Political astroturfing across the world

Franziska B. Keller[*] David Schoch[†] Sebastian Stier[‡] JungHwan Yang[§]

Paper prepared for the Harvard Disinformation Workshop Update

1 Introduction

At the very latest since the Russian Internet Research Agency's (IRA) intervention in the U.S. presidential election, scholars and the broader public have become wary of coordinated disinformation campaigns. These hidden activities aim to deteriorate the public's trust in electoral institutions or the government's legitimacy, and can exacerbate political polarization. But unfortunately, academic and public debates on the topic are haunted by conceptual ambiguities, and rely on few memorable examples, epitomized by the often cited "social bots" who are accused of having tried to influence public opinion in various contemporary political events.

In this paper, we examine "political astroturfing," a particular form of centrally coordinated disinformation campaigns in which participants pretend to be ordinary citizens acting independently (Kovic, Rauchfleisch, Sele, & Caspar, 2018). While the accounts associated with a disinformation campaign may not necessarily spread incorrect information, they deceive the audience about the identity and motivation of the person manning the account. And even though social bots have been in the focus of most academic research (Ferrara, Varol, Davis, Menczer, & Flammini, 2016; Stella, Ferrara, & De Domenico, 2018), seemingly automated accounts make up only a small part – if any – of most astroturfing campaigns. For instigators of astroturfing, relying exclusively on social bots is not a promising strategy, as humans are good at detecting low-quality information (Pennycook & Rand, 2019). On the other hand, many bots detected by state-of-the-art social bot detection methods are not part of an astroturfing campaign, but unrelated non-political spam

[*] Hong Kong University of Science and Technology

[†] The University of Manchester

[‡] GESIS – Leibniz Institute for the Social Sciences

[§] University of Illinois at Urbana-Champaign

bots. In addition, the frequent unsystematic cross-references to the problem of so-called "fake news" (Lazer et al., 2018) exacerbates the conceptual confusion and further muddles the water in the popular discourse on the topic. In this study, we therefore propose to define the term "astroturfing campaigns" narrowly, and distinguish it from other elements of disinformation campaigns such as spam bots or "fake news." We argue that these campaigns can be detected by searching for coordination patterns among groups of accounts, instead of looking at suspicious activity patterns of individual accounts.

In order to validate methods for the detection of astroturfing that we presented in an earlier study (Keller, Schoch, Stier, & Yang, 2019), we cannot rely on manual inspection, which may still be plausible for detecting social bots. We instead rely on rare "ground truth" information that unambiguously identifies accounts involved. The few studies that have taken a similar approach have only examined individual cases, such as China (King, Pan, & Roberts, 2017), South Korea (Keller, Schoch, Stier, & Yang, 2017; Keller et al., 2019) or the Russian influence campaign in the U.S. (Badawy, Ferrara, & Lerman, 2018; Linvill & Warren, 2018; Lukito et al., 2018). This is thus the first comparative study of multiple campaigns in different countries.

We use the data released by Twitter as part of its "Election Integrity Hub" initiative (Twitter, 2019) as "ground truth" data. In this paper, we focus on disinformation campaigns associated with ten different countries: the Russian Internet Research Agency's (IRA) attempt at changing public opinion in the U.S., Germany, and Russia, the South Korean secret services' attempt at influencing the elections in South Korea, the Chinese government's attempt at changing the framing of the Hong Kong protest and other events abroad, as well as campaigns allegedly associated with the governments of Ecuador, Iran, Spain, the UAE and Venezuela.[1] To contribute to the ongoing discussion, we present a scalable detection method for astroturfing, and reveal specific traces of such campaigns across different political and social contexts.

2 Literature Review

Most related studies focus on so-called social bots, i.e., automated accounts (Chu, Gianvecchio, Wang, & Jajodia, 2012; Ferrara et al., 2016; Stella et al., 2018; Stukal, Sanovich, Bonneau, & Tucker, 2017), and therefore deal with a related, but different phenomenon, as mentioned above. In addition, the bot detection approaches have been criticized as fundamentally flawed; a bot detection algorithm classifies, for instance, almost half of the members of the U.S. Congress as bots (Kreil, 2018). We would thus argue that the concep-

[1] Twitter has also released data on disinformation campaigns associated with Bangladesh, Saudi Arabia and Catalonia, but these consist only of a handful of accounts, and were therefore excluded from the analysis.

tual mismatch has resulted in the predominant research focusing on “bot detection”, which caused researchers to search for individual accounts with pronounced “robotic” behavior, instead of attempting to detect suspicious groups of accounts that behave in a coordinated manner.

In fact, the timing and form of coordination might be the one unique feature that distinguishes astroturfing from real grassroots engagement, while also capturing social bots involved in astroturfing campaigns. Recent studies have found coordination patterns in accounts that were harassing members of the Iranian diaspora on Instagram (Kargar & Rauchfleisch, 2019) or hijacking German Twitter debates during an election campaign (Grimme, Assenmacher, & Adam, 2018). In our previous study on the South Korean secret service’s campaign to influence the 2012 Presidential election (Keller et al., 2017, 2019), we have shown how the central coordination and organizational routines inherent to any information campaign lead to distinct patterns of message coordination in astroturfing campaigns. Using “ground truth” data – Twitter account names found on confiscated laptops and published in the court proceedings as evidence – we demonstrated that these patterns can reliably distinguish between the participants of an astroturfing campaign and ordinary Twitter users.

These patterns can be explained theoretically by referring to the principal-agent framework. The organizers of the campaign are a *principal* who tries to pursue (political) goals by incentivizing and instructing *agents* to create and share messages congruent with the principal’s goals. The purpose of astroturfing is to reach and persuade as many *regular users* as possible, which is contingent on a wide reach and “organic” appearance of the campaign. Principal-agent theory predicts that extrinsically motivated agents will try to shirk the hard work of coming up with original contributions to the campaign. This is particularly likely in the common situation where they control multiple accounts at the same time. Unless the principal can establish an expensive system of close monitoring, the accounts controlled by the same individual will have very similar activity patterns. In addition, information campaigns by definition require participants to send out similar messages at a similar times. Agents react to centralized instructions by the principal, such as – in our Korean case – daily briefings (SeoulHigherCourt, 2015). Finally, another expectation derived from the principal-agent theory is that agents will only extend effort when they are supervised. In the South Korean case, we indeed found that astroturfing accounts and regular users have different daily, weekly and hourly activity patterns. In this paper, we move beyond this single case and explore whether these are universal patterns in astroturfing campaigns across the world.

Table 1: Statistics of all campaigns including number of tweets and accounts involved, percentage of retweets, tweets containing hashtags and URLs, and detectable accounts using co-(re)tweeting.

			Tweets			% Detectable
Campaign	Tweets	Accounts	% RT	% #	% URLs	accounts
China	13,856,454	5,241	8.49	16.92	20.92	74.11
Ecuador	700,240	1,019	3.84	30.80	11.77	55.74
Germany (IRA)	102,657	111	12.23	22.94	82.52	76.58
Iran	5,569,992	3,081	47.18	42.63	37.15	51.95
Russia (IRA)	3,953,675	1,039	40.28	18.03	59.02	86.24
South Korea	194,190	1,002	48.10	4.74	34.30	79.40
Spain	56,712	259	15.57	59.10	9.08	67.57
UAE	1,540,428	4,519	4.43	56.88	26.97	72.29
USA (IRA)	4,606,393	2,382	37.15	37.60	40.75	68.43
Venezuela	9,946,768	1,951	20.74	41.75	70.12	61.92

3 Research Design

3.1 Data

Starting in October 2018, Twitter has released data sets of tweets by accounts that were involved in hidden information campaigns across different continents over the last decade. They target different audiences domestic and abroad, and are conducted in different languages. Building up on our previous research on South Korea, we expand our research to nine of these cases in addition to our existing study on South Korea.[2]

The data shows that the campaigns employed a variety of strategies (Table 1). Some campaigns relied mostly on retweeting, such as the South Korean, the Iranian and the IRA campaigns. Others, such as Spain and the UAE, made extensive use of hashtags in an attempt to get them to trend. Campaigns also differ in their propensity to link to outside material, i.e., in how often they share newspaper articles or links to social contents like YouTube videos.

3.2 Methods

In our previous work, we developed a methodology for the detection of astroturfing that exploits the principal-agent problems of such campaigns. The behavioral patterns caused

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

by strategic coordination are difficult to mask; thus, they can be used as indicators to find the agents associated with a broader disinformation campaign.

Our social network analysis relies on three different operationalizations of coordination patterns. The most well-known form of coordination is *retweeting*: as we have shown in the South Korean case, a large proportion of the retweets posted by astroturfing accounts tend to come from other campaign accounts. For identification purposes, we have chosen a threshold of 50% of all retweets shared by a given account coming from other campaign accounts. A second form of coordination is *co-tweeting*, the act of two accounts posting the same message within a short (i.e. 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. Finally, when two accounts retweet the same message within a one minute window, we construct a tie in between them to create the *co-retweeting network* of the astroturfing campaign. 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 the given thresholds and time windows for convenience sake, but have shown in earlier research that these can be varied without chaning the main results (Keller et al., 2019).

4 Results

The campaigns under study here display a set of patterns similar to those in the South Korean case, as the bottom half of Figure 1 shows: Most of the campaign tweets are posted during regular office hours, and not in the evening, when regular users tend to be more active (at least in South Korea, see Figure 2). Astroturfing tweets are also less likely to be posted during the day(s) off in the instigating country, i.e., Saturday and Sunday in the case of South Korea, Venezuela and the campaigns alleged to be sponsored by the Russian government (Germany, U.S., and Russia), or Friday in the case of campaigns associated with Muslim-majority countries (Iran and the UAE).

But of course, not every account that is mainly active during business hours is an astroturfing account. More telling in that regard are the networks based on message coordination. Figure 2 therefore displays one of these networks, the co-tweet network, for each campaign.

Indeed, as Table 1 indicates, a large proportion of the astroturfing accounts identified by Twitter engage in the practice of sending out suspiciously coordinated messages through co-tweeting, retweeting the same third party account within a short time frame, or by simply retweeting other campaign accounts. The column *% Detectable accounts* presents the percentage of the campaign accounts that engage in at least one form of mes-

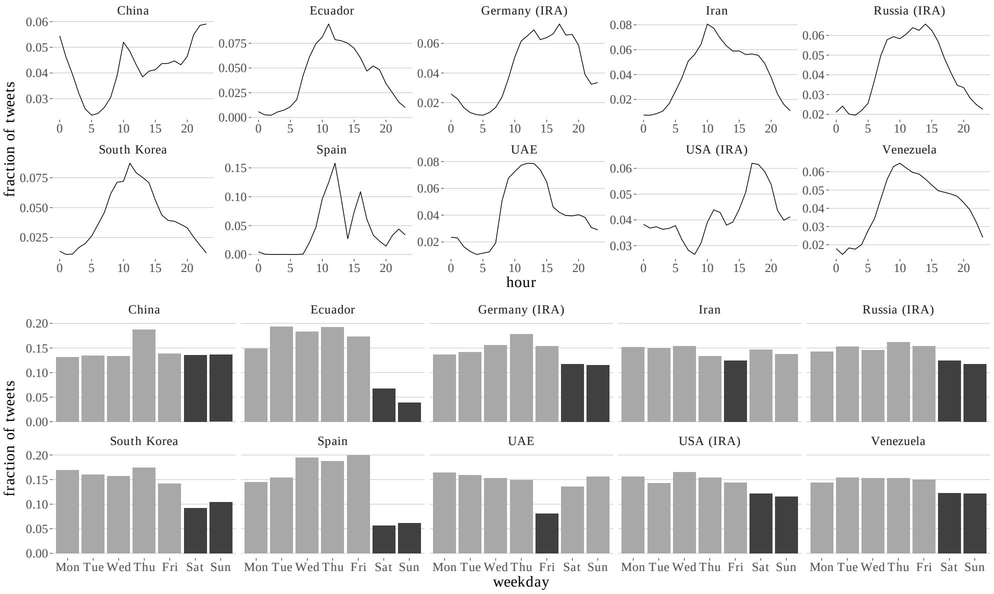

Figure 1: Top: Fraction of tweets per hour (time zone according to instigating country). Bottom: Fraction of tweets per weekday (time zone according to instigating country). Black bars emphasize weekends and Fridays for Muslim countries.

sage coordination behavior. That figure is consistently above 50%, and reaches almost 80% in several cases. With our simple method, relying on only these three message coordination feature, one would thus be able to detect a majority of the accounts involved in each campaign.

But in order to find out to what extent the coordinated behavior of astroturfing agents differs from "normal" behavior found on Twitter, we need a baseline measure to compare. Having no systematic baseline will make researchers draw conclusions based on descriptive findings relying on few memorable exemplars. For this purpose, we constructed an appropriate sample of regular users that were as similarly active as astroturfing accounts. So far, we have only collected such samples for South Korea and Germany.

For both countries, we did find that astroturfing leaves distinct temporal patterns. In South Korean case, most notable is the sudden inactivity of campaign accounts after December 11th, when the secret astroturfing campaign was publicly revealed (see Figure 3). But astroturfers also had a unique profile when aggregating the number of posts at hourly and weekday levels, as discussed above. Even more tellingly, we could show that a comparable set of regular users does, on average, not post a single co-tweet or co-retweet, and is significantly less likely to retweet each other (Keller et al., 2019). In German case, we picked random Twitter accounts that exhibit similar levels of activities of astroturfing

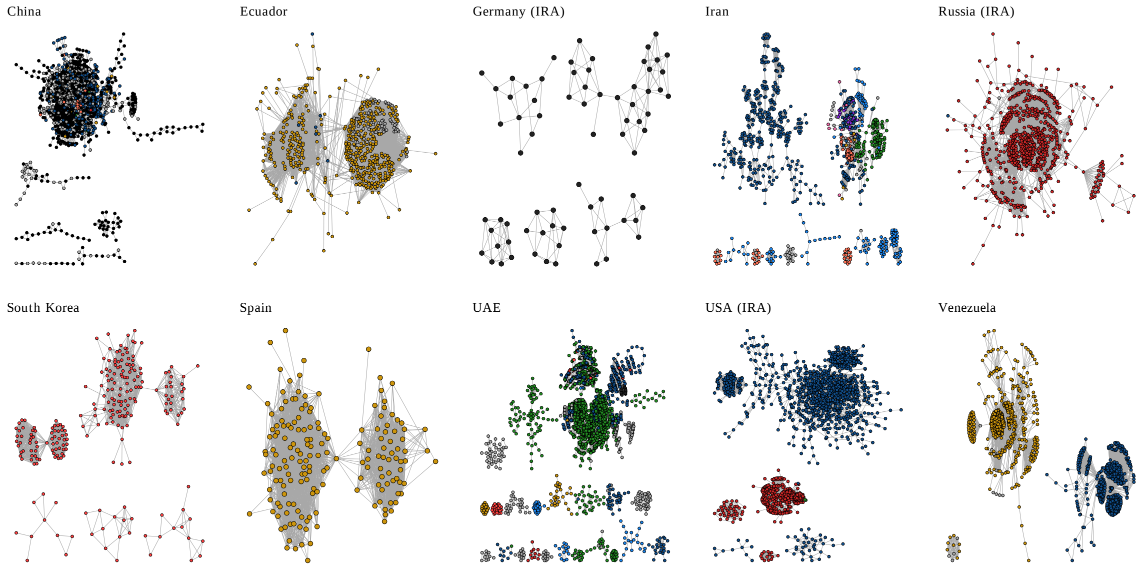

Figure 2: Co-tweet networks. Two accounts are connected if they send exactly the same tweet within one minute. Node color corresponds to most frequently used language by each account.

accounts between 2015 and 2018 (see Figure 4). Comparing to the random accounts that show consistent activitiy patterns throughout four years, the accounts involved in astroturfing campaigns has several moments when their activity intensity bursts, which shows a clear coordination pattern among the account holders.

We are in the process of drawing random samples from all of the countries in this study to investigate the extent to which such temporal patterns correspond with the time zones or daily working times of instigators. First results are promising for Germany and the U.S., where the activity of instigating Russian IRA agents not only form network clusters (Figure 2), but are also active during a different time windows than regular German or U.S. Twitter users.

5 Conclusion

This paper presented the most comprehensive investigation of different political astroturfing campaigns on Twitter to date. Our analysis covered various political and cultural contexts and spanned multiple continents. Despite this heterogeneity, we found remarkably similar patterns in all astroturfing campaigns. With our methodological approach to detect astroturfing campaigns we flag accounts that coordinate their messages, instead of

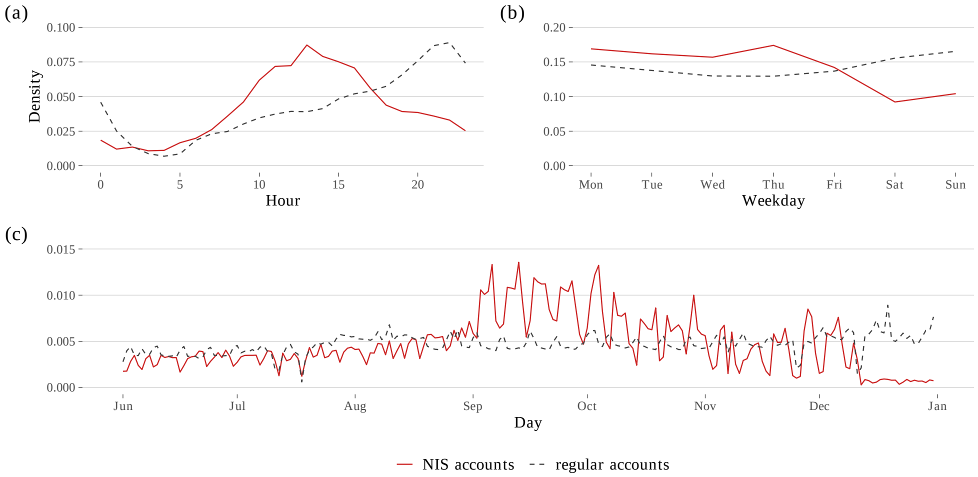

Figure 3: Distribution of tweets sent by NIS accounts and regular accounts each hour of the day (a), on each weekday (b) and on each day during the research period (c).

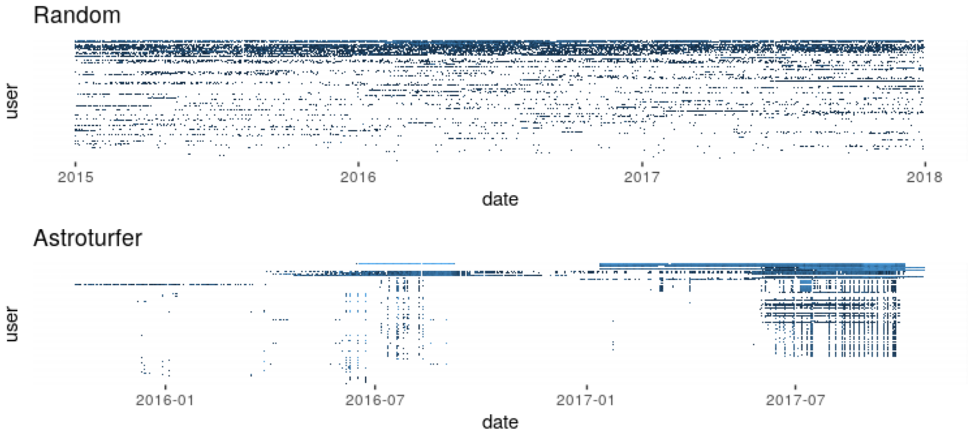

Figure 4: Activity intensity of the accounts associated with German astroturfing campaigns and the random German accounts.

focusing on individual account features like heavy automation. In the different campaigns examined, this criterion alone would have been able to identify between 52% and 86% of participating accounts. 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. They therefore invest little time in creating distinctive online personas or varying the behavior among the accounts they control. They tend to be active only when they are directly supervised by the organizers of the campaign (the principal), and have little reason to tweet once they leave their workplace.

We argue that 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. While the comparison with regular South Korean Twitter users suggests that the level of coordination exhibited by the South Korean astroturfing campaign is distinctive, it is of course possible that regular users in other countries naturally co-tweet or co-retweet at a higher frequency. Maybe the distinction between campaign accounts, which tended to tweet during office hours, and regular users, who were more active after work and on the weekend, is less clear in other cases. This question can be adjudicated by investigating to which extent the *overall activity* of astroturfers' diverges from regular Twitter users in the targeted countries. Therefore, we are currently acquiring historical data from randomly sampled Twitter users stratified by the analyzed countries and the overall level of activity so that we can construct benchmarks for the evaluation of our findings.

References

Badawy, A., Ferrara, E., & Lerman, K. (2018). *Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign.* Retrieved from https://arxiv.org/abs/1802.04291

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2012). Detecting automation of Twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing*, *9*(6), 811-824. doi: TDSC.2012.75

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, *59*(7), 96–104. doi: 10.1145/2818717

Grimme, C., Assenmacher, D., & Adam, L. (2018). Changing perspectives: Is it sufficient to detect social bots? In *International Conference on Social Computing and Social Media* (pp. 445–461).

Kargar, S., & Rauchfleisch, A. (2019). State-aligned trolling in iran and the double-edged affordances of instagram. *New Media & Society*, *21*(7), 1506–1527. doi: 10.1177/1461444818825133

Keller, F., Schoch, D., Stier, S., & Yang, J. (2017). 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). Menlo Park, CA: The AAAI Press.

Keller, F., Schoch, D., Stier, S., & Yang, J. (2019). Political astroturfing on Twitter: How to coordinate a disinformation campaign. *Political Communication*.

King, G., Pan, J., & Roberts, M. E. (2017). How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. *American Political Science Review*, *111*(3), 484–501. doi: 10.1017/S0003055417000144

Kovic, M., Rauchfleisch, A., Sele, M., & Caspar, C. (2018). Digital astroturfing in politics: Definition, typology, and countermeasures. *Studies in Communication Sciences*(1). doi: 10.24434/j.scoms.2018.01.005

Kreil, M. (2018). *The social bot research of Oxford and Co. is flawed.* Retrieved 19.01.2018, from https://data.info.graphics/blog/2018/12/21/social-bot-research-is-flawed/

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., … Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. doi: 10.1126/science.aao2998

Linvill, D. R., & Warren, P. L. (2018). Troll factories: The Internet Research Agency and state-sponsored agenda building.

Lukito, J., Wells, C., Zhang, Y., Doroshenko, L., Kim, S. J., Su, M.-H., … Freelon, D. (2018). *The Twitter exploit: How Russian propaganda infiltrated U.S. news.* Retrieved 20 March 2018, from https://uwmadison.app.box.com/v/TwitterExploit

Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, *116*(7), 2521–2526. Retrieved from https://www.pnas.org/content/116/7/2521 doi: 10.1073/pnas.1806781116

SeoulHigherCourt. (2015). *Case ID: 2014No2820.*

Stella, M., Ferrara, E., & De Domenico, M. (2018). Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences*, *115*(49), 12435–12440. Retrieved from https://www.pnas.org/content/115/49/12435 doi: 10.1073/pnas.1803470115

Stukal, D., Sanovich, S., Bonneau, R., & Tucker, J. A. (2017). Detecting bots on Russian political Twitter. *Big Data*, *5*(4), 310–324.

Twitter. (2019). Elections integrity. data archive. Retrieved from https://about.twitter.com/en_us/values/elections-integrity.html#data