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 17, 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 **1**

3 Social network analysis of the IRA retweet network **1**

3.1 Data: An Ego-Centered Retweet Network 2

3.2 Methods 4

4 Analysis **7**

5 Discussion **8**

6 Conclusion **8**

7 Code (not included in word count) **8**

References **9**

1 Introduction

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 regularity used in conflicts or during war. "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 online) *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 staring 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 Social network analysis of the IRA retweet network

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 one (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 seen as emerging out of their relations and not vice versa, as in most other quantitative analyses (Wasserman/Faust 1994: 8). This allows us to ignore the question of the actual people 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 handles[1], 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 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

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

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

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. 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.

To further extend our data and our scope of analysis, we are adding additional qualitative information on the IRA trolls, provided by Darren Linvill and Patrick Lee Warren (2018) via the online news outlet FiveThirtyEight (citation). 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, those being *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 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.

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 specifica-

tion 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.

Lastly, 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. That is why it seems to be the right approach to include all of the Tweets in the analysis.

Table 1:

User	N	Senders	Receivers
Troll	404	182	222
Non-Troll	36,485	0	36,485
Total	36,889	333	36,707

Table 2: Troll Statistics

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

3.2 Methods

To understand the IRA's division of labor between the accounts, we begin with testing weather 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

[3]For 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.

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 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_s 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 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

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

[4] A 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.

where x_{ij} 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 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 Analysis

In the preceding chapter, the agenda and methods of our analysis have been introduced. In this chapter, we will present the results of these measurements. First, the results of the modularity community detection algorithm are illustrated and new communities are established, based on these findings. Secondly, graph densities are investigated for these communities.

We run a modularity-based community detection, following the Louvain Method (Blondel et al. 2008). Figure 1 shows the distribution of algorithm communities among the qualitative account categories created by Linvill/Warren (2018). For further analysis, we want to 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 overall method of SNA. To not exceed the scope of this paper, we narrow down our analysis by merging all minor communities below an arbitrary threshold size of 50, since they presumably bear more risk of yielding unrepresentative findings. We then apply the qualitative categorization to our newly distincted communities. In Figure 1, 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 hashtager community is 90% captured by one community, which will be our new hashtager troll community. 90% of the right troll category consist of two major communities with shares of 52% and 38%, which we merge to one new right troll community. We justify this merger by arguing that these two communities could represent distinct right-wing groups underneath a general right-wing community (e.g. Conspiracy Theorist vs. Alt Right User). (Kaiser/Rauchfleisch 2018). We argue, that for these categories, which are overlapping with at least 90%, there is sufficient congruence to continue using the qualitative attributes. The non-english troll category consists of multiple communities, with a biggest community share of 27%. Non of the non-english or other categories passes our threshold community size of 50, which leaves us with four distinct communities: *Right Troll*, *Left Troll*, *Hashtager* and *Other*. It is important 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/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 2, the distribution of account categories among our new communities is depicted, which illustrates this potential bias.

Next, the graph densities are calculated for the subgraphs of our newly determined communities. Figure 3 shows the graph densities for the three communities for both directed and undirected versions of each subgraph. As a mode for creating an undirected graph from a directed one, we chose too 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. This avoids having multiple ties in one relationship. 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 hashtager community is the most dense by a significant margin. Approximately 17.7% 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 consitute 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.

5 Discussion

6 Conclusion

7 Code (not included in word count)

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).

Linvill, 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.

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.

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.