Coordination patterns reveal online political astroturfing across the world

Schoch, David; Keller, Franziska B.; Stier, Sebastian; Yang, JungHwan

Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - Projektnummer 491156185 / Funded by the German Research Foundation (DFG) - Project number 491156185

Empfohlene Zitierung / Suggested Citation:
Schoch, D., Keller, F. B., Stier, S., & Yang, J. (2022). Coordination patterns reveal online political astroturfing across the world. *Scientific Reports*, *12*, 1-10. https://doi.org/10.1038/s41598-022-08404-9

scientific reports

Check for updates

OPEN

Coordination patterns reveal online political astroturfing across the world

David Schoch[1]✉, Franziska B. Keller[2], Sebastian Stier[1] & JungHwan Yang[3]

Online political astroturfing—hidden information campaigns in which a political actor mimics genuine citizen behavior by incentivizing agents to spread information online—has become prevalent on social media. Such inauthentic information campaigns threaten to undermine the Internet's promise to more equitable participation in public debates. We argue that the logic of social behavior within the campaign bureaucracy and principal–agent problems lead to detectable activity patterns among the campaign's social media accounts. Our analysis uses a network-based methodology to identify such coordination patterns in all campaigns contained in the largest publicly available database on astroturfing published by Twitter. On average, 74% of the involved accounts in each campaign engaged in a simple form of coordination that we call co-tweeting and co-retweeting. Comparing the astroturfing accounts to various systematically constructed comparison samples, we show that the same behavior is negligible among the accounts of regular users that the campaigns try to mimic. As its main substantive contribution, the paper demonstrates that online political astroturfing consistently leaves similar traces of coordination, even across diverse political and country contexts and different time periods. The presented methodology is a reliable first step for detecting astroturfing campaigns.

At the very latest since the Russian Internet Research Agency's (IRA) intervention in the U.S. presidential election 2016, scholars and the broader public have become wary of online disinformation campaigns[1]. Such campaigns often aim to deteriorate the public's trust in electoral institutions or the government's legitimacy – or try to shore up support for authoritarian governments. Some are successful in that regard, as experimental research has demonstrated[2,3]. While there is disagreement in large-scale observational studies regarding the prevalence of untrustworthy online information in citizens' information diets[4–6], there is a consensus that disinformation is a major problem, whether it emerges organically from communities or is pushed by systematic campaigns.

However, current scholarship on disinformation campaigns is largely focused on the detection of automated accounts, so-called social bots[7–9], even though it has been shown that such accounts make up only a small part of contemporary astroturfing campaigns[10] and the validity of the bot-detection methods is in question[11]. To fill this research gap, our study focuses on "political astroturfing", i.e., centrally coordinated disinformation campaigns in which participants pretend to be ordinary citizens who act independently[12]. The accounts that are associated with political astroturfing may or may not be automated social bots and may or may not spread "fake news"[13], misinformation, or disinformation[14–18] but they do deceive the audience by disguising their identity and hiding the motivations of the account owner.

We argue that these campaigns can be more accurately detected by searching for *centralized coordination patterns among groups of accounts* instead of looking at *"suspicious" activity of individual accounts*. After all, political astroturfing is an organized campaign activity that entails a coordination of multiple accounts. Our past research on astroturfing by the South Korean secret service in 2021 demonstrated that the central coordination and organizational routines inherent to an information campaign allows researchers to distinguish between campaign agents and ordinary Twitter users[10,19]. Similarly, recent studies examined coordination patterns of the accounts that were harassing members of the Iranian diaspora on Instagram[20] and hijacking German Twitter debates during an election campaign[21]. To examine whether the relatively simple coordination patterns described in previous research can be used to detect a wide variety of astroturfing campaigns in more recent years, we apply the method developed in Keller et al.[10] to all astroturfing campaigns revealed by Twitter. We then demonstrate the generalizability of the detection method by testing its performance on more recent campaigns that engaged in different strategies, targeted different audiences, and used different languages. Overall, we argue

[1]GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany. [2]University of Bern, Bern, Switzerland. [3]University of Illinois at Urbana-Champaign, Champaign, USA. ✉email: david.schoch@gesis.org

that identifying and restricting astroturfing campaigns is important for leveling the playing field for political debates in democratic societies.

Theory and research approach

Our central argument is that studying the timing and centralization of message coordination helps locate the activity of a set of accounts on an empirically observable spectrum of group-based behaviors on social media. Such a spectrum ranges from "uncoordinated messages of unrelated users" to "centrally coordinated information campaign". Astroturfing – in particular if campaigns employ unsophisticated social bots – will be placed on the latter end of the spectrum that exhibits a strong group-based coordination. One theoretically relevant question is the location of grassroots movements in this space. Grassroots movements exhibit some coordination but are less synchronized in the timing and content of their messages, because their participants respond organically to cues sent by their peers instead of centralized instructions. Therefore, they will appear in the middle of the spectrum. As a consequence, the centrally coordinated organization of astroturfing should leave different empirical traces than grassroots campaigns.

We ground our explanation of these patterns in social science theory, more specifically, the principal–agent framework as employed in political science[22], where it has been used to explain the pitfalls of ground campaign organization during elections[23]. Applying the framework to astroturfing, we argue that the organizers of an astroturfing campaign are *principals* who try to pursue (political) goals by instructing and incentivizing *agents* to create and share messages congruent with the campaign's goals. Because one of the purposes of astroturfing is to reach and change the behavior of as many *regular users* as possible, the success of the campaign is contingent on a wide reach and an organic appearance of the campaign. However, according to principal–agent theory, reaching this more complex goal is difficult because of the misalignment between the principal's and the agents' preferences: agents thus need to be extrinsically motivated and will try to shirk[22], e.g. by creating similar or identical accounts and content instead of coming up with original contributions to the campaign. Unless the principal can establish an expensive system of close monitoring, the agents will continue to hold an information advantage over the principal: they know how much effort they exerted in creating convincing online personas (oftentimes, little), whereas the principal does not.

We thus would expect campaign accounts to post or re-post similar or identical messages within a short time window, something we call *co-tweeting* and *co-retweeting*, and coordinate with a large number of other campaign accounts. This might be similar to grassroots campaigns' operations, but their activities do not follow centralized instructions, and are therefore more likely to post similar content in a cascading fashion over an extended time period, and with greater variation in the content as they engage in more localized coordination with a few friends, imitating and varying the content they see online. Finally, grassroots campaigns may rely heavily on Twitter's decentralized mechanism for spreading the message, *retweeting*. But as astroturfing campaigns use retweeting as well[10], this is unlikely to be a useful feature to distinguish the two. Finally, principal–agent theory predicts that agents will only extend their efforts when they are supervised, which might result in *unique temporal activity patterns*, e.g., posting only during regular office hours. We expect that this principal–agent constellation results in universal patterns that appear in astroturfing campaigns in multiple countries across the world.

Figure 1 shows a schematic representation of our research design. We use the complete data released by Twitter as part of its *Information Operations Hub* initiative[24] up until February 2021 as "ground truth" data to show that similar coordination patterns appear in almost all cases. We also include a campaign that escaped Twitter's attention, namely the South Korean secret service's attempt at influencing the national elections in 2012, bringing our total population of astroturfing campaigns to 46. We then concentrate the focus of the study on the 33 campaigns that produced at least 50,000 tweets. In order to validate methods for the detection of astroturfing, we compare the behavior of astroturfing accounts to two "comparison samples" that represent groups of users that a given campaign likely tries to mimic and influence. As retreiving the tweets for such systematic samples is time and resource intensive, we select four distinct campaigns targeting audiences in six countries for an in-depth study: the Russian Internet Research Agency's (IRA) attempt at polarizing public opinion in the U.S. and Germany, and shoring up regime support in Russia, the Chinese government's attempt at changing the framing of the Hong Kong protest, a campaign cheerleading the government of Venezuela and the above mentioned South Korean case. Further details on case selection and sample construction can be found in the Methods section.

Defining adequate comparison groups is essential in this detection process because (decentralized) coordination can also happen as part of organic discussions among specific issue publics and grassroots movements. To demonstrate the added value of our research in that regard, we conducted a literature review of related research aiming to detect astroturfing (see Table S1 in the Supplementary Information [SI]). We excluded (1) studies that aim to detect social bots, as this research does not aim to reveal specific astroturfing campaigns but general automation patterns instead, and (2) studies that merely use the data released by Twitter for an analysis of astroturfing without aiming to predict which accounts are part of a specific campaign. The latter category includes a number of studies that examined the character and reach of individual astroturfing campaigns in various contexts such as China[25], Saudia Arabia[26] or the Russian influence campaign in the U.S.[27–30]).

Among the research listed in Table S1[10,31–34], the studies most closely related to ours are Vargas et al.[34] and Alizadeh et al.[31], both of which use some of the same data to distinguish between astroturfing accounts and regular users. In particular, Vargas and colleagues built on our earlier work[10,19] with the goal of constructing a classifier to detect astroturfing campaigns more generally[34]. However, they chose baselines that represent neither genuine grassroots movements nor the specific issue publics targeted by the campaigns, but instead three institutionalized English-speaking elite communities (members of the U.S. Congress, the UK Parliament and academics). Not only do these communities differ from the country-specific audiences engaged in the topics targeted by an astroturfing campaign, but the U.S. and UK parliamentarians' accounts are often run by a group of staffers who

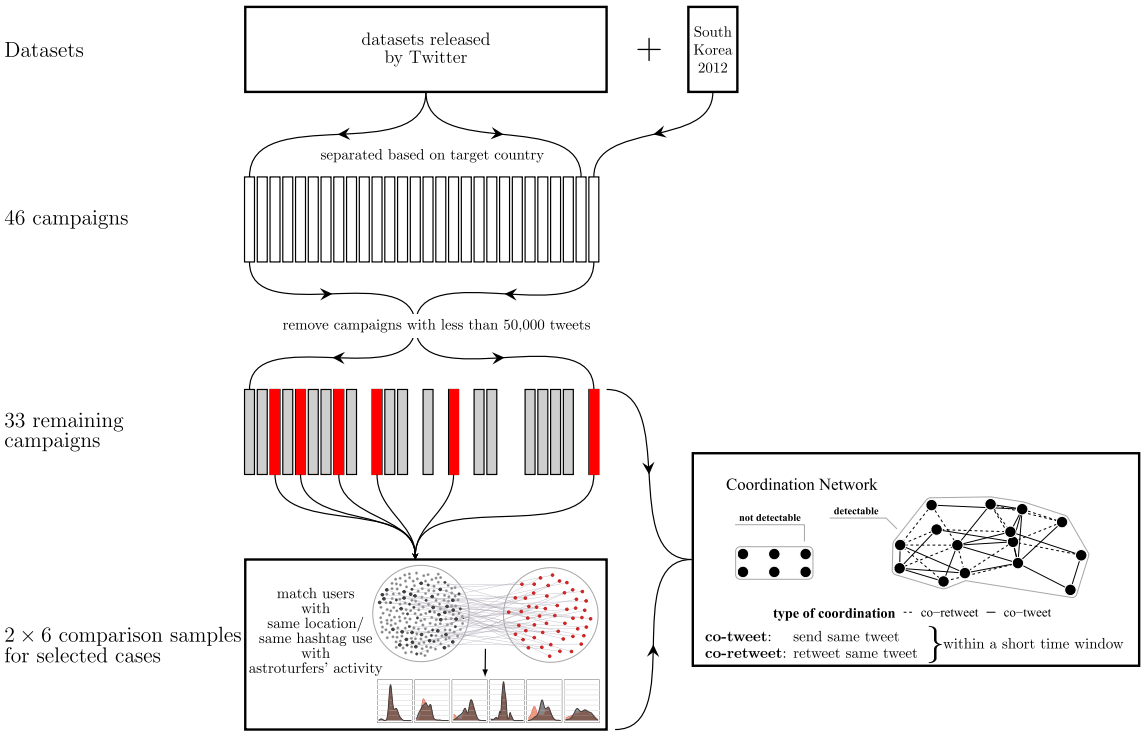

Figure 1. Research design. Datasets, case selection and analysis approach.

work for the politicians. These political elites' accounts are therefore unlikely to act like an account owned by an ordinary citizen. As a second baseline, they used a snowball sample of random users without any relation to the conversations in the target countries at the time the astroturfing campaigns were active.

The paper by Alizadeh et al. focusing on English-language astroturfing campaigns uses a subset of the Twitter data to build an algorithm that can detect other astroturfing accounts in later time periods, on other platforms, or later campaigns initiated by the same actor[31]. In other words: their approach requires a set of astroturfing accounts already identified using alternative methods. They compare the known astroturfing accounts to a random sample of regular and politically interested users. While the latter may seem like an appropriate comparison group, the authors define politically interested users as those users who follow at least three politicians – the sample therefore may contain bots designed to boost follower counts and does not consist of accounts that engage in the debates that the astroturfing campaign tries to influence.

We construct more natural comparison samples that reflect specific contexts of each astroturfing campaign, such as country and campaign-related topics and keywords. We randomly sampled users located in the targeted country that engage in the discussions that the astroturfing campaigns try to influence to compare their activity patterns with those of astroturfing. We also take the level of activities into consideration when we design comparison samples, such that the accounts in the astroturfing campaigns and the comparison groups are comparable. Taken together, our paper presents a scalable method for detecting astroturfing campaigns and validates the findings against the very accounts the campaign tries to mimic. This universal approach does not require any training data and performs well without human inputs when detecting groups of suspicious accounts in all previously revealed instances of astroturfing on Twitter. With that, the study contributes to methodological approaches for the detection of disinformation and reveals surprising similarities in astroturfing campaigns, even across heterogeneous political and social contexts.

Results

Journalistic investigations of astroturfing campaigns indicate that campaign participants are often hired and work in shifts (see, e.g. the New York Time's reporting on the Russian trolls[35]) or during regular office hours. This might result in activity patterns that are consistent across all campaigns, but distinct from those of regular users. Figure 2 shows that a large fraction of the astroturfing campaign tweets (top) are posted during regular office hours, and not in the evening, when regular users (bottom) tend to be more active. Astroturfing tweets are also less likely to be posted during the day(s) off in the target country. These results hold irrespective of whether the comparison sample are regular users based in the target country, or users sampled because they – in addition – are part of the issue public the campaign tries to infiltrate, i.e. they use the same hashtags as the astroturfing campaign in the specific time window (see Methods section and additional evidence in SI S4). The pattern also generalizes across all datasets released by Twitter: activity drops on Saturday and Sunday in the case of countries with a Christian cultural background, or Friday in the case of campaigns associated with Muslim-majority countries (e.g., Iran and the UAE, see SI S4).

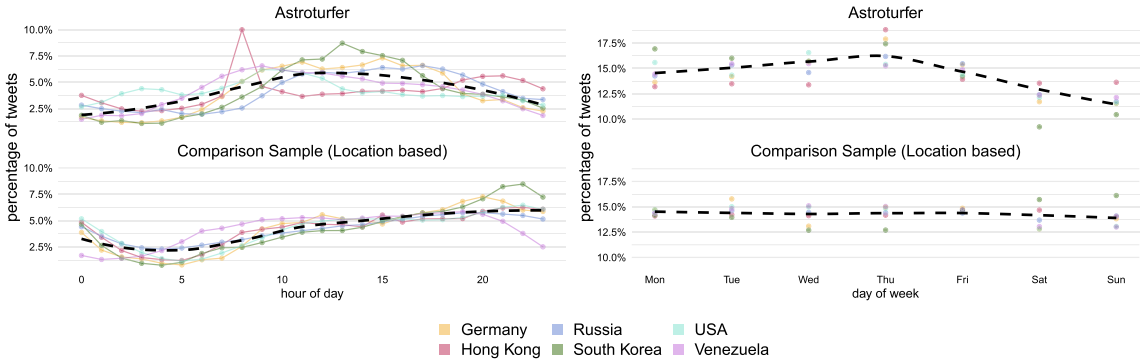

Figure 2. Astroturfing campaigns are most active during office hours and working days. Comparison of hourly (left) and weekly (right) activity of astroturfing campaigns and location-based case-specific random samples. Dashed line indicates the average activity.

Not every account that is predominantly active during business hours is an astroturfing account. More telling in that regard are patterns indicating that a synchronous coordination is taking place. To descriptively chart traces of coordination, we use the heat maps of their daily activity. Just by ordering all accounts according to their level of activity (i.e., how many tweets they posted) in SI S5, it becomes clear that all astroturfing campaigns contain groups of accounts that increase or cease their activity at similar time points. This is another pattern related to astroturfing campaigns' central coordination, in which participants receive simultaneous instructions to increase their activity or create new accounts, or due to shirking, i.e., the agents' usage of desktop applications that allow for posting from several accounts simultaneously.

These patterns can be broken down to the level of individual tweets by creating networks based on message coordination, specifically if accounts tweet or retweet the exact same message within a short time window. We use such instances of "co-tweeting" and "co-retweeting" to link two accounts and create networks among the accounts examined (see section Methods for details). To illustrate the approach, Fig. 3 shows the combined co-tweet and co-retweet network for a more complex campaign that was targeting heterogeneous audiences. Even accounts belonging to the infamous IRA campaign in the U.S. that targeted both a right-wing (Trump supporters) and a left-wing (Black Lives Matter) audience[29] still formed one large network component. A closer inspection of the content shared by both sides indicates that accounts pretending to be BLM activists and those posing as black gun supporters for instance found common grounds around the hashtag #OscarsSoWhite.

Figure 4 displays the combined co-tweet and co-retweet network for the IRA campaign in the U.S. (b), and for its baseline group of comparison users (c), using our preferred temporal threshold, one minute. In that particular case and based on that particular network or decision rule alone, we would detect more than 80 per cent of the astroturfing accounts, while the number of false positives hovers around 1 per cent (d). As the Figs. in SI S6 show, the accounts involved form large connected network components in every single campaign, while regular accounts form no or rather small network components. Finally, the bar charts in panel (a) at the bottom of Fig. 4 demonstrate that the differences in message coordination between the astroturfing accounts and regular users remain large, irrespective of the time window used.

While there are at least three ways in which accounts could coordinate their messages – co-tweeting, co-retweeting and retweeting – we find that a combination of the first two most reliably detects centrally coordinated campaigns. Detecting accounts based on retweeting results in the highest number of false positives across the three measures, in particular in the Venezuelan random samples (see the full network Fig. in SI S6). This is not surprising, as retweeting is the most decentralized form of coordination: individual accounts can make an independent decision to share an interesting tweet and in that process participate in an issue public. While to a lesser degree this is also true for co-retweeting, repeated co-tweeting still seems implausible in the absence of (centralized) coordination in which accounts are being told what to tweet about or in which one actor controls multiple accounts. Another reason why it is preferable to use co-tweeting and co-retweeting is that detection based solely on retweeting requires prior knowledge about at least some campaign participants, which renders the method impractical for the real-time detection of astroturfing. Therefore, for summarizing the performance of our methodology, we rely on the co-tweet and co-retweet networks which can universally be constructed from any Twitter dataset.

Figure 5 shows the percentage of detected astroturfing accounts for different time windows during which a pair of accounts either co-tweeted or co-retweeted together. That means an account is considered "detected" if it is not an isolate in either of the two networks. Just using the two metrics and a threshold of one minute allows for the identification of a high share of astroturfing accounts, while the false positive rate, i.e., the number of falsely flagged regular users, only increases when significantly relaxing the temporal threshold. The most reasonable threshold seems to be below 8 hours, i.e., less than the duration of a workday. This pattern is in line with media and insider reports suggesting that agents in astroturfing campaigns often receive instructions on a daily basis.

To add an organic grassroots campaign as an additional baseline, we collected data from a recent German (COVID-19 vaccination) information campaign, #allesindenArm as yet another benchmark sample. Under the hashtag #allesindenArm (roughly translated as "get a jab") influencers and regular social media users posted personal messages to motivate other users to get vaccinated. There is no indication that this campaign was

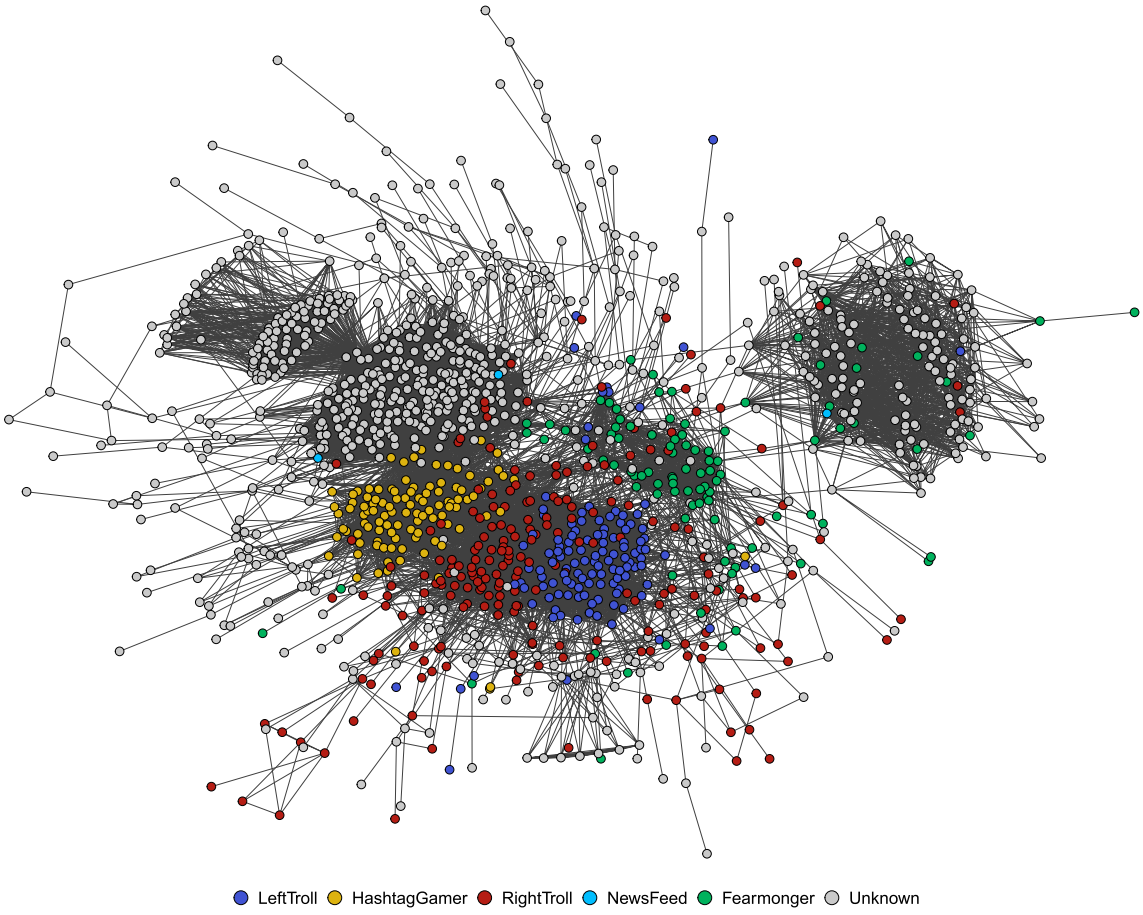

Figure 3. Astroturfing campaigns that target diverse audiences (here: Black Lives Matter activists in blue and Trump Republicans in red) are still connected in a message coordination network. Combined co-tweet and co-retweet network among IRA accounts that tweet in English. Node color is taken from a typology of Linvill and Warren, who manually classified IRA accounts according to their content[29].

supported by astroturfing and we assess it as a genuine grassroots campaign. Despite some (open) coordination among participating users, the results in SI S7 show that the level of message coordination via co-retweeting and cotweeting is roughly equivalent to the random samples and therefore clearly distinguishable from the level of coordination in an astroturfing campaign. We also show that the temporal patterns of this genuine campaign (daily and hourly frequency of tweeting) resemble that of the accounts in the random samples.

After validating the approach against comparison user baselines, we finally apply the methodology to all astroturfing campaigns published by Twitter. Figure 6 shows that on average, 74% of astroturfing accounts were detected. The variation across campaigns cannot easily be explained by a clear geographical or geopolitical pattern. This indicates that specific campaigns differ in how well they navigate principal–agent problems.

Conclusion

This paper presented the most comprehensive investigation of political astroturfing campaigns across the globe. Our analysis spanned various political and cultural contexts while covering multiple continents and time periods. Despite this heterogeneity, we found remarkably similar patterns in all astroturfing campaigns using the detection method by Keller et al.[10]. The findings suggest that astroturfing exhibits universal features that could help researchers, social media companies, and citizens to identify these types of disinformation. Our theoretical framework relates this to principal–agent theory: unlike the participants of genuine grassroots movements, astroturfing agents are not intrinsically motivated. Therefore many agents involved in astroturfing campaigns invest little time in creating distinctive online personas or varying their behavior across the accounts they control. Such patterns are difficult to camouflage, because message coordination is inherent to any information campaign, and resources to mitigate principal–agent problems are usually limited.

In contrast to much previous literature, our methodological approach detects astroturfing campaigns based on the patterns of coordinated group efforts to produce messages instead of focusing on individual account features, such as heavy automation. By relying on a parsimonious set of metrics, we also provide a transparent methodology that can be universally applied. We argue that this is an improvement over machine learning classifiers that might achieve a better performance as they are typically trained on a large number of case-specific features, but fail when trying to identify out-of-sample accounts[34]. For six cases, we constructed comparison

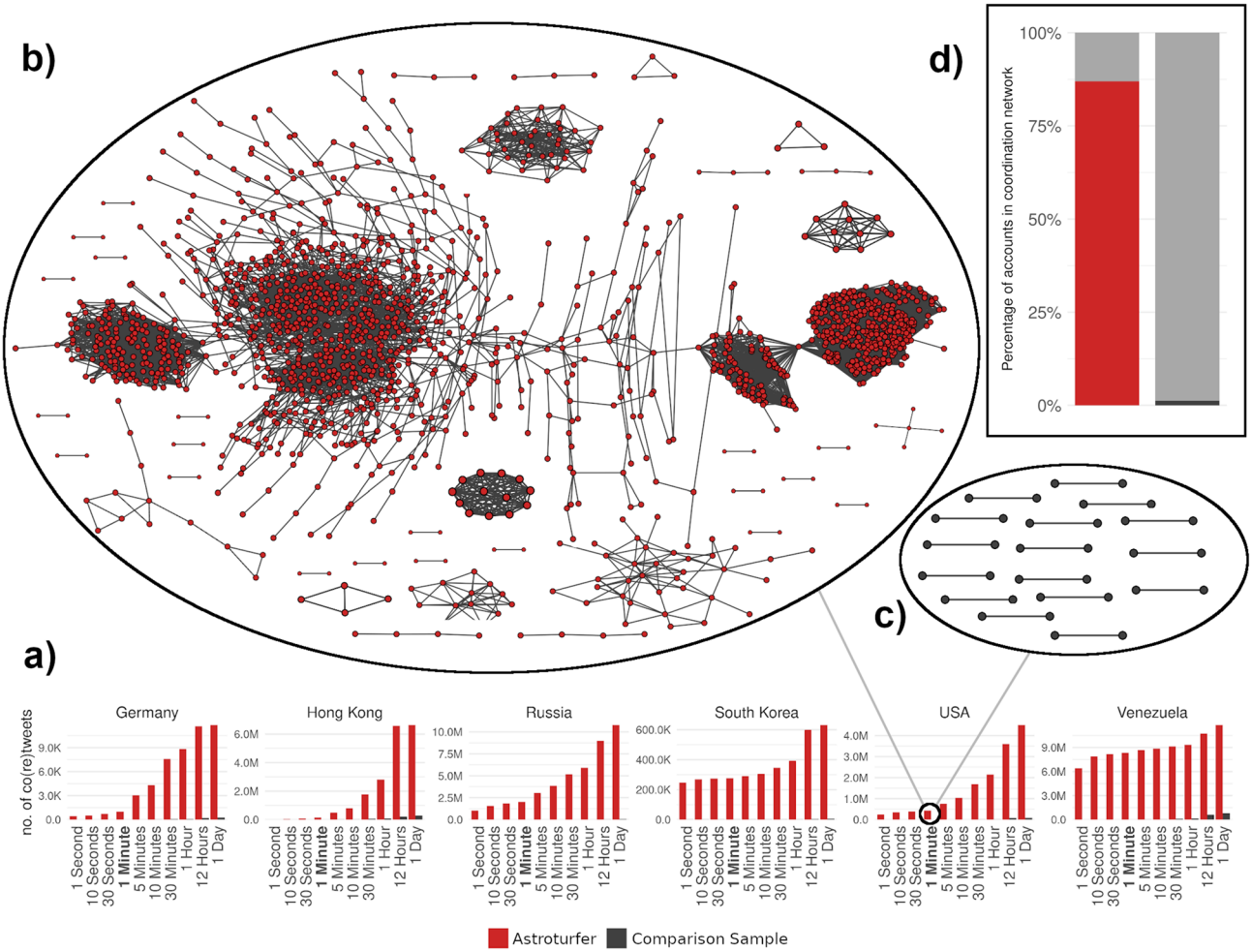

Figure 4. Message coordination in astroturfing campaigns (red) is more common and involves more accounts than in a comparison group of regular Twitter users (black). (**a**) Comparison of the number of co-(re)tweets among accounts with varying temporal threshold (i.e. how far apart two tweets are allowed to be in order to still be considered a co-(re)tweet. (**b**) Co-(re)tweet network of the U.S. campaign. Two accounts are connected if they (re)tweeted the same content within one minute. (**c**) Same as b) but for an equally sized set of comparison users located in the U.S. (**d**) Percentage of accounts (astroturfing or comparison group) appearing in the networks shown in (**b**) and (**c**). Missing accounts to 100% are isolated in the network, meaning that they do not co-(re)tweet with any other account in the respective sample.

groups of users that were engaging in salient political conversations in each of the target countries while the astroturfing campaigns were unfolding. It is important to underscore that our comparison groups were based on an account-level data collection aimed to find comparable accounts that produced tweets at similar rates with the astroturfing accounts. These comparisons consistently show that unique coordination patterns are evident only among astroturfers.

One limitation of our validation procedure is that Twitter might have used similar techniques for detecting astroturfers in the first place. To avoid circumvention, Twitter does not reveal the exact methodology used to detect the campaigns it removes from the platform. In statements accompanying the data releases, Twitter mentions information provided by industry partners and even law enforcement, as well as account-based features only the company has access to, such as IP addresses. We also note that data on the Iranian campaign was released in batches several months apart, even though the accounts contained in it engaged in co-tweeting across those batches. This leads us to believe that Twitter may (at least at that time) not have relied on too similar an approach as we propose. To safeguard against this potential endogeneity, our analysis also relied on data from a South Korean astroturfing campaign, where account names were directly retrieved from laptops of secret service agents.

The major advantage of the methodology is that it can be applied to almost any social media platform – as long as the platform allows users to post content (and therefore two accounts to post the same content within a certain time window) and to share content posted by other users (and therefore allow two accounts to simultaneously share it). Although this study did not analyze data from other social media platforms due to various constraints in data acquisition, it is crucial that researchers investigate which patterns universally apply to astroturfing across multiple social media platforms to further advance the findings of our study.

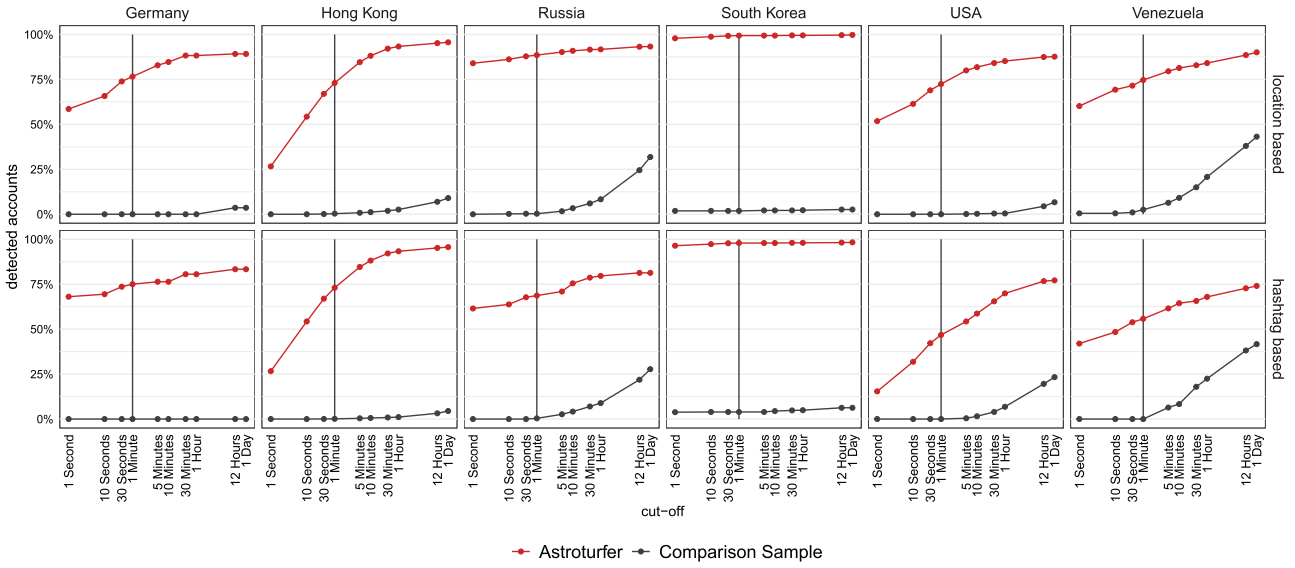

Figure 5. Astoturfing accounts are much more likely to coordinate their activity than comparison users for any given threshold. Astroturfing accounts (red) and comparison user accounts (black) classified as astroturfing accounts, location-based (top row) and hashtag-based (bottom row). Based on appearance in either the co-tweet or co-retweet network, depending on different temporal thresholds.

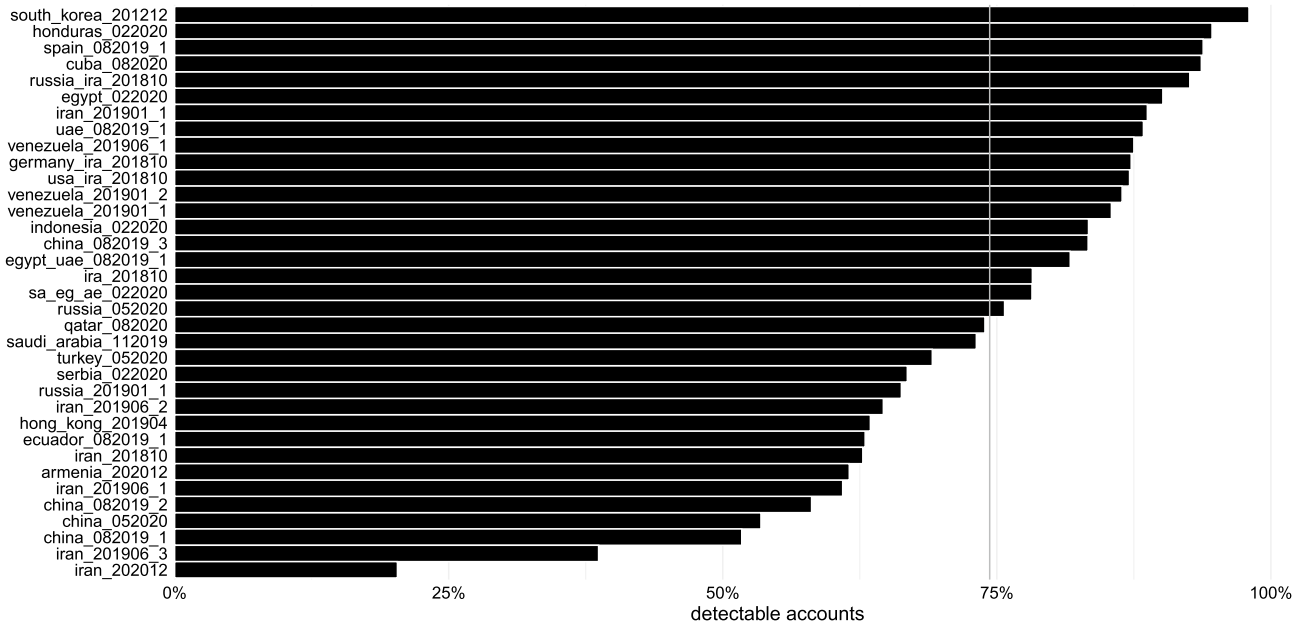

Figure 6. The methodology detects on average 74% of all accounts involved in a campaign. Detected astroturfing accounts based on appearance in either the co-tweet or co-retweet network (one-minute threshold) for all campaigns released by Twitter with more than 50,000 tweets. Campaign labels reflect information provided by Twitter in their data releases. Grey line indicates the mean.

Methods

The study was approved by the Human Participants Research Panel at the Hong Kong University of Science and Technology (G-HKUST601/19, HPR #382).

Data. Starting in October 2018, Twitter has released data sets of tweets by accounts that it deemed to have been involved in hidden information campaigns. These campaigns occurred across different continents over the last decade, targeted different audiences domestic and abroad, and were conducted in different languages (the full overview is in SI S2). Since constructing meaningful baselines is resource-consuming, we chose cases to cover a wide range of campaigns, after excluding campaigns with less than 50,000 tweets.

Some campaigns targeted one language community only, while others targeted different language communities in different countries, such as the Iranian or the IRA campaign. The campaigns also had, judging from

journalistic accounts thereof, different goals: the IRA campaign in the USA and Germany targeted at least two different communities on opposite ends of the political spectrum, while the campaign targeting its own country had the single focus of applauding the government. The Chinese campaign targeting Hong Kong had a similarly unified message: to vilify the protesters and some other perceived enemies of the PRC. For our analysis, we therefore selected the IRA campaign and split it into three different campaigns based on the three most prevalent languages: one targeting the US and its election, one Russia itself, and one targeting Germany. There are also multi-lingual campaigns that target a polyglot audience: we picked the example of the Chinese campaign influencing the Mandarin/Cantonese/English-speaking audience abroad and in Hong Kong. Among the mono-lingual campaigns, we selected the Venezuelan campaign to also include South America in our global coverage. We also examine a set of astroturfing accounts that was not revealed by Twitter, but in South Korean court documents[10].

While our detailed analysis requires comparison data from a systematically constructed set of random users, we show in SI S4 that the descriptive temporal patterns are similar in all campaigns revealed by Twitter.

Network analysis. We build on a methodology for the detection of astroturfing that exploits the principal–agent problems of such campaigns[10]. The behavioral patterns caused by strategic coordination are difficult to mask; therefore, they can be used as general indicators to find the agents associated with a broader disinformation campaign.

We distinguish three different measures of coordination patterns[10]. The most well-known form of coordination is *retweeting*: a large proportion of the retweets posted by astroturfing accounts tends to come from other campaign accounts. But as retweeting is also common among genuine grassroots campaigns, this measure creates the largest number of false positives, misidentifying potentially genuine grassroots movements as astroturfing, or including genuine converts of the campaign. A second form of coordination is *co-tweeting*, the act of two accounts posting the same message within a short (here, one minute) time window. This type of coordination is most likely to distinguish regular users from astroturfing accounts, as the former rarely post the same original message at the same time. But astroturfing accounts that exclusively amplify messages via retweeting will not get captured with this method. In order to capture this common type of behavior, we use *co-retweeting networks*: when two accounts retweet the same message within a one minute window, we construct a tie between them – but only if this occurs at least ten times. This behavior should be particularly widespread in campaigns that focus on boosting the visibility of their own and third-party accounts with little effort in creating original content.

We chose thresholds and time windows based on previous research where it became apparent that these can be varied without changing the main results[10]. Robustness tests with different temporal thresholds can be found in SI S5. In all campaigns, we prune the set of accounts to only include accounts that tweet more than ten times in the whole time period.

Sampling user accounts for comparison. Finding the right criteria for a baseline to compare the campaign accounts to is not easy. Our comparison groups are two conditional random samples of Twitter users designed to look as similar to the group of users that the astroturfing campaign tries to mimic, making the distinction between astroturfing and regular users as challenging as possible. A random sample of all Twitter users, is likely to yield trivial results: a group of accounts from different parts of the world tweeting in different languages and different time zones is unlikely to tweet many similar messages at the same time. In addition, Twitter activity is distinctly right-skewed, and a random selection of accounts will yield a large number of inactive or barely active accounts. Accounts that do not tweet of course also do not coordinate their messages.

We therefore argue that the most appropriate comparison group is a set of accounts sampled from the users that the astroturfing campaign tries to emulate, stratified by level of activity. We collected our data from the commercial social media monitoring company Brandwatch. Brandwatch infers the country of a Twitter account based on geo-coordinates of tweets, the self-reported location information in user profiles, the time zone defined by users, top-level domains and geo-IPs of the links shared by a user as well as the language in tweets. Using geo-located tweets as benchmark, the company developed algorithms able to assign 90% of non-geolocated tweets to a country with that procedure. We start by drawing a random sample of tweets posted by accounts that Brandwatch's algorithm locates in the targeted country (i.e., in the case of the IRA's 2016 U.S. Presidential election campaign, users located in the U.S.). We then collected a random sample of the accounts appearing in this sample, collected all their tweets in the relevant time period and calculated their activity level (i.e., number of tweets posted). We matched each astroturfing account by activity level with an account from this random sample (see SI S2 for an evaluation of the matching procedure). We call this comparison sample the *location-based comparison sample*.

However, this procedure may still yield a group of random accounts so dissimilar that the exercise of distinguishing them from the campaign accounts is trivial. They could, for instance, be similarly active sports or pop music fans. This is why Alizadeh et al. use politically interested users as a comparison group, and identify those users by whether they follow known political figures[31]. But such politically interested users can hold a variety of political views and participate in many different debates. We therefore argue that a more rigorous test for any detection method is a sampling based on members of "issue publics": for each campaign, we identify at least one period in which it used a handful of hashtags with increased frequency for at least a few days. We take this as evidence that the campaign tried to infiltrate a specific debate and the community forming around it. We then collect a random sample of tweets also mentioning these hashtags during the same time frame – in addition to also being based in the appropriate geographic location – and select random users from that sample to match the astroturfing accounts' activity. We call this sample the *hashtag-based comparison sample*. This is a more appropriate comparison group than communities already existing within an institutional setting, such as members of the U.S. Congress, the British Parliament or academics[34]. After all, an astroturfing campaign does not pretend

to be a well-established community, but regular users who happen to suddenly take interest in the same topic. SI S3 shows that the matching by activity was successful for each case and provides an overview of the samples.

Data availability

Data files necessary to replicate the results in this article are available on OSF: https://osf.io/ms5ue/?view_only=b020f97d49fc41b893391b0aef1bbfba.

Code availability

R scripts that replicate the results in this article are available on OSF: https://osf.io/ms5ue/?view_only=b020f97d49fc41b893391b0aef1bbfba.

Received: 13 April 2021; Accepted: 25 February 2022
Published online: 17 March 2022

References

1. Bail, C. A. *et al.* Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late (2017). *Proc. Natl. Acad. Sci.* **117**(1), 243–250. https://doi.org/10.1073/pnas.1906420116 (2020).
2. Jay, C. K. How the Pro-Beijing media influences voters: Evidence from a field experiment. (2020). URL https://www.jaykao.com/uploads/8/0/4/1/80414216/pro-beijing_media_experiment_kao.pdf.
3. Loomba, Sahil, de Figueiredo, Alexandre, Piatek, Simon J., de Graaf, Kristen & Larson, Heidi J. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nat. Hum. Behav.* **5**(3), 337–348. https://doi.org/10.1038/s41562-021-01056-1 (2021).
4. Allen, Jennifer, Howland, Baird, Mobius, Markus, Rothschild, David & Watts, Duncan J. Evaluating the fake news problem at the scale of the information ecosystem. *Sci. Adv.* **6**(14), eaay3539 (2020).
5. Guess, Andrew M., Nyhan, Brendan & Reifler, Jason. Exposure to untrustworthy websites in the 2016 US election. *Nat. Hum. Behav.* **4**, 472–480. https://doi.org/10.1038/s41562-020-0833-x (2020).
6. Gallotti, Riccardo, Valle, Francesco, Castaldo, Nicola, Sacco, Pierluigi & De Domenico, Manlio. Assessing the risks of 'infodemics' in response to COVID-19 epidemics. *Nat. Hum. Behav.* **4**(12), 1285–1293. https://doi.org/10.1038/s41562-020-00994-6 (2020).
7. Ferrara, Emilio, Varol, Onur, Davis, Clayton, Menczer, Filippo & Flammini, Alessandro. The rise of social bots. *Commun. ACM* **59**(7), 96–104. https://doi.org/10.1145/2818717 (2016).
8. Stella, Massimo, Ferrara, Emilio & De Domenico, Manlio. Bots increase exposure to negative and inflammatory content in online social systems. *Proc. Natl. Acad. Sci.* **115**(49), 12435–12440. https://doi.org/10.1073/pnas.1803470115 (2018).
9. Stukal, Denis, Sanovich, Sergey, Bonneau, Richard & Tucker, Joshua A. Detecting bots on Russian political Twitter. *Big Data* **5**(4), 310–324 (2017).
10. Keller, Franziska B., Schoch, David, Stier, Sebastian & Yang, JungHwan. Political astroturfing on Twitter: How to coordinate a disinformation campaign. *Polit. Commun.* **37**(2), 256–280. https://doi.org/10.1080/10584609.2019.1661888 (2020).
11. Rauchfleisch, A. & Kaiser, J. The false positive problem of automatic bot detection in social science research. *PLOS One* **15**(10), e0241045 (2020).
12. Kovic, M., Rauchfleisch, A., Sele, M. & Caspar, C. Digital astroturfing in politics: Definition, typology, and countermeasures. *Stud. Commun. Sci.* **18** (1)https://doi.org/10.24434/j.scoms.2020.02.005 (2018).
13. Lazer, David MJ. *et al.* The science of fake news. *Science* **359**(6380), 1094–1096. https://doi.org/10.1126/science.aao2998 (2018).
14. Benkler, Y. Faris, R. & Roberts, H. *Manipulation, Disinformation, and Radicalization in American Politics*. Oxford University Press, Network propaganda, (2018).
15. Freelon, Deen & Wells, Chris. Disinformation as political communication. *Polit. Commun.* **37**(2), 145–156. https://doi.org/10.1080/10584609.2020.1723755 (2020).
16. Krafft, P. M. & Donovan, Joan. Disinformation by design: The use of evidence collages and platform filtering in a media manipulation campaign. *Polit. Commun.* **37**(2), 194–214. https://doi.org/10.1080/10584609.2019.1686094 (2020).
17. Wardle, C. & Derakhshan, H. *Information Disorder: Toward an Interdisciplinary Framework for Research and Policymaking*, (2017). URL https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c.
18. Wilson, T. & Starbird, K. Cross-platform disinformation campaigns: Lessons learned and next steps. *Harv. Kennedy School Misinf. Rev.*, **1**(1), (2020).
19. Keller, F., Schoch, D., Stier, S. & Yang, J.. How to manipulate social media: Analyzing political astroturfing using ground truth data from South Korea. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*, pp. 564–567, (The AAAI Press, 2017) .
20. Kargar, Simin & Rauchfleisch, Adrian. State-aligned trolling in Iran and the double-edged affordances of Instagram. *New Med. Soc.* **21**(7), 1506–1527. https://doi.org/10.1177/1461444818825133 (2019).
21. Grimme, C., Assenmacher, D. & Adam, L. Changing perspectives: Is it sufficient to detect social bots? In *International Conference on Social Computing and Social Media*, pp. 445–461. Springer, (2018).
22. Miller, Gary J. The political evolution of principal–agent models. *Annu. Rev. Polit. Sci.* **8**, 203–225 (2005).
23. Enos, Ryan D. & Hersh, Eitan D. Party activists as campaign advertisers: The ground campaign as a principal–agent problem. *Am. Polit. Sci. Rev.* **109**(2), 252–278 (2015).
24. Twitter. Information Operations. Data archive, (2021). URL https://transparency.twitter.com/en/reports/information-operations.html.
25. King, Gary, Pan, Jennifer & Roberts, Margaret E. 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 (2017).
26. Barrie, C. & Siegel, A. Kingdom of trolls? Influence operations in the Saudi Twittersphere. *J. Quant. Descr.: Digit. Med.*https://doi.org/10.51685/jqd.2021.012 *(2021).*
27. Badawy, A., Ferrara, E. & Lerman, K. Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 258–265, (2018). https://doi.org/10.1109/ASONAM.2018.8508646.
28. Bail, Christopher A. *et al.* Exposure to opposing views on social media can increase political polarization. *Proc. Natl. Acad. Sci.* **115**(37), 9216–9221. https://doi.org/10.1073/pnas.1804840115 (2018).
29. Linvill, Darren L. & Warren, Patrick L. Troll factories: Manufacturing specialized disinformation on Twitter. *Polit. Commun.* **37**(4), 447–467. https://doi.org/10.1080/10584609.2020.1718257 (2020).
30. Lukito, Josephine *et al.* The wolves in sheep's clothing: How Russia's internet research agency tweets appeared in U.S. news as vox populi. *Int. J. Press/Polit.* **25**(2), 196–216. https://doi.org/10.1177/1940161219895215 (2020).
31. Alizadeh, M., Shapiro, J. N., Buntain, C. & Tucker, J. A. Content-based features predict social media influence operations. *Sci. Adv.*https://doi.org/10.1126/sciadv.abb5824 *(2020).*

32. Guarino, S., Trino, N., Celestini, A., Chessa, A. & Riotta, G. Characterizing networks of propaganda on Twitter: A case study. *Appl. Netw. Sci.*https://doi.org/10.1007/s41109-020-00286-y *(2020)*.
33. Gurajala, S., White, J. S., Hudson, B. & Matthews, J. N. Fake Twitter accounts: Profile characteristics obtained using an activity-based pattern detection approach. In *Proceedings of the 2015 International Conference on Social Media & Society - SMSociety '15*, pp. 1–7, (ACM Press, 2015). https://doi.org/10.1145/2789187.2789206.
34. Vargas, L., Emami, P. & Traynor, P. On the detection of disinformation campaign activity with network analysis, (2020). URL https://arxiv.org/abs/2005.13466.
35. Chen, A. The Agency. From a nondescript office building in St. Petersburg, Russia, an army of well-paid "trolls" has tried to wreak havoc all around the internet — and in real-life American communities., (2015). https://www.nytimes.com/2015/06/07/magazine/the-agency.html.

Acknowledgements
We thank participants at the Harvard Disinformation Workshop 2019 and the German International Relations Conference 2020 for helpful comments. We are grateful to Curt Donelson at The Social Media Macroscope at the University of Illinois at Urbana-Champaign for providing access to Twitter data and research assistants Matthew Pitchford and Wallace Golding for their contributions to the data collection.

Author contributions
D.S. analyzed data, D.S. and J.Y collected data, D.S., F.K., S.ST., and J.Y. designed research, performed research, and wrote the paper.

Funding
Open Access funding enabled and organized by Projekt DEAL.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-08404-9.

Correspondence and requests for materials should be addressed to D.S.

Reprints and permissions information is available at www.nature.com/reprints.

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