Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)

How to Manipulate Social Media: Analyzing Political Astroturfing Using Ground Truth Data from South Korea

Franziska B. Keller*
HKUST
fbkeller@ust.hk

David Schoch
University of Konstanz
david.schoch@uni.kn

Sebastian Stier
GESIS Cologne
sebastian.stier@gesis.org

JungHwan Yang
Univ. of Wisconsin-Madison
junghwan.yang@wisc.edu

Abstract

This project investigates political astroturfing, i.e. hidden propaganda by powerful political actors aimed at mimicking grassroots activity, on social media. We focus on Twitter accounts used by the South Korean secret service to influence the 2012 presidential elections in favor of the eventual winner, Park Geun-hye. Two independent cluster analyses based on activity patterns of the Twitter accounts and textual features of tweets reliably reveal that there are three groups of NIS accounts, including one group that engages mostly in retweeting, and another group focused on posting news articles with a link. We show that these groups reflect different strategic agendas and correspond to several secret service agents identified in the court documents. We argue that these patterns of coordinated tweeting are consistent with predictions derived from principal-agent theory, and should therefore appear in other astroturfing campaigns as well.

Introduction

Journalists' reports and recent research have spurred fears that political actors may use social media to manipulate public opinion, in particular during the run-up to democratic elections. Candidates in the most recent U.S. presidential election, for instance, were accused of trying to appear more popular on Twitter by using automated followers, and their supporters of spreading so called fake news.[1] Due to the anonymity afforded by the internet, such *political astroturfing* is usually covert, sophisticated, and hard to distinguish from genuine grassroots support.

We study this phenomenon using a unique dataset of political tweets collected in the South Korean 2012 presidential election campaign, during which the National Intelligence Service (NIS) waged a covert social media campaign in favor of the eventual winner, Park Geun-hye. Unlike most research on political astroturfing (Ferrara et al. 2014; Hegelich and Janetzko 2016; Howard and Kollanyi 2016; Ratkiewicz et al. 2011), we focus on non-automated Twitter accounts: accounts where the content is directly entered by human beings most of the time. Such accounts are more likely to pass as ordinary users, and therefore harder to identify and presumably more likely to actually sway the opinion of regular social media users. Previous research on the topic of automated and human-generated astroturfing mostly relied on machine learning algorithms to identify suspicous accounts (Ratkiewicz et al. 2011) or set arbitrary activity thresholds (Howard and Kollanyi 2016). Except for two recent papers (Hegelich and Janetzko 2016; King, Pan, and Roberts forthcoming), there was thus no "ground truth" (or external verification) of the accounts involved in such a campaign. We, however, use two lists of NIS Twitter accounts published in related court proceedings (SeoulHigherCourt 2015; SeoulDistrictCourt 2014).

As our main contributions, we (i) present the first analysis of human-generated astroturfing based on ground truth data on Twitter; (ii) use combined cluster analysis on temporal patterns, account features and textual information to identify the structure of the campaign; (iii) show that a particular astroturfing campaign may use different types of accounts. In future research, we will use the patterns identified to help distinguish astroturfing accounts from regular users.

Data

Our dataset contains approximately 56 million tweets mentioning a wide range of political keywords related to the 2012 presidential election campaign. The data was collected in real-time from 1 June to 31 December 2012 (Song, Kim, and Jeong 2014). The tweets were posted by over one million unique Twitter accounts, some of which were controlled by the NIS. After the opposition discovered this campaign, the prosecutor investigated and published a list of 1,008 Twitter handles used in the NIS astroturfing campaign.[2] Only 170 of these accounts show up in our dataset, of which 50 are only mentioned or retweeted. We suspect that the remaining accounts are trying to "guide" public opinion on other topics, in particular that of North Korea.

These 120 accounts are responsible for 132,154 tweets in our dataset, with the most active account tweeting more than 9,000 times during our research period. They are asso-

* All authors contributed equally to this paper.

[1] http://www.politico.com/story/2016/09/donald-trump-twitter-army-228923

[2] The authors would like to thank Professor Min Song (Yonsei University) for help with the data, the editor of Newstapa, Ki-Hoon Choi, for sharing the court records and providing background information, and Hyeonjong Min for excellent research assistance.

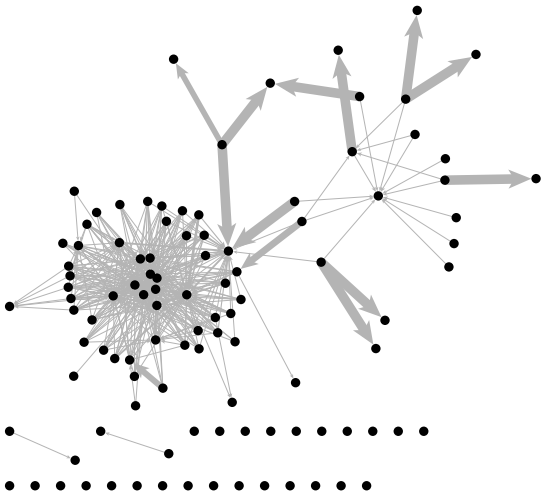

Figure 1: Retweet network of NIS accounts appearing in our dataset. Links going from retweet- to original tweet sender. Link width proportional to number of retweets.

ciated with 106 unique Twitter account IDs, as some of the accounts change their user names during the period of observation. Our approach thus takes the unique Twitter account ID as the unit of analysis and not the account name.

A crawl via the Twitter REST API revealed that of the 106 astroturfing account IDs, only six still existed on 2 January 2017. But even those only sent eight more tweets after 11 December 2012, when the NIS campaign got revealed publicly by the political opposition and journalists.

Figure 1 displays the retweet network among these NIS accounts. The retweet function is the primary mechanism responsible for the rapid information diffusion typical for Twitter (Kwak et al. 2010), and social bots are often used to increase a campaign's outreach by multiplying the message. But many NIS accounts do not retweet or get retweeted at all and remain isolated. The highest retweet count between two NIS accounts is 50, corresponding to 0.23 retweets a day in the investigated time period. This does therefore not appear to be a very well-coordinated or -organized campaign.

Activity-based cluster analysis

To uncover account specific patterns, we use a block clustering approach to simultaneously cluster NIS accounts and their time dependent activities (Hartigan 1972). We turn the daily counts of tweets per NIS account into a matrix x with $n = 106$ rows for each account and $r = 214$ columns for each day between 1 June 2012 and 31 December 2012. An entry x_{ik} is equal to one if account i tweeted at least once on day k and zero otherwise. An expectation-maximization algorithm (EM) based on a block mixture model is then used to determine the clustering (Govaert and Nadif 2005). The block mixture model has the following probability density function:

$$f(x;\theta) = \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} \prod_i p_{z_i} \prod_j q_{w_j} \prod_{i,j} \varphi(x_{ij};\alpha_{z_i w_j}).$$

The variables are as follows:

- z represents the partition of rows in g clusters, i.e. for each row i, $z_i \in \{1,\dots,g\}$. Equivalently, w represents the partition of columns in m clusters.
- $\mathcal{Z}$ and $\mathcal{W}$ denote the set of all possible partitions z and w.
- $p = (p_1,\dots,p_g)$ and $q = (q_1,\dots,q_m)$ are vectors of probabilities p_k and q_l that a row and a column belong to the kth row and to the lth column cluster respectively.
- $\varphi(x;\alpha_{kl})$ is a probability density function.
- θ is the vector of parameters, i.e. $\theta = (p, q, \alpha_{11},\dots,\alpha_{gm})$

To estimate the parameters, we maximize the classification log-likelihood, defined as

$$L_C(z,w;\theta) = L(\theta;x,z,w) = \log(f(x,z,w;\theta)$$

The conditional expectation of the classification log-likelihood with a previous estimate $\theta^{(c)}$ is given by

$$\begin{aligned} Q(\theta,\theta^{(c)}) = &\sum_{i,k} P(z_{ik}=1|x,\theta^{(c)}) \log p_k \\ &+ \sum_{j,l} P(w_{jl}=1|x,\theta^{(c)}) \log q_l \\ &+ \sum_{i,j,k,l} P(z_{ik}w_{jl}=1|x,\theta^{(c)}) \log \varphi(x_{ij};\alpha_{kl}) \end{aligned}$$

The maximization is carried out iteratively, alternating between maximizing $Q(\theta,\theta^{(c)})$ contingent on row clusters and column cluster respectively.

Figure 2 shows the optimal clustering of NIS accounts with $g = 3$ and $m = 3$, according to the Integrated Classification Likelihood (ICL), together with their daily activity. The clustering reveals three groups, each engaging in different tweeting behaviors at different times. The green cluster almost exclusively retweets, starts its activity on 1 September, and stops retweeting (like all other NIS accounts) after 11 December, the day the opposition party drew the prosecution's attention to the NIS activities. Green cluster accounts post on average 0.89 retweets per day (and about double that amount in their main activity period), while the other accounts retweet less than a quarter of that amount on average.

The blue cluster contains what we suspect to be so-called "cyborg" accounts (Chu et al. 2012): accounts that are at least partially automated and post a large number of tweets even throughout the night. They post on average more than 20 original tweets per day - but even their most active account would not meet the bot-threshold of 50 posts a day suggested by Howard et al. (Howard and Kollanyi 2016). Accounts in the other two clusters post less than one tweet on average. Accounts in the blue cluster pick up relevant headlines shortly after a news web site posts them, and disseminate the title and the link to the story. As a result, around 90% of their tweets contain a shortened URL. The red cluster, finally, contains not very active accounts and those that start tweeting only during November.

Text-based cluster analysis

We next use automated text analysis to understand what the different clusters are tweeting about. From among the words

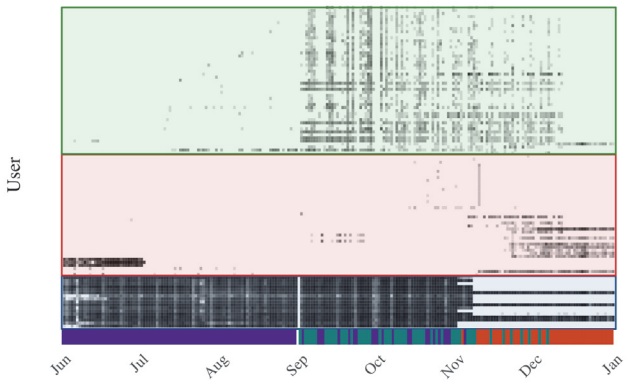

Figure 2: Activity patterns of NIS accounts and clusters obtained by block clustering. Shade of grey in each cell indicates the tweeting frequency by the give account and day. Inset at the bottom shows clustering of dates.

used by the NIS accounts, we first removed URLs, hashtags, @-mentions and retweets of other Twitter accounts. Then we constructed a stemming dictionary to remove postpositions and functional characters in Korean language. Next, we transformed the text into a quantitative feature by tokenizing and creating a user-term matrix u where rows correspond to the 106 NIS accounts and columns represent the remaining 309,229 terms used in all 132,154 tweets. We removed 308,112 terms that appeared in less than 0.1% of the tweets from the user-document matrix, since such rare terms only increase the dimension of the feature vector without contributing to predictive power. The number of remaining terms in the document after removing sparse terms is 1,117.

The user-term matrix is then used to group the NIS accounts according to their frequently used terms, i.e. to distinguish accounts according to the topics they discuss. The similarity of frequently used terms of two users i and j is calculated with the cosine similarity

$$sim_{ij} = \cos(\beta_{ij}) = \frac{\sum_k u_{ik} u_{jk}}{\sqrt{\sum_k u_{ik}^2}\sqrt{\sum_k u_{jk}^2}}.$$

Using a hierarchical clustering approach, we again group the NIS accounts into three groups. Table 1 shows a comparison of the activity-based and text-based clustering approaches.

		activity		
		green	blue	red
text	yellow	48	1	23
	brown	0	16	4
	orange	0	0	14

Table 1: Confusion Matrix comparing text based and activity based clusterings. Group names refer to colors used in Figure 3.

The green and blue activity-based clusters are almost perfectly preserved by the text-based clustering. That is, not only do the accounts have similar activity patterns, they also tweet about similar topics. In contrast, the accounts from the red activity-based cluster are spread across all text-based clusters, i.e. they have a more evenly distributed vocabulary.

The results of a *term frequency–inverse document frequency* (TF-IDF) analysis indicate that the red cluster differs from the other two clusters in meaningful ways. Interestingly, the word with the highest TF-IDF value in the red cluster is "lol", indicative of less formal languate than that used in news headlines featured in the blue and the green clusters. In addition, highly distinguishing words in the red cluster are heavily ideologically charged and the ones that are commonly used for name-calling liberals, such as "acting like a communist", "left-wing media", and terms connected with North Korea. The words defining the blue cluster of potential cyborgs, on the other hand, are mainly generic political terms, and words associated with online news, such as "click to read" and the name of a news organization, "Newsis". The green cluster also includes many political keywords and names of news organizations.

It thus appears as if the agents engaged in this political astroturfing campaign do not form a uniform mass. Instead, they are assigned to groups with different tasks: some spread relevant news articles (the blue cluster), while others amplify messages created by other NIS accounts or regular users sympathetic to the cause (the green cluster). The red cluster posts very straightforward and ideologically slanted messages attacking non-conservative politicians and liberal presidential candidates.

Comparison of cluster results

Automated astroturfing campaigns often have their bots post the same message or the same keywords at the same time, in the hope of setting the online discussion agenda and influencing what appears in the "trending now" section. As it turns out, it is fairly common for NIS accounts to tweet in the same minute: 99 of 109 have posted at least one tweet in the exact same minute as another NIS account, the average pair "co-tweets" more than 50 tweets (maximum: 5,027 instances). This is particularly common for the very active accounts in the blue cluster. A manual inspection of such instances reveals that they are indeed almost always identical tweets.

Figure 3 displays this "coordinated tweeting network". The placement of the nodes is determined by the backbone layout as implemented in visone. In Panel (a), the nodes are colored according to the activity-based clustering in Figure 2. Nodes of the same color appear placed together, indicating the similarity of not only their daily activity pattern, but also that on a much more fine-grained time scale. A similar picture emerges in Panel (b), in which node colors are assigned according to the text-based clustering.

For an explanation of this consistent pattern, we turn to our ground truth, the court proceedings describing how 30 different NIS agents managed 1,008 Twitter accounts. In Panel (c), we find that some agents are indeed responsible for distinct clusters, such as the clique in the top left corner. The same is true for the most active accounts (blue in (a), brown in (b)) at the center of the network. Our results suggest that the agent responsible "copy and pastes" the same

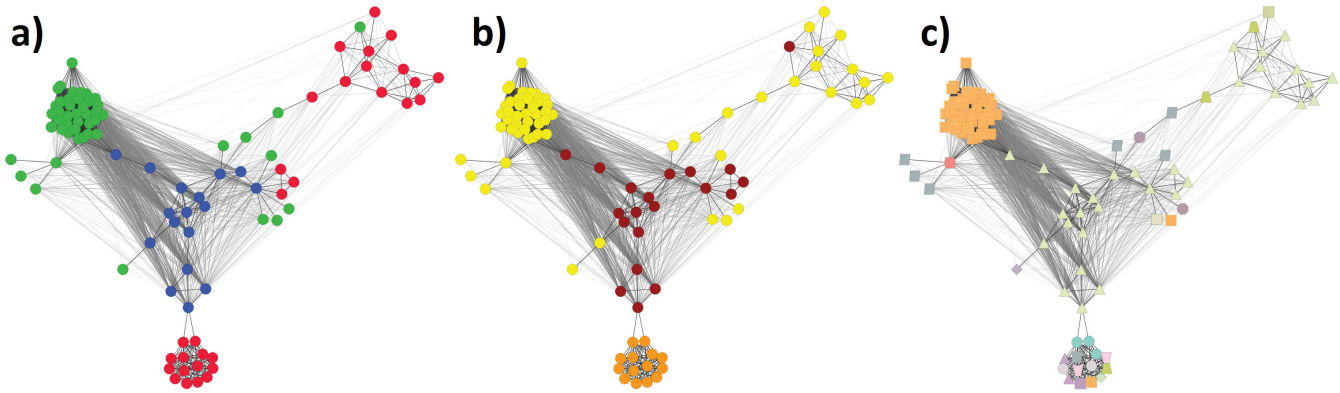

Figure 3: Coordinated tweeting among NIS accounts. A link between two accounts is present if they posted at the same time at least 5 times. Node attributes are assigned according to (a) activity-based clusters; (b) text-based clusters; (c) NIS agent information from the court records.

or similar messages into his or her multiple accounts. The exception is a cluster at the bottom which is operated by a heterogeneous set of accounts. However, their cooperation is limited to coordinated tweeting during one day only, so it is possible that one agent was taking over a group of accounts for just that one occasion.

Conclusion

This paper analyzed an astroturfing campaign attempting to manipulate election-related conversations on social media. Two independent cluster analyses based on activity patterns and tweet contents grouped the accounts responsible into very similar clusters. An in-depth text analysis showed that differences between the three clusters depict different persuasion strategies. Finally, the information in the court proceedings linking agents with accounts suggests that clusters are primarily detectable because an agent operates most of the accounts with similar behavior.

Our study indicates that while actors involved in an astroturfing campaign behave in a similar manner, division of labor may also create different patterns within the same campaign. This means that "one size fits all" approaches to astroturfing detection will certainly miss its different facets. In future research, we will use the patterns discovered in this paper to attempt to distinguish NIS accounts from regular accounts and identify additional NIS accounts.

While each astroturfing campaign likely leaves different traces, there is an underlying principal-agent logic behind them: the human astroturfers react to incentives, central commands, and the bureaucratic structure surrounding them – which is why astroturfing accounts of one campaign tend to act in similar and repetitive ways. We're therefore confident that pattern detection grounded in principal-agent theory is well-suited for other cases as well.

References

Chu, Z.; Gianvecchio, S.; Wang, H.; and 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.

Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; and Flammini, A. 2014. The rise of social bots. *arXiv preprint arXiv:1407.5225*.

Govaert, G., and Nadif, M. 2005. An EM algorithm for the block mixture model. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(4):643–647.

Hartigan, J. A. 1972. Direct clustering of a data matrix. *Journal of the American Statistical Association* 67(337):123–129.

Hegelich, S., and Janetzko, D. 2016. Are social bots on Twitter political actors? Empirical evidence from a Ukrainian social botnet. In *Proceedings of the 10th International AAAI Conference on Web and Social Media*, 579–582. Palo Alto, CA: AAAI Press.

Howard, P. N., and Kollanyi, B. 2016. Bots, #Strongerin, and #Brexit: Computational propaganda during the UK-EU Referendum.

King, G.; Pan, J.; and Roberts, M. E. forthcoming. How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. *American Political Science Review*.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*. New York, NY: ACM. 591–600.

Ratkiewicz, J.; Conover, M.; Meiss, M.; Gonçalves, B.; Flammini, A.; and Menczer, F. 2011. Detecting and tracking political abuse in social media. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 297–304. Palo Alto, CA: AAAI Press.

SeoulDistrictCourt. 2014. Case ID: 2013GoHap577, 2013GoHap1060.

SeoulHigherCourt. 2015. Case ID: 2014No2820.

Song, M.; Kim, M. C.; and Jeong, Y. K. 2014. Analyzing the political landscape of 2012 Korean presidential election in Twitter. *IEEE Intelligent Systems* 29(2):18–26.