

University of Copenhagen
Department of Political Science
Social Network Analysis (ASTK18106U)
Yevgeniy Golovchenko

Structure and Strategy of the Internet Research Agency on Twitter

Ole Fechner, Johannes Kopf

Copenhagen, December 20, 2018

Boddinstr. 15
12053 Berlin
ole.fechner@fu-berlin.de

MA Political Science

Moselstr. 50
50674 Cologne
jkopf@posteo.de

MA Political Science

Contents

1	Introduction	1
2	Agenda-building and astroturfing	2
3	The social network analysis (SNA)	3
3.1	Data: An ego-centered retweet network	3
3.2	Methods	5
4	Analyzing the IRA retweet network	8
4.1	Findings	8
4.2	Discussion	16
5	Conclusion	19
	References	21

List of Tables

1	User Statistics	4
2	Troll Statistics	5
3	Outdegree	12
4	Outdegree weighted	13
5	Indegree	14
6	Indegree weighted	15

List of Figures

1	Retweet network	9
2	Distribution of account categories in communities	10
3	Distribution of communities in account categories	11
4	Graph densities	12
5	k-core and k-in-core	16
6	Outdegree in the k-core	17

1 Introduction

“Frankly, the United States is under attack”, said Dan Coats, director of U.S. National Intelligence in an interview about the influence of Russian troll accounts in the 2016 U.S. presidential elections (Popken 2018). What seems to be a drastic statement has been emphasized by Twitter committing a list of 2,857 supposedly Russian troll accounts connected to the state-sponsored Internet Research Agency (IRA) to U.S. Congress, who have been active on Twitter for much longer than the 2016 elections. Coats expects Russia “to continue using propaganda, sympathetic spokesmen, and other means of influence to [...] exacerbate social and political fissures in the United States” (Popken 2018). With its low costs, low risk and the difficulties of tracing it back, Twitter appears especially attractive as an instrument for gaining influence in foreign countries. The degree of professionalism should thereby not be underestimated, with division in day and night shifts, training in English grammar per US holidays and quotas to meet to monitor the activity of the troll workers (Kirby 2018). This suggests an intentional strategy of troll behavior, an orchestrated act of state-institutional interference with a foreign election that urges for research on the matter. Given this relevance of the matter, research on the activity of Twitter trolls suggests shedding some light on an issue of growing importance for everyday politics.

With the recency of the IRA’s discovery, only few researchers have approached the topic and actually published their results. (Golovchenko et al. 2018: cf.) for an analysis of Twitter information warfare over Ukraine, or Stewart et al. 2018 for an analysis of trolls and polarization in a retweet network). Linvill and Warren 2018 present an early analysis using a sequential mixed method design, first qualitatively categorizing the Twitter accounts into groups as for example left or right troll and subsequently examining those account’s activity over time. This paper seeks to build upon Linvill and Warren’s work by further investigating potential differences in troll behavior between various communities of trolls. As a research question, we want to ask, how does IRA troll account behavior differ between the Twitter communities they engage in? As a theoretical foundation, we rely on the concepts of agenda-building and astroturfing, assuming strategic activity and behavior of Russian trolls in trying to gain influence on Twitter. We are using dataset of 147,428 retweets by 453 trolls, which includes a large share of trolls who have been qualitatively categorized by Linvill and Warren (Linville/Warren 2018). We are applying methods of Social Network Analysis (SNA), building a ego-centered retweet network using retweets as ties between different Twitter accounts as nodes. First, we are running community detection algorithms to cross-validate Linvill and Warren’s 2018 findings. We find that there is a sufficiently good fit of the detected

communities with the qualitative categories. Secondly, we analyze the graph densities in the different communities finding strong differences between communities. Third, we are computing out- and indegrees for our nodes, again finding support for different behavior across communities of Twitter accounts. As a last measure, we are computing different k-cores, finding an overall most active and inter-connected community of right trolls. We then discuss the findings with a focus on the mechanics of information spreading and the limitations of our SNA approach with regard to the network being ego-centered.

2 Agenda-building and astroturfing

The process of trying to move an actors agenda to the agenda of other actor's, especially policymakers, is defined as *agenda-building* (Linvill/Warren 2018: 3). This can also be extended to the question of how the public views certain issues, usually by analyzing media coverage of those issues: "Agenda-building research examines how certain groups, such as those in politics and business, influence what issues journalists cover as well as how the public views issues" (Parmelee 2014: 434). Since the rise of social media platforms like Twitter and Facebook, agenda-building takes place in those environments. This is due to journalists drawing heavily on Twitter for their job and, on the other hand, research shows that Twitter is the most popular social media platform for participating in political discussions, which from there are often taken to other media (Parmelee 2014: 435, 437). Influencing the citizens of another country through the use of media is nothing new, rather it is used in conflicts or during war regularly. "However, Russia's work on social media has taken agenda-building efforts by nations into a new context" (Linvill/Warren 2018: 3).

Closely linked to agenda-building is a second phenomenon called (political or/and on-line) *astroturfing*, which can be characterized as the "creation of a false or exaggerated impression of grassroots support" (Harcup 2014). It describes the strategic and coordinated approach of a group with the aim to create the impression of a certain public opinion, that might not exist in that way. On social media, those groups use many different accounts that post and interact with regular users to create the desired impression. For the purpose of this paper, we see astroturfing as a strategy of agenda-building. The anonymity provided by platforms like Twitter, as well as the covert structure of those groups, make them very hard to discover (Yang et al. 2017: 564). This exploratory study will therefore take the structure of the IRA as a starting point, instead of the presumed agenda behind the organization. Social Network Analysis, which will be introduced in the next section, offers an excellent tool box to conduct this task.

3 The social network analysis (SNA)

To analyze the structure and possible strategies of the IRA, we will take Social Network Analysis (SNA) as the method of choice. The data-driven character of SNA – the notion of people relating to each other, and the significance put into the structure and strength of those relations being almost the only theoretical assumptions (Golovchenko et al. 2018: 982) – makes it especially viable for our exploratory approach. SNA conceptualizes the linkages between actors as “channels for transfer or ‘flow’ of resources”, while the actors themselves are seen as “interdependent rather than independent, autonomous units” (Wasserman/Faust 1994: 4). Therefore, the different actor attributes are theorized as emerging out of their relations and not vice versa, as in most other types of quantitative analyses (Wasserman/Faust 1994: 8). This allows us to ignore the question of the actual people and their intentions behind different accounts and tell something about the IRA and its structure, as well as its strategy, as a whole. The most important relation to analyze how the IRA uses astroturfing as a strategy for agenda-building is the distribution of information, in this case via retweets. Accordingly, the foundation of this paper will be a social network consisting of Twitter accounts as nodes and retweets as edges (or linkages).

3.1 Data: An ego-centered retweet network

We will use a dataset, published by NBC News (2018), consisting of 203,451 Tweets by 453 accounts between July 2014 and September 2017, which were linked to the IRA by an official document handed over to US Congress by Twitter. Twitter justifies this linking by referring to “third party sources”, which makes it impossible to reconstruct or evaluate their method. Therefore, we have to assume those accounts’ links to the IRA to be correct, as it is the best evaluation available.

To create the social network, we need to clean the data: First, we drop any of the tweets that are not retweets, which leaves us with 147,428 retweets by 453 troll accounts. Second, 120 trolls, who did not retweet at all, were dropped accordingly. Our dataset now consists a set of 333 unique troll Twitter handles¹, who are retweeting others. A variable stating the unique User ID is used to validate uniqueness of the users, showing that there are indeed no duplicate User IDs in the dataset. Since, we only have information about who these 333 accounts did retweet, but not by whom they were retweeted, the network is ego-centered around the group of 333 trolls. Of these 333 trolls, 151 (ca 45%) retweeted others and were themselves retweeted by other trolls, thus being both sender and receiver. 182 (ca

¹A Twitter handle is the screen name of a Twitter account that can be changed by the users.

55%) trolls only retweeted others, but were not retweeted themselves, making them only senders.² A third group of 71 trolls was found by looking at who was retweeted by the original group of trolls in the data, thus increasing the number of trolls in our data to a group of 404. These 71 trolls were retweeted by others, though did not send retweets themselves. Finally, the big body of users in the dataset consists of 36,485 users, who are retweeted by the trolls, but are not themselves categorized as trolls by Twitter. Table 1 gives a first overview over our dataset. Overall, there are 36,889 unique twitter users in the dataset, 404 classified trolls and 36,485 non-trolls. The retweets contain relational information about one user retweeting another, therefore creating a directional edge between the two. Thus, the graph is a directed, ego-centered network of 404 IRA accounts with 36,889 unique Twitter accounts as nodes and edges representing retweets from the sender to the receiver.

User	N	Senders	Receivers
Troll	404	333	222
Non-Troll	36,485	0	36,485
Total	36,889	333	36,707

Table 1: Distribution of troll accounts and retweet sender and receiver in the data.

To further extend our data and our scope of analysis, we are adding qualitative information on the IRA trolls, provided through a dataset by Darren Linvill and Patrick Lee Warren (2018) via the online news outlet FiveThirtyEight (Roeder 2018). Linvill and Warren conduct a qualitative analysis, categorizing a sample of 1,133 IRA troll accounts by examining the tweet content and the account names, applying a temporal analysis of the trolls tweeting behavior after. They “identified five categories of IRA-associated Twitter handles, each with unique patterns of behaviors: *Right Troll*, *Left Troll*, *Newsfeed*, *Hashtag Gamer*, and *Fearmonger*” (Linvill/Warren 2018: 6). In addition, there are three categories, which are not used within their analysis: *Non-English*, *Commercial* and *Unknown*. The categories *Right Troll* and *Left Troll* need little explanation, as they include users who broadcasted right-leaning populist and socially liberal messages. *Hashtag Gamers* are users who are playing word games on Twitter, mostly non-political, though sometimes including left- or right-leaning messages. *Newsfeed* Trolls are posing as local US News Agencies, mostly linking to legitimate news content, often with a pro-Russian perspective. *Fearmongers* spread news of crisis events such as Tweets about salmonella infections. The *Non-English* troll category includes

²We define a retweet sender as the person retweeting an original Tweet by another person, who, accordingly, is the retweet receiver.

users who tweeted in other languages than English, predominately Russian, some German and little French and Spanish. *Commercial* Trolls are not included in our dataset. Finally, users were categorized as *Unknown*, if they could not be assigned to other categories for lack of information in their tweets. These categories will be included in this paper’s analysis, since might be interesting to see how they interact with the other trolls. We are appending Linvill and Warren’s account categories to our data, finding categories are available for 394 of the 404 troll handles in our dataset, meaning that around 98% of the trolls in our dataset are categorized. Table 2 shows the distribution of the categories among our dataset.

Category	N	Senders	Receivers	Average followers
Right	101	75	90	4649
Left	110	104	48	1783
Hashtag Gamer	61	43	60	3021
Non-English	106	100	7	2127
Newsfeed	11	1	10	16446
Fearmonger	4	0	4	0
Unknown	11	10	3	3306
Total	404	333	222	4476

Table 2: Distribution of account types categorized by Linvill and Warren (2018)

As another attribute, we are appending information on the count of followers of the troll accounts from a second dataset provided by NBC news. The information on the count of followers provided only includes one figure and does not vary over time, without specification of when these follower counts were obtained. We will assume that they are at least to some degree representative and use them as a heuristic. Table 2 gives the average number of followers for each account category. Finally, we are using the full time period of retweets, from July 2014 to September 2017. This paper is not interested in a time period preceding a specific event, like an election or a specific trending discussion, but rather strategic behavior of the trolls in general, Thus making the right approach to include all of the Tweets in the analysis.

3.2 Methods

To understand the IRA’s division of labor between the accounts, we begin with testing whether Linvill and Warren’s qualitative findings of the different account types would be reproducible via using network properties. For this, we will draw on a community detection

algorithm based on modularity, known as the Louvain Method (Blondel et al. 2008).³ Communities are mesoscopic structures of a graph, that consist “of a group of nodes that are relatively densely connected to each other but sparsely connected to other dense groups in the network” (Porter et al. 2009: 1083). The Louvain algorithm is based on modularity, “a scalar value between -1 and 1 that measures the density of links inside communities as compared to links between communities” (Blondel et al. 2008: 2; cf. also Porter et al. 2009: 1089). It approximately maximizes the modularity for each node, thus identifying communities. Finally we will compare the communities with Linvill and Warren’s account types, to see weather we can better understand the identified community structure.

One important measure for the cooperation of the accounts is graph density, the proportion of all possible edges that are present in the graph (Wasserman/Faust 1994: 101). Since the network is ego-centered, it only makes sense to calculate density for the troll subgraph. We will compute density for the directed graph, as well as for its undirected version. To create the undirected graph, we keep a tie for every relationship that is only defined with a single directional tie, but also to collapse all reciprocal ties into one tie in the undirected graph, thus avoiding multiple ties in one relationship. We will then compare densities of the different communities, which can give an indication weather the IRA has distinctive strategies for those groups. Following Wassermann and Faust (1994: 102) we calculate the density of the different community subgraphs Δ_c as:

$$\Delta_c = \frac{2L_c}{g_c(g_c - 1)},$$

where L_c is the number of edges present in the community subgraph, and g_c is the number of nodes in said graph. In communities with higher density, the accounts are working more together, meaning they spread information further through retweeting, while accounts in less dense communities rely more on information they put into the network themselves.

Identifying the most important accounts helps to understand the structure of the IRA. Operationalizing importance in SNA is usually done through centrality and prestige indices, whereas not only the chosen actors are considered prestigious, but also those doing the choosing (Wasserman/Faust 1994: 170). Since the network is ego-centered around the troll accounts, the indices have limited reach in their interpretation, but still deliver important insights. To show the most central accounts, we will calculate outdegree. The index can only be computed for IRA accounts, since they are the only retweeting in our dataset. We

³For our network, this algorithm produced the best results. The Infomap algorithm (cf. Rosvall/Bergstrom 2008) results in one big community containing ca 98% of the nodes.

will calculate the outdegree weighted and unweighted respectively, the former results in the accounts that retweet the majority of different accounts, the latter in those retweeting the most.⁴ The outdegree of a node $d_{out}(n_i)$ is computed as

$$d_{out}(n_i) = \sum_j x_{ij},$$

where x_{ij} ⁵ is an edge from i to j (cf. Wasserman/Faust 1994: 178). For the weighted outdegree x_{ij} is multiplied by the value of the edge. Prestige, on the other hand, will be calculated through indegree, which results in those accounts retweeted the most and can therefore tell about what accounts the IRA draws on. It will also be computed for the weighted and unweighted graph separately, which shows the accounts retweeted by the majority of different accounts and those retweeted the most in general respectively. The indegree of a node $d_{in}(n_i)$ is computed simply as

$$d_{in}(n_i) = \sum_j x_{ji},$$

where x_{ji} is an edge from j to i (cf. Wasserman/Faust 1994: 202). For the weighted outdegree x_{ji} is multiplied by the value of the edge.

Finally, we will look at the most engaged users in the network to identify the so-called super-spreaders of information. In SNA, the ability to spread information better than other individuals is ascribed to their unique location in the network, which makes identifying those a prime task of analyzing networks used for false information (Pei et al. 2015: 1). Pei et al. (2015) show that, compared to other approaches, the k-core method predicts the super-spreaders most accurately. “A k-core is a maximal subset of the network where all nodes are connected to at least ‘k’ number of other nodes: so ‘k’ can be any whole number” (Golovchenko et al. 2018: 983). Hence, the k-core specifies a subgraph via the number of ties that must be present between nodes, it is based on the nodal degree. We can define a subgraph G_s as a k-core if

$$d_s(i) \geq k \text{ for all } n_i \in N_s,$$

where $d_s(i)$ is the (in-)degree of node i and N_s the set of nodes in the subgraph (Wasserman/Faust 1994: 266). Once again, the network being ego-centered limits the validity of the results, since every non-troll in reality has a higher degree through retweets from

⁴A weighted graph is a graph in which each edge carries a value (Wasserman/Faust 1994: 140), in this case the number of retweets between two accounts. In an unweighted graph, each edge has the value 1.

⁵Note that in- and outdegree are calculated for directed graphs, hence $x_{ij} \neq x_{ji}$.

other non-troll accounts and retweeting themselves. We can therefore only identify the most important IRA accounts regarding their own network. To account for this, we will also compute the k -in-core through the indegrees. The idea is that information from one account is spread quickly, if this account is retweeted a lot and the accounts retweeting it also get retweeted a lot.

4 Analyzing the IRA retweet network

In the preceding chapter, the data and methods that form the basis of our analysis have been introduced. Figure 1 shows the entire retweet network consisting of three main clusters. We will focus our analysis on those clusters in four steps. First, the results of the modularity community detection algorithm are illustrated and new communities are established, based on our findings, as well as the qualitative coding of Linvill and Warren. Second, graph densities are compared between these communities. Third, degree centrality indices are analyzed for the whole graph to identify the most active IRA accounts and the most prestigious accounts. Finally, we will take a look at the k -cores, to detect information super-spreaders and the connection between the clusters. In the last section, we discuss the findings regarding the strategy and structure of the IRA.

4.1 Findings

We run a modularity-based community detection, following the Louvain Method (Blondel et al. 2008). Figure 2 shows the distribution of algorithm communities among the qualitative account categories created by Linvill and Warren (2018). For further analysis, we rely on communities that were computed with the Louvain Method, since they originate from the structure of ties within our dataset and this appears to be more consistent with the methodology of SNA. To not exceed the scope of this paper, we narrow down our analysis by merging all minor communities below an arbitrary threshold of 50 nodes, since they bear a larger risk of yielding unrepresentative findings. We then apply the qualitative categorization to our newly distincted communities. In Figure 2, we see that some categories overlap strongly with distinct communities, whereas other categories show a rather mixed composition in our data. 94% of the left troll category is captured with one community, which will be our new left troll community. The hashtagger community is 90% captured by one community, which will be our new hashtagger troll community. 90% of the right troll category consist of two major communities with shares of 52% and 38%, which we

to note, that by using this set of communities, we are treating some nodes as e.g. right trolls, even though they have been categorized as left trolls by Linvill and Warren (2018). This bears the risk of having biased results. Though, with the given incongruence of the account categorization with our detected communities, having some potential bias appears inevitable. In Figure 3, the distribution of account categories among our new communities is depicted, which illustrates this potential bias.

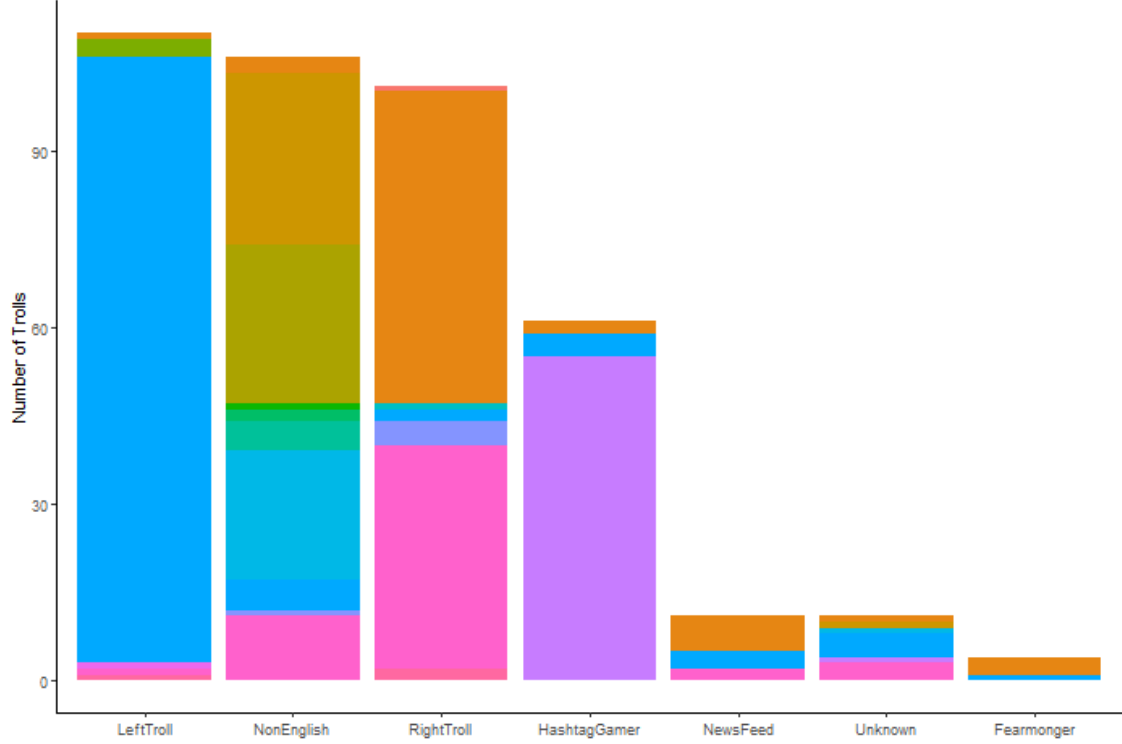


Figure 2: Distribution of account categories by Linvill and Warren. Color reflects account categories among communities detected by the Louvain Method.

Next, the graph densities are calculated for the subgraphs of our newly determined communities. Figure 4 shows the graph densities for the three communities for both directed and undirected versions of each subgraph. First of all, we see that the undirected subgraphs' densities are approximately twice as high as directed ones, with only minor deficits. With our mode of creating an undirected graph, we can interpret this that way, that there are not many reciprocal ties in the subgraphs. If there was a considerable amount of reciprocal ties, they would cause a higher density of the directed subgraph in comparison to the undirected one ($>50\%$).

Examining the magnitude of densities across different community subgraphs, we find that the hashtagger community is the most dense by a significant margin. Approximately 17.7%

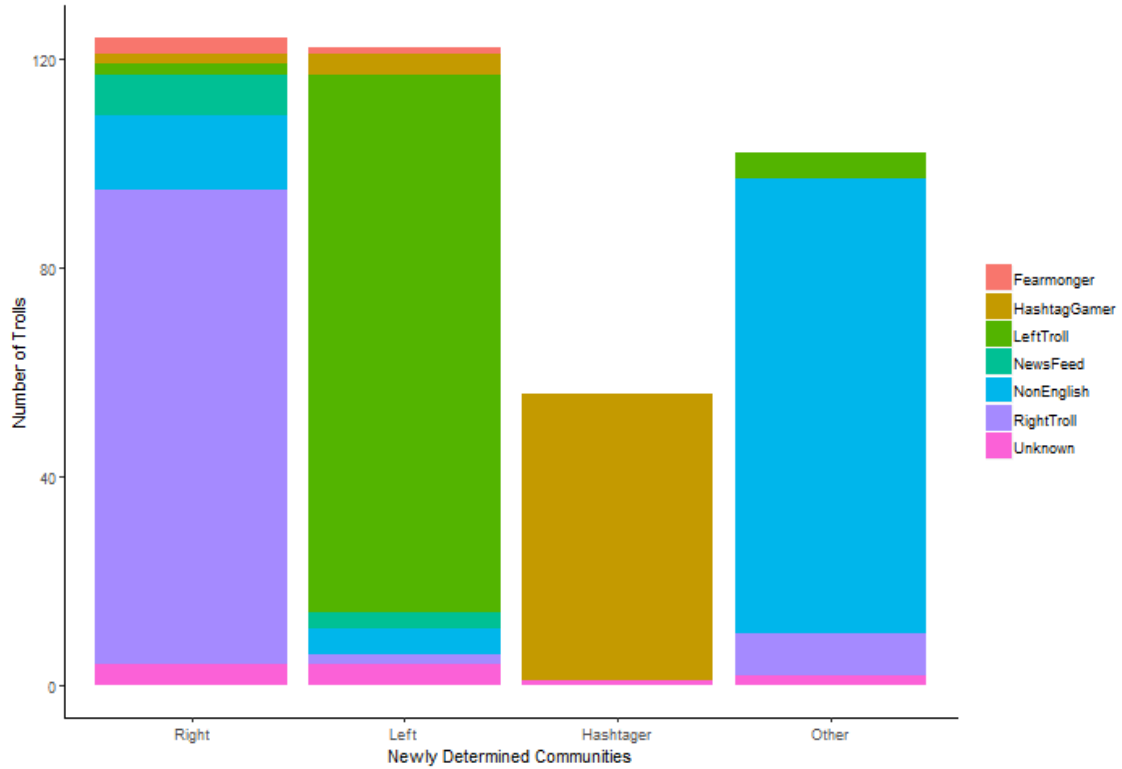


Figure 3: Distribution of account categories among communities detected by the Louvain Method. Color reflects account categories by Linvill and Warren.

of all possible ties are present for the undirected subgraph. This could relate to the nature of this community. As laid out above, hashtag gamers retweet themselves playing a word game, which means that it appears to constitute this community’s cohesion to retweet one another frequently. Right trolls have the second largest density with 7.4%, which is more than six times larger than the left community’s density of 1.2%, looking at the numbers for the undirected graphs.

We now analyze the degree centrality of our graph starting with the outdegree, which we use as a measure for centrality and the weighted outdegree, which gives additional information on the activity of the user. Because of our network is ego-centered around the central group of trolls, only troll accounts have outdegrees. Table 3 shows the top 10 users ranked by outdegree.

The first ranked user *ameliebaldwin* sticks out with an outdegree of 4,896, meaning that the user retweeted nearly 5,000 different users. The weighted outdegree of *ameliebaldwin* is 9,243, which includes multiple retweets of the same user. Since the weighted outdegree is significantly higher than the unweighted one, this means that *ameliebaldwin* retweeted a large number of users only a few times, rather than retweeting the same small number of

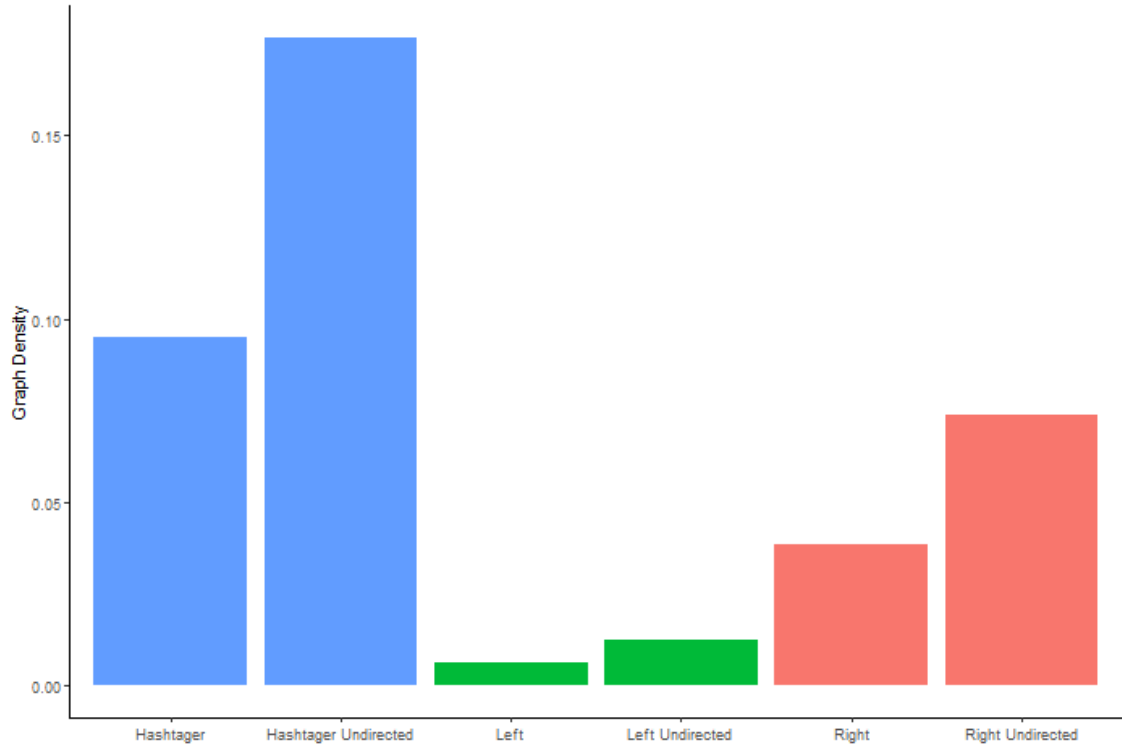


Figure 4: Graph densities in the three biggest communities.

User	Community	Category	Followers	$d_{out}(n_i)$	$d_{out}(n_i)$ weighted
ameliebaldwin	Other	RightTroll	2,464	4,896	9,243
patriotblake	Other	RightTroll	2,035	2,856	4,106
hyddrox	Right	RightTroll	2,225	2,650	6,788
giselleevns	hashtagger	Unknown	24,344	2,282	5,403
cookncooks	Other	RightTroll	1,468	2,153	2,893
emileewaren	Other	RightTroll	1,909	2,116	2,891
dorothiebell	Right	RightTroll	1,893	2,010	2,874
baobaeham	Left	LeftTroll	1,032	1,802	3,215
michellearry	Right	RightTroll	3,229	1,611	2,677
nickluna	Right	RightTroll	1,457	1,597	2,825

Table 3: Top 10 users ranked by outdegree

users over and over again. The category and community allocation disagree for *ameliebaldwin*. While the qualitative analysis by Linvill/Warren 2018 found the user to be a right troll, the community detection algorithm did not find it to belong one of the major right troll communities. This exact pattern of disagreement reappears for three other users in Table 3.

Subsequent to *ameliebaldwin*, there is a group of accounts retweeting more than 2,000 different users, which is mainly dominated by right trolls following the categorization, but also with partial disagreement, following the above-mentioned pattern. The fourth ranked user *giselleevns* sticks out for having the most followers in the ranking with 24,344. With a mean follower count of 3,022 and a maximum of 61,109, *giselleevns* appears to be one of the most followed accounts in the dataset. Though, it needs to be noted, that information on followers is only available for troll accounts in our dataset. An exception to the right-wing dominated top-ten is user *giselleevns*, who was allocated to the hashtagger community.⁶ When ranking the account for weighted outdegree the picture slightly changes, as can be seen in Table 4. While some accounts like *ameliebaldwin* and *giselleevns* remain in high-ranked positions, six left-wing troll accounts are now in the ranking, for which category and community allocation mostly agrees.

User	Community	Category	Followers	$d_{out}(n_i)$	$d_{out}(n_i)$ weighted
ameliebaldwin	Other	RightTroll	2,464	4,896	9,243
hyddrox	Right	RightTroll	2,225	2,650	6,788
giselleevns	hashtagger	Unknown	24,344	2,282	5,403
patriotblake	Other	RightTroll	2,035	2,856	4,106
mrcldepratt	Left	LeftTroll	914	1,583	3,262
brianaregland	Other	LeftTroll	768	1,360	3,259
baobaeham	Left	LeftTroll	1,032	1,802	3,215
datwiseenigga	Left	LeftTroll	904	1,540	3,196
willisbonnerr	Left	LeftTroll	571	1,563	3,155
melanymelanin	Left	LeftTroll	963	1,079	3,071

Table 4: Top 10 users ranked by weighted outdegree

The highest ranked left troll account *mrcldepratt* has a weighted outdegree of 3,262 and an unweighted one of 914. Compared to the top ranked right-wing troll accounts, the left troll accounts approximately only have a maximum of half of the unweighted outdegree, but are much closer on the weighted outdegree. This means that the left trolls retweet a smaller number of accounts more often than right trolls do. This relates to the lower graph density for the left troll community that was computed above.

When examining the accounts with the highest indegrees, contrary to outdegree, both trolls and non-troll accounts are included. We use indegree as a measure for prestige, as it shows how often a user is chosen to be retweeted by the trolls. Table 5 shows the ten

⁶A finding that contradicts Linvill and Warren 2018, who could not attribute the account to any category.

highest ranked accounts for unweighted indegree. Here, mostly non-troll accounts are present. Among them, we find accounts from popular U.S. politicians like President Donald Trump or Hillary Clinton, but also accounts of official media outlets like *The Hill* or *Fox News*. The account *thehill* has the highest unweighted indegree with 102, which represents the number of unique trolls who retweeted him in the data. The indegree quickly declines from 102 for rank one to 53 for rank ten. One account that sticks out is *ten_gop*, who is the only troll among the list. It is categorized as a right-wing troll by both Linvill and Warren 2018 as well as the community detection algorithm. Further investigation shows that *ten_gop* pretended to be an official Republican Party account, which fits his ranking among the other official accounts.

User	Community	Category	$d_{in}(n_i)$	$d_{in}(n_i)$ weighted
thehill	Non-Troll	Non-Troll	102	358
realdonaldtrump	Non-Troll	Non-Troll	100	544
wikileaks	Non-Troll	Non-Troll	82	247
blicqer	Non-Troll	Non-Troll	69	2,207
hillaryclinton	Non-Troll	Non-Troll	61	98
joyannreid	Non-Troll	Non-Troll	58	267
prisonplanet	Non-Troll	Non-Troll	56	462
jamilsmith	Non-Troll	Non-Troll	55	118
ten_gop	Right	RightTroll	53	430
foxnews	Non-Troll	Non-Troll	53	336

Table 5: Top 10 users ranked by indegree.

Table 6 shows the top ten ranked accounts for weighted indegree. Similarly to unweighted indegree, official politician or news outlet accounts dominate the picture. Weighted indegree appears to not correlate strongly with the unweighted indegree for the top ten ranked accounts, wherefore the order changed substantially. Four new accounts appear in the ranking compared to Table 5. These accounts have a relatively low number of trolls retweeting them, but in a relatively high frequency. For example, the second ranked user *conservatexian* has only 30 different users retweeting him 1,082 times, which results in an average of ca 37 retweets per troll. For the account *nine_oh* the average is even higher, with ca 45 retweets per troll. For comparison, *realdonaldtrump* has an average of ca 5 and the only right troll in the ranking *ten_gop* a average of ca 8.

For the k-core we limited the network to the core of $k = 10$, for the k-in-core to $k\text{-in} = 4$, meaning that the users of the k-10-core have been retweeted and/or retweet others at least 10 times and the users of the k-4-in-core have been retweeted at least 4 times,

User	Community	Category	$d_{in}(n_i)$	$d_{in}(n_i)$ weighted
blicqer	Non-Troll	Non-Troll	69	2,207
conservatexian	Non-Troll	Non-Troll	30	1,082
realdonaldtrump	Non-Troll	Non-Troll	100	544
nine_oh	Non-Troll	Non-Troll	11	500
prisonplanet	Non-Troll	Non-Troll	56	462
zaibatsunews	Non-Troll	Non-Troll	16	451
gerfingerpoken	Non-Troll	Non-Troll	46	434
ten_gop	Right	RightTroll	53	430
bizpacreview	Non-Troll	Non-Troll	17	401
beforeitsnews	Non-Troll	Non-Troll	8	399

Table 6: Top 10 users ranked by weighted indegree.

which is the maximum value for k-in in the dataset.⁷ The original graph consists of 36,889 nodes with 147,428 edges, of which the k-10-core contains 1257 (3.4%) nodes and 18214 (12.3%) edges, and the k-in-4-core 1754 (4.8%) nodes and 11988 (8.1%) edges respectively. Although the k-in-4-core contains some nodes more, it has far less edges, which is shown in Figure 5, that depicts both subgraphs. It is immediately apparent, that the k-core still contains the basic structure of the network with its three main clusters, while the k-in-core only consists of one large cluster with mainly right-wing accounts and one very small hashtagger cluster. This is contingent with the findings regarding degree centrality, which showed that the right-wing trolls have a larger unweighted indegree, while the results are more equal for the unweighted degree. This is no surprise considering the k-in-core is calculated through the indegree. Interestingly, the k-10-core looks very different.⁸ Here, the basic structure of the entire network is still intact, which indicates a lot of left and hashtagger accounts retweeting right-wing IRA accounts in the k-10-core. Otherwise, we would see more of the former in the k-in-core, but as they only have a large outdegree (because of retweeting relatively more right-wing accounts) the two subgraphs are structured this way. In other words, IRA accounts retweet other trolls mainly from the left and hashtagger clusters, while right-wing accounts mainly retweet themselves.

Since the k-in-core consists mostly of right-wing accounts, a further analysis of the k-10-core seems to be more fruitful. The accounts in the k-10-core are those with the best

⁷We disregarded self loops, since they cannot count as new information and are therefore irrelevant for analyzing information spreading.

⁸A quick analysis in Gephi via filtering gives one cluster as a result only for the k-18-core, when 271 nodes are left.

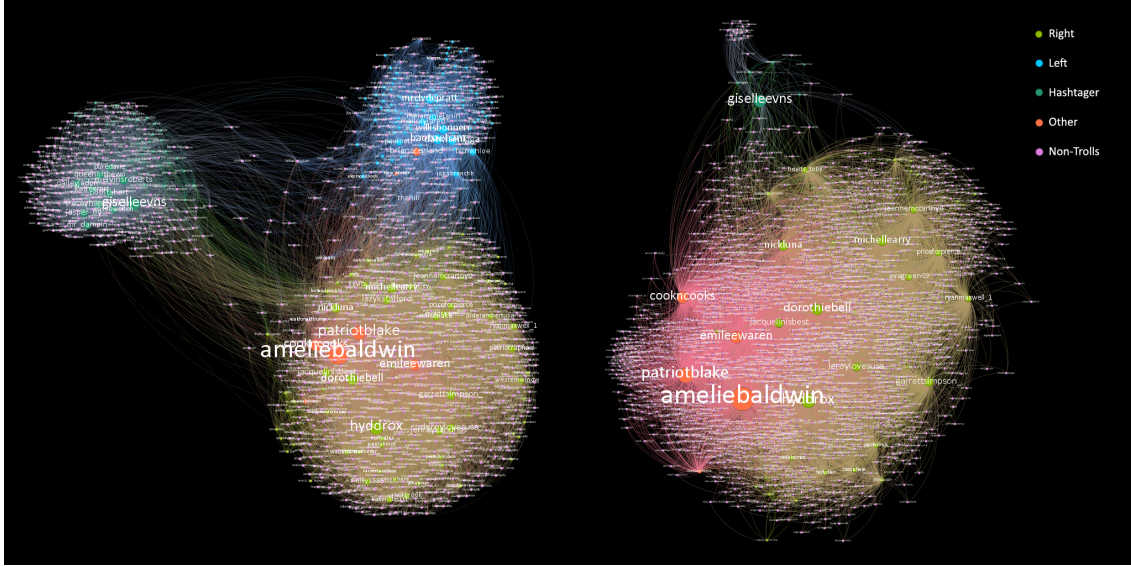


Figure 5: k-10-core and k-in-4-core. Node and label size represent outdegree centrality, colors reflect community.

positions in the network regarding the ability to spread information. We therefore proceed to examine the most active accounts, since they can be considered as the most influential. Figure 6 provides an overview of the 30 accounts with the highest outdegree within the k-10-core, and thus the profiles who retweeted the relevant accounts most often. The profile *ameliebaldwin* has by far the largest outdegree, which puts this account at the central position of the core. Interestingly, of the top 6 profiles, 4 were categorized as *Other* by the community detection algorithm. Looking on Figure 5, they all are very close to the right-wing community (the same is true for the k-in-core). Linvill and Warren categorized most of them as right-wing as well, which indicates some inaccuracy of the community detection algorithm, i.e. it is too sharp, thus detecting too many communities. Seeing them as right-wing shows a very clear dominance of right-wing profiles concerning information super-spreaders, which is consistent with our previous findings.

4.2 Discussion

Our analysis shows that the IRA did not only conduct astroturfing for one single community, but for at least three distinct social groups on Twitter. Targeting all of them ensured a way broader influence, especially considering the selection of left- and right-wing communities, suggesting a strategy to enhance an already present divide in society, that is reflected on Twitter. We can therefore verify Linvill and Warren’s qualitative findings out of the structure

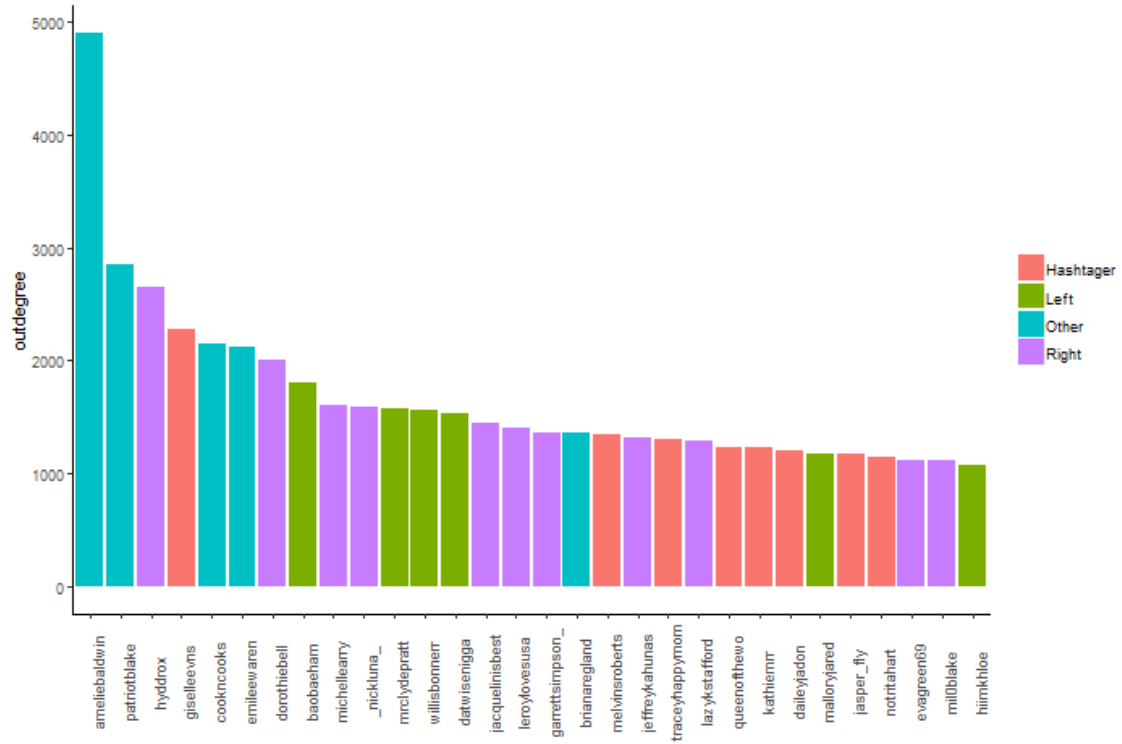


Figure 6: Top 30 user regarding outdegree in the k-10-core.

of the network itself. The community detection results in a distribution of the right-wing cluster into two smaller communities. We can only speculate this being due to the right-wing's division into subgroups e.g. conservatives and alt-right. Another possibility is the relatively small size or the ego-centered character of our network leading to a too fine result of the algorithm. An interesting finding is the IRA's targeting of the *Hashtag Gamer* community, since they are not known for their wide political influence. Again, we can only speculate: Since they are a relatively well connected group, therefore information can be spread well through their channels. Adding Linvill and Warren's findings, the IRA might have tried to politicize this group to widen the scope of their astroturfing to a big, well-connected, but not very politically active twitter community.

Comparing the three communities, the IRA put a definite focus on the right-wing community. Right troll accounts are retweeting more regarding the mere quantity, as well as the range of different accounts retweeting. Accounts in the left community are less well connected and tend to repeatedly retweet the same accounts very often. It is unclear, whether this is intentional IRA behavior or is based in the way those communities function differently in the way they interact within, although we can assume that it is in fact intentional, since knowing from the statements of Lyudmila Savchuk (Savchuk 2018), an investigative

journalist who worked for a couple of months undercover for the IRA, that every tweet was seen through before posting. The *Hashtag Gamers* are the most connected community, but this is most likely due to the nature of the hashtag-games this community is engaged in.

The IRA's focus on the right-wing community is reinforced through those accounts mainly retweeting themselves, while user from the other communities do retweet right-wing accounts. Again, we can only assume, this is intentional and not based in the community structure. Whether intentional or not, the IRA was definitely more active and able to spread information broader within and from the right-wing community. The analysis of the k-10-core and the k-in-4-core locates the super-spreaders of information mostly in the right-wing community as well, which is no surprise considering the fact that this community is retweeted by everyone else. The most active troll account in the most connected core of the network is by far *ameliebaldwin*. It is unclear why exactly this account has such a key position, while the other super-spreaders are relatively close to each other, which suggests a division of labor between a couple of accounts. Although the focus is on the right-wing community, left-wing and *Hastag Gamer* accounts are not unimportant, some of them occupying considerable central positions, just in total in a lesser degree than right-wing trolls.

The most prestigious accounts, meaning the accounts retweeted by the most trolls, range from accounts of important politicians like President Trump or Hillary Clinton through big news channels like *The Hill* or *Fox News* to political journalists like Joy Reid or Jamil Smith. This makes sense, since those profiles have a lot of followers and people are referring on them very often. This would suggest those profiles also being the most prestigious quantitatively, however with the exception of President Trump other profiles take their spot. With *blicqer* a black activist user is by far the most retweeted profile in our dataset, the second being *conservatexian*, a right-wing user, who stand as prime examples for users of the left and right communities. Both are big profiles, but not of official government or news accounts. It suggests that the IRA rather shares information from unofficial accounts which do not employ standards regarding information checking before they tweet.

This study had to deal with some severe limitations regarding the available data, the network being ego-centered being the main obstacle, since it allows for less interpretation. While we have a lot of information on the trolls' own retweet behavior, we do not know which troll has been retweeted how much by non-troll users. This left us with fewer possibilities than with a complete non-ego-centered network (e.g. the number of regular tweets of accounts as a node attribute or ratios of how many retweets per tweet a user is receiving can enhance the analysis). We cannot presume that the indegree of a troll account is a good heuristic for the actual number of retweets an account gets, since non-troll retweet behav-

ior could be completely different from troll activity on Twitter. It is important to highlight, that a complete network of the same nodes would presumably change the structure of the graph, including centralities and community allocation. This implies, that our graph does not tell us the actual positions of the nodes in the holistic Twitter network. With a bigger and thus more representative dataset, the smaller communities of trolls we categorized as other would be represented accurately and could be analyzed. Our graph mainly is a depiction of how the group of trolls constituting our dataset interacted with each other and whose tweets they chose to spread. Outdegree centralities and the k-core measure are therefore eligible to compare trolls or groups of trolls with each other. This is why our findings on interaction or cooperation between the trolls should be valid within the scope of an ego-centered network, although we rely on the assumption that there is no substantial group of trolls missing, which would skew the results comparing the communities.

5 Conclusion

Distinguishing true from false information becomes increasingly difficult in today's social media landscape. In this paper, we try to shed some light on the structure and strategy of the organized disinformation campaign of the IRA. Situating their behavior in astroturfing as a strategy for agenda building, we showed how they target specific communities on Twitter, in order to create the impression of a certain public opinion as diametrical opposed. We have explored the division of labor of different troll accounts, as well as their positions within a retweet network. Instead of creating the impression of one social movement, the IRA is operating within different, even opposed communities, while focusing on the right-wing community. With one exception, they have a couple of accounts in central positions spreading mainly (dis-)information regarding political topics. Hence, we can confirm, that the problem of disinformation campaigns cannot be solved, only marginal, improved, by banning some highly influential accounts, even if examples like *ten_gop* might suggest otherwise. Interestingly, our results show the IRA trying to influence the well connected, but not distinct political community of *Hashtag Gamers*, which calls for attention. If organizations like the IRA specifically target non-political communities, the problem of agenda-building through disinformation campaigns is even bigger than assumed right now.

Although the dataset used in this paper posed some severe limitations – especially the dataset consisting of only tweets from the troll accounts, which resulted in the network being ego-centered, as well as it not being complete – we were able to highlight central aspects of the strategies of the IRA. Thus, we demonstrated the merits of SNA in exploring strategies of

organizations conducting disinformation campaigns, while having no possibility to learn of their intentions through other social science methods like interviews. SNA's methodological approach in taking structures as a starting point make it the prime tool for cases like this, even if the data is incomplete. That being said, to fully understand how organizations like the IRA work, much more research is needed. Especially with a more complete dataset, that does not result in just a retweet network, a SNA can deliver more valid results. This study pointed to some further questions regarding the functioning of communities on Twitter. Here, more sociological knowledge can help identifying whether the IRA's different retweet structure in different communities is strategic or rooted in those communities themselves, e.g. through an analysis of different trending hashtag discussions, which could shed light on how specific community behavior might change between policy areas. The question why some troll accounts have much more output also should be discussed. Lastly, it can be learned a lot from comparing cases. Astroturfing is no new phenomenon, accordingly compare the IRA with similar agencies can give a better understanding of how these operate. Taking one step back, SNA does not necessarily have to be conducted via a network of retweets, but could also be done with mentions.⁹ Contrary to retweets, mentions appear to be more easily used to contradict another user, which would make the ties a lot harder to define and interpret, but an interesting choice for comparing a network drawn from the same data.

All in all, influencing and agenda-building poses a big problem, especially for democratic states, since social media here reaches very far into society. Right now organizations like the IRA have the advantage, because they were able to operate out of the dark. But through in-depth research happening right now, of which this paper is only a marginal part, we can analyze their structure and strategies and thus hopefully find the necessary tools to prevent disinformation campaigns from reaching considerable impact.

⁹Mentions mean the tagging of another user's name without spreading information of their tweet.

References

- Blondel, Vincent D./Jean-Loup Guillaume/Renaud Lambiotte/Etienne Lefebvre (2008): “Fast unfolding of communities in large networks”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10.
- Golovchenko, Yevgeniy/Mareike Hartmann/Rebecca Adler-Nissen (2018): “State, Media and Civil Society in the Information Warfare over Ukraine: Citizen Curators of Digital Disinformation”. In: *International Affairs* 94.5, pp. 975–994.
- Harcup, Tony (2014): *Astroturfing*. In: *A Dictionary of Journalism*. Oxford University Press.
- Kaiser, Jonas/Adrian Rauchfleisch (2018): *Unite the Right? How YouTube’s Recommendation Algorithm Connects The U.S. Far-Right*. Medium. URL: <https://medium.com/@MediaManipulation/unite-the-right-how-youtubes-recommendation-algorithm-connects-the-u-s-far-right-9f1387ccfabd> (visited on 11/09/2018).
- Kirby, Jen (2018): *What to know about the Russian troll factory listed in Mueller’s indictment*. Vox. URL: <https://www.vox.com/2018/2/16/17020974/mueller-indictment-internet-research-agency> (visited on 11/14/2018).
- Linville, Darren L./Patrick L. Warren (2018): *Troll Factories: The Internet Research Agency and State-Sponsored Agenda Building*. Working Paper. The Social Media Listen Center, Clemson University.
- Parmelee, John H (2014): “The agenda-building function of political tweets”. In: *New Media & Society* 16.3, pp. 434–450.
- Pei, Sen/Lev Muchnik/José S. Andrade/Zhiming Zheng/Hernán A. Makse (2015): “Searching for superspreaders of information in real-world social media”. In: *Scientific Reports* 4.1.
- Popken, Ben (2018): *Twitter deleted 200,000 Russian troll tweets. Read them here*. NBC News. URL: https://www.nbcnews.com/tech/social-media/now-available-more-200-000-deleted-russian-troll-tweets-n844731?cid=sm_npd_nn_tw_ma (visited on 11/14/2018).
- Porter, Mason A/Jukka-Pekka Onnela/Peter J Mucha (2009): “Communities in networks”. In: *Notices of the AMS* 56.9, pp. 1082–1097.

- Roeder, Oliver (2018): *Why We're Sharing 3 Million Russian Troll Tweets*. FiveThirtyEight. URL: <https://fivethirtyeight.com/features/why-were-sharing-3-million-russian-troll-tweets/> (visited on 11/14/2018).
- Rosvall, M./C. T. Bergstrom (2008): "Maps of random walks on complex networks reveal community structure". In: *Proceedings of the National Academy of Sciences* 105.4, pp. 1118–1123.
- Savchuk, Lyudmila (2018): *Inside a Russian troll factory*. NBC News. In collab. with NBC News. URL: <https://www.nbcnews.com/think/video/inside-a-russian-troll-factory-1265275459562> (visited on 11/14/2018).
- Stewart, Leo G/Ahmer Arif/Kate Starbird (2018): "Examining trolls and polarization with a retweet network". In: *Proceedings of WSDM workshop on Misinformation and Misbehavior Mining on the Web (MIS2)*. New York: ACM, p. 6.
- Wasserman, Stanley/Katherine Faust (1994): *Social Network Analysis: Methods and Applications*. Structural Analysis in the Social Sciences 8. Cambridge; New York: Cambridge University Press. 825 pp.
- Yang, JungHwan/Sebastian Stier/David Schoch/Franziska Keller (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*.