

# The Spread of Propaganda by Coordinated Communities on Social Media

Kristina Hristakieva

christina.hristakieva@gmail.com

Sofia University

Bulgaria

Giovanni Da San Martino, Mauro Conti

{dasan,conti}@math.unipd.it

University of Padova

Italy

Stefano Cresci\*

s.cresci@iit.cnr.it

IIT-CNR

Italy

Preslav Nakov

pnakov@hbku.edu.qa

Qatar Computing Research Institute, HBKU

Qatar

## ABSTRACT

Large-scale manipulations on social media have two important characteristics: (i) they use *propaganda* to influence others, and (ii) they adopt *coordinated behavior* to spread propaganda and to amplify its impact. Despite the connection between them, these two characteristics have so far been considered in isolation. Here we aim to bridge this gap. In particular, we analyze the spread of propaganda and its interplay with coordinated behavior on a large Twitter dataset about the 2019 UK general election. We first propose and evaluate several measures for quantifying the use of propaganda on Twitter. Then, we investigate the use of propaganda by different coordinated communities that participated in the online debate. The combined analysis of propaganda and of coordination provides evidence about the harmfulness of coordinated communities that would not be available otherwise. For instance, it allows us to identify a harmful politically-oriented community as well as a harmless community of grassroots activists. Finally, we compare our measures of propaganda and of coordination to automation scores (i.e., the use of bots) and Twitter suspensions, revealing interesting trends. From a theoretical viewpoint, we introduce a methodology for analyzing several important dimensions of online behavior that are seldom conjointly considered. From a practical viewpoint, we provide new and nuanced insights into inauthentic and harmful online activities in the run up to the 2019 UK general election.

## ACM Reference Format:

Kristina Hristakieva, Stefano Cresci, Giovanni Da San Martino, Mauro Conti, and Preslav Nakov. 2022. The Spread of Propaganda by Coordinated Communities on Social Media. In *14th ACM Web Science Conference 2022 (WebSci '22), June 26–29, 2022, Barcelona, Spain*. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3501247.3531543>

\*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WebSci '22, June 26–29, 2022, Barcelona, Spain

© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9191-7/22/06...\$15.00

<https://doi.org/10.1145/3501247.3531543>

## 1 INTRODUCTION

Social media currently represent one of the main channels for information spread and consumption. They are increasingly used by a constantly growing part of the population to maintain active social relationships, to stay informed about socially relevant issues, and to produce content, thus giving voice to the crowds. At the same time, a large portion of online information is biased, misleading, or outright fake [15, 45]. Moreover, harmful content can be purposefully shared by malicious actors, and even by unaware users, with the aim to manipulate online audiences, to sow doubt and discord, and to increase polarization [9, 41]. The unprecedented importance of social media for information diffusion, combined with their vulnerability to organized misbehavior, sets the stage for online manipulations that can cause tremendous societal repercussions, as witnessed during the US Capitol Hill assault in January 2021 [6, 30], and with the rampaging COVID-19 vaccine misinformation [1, 19].

Despite the differences between the broad array of tactics used to carry out online manipulation, many social media campaigns share two fundamental characteristics: (i) they use *propaganda* to influence those targeted by the manipulation [4], and (ii) they adopt *coordinated actions* to amplify the spread and the outreach of the manipulation and to increase its impact [31, 42]. Given their importance for online manipulation, each of these characteristics has received scholarly attention. For example, computational linguists developed AI solutions to automatically detect the use of propaganda techniques [11, 13, 16]. Similarly, network science frameworks were proposed for detecting coordinated groups of users [33] and for measuring the extent of coordination [31]. Despite recent progress, the study of computational propaganda and coordinated behavior is still in its early stages.<sup>1</sup> As such, and in spite of the interrelationship between propaganda and coordinated behavior, so far, these two aspects have been investigated in isolation. Nonetheless, their combined analysis is promising under multiple viewpoints:

- From the propaganda viewpoint, there already exist methods for detecting the use of rhetorical techniques to influence others [13]. However, there have been no studies to detect the *intent to harm* behind propaganda campaigns [11]. Notably, coordination between users implies a *shared intent*. Thus, adding coordination information to the analysis of propaganda can contribute to bridging this gap.

<sup>1</sup><https://medium.com/1st-draft/how-to-improve-our-analysis-of-coordinated-inauthentic-behavior-a4ec62ce9bff>

- From the coordination viewpoint, there already exist methods for detecting coordinated users in social media [31, 33, 47]. However, distinguishing between *harmless* (e.g., activists, fandoms) and *harmful* (e.g., botnets, trolls) coordination is still an open challenge [44]. Propaganda implies the *aim to mislead and to manipulate*. Thus, adding information about propaganda to the analysis of coordination can help detect harmful behavior.

Our aim is to combine techniques for the analysis of propaganda and online coordination to draw nuanced insights into (i) the spread of propaganda online, (ii) the behavior of coordinated communities, and (iii) the interplay between propaganda and coordination.

**Contributions.** We analyze the – so far unexplored – interplay between propaganda and coordination in the context of societal online debates. Towards this goal, we adopt a methodological approach grounded on state-of-the-art techniques for detecting propaganda in texts and for measuring coordinated behavior on social media, and we apply it to study a recent and relevant online debate on Twitter, about the 2019 UK general election. In particular, we propose and experiment with several measures for quantifying the spread of propaganda by social media users and communities. We further carry out network analysis of coordinated online communities that participated in the electoral debate. Next, we combine our results on propaganda and coordination by comparing the spread of propaganda to the activity of coordinated communities. We also compare our results to clear signs of inauthenticity and harmfulness – namely, bot scores provided by Botometer [37] and Twitter suspensions. Our analysis provides more nuanced results compared to existing work, and it surfaces interesting patterns in the behavior of online communities that would not be visible otherwise. For instance, it allows to clearly identify and to differentiate communities that exhibit opposite behaviors, such as (i) a malicious politically-oriented community, characterized by strongly coordinated users that are involved in spreading propaganda, and (ii) a grassroots community of activists protesting for women's rights. Our main contributions can be summarized as follows:

- We explore the interplay between propaganda and coordination in online debates, which so far has received little attention.
- By cross-checking propaganda and coordination, we get better insights into malicious behavior, thus moving in the direction of identifying and studying coordinated inauthentic behaviors (CIB) as well as propaganda campaigns.
- Regarding malicious behavior, we draw insights into the interplay between propaganda/coordination and automation/suspensions.
- From a practical standpoint, our analysis reveals interesting and nuanced characteristics of several online communities, which were previously unknown.

**Significance.** Our proposed methodology and results contribute to improving our understanding of coordinated harmful behavior. Furthermore, insights such as those obtained thanks to our analysis can support platform administrators at enforcing targeted moderation interventions for curbing online harms [26, 43].

## 2 RELATED WORK

Propaganda and coordinated behavior pose peculiar challenges that mandate different methods, such as natural language understanding for the former, and network science for the latter. Thus, they have been the focus of largely disjoint efforts by different communities.

### 2.1 Coordinated Behavior

Coordinated behavior, be it authentic or not, was introduced as a concept by Facebook in 2018<sup>2</sup> and later widely adopted in studies on online manipulation. Given its recency, the computational analysis of coordinated behavior poses several challenges. Some are conceptual: What exactly is coordinated behavior? How many accounts, or how much coordination, is needed for meaningful coordinated behavior to surface? Currently, there are no agreed-upon answers,<sup>3</sup> which makes computational analysis problematic. In fact, many solutions still require a great deal of manual intervention [29, 41].

In the few recent computational frameworks, coordination was defined as an exceptional similarity between a number of users. Nizzoli et al. [31] proposed a state-of-the-art pipeline organized in six analytical steps, starting with (i) selection of a set of users to investigate, (ii) selection of a measure for the similarity between users, (iii) construction of a weighted user-similarity network, (iv) network filtering, (v) coordination-aware community detection, and finally, (vi) analysis of the discovered coordinated communities. This approach is the only one so far that has proven capable of producing fine-grained estimates of the extent of coordination in the continuous [0, 1] range, rather than a binary {0, 1} classification of coordinated vs. non-coordinated communities. Examples of methods of the latter type include [28, 32, 33, 40, 47]. Similarly to [31], Pacheco et al. [33] built a weighted user-similarity network. Then, they discarded all edges whose weight is below a threshold, and clustered the remaining network to discover coordinated communities. The drawback of this method, and similar ones [21], is the need to specify arbitrary thresholds to distinguish between coordinated and non-coordinated behavior, thus providing a binary classification of a non-binary and nuanced phenomenon. Moreover, additional arbitrariness might also arise from the network projection and filtering steps, whose choice can significantly affect coordination results [8]. Other methods do not embed a notion of coordination, but rather propose to apply community detection to weighted user-similarity networks, thus leaving the task of investigating coordinated communities for subsequent analysis [46].

Notably, in all previous work, coordination was detected or measured independently of harmfulness or authenticity. In fact, coordination does not necessarily imply malicious activities: think of online fandoms, or other grassroots initiatives, which are examples of coordinated *harmless and authentic behavior*. Vargas et al. [44] evaluated the capabilities of existing systems to distinguish between harmful and harmless coordination, finding unsatisfactory results and highlighting the difficulty of this task. To this end, our results show that the combined analysis of coordination and propaganda allows to draw insights into the harmfulness and the authenticity of online behavior, thus contributing to bridging this scientific gap.

<sup>2</sup><http://about.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/>

<sup>3</sup><http://slate.com/technology/2020/07/coordinated-inauthentic-behavior-facebook-twitter.html>

hashtag	users	tweets
#GE2019	436,356	2,640,966
#GeneralElection19	104,616	274,095
#GeneralElection2019	240,712	783,805
#VoteLabour	201,774	917,936
#VoteLabour2019	55,703	265,899
#ForTheMany	17,859	35,621
#ForTheManyNotTheFew	22,966	40,116
#ChangelsComing	8,170	13,381
#RealChange	78,285	274,254
#VoteConservative	52,642	238,647
#VoteConservative2019	13,513	34,195
#BackBoris	36,725	157,434
#GetBrexitDone	46,429	168,911
<b>total</b>	<b>668,312</b>	<b>4,983,499</b>

**Table 1:** Statistics about the data collected via hashtags.

## 2.2 Computational Propaganda

Work on propaganda detection has focused on analyzing textual documents [3, 13, 35]. See [11] for a recent survey on computational propaganda detection. Rashkin et al. [35] developed a corpus with document-level annotations with four classes (*trusted*, *satire*, *hoax*, and *propaganda*), labeled using distant supervision: all articles from a given news outlet were assigned the label of that outlet. The news articles were collected from the English Gigaword corpus, which covers reliable news sources, as well as from seven unreliable news sources including two propagandistic ones. They trained a model using word  $n$ -grams, and found that it performed well only on articles from sources that the system was trained on, and that the performance degraded quite substantially when evaluated on articles from unseen news sources. Barrón-Cedeno et al. [3] developed a corpus with two labels (i.e., *propaganda* vs. *non-propaganda*) and further investigated writing style and readability level. Their findings confirmed that using distant supervision, in conjunction with rich representations, might encourage the model to predict the source of the article, rather than to discriminate propaganda from non-propaganda. The studies by Habernal et al. [22, 23] also proposed a corpus with 1.3k arguments annotated with five fallacies that directly relate to propaganda techniques, including *ad hominem*, *red herring*, and *irrelevant authority*.

A more fine-grained analysis was done by Da San Martino et al. [13], who developed a corpus of news articles annotated with the spans of use of eighteen propaganda techniques. They asked to predict the spans of use of propaganda, as well as the specific technique being used, and they further tackled a sentence-level propaganda detection task. They proposed a multi-granular gated deep neural network that captures signals from the sentence-level to improve the performance of the fragment-level classifier and vice versa. Subsequently, an online demo Prta was made publicly available [12], and there were several shared tasks [10, 17].

A limitation of this body of work lies in the lack of methods and tools for uncovering orchestrated propaganda campaigns rather than for detecting individual posts or articles that make use of propaganda. Below, we show that our analysis of coordinated behavior contributes to reaching this goal.

account	tweets	interactions	
		retweets	replies
@jeremycorbyn	788	1,759,823	414,158
@UKLabour	1,002	325,219	79,932
@BorisJohnson	454	284,544	382,237
@Conservatives	1,398	151,913	169,736
<b>total</b>	<b>3,642</b>	<b>2,521,499</b>	<b>1,046,063</b>

**Table 2:** Statistics about the data from Twitter accounts.

## 3 DATA

The starting point for our study is the dataset from [31]. It contains 11,264,820 tweets about the 2019 UK general election, published by 1,179,659 distinct users. The tweets were collected between November 12, 2019 and December 12, 2019 (i.e., the election day) using the Twitter Streaming API. In particular, the dataset contains all tweets that use at least one of the election-related hashtags shown in Table 1. We can see in the table that the hashtags used for data collection include both partisan hashtags as well as neutral ones. The dataset further contains the tweets shared by the two main parties (labour and conservative) and their leaders, as well as the interactions (i.e., retweets and replies) with such tweets, as summarized in Table 2. The final dataset for this study is the combination of the data shown in Tables 1 and 2, and quoted retweets (not counted in the tables). The dataset from [31] is publicly available for research purposes.<sup>4</sup>

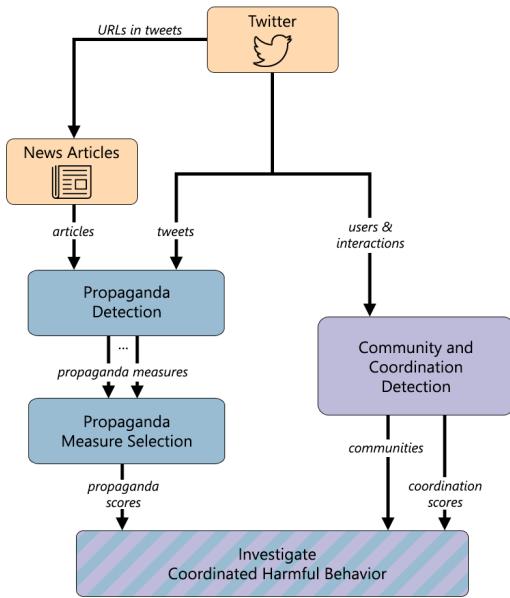
In this work, we extend the above Twitter dataset by also collecting and analyzing the textual content of all the news articles shared during the online electoral debate. To collect data about articles, we first parse the 11M tweets, looking for URLs pointing to news outlets, blogs, or other news Web sites. Out of the entire Twitter dataset, we found 35,976 distinct articles from 3,974 Web sites, that were shared 329,482 times during the data collection period. Finally, we leverage the *newspaper3k* Python package<sup>5</sup> to collect the textual content, together with some metadata about each shared article. We leverage the textual content of the shared articles and tweets to detect the use of propaganda. We further measure the similarities in the user tweeting behaviors to measure coordination.

## 4 METHODS

Figure 1 shows our methodological approach to the analysis of coordinated harmful behavior, based on two main building blocks: (i) a method for measuring coordination, and (ii) a propaganda classifier. As shown in figure, coordination is measured based on user activities and interactions, and results in a coordination score assigned to each user and in the identification of coordinated communities. Instead, propaganda scores are computed from the textual content of tweets and news articles. This process assigns a propaganda score to each user. Finally, each coordinated community is analyzed in terms of the coordination and propaganda scores of its members. The methods for measuring coordination and propaganda are described below.

<sup>4</sup><http://doi.org/10.5281/zenodo.4647893>

<sup>5</sup><http://newspaper.readthedocs.io/en/latest/>

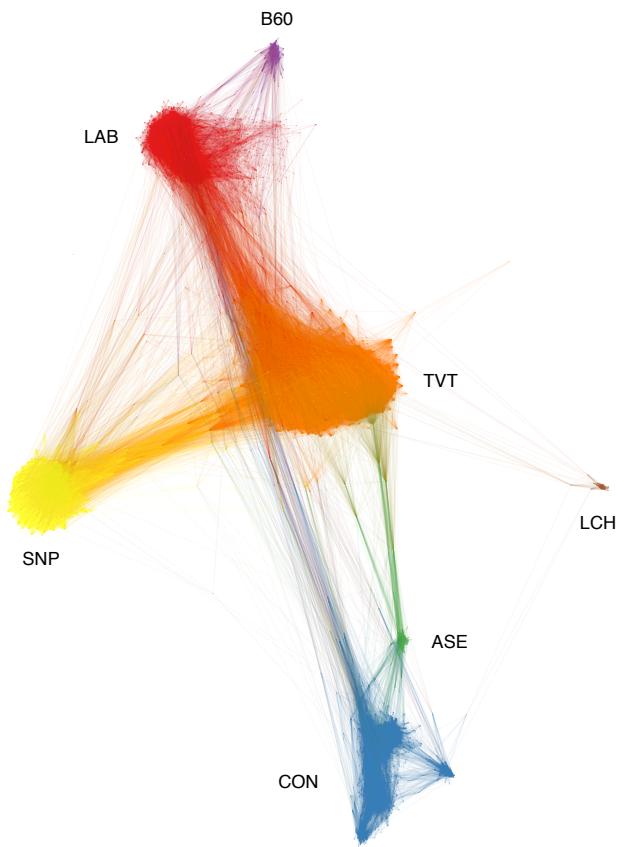


**Figure 1: Overview of our approach based on combining the analysis of propaganda and coordination for studying coordinated harmful behavior.**

#### 4.1 Measuring the Extent of Coordination

To measure the extent of online coordination, we followed network analysis approaches that have been recently proposed in state-of-the-art studies [31, 33, 47], which compute similarities between users and consider exceptional or unexpected similarities as a proxy for coordination. We specifically follow the approach in [31], as it is the only one that produces a coordination score rather than a binary label. For the user selection step, we constrained our analysis to *superspreaders*, defined as the top 1% of the users who shared the most retweets. Despite being only 10,782, superspreaders shared 3.9M tweets, which is 39% of the tweets and 44% of the retweets in our dataset. Previous work has shown that focusing on superspreaders is particularly relevant [34]. We measured the similarity between superspreaders in terms of co-retweets, in order to highlight users who frequently reshare the same messages. For each superspreader, we computed a TF.IDF-weighted vector of the tweet IDs that he/she retweeted. Using TF.IDF-weighting discounts viral tweets by influencers and popular users, while emphasizing retweets of unpopular tweets. Then, we computed the similarity between all pairs of superspreaders as the cosine similarity between their corresponding vectors, thus obtaining a weighted undirected user-similarity network. We filtered the network by computing its multiscale backbone, which allows to retain only statistically significant network structures [39]. Then, we applied the well-known Louvain community detection algorithm to group users into network communities. Finally, we applied network dismantling, which assigns a coordination score to each user in the network. We carried out the latter step by iteratively removing network edges and nodes based on a moving edge weight threshold.

At each iteration, we removed all edges whose raw weight was lower than a threshold, as well as all nodes that ended up being



**Figure 2: User-similarity network for the 2019 UK electoral debate on Twitter. The different coordinated communities that took part in the online debate are color-coded.**

disconnected from the largest connected component. The threshold increased at each iteration, until the network was completely dismantled, i.e., no more connected nodes remained. For each node, we assigned a coordination score as the threshold value that disconnected that node from the rest of the network. We normalized the coordination score in the [0, 1] range, with 1 indicating maximum coordination.

#### 4.2 Propaganda Detection

In order to assess whether an input piece of text is propagandistic, we used Proppy, a state-of-the-art propaganda detection system that achieved an  $F_1$  score of 0.83 on a reference benchmark dataset, outperforming several rival approaches [3]. It uses a maximum entropy binary classifier with  $L_2$  regularization to discriminate propagandistic vs. non-propagandistic texts. Proppy was trained on the QProp corpus, which includes 51k news articles from 94 propagandistic and from 10 non-propagandistic news sources.<sup>6</sup>

<sup>6</sup>From the Media Bias/Fact Check website: <http://mediabiasfactcheck.com>

community	users	articles			tweets		
		#	# distinct	(%)	#	# distinct	(%)
● LAB	5,213	79,157	5,861	(7.4%)	2,064,041	179,601	(8.7%)
● CON	2,279	13,277	1,781	(13.4%)	777,537	76,191	(9.8%)
● TTV	2,258	16,675	3,363	(20.2%)	690,900	62,058	(9.0%)
● SNP	491	2,735	772	(28.2%)	140,338	8,601	(6.1%)
● B60	296	3,231	789	(24.4%)	139,988	9,663	(6.9%)
● ASE	107	706	396	(56.1%)	32,887	4,723	(14.4%)
● LCH	101	150	57	(38.0%)	28,970	2,230	(7.7%)
overall	10,745	116,205	9,960	(8.6%)	3,886,382	343,750	(8.9%)

**Table 3: Statistics about the coordinated communities that took part in the 2019 UK electoral debate on Twitter: number of users and shared articles/tweets.**

Proppy represents the input text using a large number of features, including (i) TF.IDF-weighted  $n$ -grams, (ii) frequency of specific words from a number of lexicons coming from Wiktionary, LIWC, Wilson's subjectives, Hyland hedges, and Hooper's assertives, (iii) writing style features such as TF.IDF-weighted character 3-grams, readability level and vocabulary richness (e.g., Flesch-Kincaid grade level, Flesch reading ease and the Gunning fog index), Type-Token Ratio (TTR), hapax legomena and dislegomena, and (iv) the NEws LAndscape (NELA) features. The latter category includes 130 content-based features collected from the existing literature, which measure different aspects of a news article, comprising sentiment, bias, morality, and complexity [24]. The lexicon features are based on the analysis of the language of propaganda and trustworthy news, discussed in [35]. The encouraging performance achieved by Proppy on reference datasets, as well as its capacity to outperform competing approaches and systems, makes it a favorable system to adopt for propaganda analysis.

## 5 ANALYSIS AND RESULTS

Below, we analyse coordinated communities, and then we discuss the results and the limitations. We then combine these initial results with the analysis of propaganda, and we show how our combined approach helps to overcome the limitations of previous work.

### 5.1 Finding Coordinated Communities

The application of the method for investigating coordination to our dataset resulted in the user-similarity network shown in Figure 2. The network is composed of seven communities of coordinated users, depicted with different colors in the figure and analytically described in Table 3 and in the rest of this subsection:

- LAB: A large community of labourists that supported the Labour party and its leader Jeremy Corbyn, as well as traditional Labour themes such as healthcare and climate change.
- CON: A large community of conservative users. In addition to supporting the party and its leader Boris Johnson, this community was also strongly in favor of Brexit.
- TTV: A large community that included several parties, e.g., liberal democrats, who teamed up with labourists against the conservative party, a strategy dubbed *tactical voting* in the 2019 UK election.

● SNP: A medium-sized community of supporters of the Scottish National Party (SNP). These users also supported Scottish independence from the UK and asked for a new independence referendum.

● B60: A small community of “Backto60” activists. Unlike the previous communities, these users did not represent a political party involved in the election. Instead, B60 users leveraged the electoral debate to protest against a state pension age equalisation law that unfairly affected 4M women born in the 1950s.<sup>7</sup>

● ASE: A small community of conservative users. Despite sharing the same political orientation, these users were separated from the users in the CON community because, rather than supporting the conservative party, they were mainly involved in attacking the labour party. An important narrative for ASE were antisemitism allegations targeted at labourists and Jeremy Corbyn throughout the electoral debate.<sup>8</sup>

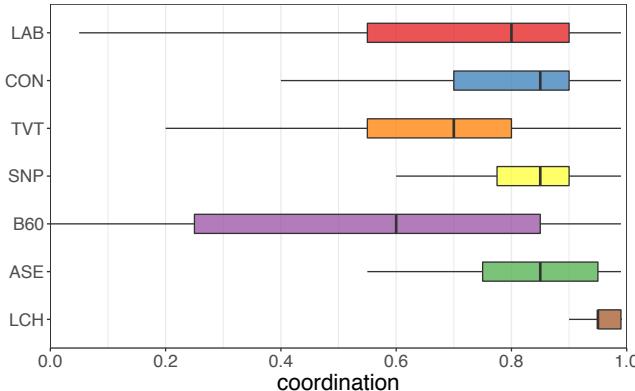
● LCH: Another small community of activists. Similarly to B60, these users were not particularly interested in the electoral debate, but rather protested against a retrospective taxation called “loan charge” that forced certain people to return unsustainable amounts, and which also resulted in several suicides.<sup>9</sup>

The communities that emerged from the analysis of our user-similarity network are consistent with the 2019 UK political landscape and with the results of previous work [25, 38]. Each of these communities had different goals, featured different narratives, and showed diverse degrees of coordination. In particular, we found both large and small communities, as shown in Figure 2. While the larger communities are related to the main political parties in UK that participated in the election (e.g., LAB, CON, TTV, and SNP), the smaller ones represent other highly coordinated users who share a common goal, such as protesting activists (e.g., B60 and LCH) and political antagonists (e.g., ASE). The analysis of Table 3 also reveals a few differences in the sharing behaviors of the different communities. Overall, larger communities seem to share less original articles and tweets, compared to the smaller groups. Moreover, previous analysis showed that these communities are characterized

<sup>7</sup><http://pensionsage.com/pa/Backto60-granted-leave-to-appeal.php>

<sup>8</sup><http://www.thetimes.co.uk/article/revealed-the-depth-of-labour-anti-semitism-bb57h9pdz>

<sup>9</sup><http://www.gov.uk/government/publications/disguised-remuneration-independent-loan-charge-review/guidance>



**Figure 3: Distribution of coordination scores for users from the different communities involved in the online debate.**

by a large negative assortativity, meaning that influential users are mostly connected to ordinary ones, and vice-versa [31].

These figures are indicative of top-down behaviors, where a small number of highly influential characters (e.g., the party leaders) drive the activities of the remaining members. In contrast, smaller communities seem to exhibit bottom-up behavior, characterized by grassroots activities and more content heterogeneity, as testified by the large % of original articles and tweets. The distribution of coordination scores for users of the different communities is shown in Figure 3. Again, different behavior and characteristics emerge for the different communities, and particularly, for the smaller ones. For instance, while B60 features users with diverse degrees of coordination, as shown by a relatively wide boxplot, LCH and SNP are much more homogeneous. B60 is also the community with the lowest degree of average coordination, in contrast to LCH and, to a lower extent, to ASE, SNP, and CON.

**Discussion.** Our results so far match the state-of-the-art in the analysis of coordinated behavior [31, 33, 46]. On the one hand, this approach allows us to obtain nuanced results in terms of coordinated communities. In fact, it allows us to detect several groups of coordinated users, both large and small, thus yielding more informative results compared to coarser analyses that only focused on the two main factions involved in an online debate (e.g., right vs. left, Democrats vs. Republicans, etc.), as done in [7, 20]. Moreover, it surfaces different patterns of coordination (e.g., top-down vs. bottom-up). This approach to the analysis of coordinated behavior allows us to obtain nuanced and fine-grained results, typical of studies that require a great deal of offline, manual investigation [2, 41], while still retaining the advantages of large-scale, automated analysis.

On the other hand, these results do not give insights into the harmfulness of the coordinated communities. In other words, it is still not possible to clearly identify which communities (if any) of those shown in Figure 2, exploited coordination for tampering with the 2019 UK electoral debate on Twitter, and which instead represent neutral or well-intentioned coordinated users. Our subsequent analysis contributes to answering this question.

## 5.2 Measuring Propaganda on Social Media

Our dataset contains two sources of textual content that can potentially convey propaganda: (i) articles and (ii) tweets. Thus, the first choice for computing propaganda scores is which items to analyze: articles or tweets. All propaganda detection systems so far – including Proppy, the one we use in our analysis – were developed for the analysis of news articles [11]. However, from Table 3, we notice that our dataset features, on average, less than one original news article per user and about 32 original tweets per user. Thus, basing our propaganda scores on articles would result in sparse and unreliable estimations. Moreover, the original tweets are authored by the users themselves, unlike news articles, which are just reshared. For these reasons, tweets seem to represent a more direct and reliable input for estimating a user's propaganda. Nonetheless, we computed propaganda scores based on both articles and tweets, and we subsequently compared and validated each of them. The outcome of this comparison and validation allowed us to identify a suitable propaganda score to use in the remainder of our study.

For computing propaganda scores based on articles, we used Proppy with the same configuration proposed by its authors in [3]. For tweets, we made adjustments to account for the inherent differences between news articles and tweets. Specifically, several machine learning features used in the textual classifiers are influenced by document length, and tweets are obviously much shorter than news articles. Thus, we did not classify single tweets, but we grouped the original tweets (i.e., without retweets) by the same author into chunks whose length was comparable to that of the articles used to train Proppy. The grouping merged tweets in chronological order, but we did not apply any filtering based on their textual content (e.g., topic). We carried out a first validation of our propaganda estimations by manually inspecting a subset of the tweets classified by Proppy, which revealed meaningful and satisfactory classifications, and supported our approach for detecting propagandistic vs. non-propagandistic tweets.

Independently on the choice of analyzing articles or tweets, we obtained a propaganda score for user  $j$  as follows:

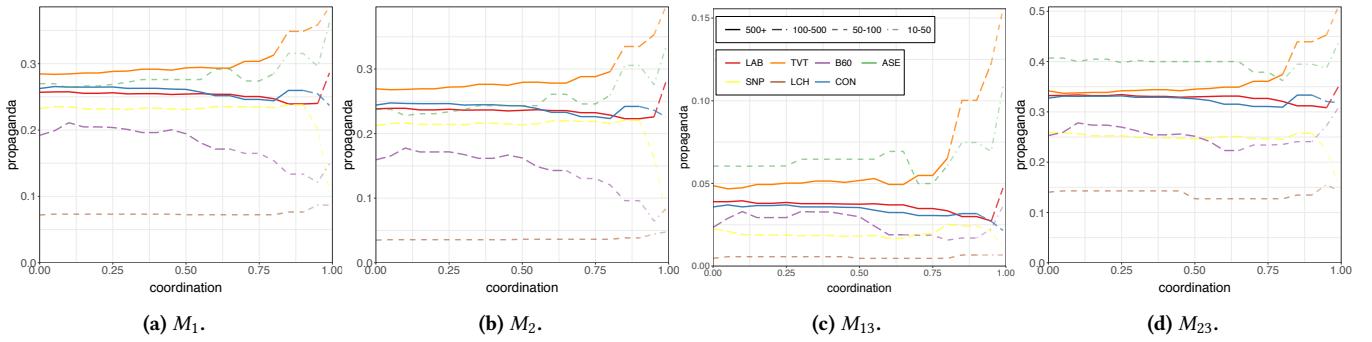
$$P_u(u_j) = \Psi(P(i_k)) \vee i_k \text{ shared by } u_j,$$

where  $\Psi$  is the user-level aggregation function,  $i_k$  are all chunks of original tweets, or all distinct news articles, shared by  $u_j$ , and  $P(i_k)$  are Proppy's classifications of such items. Finally, since we want to compare different communities, we aggregate the user scores for each community. We compute the propaganda score of the  $i$ -th community  $c_i$  as follows:

$$P_c(c_i) = \Phi(P_u(u_j)) \vee u_j \in c_i,$$

where  $\Phi$  is the community-level aggregation function.

Different aggregation functions  $\Psi$  and  $\Phi$  (e.g., mean, median, max, etc.) can be used to compute  $P_u$  and  $P_c$ , respectively. These, in addition to the choice of analyzing articles vs. tweets, result in many possible measures for computing propaganda scores. Table 4 lists some of the measures that we experimented with. In the next sections, we compare the informativeness of these measures, and we choose a suitable one for our further analysis.



**Figure 4: Examples of the relationship between propaganda and coordination for the different communities, obtained with a subset of the measures in Table 4. Independently on the measure  $M_i$ , some communities consistently appear as more propagandistic than others. Specifically, TTV and ASE are among the communities that shared the most propaganda, while SNP, B60 and LCH shared the least. Moreover, the most informative measures from Table 4, i.e.,  $M_1$  and  $M_2$ , exhibit qualitatively similar propaganda trends.**

#	item	user-level aggregation ( $\Psi$ )	community-level aggregation ( $\Phi$ )	informativeness ( $I$ )
$M_1$	🐦	median	mean	0.5491
$M_2$	🐦	majority voting	mean	0.5457
$M_3$	🐦	majority voting	ratio	0.5457
$M_4$	🐦	median	ratio	0.5442
:	:	:	:	:
$M_{10}$	▣	median	mean	0.5156
$M_{11}$	▣	majority voting	mean	0.4990
$M_{12}$	▣	majority voting	ratio	0.4990
$M_{13}$	🐦	median	median	0.4955
:	:	:	:	:
$M_{21}$	▣	median	ratio	0.4475
$M_{22}$	🐦	max	mean	0.4400
$M_{23}$	🐦	max	ratio	0.4400
$M_{24}$	▣	median	median	0.4214

🐦: tweets, ▣: articles

**Table 4: Excerpt of the propaganda measures that we compared. Measures  $M_i$  are shown in descending order of informativeness.**

### 5.3 Spread of Propaganda by Coordinated Users

**Combining coordination and propaganda.** So far, our approach provided us with three pieces of information that we can combine: (i) communities, (ii) coordination scores, and multiple (iii) propaganda scores. By combining community labels with coordination and propaganda scores, we can study the trends of propaganda as a function of coordination for each community. Let  $C_u(u_j)$  be the coordination score of the  $j$ -th user  $u_j$ . Then, the propaganda score for community  $c_i$ , as a function of coordination, is defined as follows:

$$P_c(c_i, k) = \Phi(P_u(u_j)) \quad \forall u_j \in c_i : C_u(u_j) \geq k. \quad (1)$$

In other words, Equation (1) defines how to compute a propaganda score at different coordination thresholds  $k$ , for each community. Therefore, we can assess whether the most coordinated users in each community were also the most propagandistic ones. In turn, this provides valuable information for assessing the harmfulness (or lack thereof) of the different communities.

Figure 4 shows examples of this analysis obtained by applying Equation (1) at different levels of coordination, for some of the propaganda measures shown in Table 4. In this figure, we use different line types and transparencies to indicate the number of users in each community at different levels of coordination. In fact, each community has a different cardinality, as reported in Table 3. Moreover, fewer users are considered when moving towards large coordination values, i.e., only the most coordinated ones. As an example, Figure 4a shows that, for each community, we always have more than 10 users (as well as the several hundreds of tweets that they shared), even at coordination of about 1. Moreover, we always have more than 50 users and thousands of tweets, when the coordination is 0.8 or lower. For propaganda scores derived from tweets, this ensures that the trends shown in the figure are not derived from a trivial number of tweets.

**Choosing a suitable propaganda measure.** In this paragraph, we report the results of a qualitative comparison and a quantitative evaluation of the propaganda measures from Table 4. A desirable characteristic of a propaganda measure is the capacity to distinguish propagandistic vs. non-propagandistic communities. That is, its capacity to highlight the differences between the several communities involved in the online electoral debate, with respect to the use of propaganda.

We evaluated the *informativeness* ( $I$ ) of each measure based on the differences between the propaganda trends that it produces. We quantified the difference between the propaganda trends for two communities as the possible negative linear correlation between them. Thus, for any given measure, we computed the average of the Pearson's correlations  $r$  between the propaganda trends  $P_c(c_x, k)$ ,  $P_c(c_y, k)$  of each possible pair of communities  $c_x$  and  $c_y$ :

$$\bar{r} = \frac{1}{N} \sum_{x,y=1}^N r(P_c(c_x, k), P_c(c_y, k)) \quad \forall x, y : x \neq y.$$

Then, we computed the informativeness of a measure as  $I = \frac{1-\bar{r}}{2}$ . Intuitively, if a measure yields a *positive* correlation between propaganda community trends, then  $\bar{r} \approx 1$  and  $I \approx 0$ . This means that such a measure is not able to diversify the behavior of the different communities, as reflected by the low informativeness.

community	coordination	automation	suspensions	$\delta$ (%)
LAB	-0.193	0.428 *	0.925 ***	-0.016 (-6.5%)
CON	-0.754 ***	-0.688 ***	-0.079	-0.008 (-3.0%)
TVT	0.813 ***	0.836 ***	0.844 ***	+0.074 (+26.0%)
SNP	-0.404 *	0.358	0.902 ***	-0.031 (-13.5%)
B60	-0.899 ***	-0.961 ***	0.882 ***	-0.070 (-36.7%)
ASE	0.742 ***	0.762 ***	-0.218	+0.026 (+9.7%)
LCH	0.562 ***	0.704 ***	-0.733 ***	+0.014 (+20.7%)
overall	-0.679 ***	-0.285	-0.052	-
	(a)	(b)	(c)	

\*\*\*:  $p < 0.01$ , \*\*:  $p < 0.05$ , \*:  $p < 0.1$

**Table 5: Correlation analysis between propaganda scores and (a) coordination scores, (b) automation scores, and (c) Twitter suspensions; for all communities.**

Conversely, if a measure yields a large *negative* correlation between propaganda community trends, then  $\bar{r} \approx -1$  and  $I \approx 1$ , meaning that the measure can diversify the different communities. Notably, our approach is similar but favorable over other alternatives for measuring informativeness, such as those based on mutual information, since they require additional problematic steps for estimating unknown distributions, as discussed in [27].

The last column of Table 4 reports the informativeness of each propaganda measure, with  $M_1$  being the most informative one. As shown in the table, there are relatively small differences in the informativeness of the propaganda measures that we evaluated, which ranges between 0.42 and 0.55 on a 0 to 1 scale. This means that the majority of the measures yield comparable results, as is also visible by the qualitatively similar propaganda trends in Figures 4a and 4b. This suggests that changing the measure would not drastically alter the results. Nonetheless, the topmost measures in Table 4 are relatively better (i.e., more informative) at surfacing differences between the investigated communities. Interestingly, the most informative measures in Table 4 are all based on the analysis of tweets, which reinforces our initial hypotheses and the results of our manual validation of propaganda classifications of tweets. Following these preliminary results and without loss of generality, in our subsequent analysis we adopt  $M_1$  for obtaining propaganda scores.

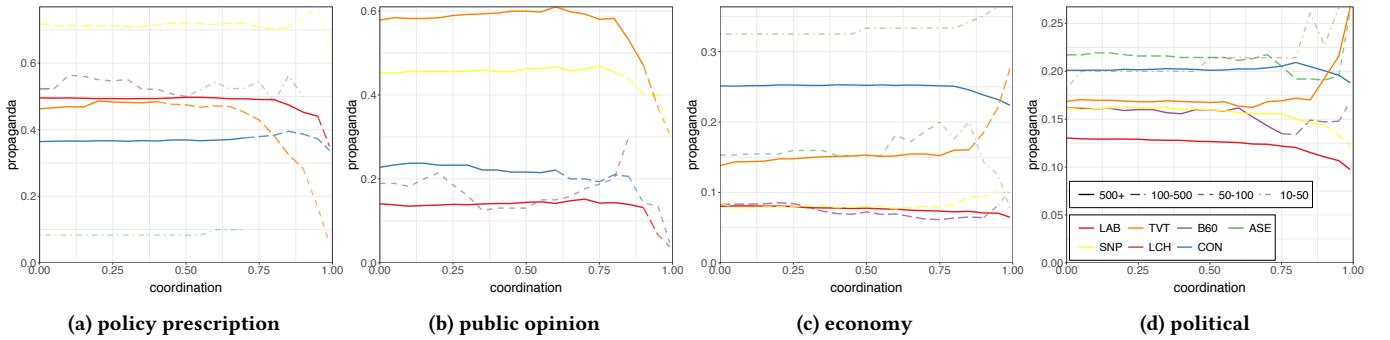
**Results.** Figure 4a shows interesting propaganda trends for some communities. First, LCH is characterized by the lowest degree of propaganda among all communities. Similarly, B60 shows a marked decreasing propaganda trend, implying that the core users of the community (i.e., the most coordinated ones) are not engaged in propaganda. Both these findings suggest harmless behavior. In other words, LCH appears to be highly coordinated, as shown in Figure 3, but harmless. B60 features diverse degrees of coordination among its users, but is nonetheless harmless. On the contrary, other communities feature increasing propaganda trends: above all, TVT and ASE. For coordination of 0.5 and above, both show increasing levels of propaganda, which supports the hypothesis of harmful communities. For the remaining three communities (i.e., LAB, CON, and SNP) the coordination appears to be mostly unrelated to propaganda.

Note that these results are overall robust to changing the measure used to compute the propaganda scores.

The above qualitative findings are further confirmed by the quantitative results reported in Table 5a. In particular, our correlation analysis shows strong, positive, and statistically significant Pearson correlations between propaganda and coordination for the TVT and the ASE communities, with  $r = 0.813$  and  $r = 0.742$ , respectively. Instead, B60 features a strong, negative, and statistically significant correlation  $r = -0.899$ . The remaining results in Table 5 are not meaningful, either because of small correlations or due to low statistical significance (for LAB, SNP, and LCH), or because of limited variation of propaganda (for CON). Regarding the latter, Pearson correlation measures the strength of the linear relationship between two variables, but not the extent of variation for either one, which is relevant in our analysis. The last column of Table 5 accounts for this aspect by measuring the variation in propaganda as  $\delta = P_c(c_i, k = 0.9) - P_c(c_i, k = 0)$ , for each community  $c_i$ . Despite featuring a marked negative and significant correlation, CON exhibits a very small variation in propaganda, with  $\delta = -0.008$ , represented in Figure 4a by a mostly flat line.

**Discussion.** Given the lack of ground truth on coordinated harmful vs. harmless behaviors, one way to qualitatively validate our analysis is by cross-checking our results with previous work and with the role of the communities in the electoral debate, as also done in previous work [31, 33]. Two communities emerged as authentic and harmless: LCH and B60. This means that, according to our proposed methodology, their activities are coordinated, but not malicious nor deceptive. In other words, they exhibit coordinated but authentic and harmless behavior. From previous work [31] and from our analysis of coordinated communities, we find that these are groups of activists protesting against unfair taxation (LCH) and in favor of women's rights (B60). Table 6 provides a detailed look at some of the tweets from B60's highly coordinated users, confirming that their focus was to promote their cause and to encourage women to exercise their right to vote. They further endorsed the Labour leader Jeremy Corbyn, who expressed support for their initiative. Hence, our methodology correctly highlighted these activists as harmless examples of grassroots coordination. In contrast, our analysis revealed that TVT and ASE featured characteristics related to harmful behaviors. Both are highly polarized communities with strong political motivations.

Regarding TVT and its highly coordinated members, Table 6 shows that the majority of their tweets and shared articles are



**Figure 5: Frame analysis of the news articles shared by the different communities. Propaganda is measured as the fraction of articles in a certain frame flagged as propagandist, out of all articles from that frame. Line-types indicate the number of tweets. Lines may interrupt if no tweets were shared by coordinated users ( $\text{coordination} \geq k$ ) of a community for a given frame.**

tweet text		article title
Join @fckboris & #RegisterToVote #GE2019 <url>		Five ways to say Fck Boris
GE2019: Five ways to say F**k Boris #GeneralElection2019#Brexit#VoteToriesOut <url>		Five ways to say Fck Boris
KARMA: Katie Hopkins forced to sell £1m home and now rents after cringe libel loss / #GE2019 #Brexit <url>		Katie Hopkins forced to sell £1m home and now rents after humiliating libel case
Here is vocal #RemainerNow and life long Tory voter @OborneTweets on #GE2019 <url>		As a lifelong Conservative, here's why I can't vote for Boris Johnson
TVT Boris Johnson unfit to be Prime Minister. Brexit damages the Henley constituency - John Howell MP knows this but continues his support- @LaureacoyleLD Laura Coyle the People's Vote recommendation <url> #TacticalVoting #PeoplesVote #Boris #GeneralElection2019		I was Boris Johnson's boss: he is utterly unfit to be prime minister
This is beyond desperate and beyond tin pot banana republic actions if true. #GE2019 <url>		General election: Farage claims No 10 offered Brexit Party candidates jobs to stand down
B60 I'm voting Labour for the final say on Brexit. Share why you're voting Labour Right #VoteLabour <url>		I'm Voting Labour...
I'm voting Labour for a million green jobs. Share why you're voting Labour #VoteLabour <url>		I'm Voting Labour...
WASPI Women won't be silenced #GeneralElection2019 <url> via @WASPI_Campaign		WASPI Women won't be silenced #GeneralElection2019
Owen Jones: 'They don't want you to vote. Defy them' #GTTO #GeneralElection2019 <url> Sent via @updayUK		Owen Jones: 'They don't want you to vote. Defy them'
Well done May from Falkirk Waspi. <url> #waspicampaign2018 #deedsnotwords #onevoice #GeneralElection2019		WASPI woman puts Boris Johnson on spot about trust after he publicly pledged to try and sort out pension row during visit to Cheltenham
B60 Now that most party manifestos have been published, join @WASPI_Campaign today at <url> and find your nearest local group at <url> to find out about our #GE2019 Toolkit so you can speak up for #WASPI in your local area #WASPIwomenvote		How to join WASPI

**Table 6: Excerpt of the activity of strongly coordinated ( $k \geq 0.9$ ) members of TTV and B60. While the TTV users attack Boris Johnson and Brexit, the B60 users encourage women to vote and support the WASPI (Women Against State Pension Inequality) campaign.<sup>10</sup> Our method labeled all TTV tweets in the table as propagandistic, and all B60 tweets as non-propagandistic.**

politically themed. Most of the time they attack Boris Johnson and the conservatives. Similarly, at the opposite side of the political spectrum, ASE's peculiarity was that of repeatedly attacking the Labour party and its leader with allegations of antisemitism. Here, our methodology highlighted aggressive communities as harmful. Lastly, LAB, CON, and SNP appear as neither markedly harmless nor harmful. This is in line with the role of these communities, since they are large communities of moderate users [31].

We further inspected the framing of the articles shared by different communities. We used the frame inventory of the Media Frames Corpus [5], and we performed automatic annotation of the frames using the Tanbih API.<sup>11</sup> Tanbih is a news aggregator platform with intelligent data analysis capabilities, including the possibility to analyze articles and news outlets based on their degree of factual reporting, propagandistic content, hyper-partisanship, political bias, and stance with respect to various claims and topics [14, 18, 50]. Figure 5 shows the analysis for four relevant frames, highlighting

striking differences across the frames, even within a single community. For example, Figure 5d shows that the *Political* frame for TTV evolves into propagandistic behavior as coordination increases. Conversely, for *Policy Prescription*, the propaganda score for the same community decreases. This suggests that the spread of propaganda is a theme-dependent phenomenon. Figure 5 also highlights that some communities deviate from the rest in terms of overall propagandistic content, such as SNP and TTV, which maintain a relatively high propaganda score of 0.45–0.60 for *Public Opinion*, compared to the remaining communities.

**Comparisons.** Here, we discuss our methodology and our results compared to previous work, highlighting the usefulness and the advantages of our approach. Several earlier attempts at detecting inauthentic and harmful campaigns only investigated coordination and synchronization between user accounts [33, 40, 46]. In this work, all groups of users exhibiting unexpected coordination were considered to be malicious [33]. Despite representing an initial solution to the task of detecting malicious campaigns, this approach has a number of drawbacks.

<sup>10</sup><https://www.waspi.co.uk/about-us-2/>

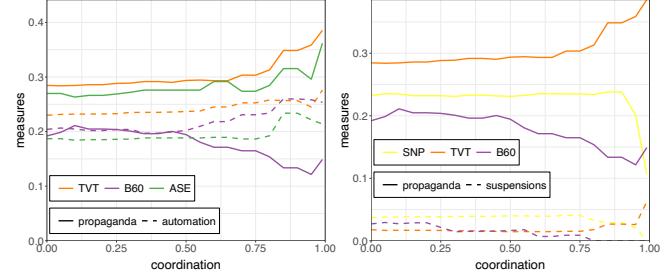
<sup>11</sup><http://app.swaggerhub.com/apis/yifan2019/Tanbih/0.8.0/>

For example, if applied to our dataset, it would have flagged the LCH community as malicious, due to its extreme degree of coordination, as can be seen in Figure 3. However, our nuanced analysis of propaganda and coordination revealed that the LCH users are protesting activists, which is a finding also confirmed by Nizzoli et al. [31]. Conversely, the TTV community features the second-lowest degree of coordination among our communities. As such, it would have been labeled as non-suspicious by previous techniques. Our analysis further revealed a strong positive correlation between propaganda and coordination for TTV users, thus uncovering their malicious intent. In summary, our results show that coordination alone does not provide enough information for assessing the real activities and intent of online communities. Instead, a methodology combining the analysis of coordination with signs of malicious intent (e.g., propaganda), such as the proposed one, can distinguish inauthentic and harmful behavior vs. authentic and harmless one.

Our findings also confirm and extend previous results about the role of small and fringe Web communities in information disorder. Zannettou et al. [48, 49] noted that fringe, polarized, and strongly motivated communities are those that exert the most influence on the Web for issues such as disinformation and online abuse, despite being relatively small. In our analysis, we obtained comparable results. Indeed, the most interesting communities (i.e., those that exhibit coordinated yet markedly harmless or harmful behavior) are small and non-mainstream, such as LCH, B60, and ASE. However, while Zannettou et al. investigated this phenomenon *across* Web platforms, here we show that the same also occurs *within* platforms.

#### 5.4 Inauthentic and Harmful Behavior

We conclude our analysis by comparing propaganda and coordination scores to other clear signs of inauthenticity and harmfulness. We leverage scores indicating automation (i.e., botness) as a proxy for inauthenticity. For each account, we use the maximum of Botometer's English and universal scores, both provided in the  $[0, 1]$  range, as its automation score [37]. While we are aware of the limitations of current bot detectors [9], including Botometer [36], the strong interest on the role played by social bots in online manipulation campaigns motivates this analysis. Similarly, we investigate the number of accounts suspended by Twitter in each community, as a proxy for harmfulness. Table 5 reports in columns (b) and (c) the correlation results between our propaganda scores versus automation and Twitter suspensions, respectively. By cross-checking strong and significant correlations against notable variations in propaganda ( $\delta$ ), we highlight interesting trends. Regarding automation, the same communities that featured a strong positive correlation between propaganda and coordination – namely, TTV and ASE – are also strongly correlated with automation scores. This means that highly coordinated users in TTV and ASE are both inauthentic and harmful, further confirming our earlier results. An unexpected result is instead obtained for B60, which features a strong negative correlation between propaganda and automation. In other words, while propaganda decreases as a function of coordination, automation scores increase. Thus, coordinated B60 users could be leveraging automation as a way to boost their online actions. Propaganda and automation trends for TTV, ASE, and B60 are shown in Figure 6a.



(a) Propaganda vs. automation. (b) Propaganda vs. suspensions.

**Figure 6: Trends of propaganda versus trends of automation scores and suspensions, for a few relevant communities. The communities are color-coded, and the line types indicate trends of propaganda, automation, and suspensions.**

Overall, TTV appears as the most harmful community throughout the online UK electoral debate, with high propaganda, high automation, and the largest share of accounts suspended by Twitter. We further measure strong positive correlation between propaganda and suspension trends for B60 and SNP. Since, for these communities, propaganda decreases with coordination, these positive correlations mean that Twitter suspensions also decrease, which is a sign of harmless behavior. This result is particularly relevant for B60, and it corroborates our previous findings. The trends of propaganda and suspensions for SNP, TTV, and B60 are shown in Figure 6b.

## 6 CONCLUSION AND FUTURE WORK

We carried out the first combined analysis of propaganda and coordination in online debates. Specifically, we applied our methodology to the 2019 UK electoral debate on Twitter, revealing (i) harmful, (ii) neutral, and (iii) well-intentioned communities that took part in the debate. Among the most harmful communities, we found “tactical voters” (TTV), who colluded against conservatives, and a small group of political antagonists, who attacked labourists and Jeremy Corbyn with accusations of antisemitism (ASE). Among the harmless coordinated communities, we uncovered groups of activists protesting against loan taxation (LCH) and in favor of women’s rights (B60). Besides providing novel and interesting insights into the communities that participated in the 2019 UK electoral debate, our results also demonstrate the need to combine analysis of coordinated user behavior and intent. Our methodology contributes to distinguishing between coordinated *harmful* and *harmless* behavior, thus overcoming one of the main limitations of earlier work.

Among the future challenges along this important research direction is the construction of a reliable ground truth for coordinated harmful and harmless behavior. This endeavor would allow shifting from the current descriptive work to predictions, by training models that can detect harmful behavior. We also plan to collect and to investigate additional information about online communities, thus going beyond the analysis of coordination and propaganda. If successful, these efforts will allow a deeper understanding of coordinated online behavior, thus enabling the possibility to rapidly intervene, and ultimately to limit the spread, the influence, and the societal impact of online information disorders.

## ACKNOWLEDGMENTS

The work is part of the Tanbih mega-project, developed at the Qatar Computing Research Institute, HBKU, which aims to limit the impact of “fake news,” propaganda, and media bias by making users aware of what they are reading, thus promoting media literacy and critical thinking.

## REFERENCES

- [1] Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, Abdulaziz Al-Homaid, Wajdi Zaghouani, Tommaso Caselli, Gijs Danoe, Friso Stolk, Britt Bruntink, and Preslav Nakov. 2021. Fighting the COVID-19 Infodemic: Modeling the Perspective of Journalists, Fact-Checkers, Social Media Platforms, Policy Makers, and the Society. In *EMNLP*.
- [2] Dennis Assenmacher, Lena Clever, Janina Susanne Pohl, Heike Trautmann, and Christian Grimme. 2020. A Two-Phase Framework for Detecting Manipulation Campaigns in Social Media. In *SCSM*.
- [3] Alberto Barrón-Cedeno, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. 2019. Propppy: Organizing the news based on their propagandistic content. *Information Processing & Management* 56, 5 (2019), 1849–1864.
- [4] G Bolsover and P Howard. 2017. Computational Propaganda and Political Big Data: Moving Toward a More Critical Research Agenda. *Big Data* 5, 4 (2017), 273–276.
- [5] Dallas Card, Amber E. Boydston, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The Media Frames Corpus: Annotations of Frames Across Issues. In *ACL IJCLP*. 438–444.
- [6] Center for an Informed Public, Digital Forensic Research Lab, Graphika, and Stanford Internet Observatory. 2021. The Long Fuse: Misinformation and the 2020 Election. *Stanford Digital Repository: Election Integrity Partnership* (2021). <https://purl.stanford.edu/tr171zs0069>
- [7] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on Twitter. In *AAAI ICWSM*.
- [8] Michele Coscia and Luca Rossi. 2019. The impact of projection and backboning on network topologies. In *IEEE/ACM ASONAM*.
- [9] Stefano Cresci. 2020. A decade of social bot detection. *Commun. ACM* 63, 10 (2020), 72–83.
- [10] Giovanni Da San Martino, Alberto Barrón-Cedeno, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *SemEval*.
- [11] Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeno, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020. A survey on computational propaganda detection. In *IJCAI*. 4826–4832.
- [12] Giovanni Da San Martino, Shaden Shaar, Yifan Zhang, Seunghak Yu, Alberto Barrón-Cedeno, and Preslav Nakov. 2020. Prta: A system to support the analysis of propaganda techniques in the news. In *ACL*.
- [13] Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019. Fine-Grained Analysis of Propaganda in News Articles. In *ACL EMNLP*.
- [14] Kareem Darwish, Peter Stefanov, Michaël Aupetit, and Preslav Nakov. 2020. Unsupervised user stance detection on Twitter. In *AAAI ICWSM*.
- [15] Roberto Di Pietro, Maurantonio Caprolu, Simone Raponi, and Stefano Cresci. 2021. *New Dimensions of Information Warfare*. Advances in Information Security, Vol. 84. Springer. 268 pages.
- [16] Dimiter Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Detecting Propaganda Techniques in Memes. In *ACL-IJCNLP*.
- [17] Dimiter Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. SemEval-2021 task 6: Detection of Persuasion Techniques in Texts and Images. In *SemEval*.
- [18] Tiziano Fagni and Stefano Cresci. 2022. Fine-Grained Prediction of Political Leaning on Social Media with Unsupervised Deep Learning. *Journal of Artificial Intelligence Research* 73 (2022), 633–672.
- [19] Emilio Ferrara, Stefano Cresci, and Luca Luceri. 2020. Misinformation, manipulation and abuse on social media in the era of COVID-19. *Journal of Computational Social Science* 3 (2020), 271–277.
- [20] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2017. Reducing controversy by connecting opposing views. In *ACM WSDM*.
- [21] Fabio Giglietto, Nicola Righetti, Luca Rossi, and Giada Marino. 2020. It takes a village to manipulate the media: coordinated link sharing behavior during 2018 and 2019 Italian elections. *Information, Communication & Society* (2020), 1–25.
- [22] Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational Argumentation Meets Serious Games. In *ACL EMNLP*.
- [23] Ivan Habernal, Patrick Pauli, and Iryna Gurevych. 2018. Adapting Serious Game for Fallacious Argumentation to German: Pitfalls, Insights, and Best Practices. In *LREC*.
- [24] Benjamin D. Horne, William Dron, Sara Khedr, and Sibel Adali. 2018. Sampling the News Producers: A Large News and Feature Data Set for the Study of the Complex Media Landscape. In *AAAI ICWSM*.
- [25] Dan Jackson, Einar Thorsen, Darren Lilleker, and Nathalie Weidhase. 2019. *UK Election Analysis 2019: Media, Voters and the Campaign*. Technical Report. Bournemouth University. <http://eprints.bournemouth.ac.uk/33165/>
- [26] Shagun Jhaver, Christian Boylston, Diyi Yang, and Amy Bruckman. 2021. Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter. In *ACM CSCW*.
- [27] Shiraj Khan, Sharba Bandyopadhyay, Auroop R Ganguly, Sunil Saigal, David J Erickson III, Vladimir Protopopescu, and George Ostrouchov. 2007. Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. *Physical Review E* 76, 2 (2007), 026209.
- [28] Thomas Magelinski, Lynnette Hui Xian Ng, and Kathleen M Carley. 2021. A Synchronized Action Framework for Responsible Detection of Coordination on Social Media. *arXiv:2105.07454* (2021).
- [29] Milad Mirbabaei, Felix Brünker, Magdalena Wischniewski, and Judith Meiners. 2021. The Development of Connective Action during Social Movements on Social Media. *ACM Transactions on Social Computing* 4, 1 (2021), 1–21.
- [30] Lynnette Hui Xian Ng, Iain Cruickshank, and Kathleen M Carley. 2021. Coordinating Narratives and the Capitol Riots on Parler. *arXiv:2109.00945* (2021).
- [31] Leonardo Nizzoli, Serena Tardelli, Marco Avvenuti, Stefano Cresci, and Maurizio Tesconi. 2021. Coordinated behavior on social media in 2019 UK general election. In *AAAI ICWSM*.
- [32] Diogo Pacheco, Alessandro Flammini, and Filippo Menczer. 2020. Unveiling Coordinated Groups Behind White Helmets Disinformation. In *WWW Companion*.
- [33] Diogo Pacheco, Pik-Mai Hui, Christopher Torres-Lugo, Bao Tran Truong, Alessandro Flammini, and Filippo Menczer. 2021. Uncovering coordinated networks on social media: Methods and case studies. In *AAAI ICWSM*.
- [34] Sen Pei, Lev Muchnik, José S Andrade Jr, Zhiming Zheng, and Hernán A Makse. 2014. Searching for superspreaders of information in real-world social media. *Scientific reports* 4 (2014), 5547.
- [35] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In *ACL EMNLP*.
- [36] Adrian Rauchfleisch and Jonas Kaiser. 2020. The false positive problem of automatic bot detection in social science research. *PLoS One* 15, 10 (2020).
- [37] Mohsen Sayyadihikandeh, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2020. Detection of novel social bots by ensembles of specialized classifiers. In *ACM CIKM*.
- [38] Shannon Schumacher. 2019. *Brexit divides the UK, but partisanship and ideology are still key factors*. Technical Report. Pew Research Center.
- [39] M Ángeles Serrano, Marián Boguá, and Alessandro Vespignani. 2009. Extracting the multiscale backbone of complex weighted networks. *PNAS* 106, 16 (2009).
- [40] Karishma Sharma, Yizhou Zhang, Emilio Ferrara, and Yan Liu. 2021. Identifying coordinated accounts on social media through hidden influence and group behaviours. In *ACM KDD*.
- [41] Kate Starbird, Ahmer Arif, and Tom Wilson. 2019. Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations. In *ACM CSCW*.
- [42] Serena Tardelli, Marco Avvenuti, Maurizio Tesconi, and Stefano Cresci. 2022. Detecting inorganic financial campaigns on Twitter. *Information Systems* (2022).
- [43] Amaury Trujillo and Stefano Cresci. 2022. Make Reddit Great Again: Assessing Community Effects of Moderation Interventions on r/The\_Donald. *arXiv:2201.06455* (2022).
- [44] Luis Vargas, Patrick Emami, and Patrick Traynor. 2020. On the detection of disinformation campaign activity with network analysis. In *ACM CCSW*.
- [45] Claire Wardle and Hossein Derakhshan. 2017. Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe* 27 (2017).
- [46] Derek Weber and Frank Neumann. 2020. Who's in the Gang? Revealing Coordinating Communities in Social Media. In *IEEE/ACM ASONAM*.
- [47] Derek Weber and Frank Neumann. 2021. Amplifying influence through coordinated behaviour in social networks. *SNAM* 11, 1 (2021), 1–42.
- [48] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the origins of memes by means of fringe Web communities. In *ACM IMC*.
- [49] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2017. The Web centipede: understanding how web communities influence each other through the lens of mainstream and alternative news sources. In *ACM IMC*.
- [50] Yifan Zhang, Giovanni Da San Martino, Alberto Barrón-Cedeno, Salvatore Romeo, Jisun An, Haewoon Kwak, Todor Staykovski, Israa Jaradat, Georgi Karadzhov, Ramy Baly, et al. 2019. Tanbih: Get To Know What You Are Reading. In *ACL EMNLP*.