in order to reveal influence attempts with social bots and co-hashtag analysis to discover discussion groups and their relations.

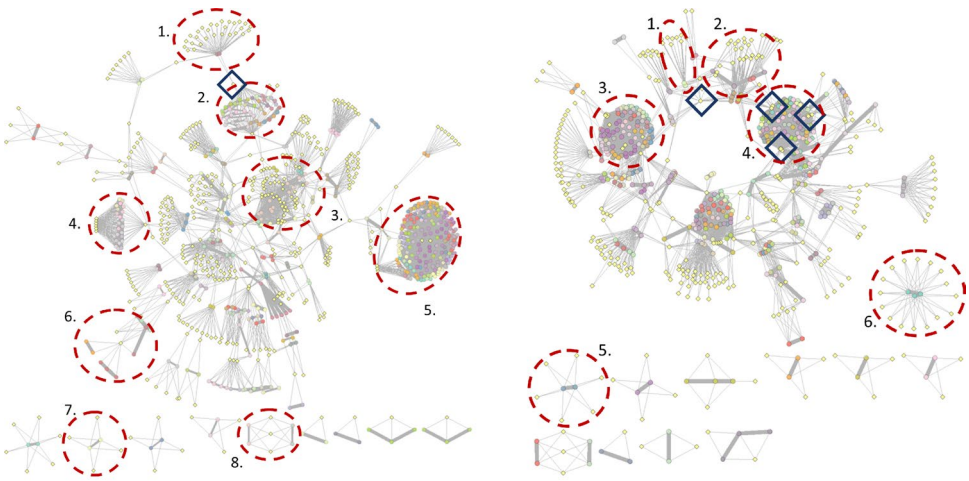
4.4.1 `#ArsonEmergency` and Australia's "Black Summer"

During the Australian summer of 2019–2020, Graham and Keller (2020) discovered inauthentic behaviour on the Twitter hashtag `#ArsonEmergency` in the first week of January 2020, which was first reported in the technology media (Stilgherrian 2020) and then in the mainstream media. Weber et al. (2020) performed a further study of activity on the hashtag for the first 18 days of January, observing both before and after the story became widely known. Analysis of the 27,546 tweets revealed two clearly polarised retweeting communities: one with 497 members supporting the narrative that arson was the cause of the bushfires and that eco-activism had prevented forest fuel load management (*Supporters*) and one with 593 members that countered the narrative, providing evidence that the fires were mostly started by natural or unintended causes (e.g. lightning and sparks from machinery) and the bushfires' ferocity was exacerbated by climate change (*Opposers*). The remaining 11,782 accounts in the dataset were referred to as *Unaffiliated*.



(a) DNC. (b) RNC.

Fig. 23 Co-retweeting HCCs detected during the August 2020 DNC and RNC (Threshold, $t = 0.1$, $\gamma = 10$ s). Nodes are HCC member accounts, sized by the number of tweets they contributed to the discussion, and joined by edges sized and labelled according to the number of times they retweeted the same tweet. The nodes are coloured by Louvain cluster for convenience, but any matching colours between the DNC and RNC subfigures has no meaning



(a) DNC. (b) RNC.

Fig. 24 Account/hashtag two-level networks of co-hashtag HCCs and the hashtags they used during the August 2020 DNC and RNC (FSA_V, $\theta = 0.3$, $\gamma = 10$ s). Circular nodes are HCC member accounts, coloured by HCC, and hashtags are yellow diamond nodes. The links between accounts are sized by their co-hashtag frequency (i.e. how often they used the same hashtag in the same time window). *visone*'s *stress minimisation* layout was used for both networks. Notable clusters have been highlighted with red dashed ovals and numbered, while particular hashtag clusters have been highlighted with blue diamonds

Differences between the communities' interaction and information sharing behaviour was apparent. *Supporters* interacted with other accounts more by using replies, quotes, and mentions more than *Opposers* or the *Unaffilated*, as well as more hashtags and 'external' URLs (i.e. referring to domains other than `twitter.com`), but they retweeted less. Notably, an analysis of the most shared URLs revealed that *Opposers* shared articles debunking the narrative exclusively, while *Supporters* shared a mixture of articles, mostly ones supporting or actively discussing their preferred narrative as well as some conspiratorial content. *Unaffiliated* accounts shared narrative-focused URLs initially, but in the latter phase of the collection they shared debunking articles nine times more often.

Analysis of the co-use of URLs (Threshold, $t = 0.1$, $\gamma = 10$ s) revealed further behavioural differences between the communities (Fig. 21). *Opposers* were more focused than *Supporters* in the URLs they shared, relying on a small set of debunking articles, which were ultimately also heavily shared by *Unaffiliated* accounts. *Supporters* were less tightly clustered around particular articles, and did share some debunking material as well as a variety of narrative-aligned articles.

Co-domain analysis (Threshold, $t = 0.1$, $\gamma = 10$ s) identified not just distinct URLs but distinct URL domains favoured by the different communities. Figure 22 shows two clusters of domains: the red one contains domains from a number of conservative and right wing media organisations, while the blue one contains academic and centre and left wing media organisations. Although *Supporters* mostly referred to the narrative-supporting domains while the *Opposers* mostly referred to the debunking domains, it is notable that members of both communities referred heavily the ABC and the Guardian, which both published articles

debunking the arson theory, often with reports from local fire fighting and law enforcement organisations. What is lost at this level of analysis is the way in which the articles were discussed when mentioned in tweets, including whether the tweets were agreeing or attacking the article content.

These co-analyses reveal how focused the polarised communities were in their information sharing activities, contributing to the argument that the targeted efforts of the *Opposer* community may have helped influence the broader *Unaffiliated* community into sharing debunking articles.

4.4.2 Twitter discussion groupings during the 2020 US political conventions

A second case study making use of these techniques relates to the search for social bots attempting to influence the online discussion surrounding the Democratic and Republican National Conventions in August 2020, at which the parties formally nominated their candidates for the Presidential Election, later that year. For a 96 hour period over each 4-day convention, tweets were filtered using RAPID (Lim et al. 2019), starting with `#demconvention` and `#rnc2020` as seed hashtags for the Democratic National Convention (DNC) and the Republican National Convention (RNC), respectively. For the three hours prior to the formal collection period, RAPID's topic tracking feature was enabled, adding hashtags that appeared frequently in the tweets observed, bolstering the filter terms for each convention:

- DNC: `#demconvention` , `#bidenharris`, `#bidenharris2020`, `#khive`, `#signsacrossamerica`, `#unitedforbiden`, and `#wewantjoe`;
- RNC: `#rnc2020` , `#rncconvention`, and `#nevertrump`.

Despite the disparity in hashtags, each dataset ultimately comprised approximately 1.5 million tweets by over 400 thousand unique users at each convention. Bots are often used to boost tweets, reaching other accounts that follow them, or by flooding hashtag communities or gaming trending algorithms (Woolley 2016; Hegelich and Janetzko 2016; Keller et al. 2019; Graham and Keller 2020; Graham et al. 2020). Social bots are specifically designed to mimic genuine human users, hiding the fact they are automated (Ferrara et al. 2016; Grimme et al. 2018). They do this to avoid detection, and in doing so can contribute to astroturfing campaigns, artificially boosting narratives while making them appear as simple popular grass roots movements.

By searching for *Boost*ing via co-retweet (Threshold, $t = 0.1$, $\gamma = 10$ s), several HCCs were identified in each convention (see Fig. 23). Analysis of the HCC members using Botometer (Davis et al. 2016) found the majority had CAP scores above 0.6, indicating a high probability that they made use of automation. Further analysis of the HCCs' content provided some indication of their agendas, and examination of their account age and posting rates enabled categorisation into official accounts (verified by Twitter), unofficial reposters (topic-focused aggregators), and accounts that gave the appearance of typical human users. These 'normal people', however, posted at very high average daily rates for years, often at far greater rates than previous automation detection methods have used (e.g. 50 tweets a day, Neudert 2018).

The largest HCC (the large blue HCC in Fig. 23b) consisted of a cluster of potential social bot accounts supporting an official political campaign account, `@TrumpWarRoom`, responsible for 2085 tweets during the Republican Convention. For each pair of members in each HCC, we considered the proportion of time that one account retweeted a tweet before the other, to determine if both accounts were potentially working together (in which case, they would be equally likely to retweet a tweet first), or if one was a 'cheerleader' for the other (in which case the cheered account would always retweet first, quickly followed by the other account). We found strong evidence that at least three of the accounts were cheerleaders for `@TrumpWarRoom`, retweeting the same tweet within ten seconds on 214, 229, and 89 occasions over the four day collection period. These particular accounts had daily tweeting rates of 78.7, 209.4 and 147.4 tweets per day for 0.9, 8.5 and 3.6 years, respectively. Given the age of these accounts, it is clear that they have successfully avoided Twitter's bot scanning processes for some considerable time.

We also applied co-hashtag analysis (FSA_V, $\theta = 0.3$, $\gamma = 10$ s) to the two datasets and plotted two-level networks of the resulting HCCs with the hashtags they used (Fig. 24). Regardless of the content, a number of structures are immediately apparent. These include:

- clusters that are bound by a few yellow diamond hashtag nodes (e.g. DNC clusters 5, 6 and 8) or lie between hashtags (e.g. DNC clusters 2 and 4);
- fan shapes that consist of a small number of accounts using a wide variety of hashtags (e.g. DNC clusters 1 and 7);
- island clusters that are bound by the hashtags they use but are isolated from the broader community which has ignored the hashtags they are using (e.g. DNC clusters 7 and 8).

The fact that the clusters are coloured according to their HCC in Fig. 24 highlights what FSA_V regards as distinct clusters are, in fact, bound together by the topics they are discussing (by the hashtags they are co-using). This indicates that there may be benefit in re-introducing the re-stitching step in FSA (Şen et al. 2016) that FSA_V avoids, or also

experimenting further with FSA itself. Using conductance cutting (Brandes et al. 2007) for cluster detection aligned better with the visible clusters, but these clusterings may be somewhat misleading, as it may combine polarised HCCs, as can be seen on closer inspection below.

Several co-hashtag clusters in Fig. 24a provide insight into the nature of parts of the online discussion.

- Cluster 5 is closely centred on two hashtags (`#goodyear` and `#ohio`) that relate to then US President Donald Trump's call for a boycott of Goodyear tires,[14] though it is unclear whether the surrounding accounts are for or against the boycott. Several hashtags linked on the left edge of the cluster indicate that some are against, as they refer to support for the then Democratic candidate Vice President Joe Biden.
- The fan-shaped cluster 1 at the top consists of two accounts that are attempting to disseminate their message across America, as each hashtag is a US state code (e.g. `#ga` for Georgia) or a minority group (e.g. `#latinos`). These hashtags are all apparently unique, apart from the one highlighted just below cluster 1 surrounded by a blue diamond (`#blm`) linking cluster 1 to cluster 2, and the one to the left (another state code, `#nc`, for North Carolina).
- Cluster 2 binds a number of HCCs spanning two relatively disjoint hashtags, one being `#vote` (below the cluster) and the other being the name of a musician who had recently encouraged his fans to vote.
- Cluster 3 is more diffuse than the others and appears to relate to a discussion of data science and big data in the context of the election campaign.
- Cluster 4 appears to join a number of potentially opposed HCCs, as they refer to `#trump2020landslide` and `#snowflakes` as well as `#epstein`[15] and `#trumpvirus` (a condemnation of the Trump administration's handling of the response to the COVID-19 pandemic), the final hashtag which links the cluster into the broader community.
- The island clusters 7 and 8 are focused on groups of particular politicians, which were not picked up by the broader community: Republicans who had pledged to vote for the Democratic presidential candidate and US Congress members known to campaign for social equality, respectively.

The links between the clusters are sometimes deceptive. Already, we observed that some single clusters include polarised HCCs, however it is also possible to see internally (politically) consistent clusters that are linked but also contrary in their views. DNC cluster 1 (in Fig. 24a) is linked to the left by `#nc` to another left-leaning cluster (calling for gun control), which itself is linked to the left by `#america` to another small cluster, which is clearly right-leaning (one of its hashtags is `#voteredtosaveamerica`). These visualisations may highlight how HCCs can be merged, but care must be taken when interpreting them.

Analysis of the RNC co-hashtag HCCs and their hashtags in Fig. 24b offers further examples of these observations and offers new insights. Clusters 1 and 2 are joined by the blue diamond-highlighted hashtag, `#blacklivesmatter`, but cluster 1 is a detractor group (using `#alllivesmatter`) while cluster 2 is a supporter group using several Black rights-related hashtags. Cluster 4 discusses riots following Black Lives Matter protests in Kenosha, Wisconsin, however, while the two sets of hashtags highlighted at the top of the cluster relate mostly to current events (e.g. `#kenosha` and `#covid19` on the left, and `#kenoshariots` and `#thursdaythoughts`, plus `#walkaway`, which links to a small fan, as it is a pro-Republican statement to avoid conflict), the hashtags at the bottom of the cluster are more clearly right wing or conservative in nature, referring to a relevant media organisation, `#kag2020` (Keep America Great, a pro-Trump slogan) and `#ccot` (Christian Conservative on Twitter). Whereas cluster 4 in Fig. 24a includes polarised HCCs, the placement of the hashtag nodes they are linked to offers no guidance on how they might be separated. Cluster 4 in Fig. 24b indicates that an alternative layout algorithm may aid analysis. Cluster 6 represents a concerted anti-Trump effort with many attacking hashtags, but the isolation of the HCC at the cluster's centre makes it clear that not many of the others tweeting during the RNC took its lead. Cluster 5 is an effort to draw attention to an instance of police brutality, which also did not gain traction with the broader co-hashtag community.

5 Conceptual comparison and critique

Methods to discover coordinated behaviour by inferring links between accounts based on related interactions is not unique. Cao et al. (2015) and Giglietto et al. (2019) identified groups of accounts based on the URLs they shared in common, while Lee et al. (2013), Keller et al. (2019), Dawson and Innes (2019) and Graham et al. (2020) relied

[14] The Goodyear factory in Ohio banned clothing with political messaging, including the Trump campaign's MAGA caps, during the election campaign: https://www.abc.net.au/news/2020-08-20/donald-trump-calls-for-goodyear-boycott-over-alleged-maga-ban/12577372.

[15] Jeffrey Epstein was a billionaire arrested for sex crimes before dying in custody, however he was known to Donald Trump, and therefore this hashtag's use can be seen as an attack on his political campaign: https://www.forbes.com/sites/lisettevoytko/2020/10/18/spider-book-excerpt-how-trumps-presidency-helped-expose-jeffrey-epstein/.

on the similarity of the content posted by accounts to do the same. Giglietto et al. explicitly added a temporal element by considering potential links only between accounts that share a URL within a constrained time frame. Their "rationale is that, while it may be common that several entities share the same URLs, it is unlikely, unless a consistent coordination exists, that this occurs within the time threshold and repeatedly."[16] To the knowledge of the authors, only three other proposed approaches appear to generalise the idea to allow links between accounts to be inferred based on a variety of behaviours common to the major OSNs: Pacheco et al. (2021), Graham et al. (2020), and Nizzoli et al. (2021).

Pacheco et al.'s method creates strong ties between accounts that share similar behavioural traits. Behavioural traits are extracted from social media data (e.g. hashtags or URLs) and, together with the accounts using them, a bigraph is created, similar to our account/reason networks. A weighted account network is projected from this bipartite network, linking accounts that have edges to the same trait node. The more shared traits, the heavier the edge between accounts. Finally, the account network undergoes cluster analysis specific to the nature of coordination sought. In their examples, Twitter accounts linked by using the same account handle are divided into clusters by virtue of the connected component in which they appear. A second example examining share market "pump and dump" scams links accounts based on the similarity of the text they post, using *text frequency/inverse document frequency* (TF-IDF), and then clusters are discovered by simply filtering out edges with a final weight less than 0.9. A third example connects accounts that use multiple hashtags in the same order in their tweets. The approach was employed searching for co-retweeting communities spreading propaganda attacking the Syrian White Helmet movement by linking accounts that retweeted tweets within 10 s Pacheco et al. (2020).

In contrast, Graham et al.'s "coordination network toolkit"[17] (CNT) is written in Python (as is ours), and relies on a database populated with information extracted from tweets to carry out searches for: coordinated retweeting (retweeting the same tweet); co-tweeting (tweeting identical text); co-*similarity* (tweeting similar text); co-linking (sharing the same URL); and co-replying (replying to the same tweet). The database implementation uses an inner join to improve the performance of searching for evidence of coordination between pairs of accounts (which, similar to our approach, requires pairwise comparison of all accounts in the dataset). This implementation would need to be modified to suit a streaming data source, but could theoretically be applied to data from a variety of OSNs as it employs a technique similar to our Steps 1 and 2.

The approach of Nizzoli et al. (2021) is very similar to ours, however it explicitly begins by selecting a set of users of interest, whereas we begin with a corpus of posts and our set of users is defined by those present in it. Nizzoli et al. make clear that the users may be defined by using the corpus in the same way at the outset, or may be otherwise nominated by virtue of being superproducers or superspreaders or followers of a prominent account. They also introduce a filter step before the extraction of HCCs. Pacheco et al. (2021) filter their user similarity network with an arbitrary filter, which, as pointed out by Nizzoli et al., results in a binary classification of coordinating and non-coordinating users, but importantly disregards the effect of the network structure. Instead, Nizzoli et al. rely on multiscale filtering approaches for complex networks, which retain network structures (not just individual edges) based on statistical significance. Furthermore, these can be scaled to retain more or less of the network, permitting examination of the 'degree' of coordination, not just a binary answer to whether or not it is present. They propose an iterative algorithm at this point for detecting clusters of coordinating users, which makes use of an increasingly strict definition of user similarity (i.e. coordination) and each time relies on the communities found in the previous step as the starting point, guaranteeing they are kept in some form. This makes it possible to track communities at different levels of coordination, similar to how k core decomposition provides insight into how deeply particular nodes and structures are embedded within a network. Finally, they apply a validation step, studying the resulting networks with network measures, and text analysis of the posts of the HCCs, but all as a function of the resolution at which the HCCs were detected. The FSA_V algorithm is our alternative to their filtering and cluster detection steps. The ability for Nizzoli et al. to examine different degrees of coordination is a distinguishing factor, however they also (just like Pacheco et al. 2021) must decide beforehand what similarity measure to connect users with—this is equivalent to the behaviours that underpin the coordination strategies we discussed in Sect. 2, however they make the point that the similarity measure may involve any relevant information about the user profiles, not just their behaviour within the corpus. The temporal aspect of the coordination is not discussed, presumably as it is assumed to be a component of the user similarity measure.

Giglietto et al.'s CoorNet R package does not allow specification of a time window directly, but instead uses a proportion threshold to determine what to regard as an anomalously small but active time window, and thus requires access to an entire dataset. It is designed to study Coordinated Link Sharing Behaviour (Giglietto et al. 2020a) and thus only considers URLs in posts, however, it accepts URLs from a variety of sources, including via CrowdTangle[18] and MediaCloud.[19]

[16] Quoted from the README of Giglietto et al.'s open-source code (as of 2021-01-19): https://github.com/fabiogiglietto/CooRnet.

[17] https://pypi.org/project/coordination-network-toolkit/.

[18] https://www.crowdtangle.com/.

[19] https://www.media.mit.edu/projects/media-cloud/overview/.

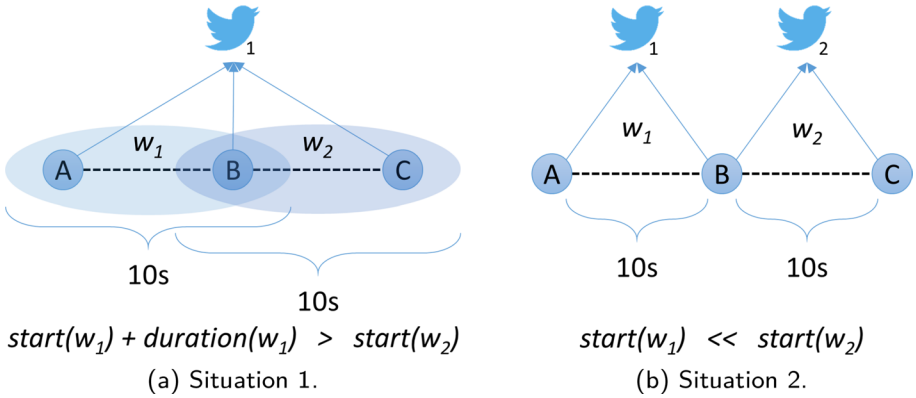


Fig. 25 The semantics of edges in LCNs requires clarification. If A is connected to B and B is connected to C, it may be due to the events in Situation 1 or Situation 2, but any inference of a relationship between A and C will be different depending on which it actually occurred. Each window, w_x, starts at timestamp start(w_x) and has duration duration(w_x). An arrow between, e.g. A and the Twitter bird 1 implies A retweeted tweet 1 within a specific window

Our method is similar to all of these but is described in greater detail, relies upon a discrete window-based approach to apply temporal constraints, and we provide and evaluate a novel cluster extraction algorithm, and an open-source implementation is available. By applying time constraints in discrete windows, connections may be missed across windows, but this makes it easier to apply in near real-time streaming settings. If one were to infer connections between accounts as each new tweet is posted, it could create a potentially significant, ongoing processing cost depending on the number of unique accounts observed in the current time window. As new posts arrive, new nodes may need to be added to the account network, while others may need to be removed, along with their adjacent edges (which, it is important to recall, represent indirect evidence of coordination, not the individual timestamped interactions as one might find in a social network based on direct retweets, mentions or replies). Furthermore, this constantly updated account network must be complete, i.e. edges should always be added in case the evidence they represent may be consolidated by future posts.

If the choice of time window is very short (e.g. 10 s, as per Pacheco et al. 2020), and LCNs from adjacent windows are aggregated (as per our method), the absence of a truly sliding window like Graham et al.'s may not significantly affect results, as ongoing high levels of coordination will appear over multiple windows. In contrast, if the time window is longer (e.g. five or more minutes), then the hard boundary between windows may cause coordinated activities to be missed. The question is, then, what kind of coordination is being sought. Teams of bots tweeting or retweeting the same tweet within small time frames will be vulnerable to detection, however a deliberate covert human team with sockpuppet accounts may escape detection (at least initially) by varying the time frame over which retweets are posted (e.g. spread them unevenly over an hour or more), but if the same accounts cooperate for extended periods, our method will find them once their activities are aggregated. One type of coordination that is very difficult to detect is single event boosts of a post: e.g. when, say, 1000 paid accounts retweet or reply to a single tweet or comment on an online review. In a large discussion, 1000 tweets will not stand out, but, depending on how connected the paid accounts are to the broader discussion, they may spread the content a considerable distance through the network. Furthermore, gaming OSN trending algorithms may not be difficult,[20] and even a thousand retweets may result in a valuable degree of influence in comparatively smaller communities (e.g. Australia).

As a final comment, all methods discussed in this section are suited to post-collection analysis. Graham et al.'s relies on the power of database systems to build the LCN but avoids exploring clustering analysis for HCCs. Giglietto et al.'s relies on R's expressivity and filtering based on anomaly detection, while our implementation uses Python and batch mode processing to enable flexibility in the choice of cluster analysis technique. Pacheco et al.'s implementation is in Python, but has been applied to very large datasets, and so may also rely on a high performance language (e.g. Java) or distributed processing platform, such as Hadoop. Nizzoli et al. do not mention the availability of their implementation, only that their test dataset will be forthcoming.

Our paper is the only one of these to address the concept of searching for multiple coordination criteria, and how to treat the combination of their evidence, and the attendant complications explored in Step 4 of Sect. 3.1. Magelinski et al. (2021) have proposed second-order interactions to

[20] OSN gaming efforts of the form "Let's get X trending" are quite common in Australia, e.g. https://twitter.com/Timothyjgraham/status/1351742513044807680.

address this, but only in combination (e.g. connect accounts using the same hashtag+URL in tweets). In fact, the other papers primarily treat the coordination criteria (i.e. user similarity measure) as entirely dependent on the current investigation and no generalisation of the concepts is discussed.

6 Future work

The most important directions to take this work relate to the following aspects:

1. ***Temporal analysis.*** Temporal analysis of the evolution of HCCs and their influence on the broader discussion over time will provide insight into how the HCCs operate and achieve their goals, and perhaps will reveal distinct classes of HCCs and the strategies they employ.
2. ***The semantics of LCN edges.*** Further theoretical research relying on the temporal aspect is also required to determine the real meaning of edges in the LCN. Figure 25 provides an illustration of where there is ambiguity. If accounts A, B, and C all retweet tweet 1 within a single time window, or at least overlapping time windows, then we join A to B and B to C in the LCN, and there is a reasonable assumption that A and C may be related somehow. This is less clear when A and B retweet tweet 1 and B and C retweet tweet 2, both in different time windows; it is much less reasonable to assume a relationship between A and C in that case (especially if the time windows are far apart), but both situations result in the same LCN structure: A connects to B and B connects to C. This is now under investigation (Weber and Falzon 2021).
3. ***Coordinated amplification strategies.*** Not all coordination occurs in short periods of time nor in adjacent time windows. For example, an unscrupulous political campaign may purchase the services of 1000 bots—behaviour identified by Dawson and Innes (2019)—which are then told to retweet a single campaign post once or to reply to an opponent's post with attacks at a particular time to maximise the political effect. Alternatively, a small group of trolls may post the same tweets harassing a public figure each day at 5pm on weekends but not weekdays. Considering how to address broader definitions of coordination will be an ongoing challenge as OSNs change their features and people find new ways to use and abuse them. Additionally, higher abstractions of coordination, as observed in forensic studies of online influence campaigns (Benkler et al. 2018; Jamieson 2020; Singer and Brooking 2019; Nimmo et al. 2020) present further challenges for automated detection systems.
4. ***Distinguishing authentic and inauthentic behaviour.*** The issue of astroturfing (e.g. Metaxas and Mustafaraj 2012) brings this into sharp relief: some campaigns are genuine grass roots movements driven by a broad desire to see policies change (e.g. campaigning to address climate change), however some are artificially organised, aimed at gaming OSN trending algorithms to spread their narrative further and to give the appearance of genuine wide public support (e.g. efforts to convince US Congress to release a politically controversial FBI memo, McKew 2018). As covert campaigns become more sophisticated and numerous, it will become more important for OSNs, law enforcement and relevant agencies to focus their efforts on the relevant malicious activities and to be able to discriminate harmless fan campaigns from harmful disinformation campaigns. Others have noted that this problem remains unresolved at this time (Vargas et al. 2020).
5. ***Process improvement.*** Improving the implementation of the process, and how HCC extraction is performed, and how validation is conducted will be an ongoing activity. As demonstrated by Pacheco et al. (2020, 2021) and Graham et al. (2020), different types of community extraction will suit different types of coordination strategy, just as will the choice of strategy to search for (e.g. pollution or boost). Introducing a genuine sliding window (cf. our distinct, adjacent windows) will prevent missing further instances of coordination, but modification of the approach will be required to apply it in a near real-time setting (cf., Carnein et al. 2017; Assenmacher et al. 2020). Finally, to bring some statistical robustness to the question of validity, there are measures that can be used to determine if the accounts in HCCs engage in statistically significant greater or lesser levels of, say, retweeting than the general population. Broniatowski (2021) has very recently offered a contribution to this challenge. These measures will help determine in what ways HCC behaviours differ, but will leave unresolved the question of intent and the authenticity of that behaviour.

7 Conclusion

As coordinated online influence activities grow in sophistication, so must our automation and campaign detection methods also improve in order to expose the accounts covertly engaging in "orchestrated activities" (Grimme et al. 2018). We have described several strategies for coordinated amplification, their purpose and execution methods, and demonstrated a novel pipeline-based approach to finding sets of accounts engaging in such behaviours in two politically relevant Twitter datasets. We have also explained

and provided examples of how our method is conceptually applicable to a range of OSNs based on common features and functionality. Using discrete time windows, we temporally constrain potentially coordinated activities, successfully identifying groups operating over various time frames. Guided by research questions posed in Sect. 1, our results were validated by using a variety of techniques, including developing three one-class classifiers to compare the HCCs found in two relevant datasets, plus a randomised one, with HCCs from a ground truth subset. Two case studies of contentious online discussion were also presented, in which our technique was applied to reveal insights into the activity of polarised groups in one and the activity of social bots and bot-like accounts in the other. The algorithmic complexity of our approach was discussed, as well as comparison with several similar contemporary approaches.

This technique provides a valuable addition to the suite of analytical tools used in deep forensic investigations of SIOs, such as Benkler et al. (2018), Jamieson (2020) and Nimmo et al. (2020), as well as law enforcement and open-source investigation groups—in particular, this technique can help reveal entities that deliberately avoid direct connections to hide their cooperation.

The temporal analysis of HCC evolution and their impact on the broader discussion, theoretical questions of the semantics of edges in LCNs, the ability to distinguish between authentic and inauthentic coordinated behaviour, improvement of HCC extraction and validation techniques and application in near real-time processing environments all provide opportunities for future research in this increasingly important field.

Author contributions Both authors contributed to the conception of the study, and Derek Weber performed the data collection, analysis and writing of the first draft. Both authors read and approved the final manuscript.

Availability of data and material The data (the identifiers of tweets only, as per Twitter's terms and conditions) used in this study are available at https://github.com/weberdc/find_hccs. The collection filter terms and the handles of political accounts are available on request.

Code availability The data manipulation and analysis software written for this study is available at https://github.com/weberdc/find_hccs.

Declarations

Conflict of interest All authors declare that they have no conflict of interest.

Consent for publication All authors consent to this work being published.

Ethics approval All data were collected, stored, processed and analysed according to the ethics protocol H-2018-045, approved by the University of Adelaide's human research and ethics committee.

References

Adjali O, Besançon R, Ferret O, Borgne HL, Grau B (2020) Multimodal entity linking for tweets. In: Lecture notes in computer science. Springer, pp 463–478. https://doi.org/10.1007/978-3-030-45439-5_31

Alizadeh M, Shapiro JN, Buntain C, Tucker JA (2020) Content-based features predict social media influence operations. Sci Adv 6(30):eabb5824. https://doi.org/10.1126/sciadv.abb5824

Assenmacher D, Adam L, Trautmann H, Grimme C (2020) Towards real-time and unsupervised campaign detection in social media. In: FLAIRS Conference. AAAI Press

Bacco CD, Power EA, Larremore DB, Moore C (2017) Community detection, link prediction, and layer interdependence in multilayer networks. Phys Rev E 95(4):042317. https://doi.org/10.1103/physreve.95.042317

Badawy A, Ferrara E (2018) The rise of Jihadist propaganda on social networks. J Comput Soc Sci 1(2):453–470. https://doi.org/10.1007/s42001-018-0015-z

Bedru HD, Yu S, Xiao X, Zhang D, Wan L, Guo H, Xia F (2020) Big networks: a survey. Comput Sci Rev. https://doi.org/10.1016/j.cosrev.2020.100247

Benkler Y, Farris R, Roberts H (2018) Network Propaganda. Oxford University Press, Oxford. https://doi.org/10.1093/oso/9780190923624.001.0001

Bessi A, Ferrara E (2016) Social bots distort the 2016 U.S. Presidential election online discussion. First Monday. https://doi.org/10.5210/fm.v21i11.7090

Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 10:P10008. https://doi.org/10.1088/1742-5468/2008/10/p10008

Borgatti SP, Mehra A, Brass DJ, Labianca G (2009) Network analysis in the social sciences. Science 323(5916):892–895. https://doi.org/10.1126/science.1165821

Brandes U, Gaertler M, Wagner D (2007) Engineering graph clustering: models and experimental evaluation. ACM J Exp Algorithm 12:1–26. https://doi.org/10.1145/1227161.1227162

Broniatowski DA (2021) Towards statistical foundations for detecting Coordinated Inauthentic Behavior on Facebook. Techreport Preprint, Institute for Data, Democracy and Politics—The George Washington University. Last accessed on 2021-05-14 at https://iddp.gwu.edu/towards-statistical-foundations-detecting-coordinated-inauthentic-behavior-facebook

Brooking ET, Singer PW (2016) War goes viral: How social media is being weaponized across the world. The Atlantic https://www.theatlantic.com/magazine/archive/2016/11/war-goes-viral/501125/

Burgess J, Matamoros-Fernández A (2016) Mapping sociocultural controversies across digital media platforms: one week of #gamergate on Twitter, YouTube, and Tumblr. Commun Res Pract 2(1):79–96. https://doi.org/10.1080/22041451.2016.1155338

Cao C, Caverlee J, Lee K, Ge H, Chung J (2015) Organic or organized?: Exploring URL sharing behavior. In: CIKM, ACM, pp 513–522. https://doi.org/10.1145/2806416.2806572

Carnein M, Assenmacher D, Trautmann H (2017) Stream clustering of chat messages with applications to Twitch streams. ER Workshops, Springer, LNCS, vol 10651, pp 79–88. https://doi.org/10.1007/978-3-319-70625-2_8

Carvin A (2012) Distant witness: social media, the Arab spring and a journalism revolution. CUNY Journalism Press, New York, NY
Chavoshi N, Hamooni H, Mueen A (2017) Temporal patterns in bot activities. In: WWW (Companion Volume), ACM, pp 1601–1606. https://doi.org/10.1145/3041021.3051114
Chen A (2015) The Agency. The New York Times Magazine https://www.nytimes.com/2015/06/07/magazine/the-agency.html
Chu Z, Widjaja I, Wang H (2012) Detecting social spam campaigns on Twitter. In: ACNS, LNCS, vol 7341. Springer, pp 455–472. https://doi.org/10.1007/978-3-642-31284-7_27
Cresci S (2020) A decade of social bot detection. Commun ACM 63(10):72–83. https://doi.org/10.1145/3409116
Cresci S, Pietro RD, Petrocchi M, Spognardi A, Tesconi M (2017) The paradigm-shift of social spambots. In: WWW (Companion Volume), ACM, pp 963–972. https://doi.org/10.1145/3041021.3055135
Damashek M (1995) Gauging similarity with n-grams: language-independent categorization of text. Science 267(5199):843–848. https://doi.org/10.1126/science.267.5199.843
Datta S, Adar E (2019) Extracting inter-community conflicts in Reddit. In: ICWSM, AAAI Press, pp 146–157. https://aaai.org/ojs/index.php/ICWSM/article/view/3217
Davis CA, Varol O, Ferrara E, Flammini A, Menczer F (2016) BotOrNot: a system to evaluate social bots. In: WWW (Companion Volume). ACM, pp 273–274. https://doi.org/10.1145/2872518.2889302
Dawson A, Innes M (2019) How Russia's Internet Research Agency built its disinformation campaign. Polit Q 90(2):245–256. https://doi.org/10.1111/1467-923x.12690
Fang Y, Huang X, Qin L, Zhang Y, Zhang W, Cheng R, Lin X (2019) A survey of community search over big graphs. VLDB J 29(1):353–392. https://doi.org/10.1007/s00778-019-00556-x
Ferrara E (2017) Disinformation and social bot operations in the run up to the 2017 French presidential election. First Monday. https://doi.org/10.5210/fm.v22i8.8005
Ferrara E, Varol O, Davis C, Menczer F, Flammini A (2016) The rise of social bots. Commun ACM 59(7):96–104. https://doi.org/10.1145/2818717
Fisher A (2018) Netwar in Cyberia: Decoding the media Mujahadin. CPD Perspectives Paper 5, USC Center on Public Diplomacy. https://www.uscpublicdiplomacy.org/sites/uscpublicdiplomacy.org/files/Netwar%20in%20Cyberia%20Web%20Ready_with%20disclosure%20page%2011.08.18.pdf
Giglietto F, Righetti N, Marino G (2019) Understanding coordinated and inauthentic link sharing behavior on Facebook in the run-up to 2018 general election and 2019 European election in Italy. SocArvix https://doi.org/10.31235/osf.io/3jteh
Giglietto F, Righetti N, Rossi L, Marino G (2020a) Coordinated link sharing behavior as a signal to surface sources of problematic information on Facebook. In: SMSociety, ACM. https://doi.org/10.1145/3400806.3400817
Giglietto F, Righetti N, Rossi L, Marino G (2020b) It takes a village to manipulate the media: coordinated link sharing behavior during 2018 and 2019 Italian elections. Inform Commun Soc. https://doi.org/10.1080/1369118x.2020.1739732
Graham T, Ackland R (2017) Do socialbots dream of popping the filter bubble? The role of socialbots in promoting participatory democracy in social media. In: Gehl RW, Bakardjieva M (eds) Socialbots and their friends: digital media and the automation of sociality. Routledge, London, chap 10, pp 187–206
Graham T, Keller TR (2020) Bushfires, bots and arson claims: Australia flung in the global disinformation spotlight. The Conversation. https://theconversation.com/bushfires-bots-and-arson-claims-australia-flung-in-the-global-disinformation-spotlight-129556. Accessed 07 Feb 2020
Graham T, Bruns A, Zhu G, Campbell R (2020) Like a virus: the coordinated spread of coronavirus disinformation. Tech. rep., Centre for Responsible Technology, The Australia Institute. https://apo.org.au/node/305864
Grimme C, Preuss M, Adam L, Trautmann H (2017) Social bots: human-like by means of human control? Big Data 5(4):279–293. https://doi.org/10.1089/big.2017.0044
Grimme C, Assenmacher D, Adam L (2018) Changing perspectives: is it sufficient to detect social bots? In: HCI (13). Springer, LNCS, vol 10913, pp 445–461. https://doi.org/10.1007/978-3-319-91521-0_32
Gupta S, Kumaraguru P, Chakraborty T (2019) MalReG: Detecting and analyzing malicious retweeter groups. In: COMAD/CODS. ACM, pp 61–69. https://doi.org/10.1145/3297001.3297009
Hegelich S, Janetzko D (2016) Are social bots on Twitter political actors? Empirical evidence from a Ukrainian social botnet. In: ICWSM. AAAI Press, pp 579–582. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13015
Hine GE, Onaolapo J, Cristofaro ED, Kourtellis N, Leontiadis I, Samaras R, Stringhini G, Blackburn J (2017) Kek, cucks, and God Emperor Trump: a measurement study of 4chan's politically incorrect forum and its effects on the Web. In: ICWSM. AAAI Press, pp 92–101. https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15670
Holme P, Saramäki J (2012) Temporal networks. Phys Rep 519(3):97–125. https://doi.org/10.1016/j.physrep.2012.03.001
Howard PN, Kollanyi B (2016) Bots, #StrongerIn, and #Brexit: Computational propaganda during the UK-EU referendum. Research Note 2016.1, Oxford, UK: The Computational Propaganda Research Project. http://comprop.oii.ox.ac.uk/wp-content/uploads/sites/89/2016/06/COMPROP-2016-1.pdf
Jamieson KH (2020) Cyberwar: how Russian hackers and trolls helped elect a president: what we don't, can't and do know. Oxford University Press. https://doi.org/10.1093/oso/9780190058838.001.0001
Karell D, andrew linke, Holland EC, (2021) Right-wing social media and unrest correspond across the United States. SocArXiv https://doi.org/10.31235/osf.io/pna5u
Keller FB, Schoch D, Stier S, Yang J (2017) How to manipulate social media: Analyzing political astroturfing using ground truth data from South Korea. In: ICWSM. AAAI Press, pp 564–567. https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15638
Keller FB, Schoch D, Stier S, Yang J (2019) Political astroturfing on Twitter: how to coordinate a disinformation campaign. Polit Commun 37(2):256–280. https://doi.org/10.1080/10584609.2019.1661888
King G, Pan J, Roberts ME (2017) How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. Am Polit Sci Rev 111(3):484–501. https://doi.org/10.1017/S0003055417000144
Kosinski M, Stillwell D, Graepel T (2013) Private traits and attributes are predictable from digital records of human behavior. PNAS 110(15):5802–5805. https://doi.org/10.1073/pnas.1218772110
Kumar S, Hamilton WL, Leskovec J, Jurafsky D (2018) Community interaction and conflict on the web. In: WWW, ACM, pp 933–943. https://doi.org/10.1145/3178876.3186141
Kuran T, Sunstein CR (1999) Availability cascades and risk regulation. Stanford Law Rev 51(4):683. https://doi.org/10.2307/1229439
Lee K, Caverlee J, Cheng Z, Sui DZ (2013) Campaign extraction from social media. ACM Trans Intell Syst Technol 5(1):9:1-9:28. https://doi.org/10.1145/2542182.2542191
Lim KH, Jayasekara S, Karunasekera S, Harwood A, Falzon L, Dunn J, Burgess G (2019) RAPID: real-time analytics platform for interactive data mining. In: ECML PKDD 2018. Springer, LNCS, vol 11053, pp 649–653. https://doi.org/10.1007/978-3-030-10997-4_44

Magelinski T, Ng LHX, Carley KM (2021) A synchronized action framework for responsible detection of coordination on social media. CoRR abs/2105.07454

Malone TW, Crowston K (1994) The interdisciplinary study of coordination. ACM Comput Surv 26(1):87–119. https://doi.org/10.1145/174666.174668

Mariconti E, Onaolapo J, Ahmad SS, Nikiforou N, Egele M, Nikiforakis N, Stringhini G (2017) What's in a name?: Understanding profile name reuse on Twitter. In: WWW, ACM, pp 1161–1170. https://doi.org/10.1145/3038912.3052589

Mariconti E, Suarez-Tangil G, Blackburn J, Cristofaro ED, Kourtellis N, Leontiadis I, Serrano JL, Stringhini G (2019) "You know what to do": Proactive detection of YouTube videos targeted by coordinated hate attacks. PACMHCI 3(CSCW), pp 1–21. https://doi.org/10.1145/3359309

Mazza M, Cresci S, Avvenuti M, Quattrociocchi W, Tesconi M (2019) RTbust: Exploiting temporal patterns for botnet detection on Twitter. In: WebSci. ACM, pp 183–192. https://doi.org/10.1145/3292522.3326015

McGregor A (2014) Graph stream algorithms: a survey. ACM SIGMOD Rec 43(1):9–20. https://doi.org/10.1145/2627692.2627694

McKew MK (2018) How Twitter bots and Trump fans made #ReleaseTheMemo go viral. Politico https://www.politico.eu/article/how-twitter-bots-and-trump-fans-made-releasethememo-go-viral/amp/, 4th February 2018

Metaxas PT, Mustafaraj E (2012) Social media and the elections. Science 338(6106):472–473. https://doi.org/10.1126/science.1230456

Miller G (2018) The apprentice: Trump, Russia, and the Subversion of American Democracy. William Collins, London

Mordelet F, Vert JP (2014) A bagging SVM to learn from positive and unlabeled examples. Pattern Recogn Lett 37:201–209. https://doi.org/10.1016/j.patrec.2013.06.010

Morstatter F, Shao Y, Galstyan A, Karunasekera S (2018) From *Alt-Right* to *Alt-Rechts*: Twitter analysis of the 2017 German Federal Election. In: WWW (Companion Volume). ACM, pp 621–628. https://doi.org/10.1145/3184558.3188733

Mueller R (2018) Indictment, United States v. Internet Research Agency LLC et al. US District Court for the District of Columbia case no. 18-cr-00032-DLF(docket entry 1), docket entry 1, Feb. 16, 2018, case no. 18-cr-00032-DLF, U.S. District Court for the District of Columbia

Nasim M, Nguyen A, Lothian N, Cope R, Mitchell L (2018) Real-time detection of content polluters in partially observable Twitter networks. In: WWW (Companion Volume). ACM, pp 1331–1339. https://doi.org/10.1145/3184558.3191574

Neudert LMN (2018) Germany: A cautionary tale. In: Woolley SC, Howard PN (eds) Computational propaganda: political parties, politicians, and political manipulation on social media. Oxford University Press, chap 7, pp 153–184. https://doi.org/10.1093/oso/9780190931407.003.0008

Nimmo B, François C, Eib CS, Ronzaud L, Ferreira R, Hernon C, Kostelancik T (2020) Exposing secondary infektion. Report, Graphika. https://secondaryinfektion.org/

Nizzoli L, Tardelli S, Avvenuti M, Cresci S, Tesconi M (2021) Coordinated behavior on social media in 2019 UK general election. In: ICWSM. AAAI Press, pp 443–454. https://ojs.aaai.org/index.php/ICWSM/article/view/18074

Pacheco D, Flammini A, Menczer F (2020) Unveiling coordinated groups behind White Helmets disinformation. In: WWW (Companion Volume). ACM, pp 611–616. https://doi.org/10.1145/3366424.3385775

Pacheco D, Hui P, Torres-Lugo C, Truong BT, Flammini A, Menczer F (2021) Uncovering coordinated networks on social media: methods and case studies. In: ICWSM. AAAI Press, pp 455–466. https://ojs.aaai.org/index.php/ICWSM/article/view/18075

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, VanderPlas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12(85):2825–2830

Radicioni T, Pavan E, Squartini T, Saracco F (2020) Analysing Twitter semantic networks: the case of 2018 Italian elections. CoRR arXiv:abs/2009.02960

Ratkiewicz J, Conover MD, Meiss MR, Gonçalves B, Flammini A, Menczer F (2011) Detecting and tracking political abuse in social media. In: ICWSM, AAAI Press. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2850

Rizoiu MA, 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. In: ICWSM. AAAI Press, pp 300–309. https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17886

Şen F, Wigand R, Agarwal N, Tokdemir S, Kasprzyk R (2016) Focal structures analysis: Identifying influential sets of individuals in a social network. Soc Netw Anal Min 6(1):17:1-17:22. https://doi.org/10.1007/s13278-016-0319-z

Shearer E, Grieco E (2019) Americans are wary of the role social media sites play in delivering the news. Report, Pew Research Center. https://www.journalism.org/2019/10/02/americans-are-wary-of-the-role-social-media-sites-play-in-delivering-the-news/

Singer PW, Brooking ET (2019) Likewar: The Weaponization of Social Media. Mariner Books

Starbird K, Wilson T (2020) Cross-platform disinformation campaigns: lessons learned and next steps. Harvard Kennedy School Misinformation Review. https://doi.org/10.37016/mr-2020-002

Starbird K, Arif A, Wilson T (2019) Disinformation as collaborative work: surfacing the participatory nature of strategic information operations. PACMHCI 3(CSCW), pp 1–26. https://doi.org/10.1145/3359229

Stilgherrian (2020) Twitter bots and trolls promote conspiracy theories about Australian bushfires. ZDNet. https://www.zdnet.com/article/twitter-bots-and-trolls-promote-conspiracy-theories-about-australian-bushfires/

The Soufan Center (2021) Quantifying the Q conspiracy: a data-driven approach to understanding the threat posed by QAnon. Special report, The Soufan Center. https://thesoufancenter.org/research/quantifying-the-q-conspiracy-a-data-driven-approach-to-understanding-the-threat-posed-by-qanon/

Tversky A, Kahneman D (1973) Availability: a heuristic for judging frequency and probability. Cogn Psychol 5(2):207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Vargas L, Emami P, Traynor P (2020) On the detection of disinformation campaign activity with network analysis. In: CCSW@CCS. ACM. https://doi.org/10.1145/3411495.3421363

Varol O, Ferrara E, Menczer F, Flammini A (2017) Early detection of promoted campaigns on social media. EPJ Data Sci 6(1):13. https://doi.org/10.1140/epjds/s13688-017-0111-y

Verma V, Aggarwal RK (2020) A comparative analysis of similarity measures akin to the Jaccard index in collaborative recommendations: empirical and theoretical perspective. Soc Netw Anal Min https://doi.org/10.1007/s13278-020-00660-9

Vo N, Lee K, Cao C, Tran T, Choi H (2017) Revealing and detecting malicious retweeter groups. In: ASONAM. ACM, pp 363–368. https://doi.org/10.1145/3110025.3110068

Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Science 359(6380):1146–1151. https://doi.org/10.1126/science.aap9559

Wasserman S, Faust K (1994) Social network analysis: methods and applications, vol 8. Cambridge University Press, Cambridge. https://doi.org/10.1017/cbo9780511815478

Weber D (2019) On coordinated online behaviour. Poster presented at the Fourth Australian Social Network Analysis Conference, ASNAC'19, 27–29 November, Adelaide, Australia. https://www.slideshare.net/derekweber/on-coordinated-online-behaviour

Weber D, Falzon L (2021) Temporal nuances of coordination networks. CoRR arXiv:abs/2107.02588

Weber D, Neumann F (2020) Who's in the gang? Revealing coordinating communities in social media. In: ASONAM. IEEE, pp 89–93. https://doi.org/10.1109/asonam49781.2020.9381418

Weber D, Nasim M, Falzon L, Mitchell L (2020) #ArsonEmergency and Australia's "Black Summer": Polarisation and misinformation on social media. MISDOOM. Springer, LNCS, vol 12259, pp 159–173. https://doi.org/10.1007/978-3-030-61841-4_11

Woolley SC (2016) Automating power: social bot interference in global politics. First Monday. https://doi.org/10.5210/fm.v21i4.6161

Woolley SC, Guilbeault DR (2018) United States: manufacturing consensus online. Oxford University Press, Oxford, vol 8, pp 185–211. https://doi.org/10.1093/oso/9780190931407.001.0001

Woolley SC, Howard PN (2018) Computational propaganda: Political parties, politicians, and political manipulation on social media. Oxford University Press. https://doi.org/10.1093/oso/9780190931407.001.0001

Wu T, Wen S, Xiang Y, Zhou W (2018) Twitter spam detection: survey of new approaches and comparative study. Comput Secur 76:265–284. https://doi.org/10.1016/j.cose.2017.11.013

Yu W (2021) A framework for studying coordinated behaviour applied to the 2019 Philippine midterm elections. In: ICICT. https://archium.ateneo.edu/discs-faculty-pubs/207/

Zhao Q, Erdogdu MA, He HY, Rajaraman A, Leskovec J (2015) SEISMIC: A self-exciting point process model for predicting tweet popularity. In: KDD, ACM, pp 1513–1522. https://doi.org/10.1145/2783258.2783401

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations