A Synthesis on Social Media and the Social Bots that Inhabit Them

By Stephen Scarano

The imitation of human behavior remains a fixation in popular media, academic curiosity, and the culture at large; however, strides in accessibility–in tandem with powerful economic and political incentives–drive its expansion into social media timelines.

I – Terminology and Background

Bots, automated accounts developed and set free to post, are neither particularly new nor inherently malicious [3]. Plenty extend welcome services to their respective communities by aggregating data, providing aid, assisting research or disseminating time-sensitive information. As an example, it is common to find corporations soliciting bots for online customer service interactions [16]. In many cases, we could potentially identify a chatbot of this kind as a *social bot*.

While there does not currently exist an academic consensus on the term, social bots are widely considered to be accounts or algorithms attempting to be perceived as human [16]. A news aggregator on Twitter is (likely) not designed to mislead others to its identity, but a company chat bot, for instance, may. It is the worth noting that not all social bots are malicious and not all conventional bots are benign: the former may exist for research, satirical, or public arts purposes while the latter has taken the form of spamming, scamming, and malware. Benign bots are diverse and situational, but malicious ones may be divided into two main disciplines: *commercial bots*, comprised of the aforementioned malware or spam algorithms, and *sybils*, automatons intended to perform a false identity [16]. As this paper is most concerned with the implications of sybils in particular, it is worth specific note to define them as:

(a) Imitating Human Behavior
(b) Inherently Malicious

For summary, we may categorize all automated accounts on the axes of imitation and intent [16] as follows:

	Malicious Intention	**Neutral Intention**	**Benign Intention**
High Imitation of Human Behavior	Astroturfing bots, Social botnets in political conflict, organization or company infiltration bots, sybils, impersonation/ doppelgänger bots	Humorist Bots	Chat Bots
Low Imitation of Human Behavior	Spam, Pay bots	Nonsense Bots	News, Recruitment, Public Dissemination, Earthquake Warning, Editing, Anti-Vandalism

Intention, however, must not be confused for consequence: the relatively recent rollout of complex bot algorithms has given rise to striking disruptions in political and economic institutions. In 2010, the Dow Jones suddenly plunged nearly 10%, the largest one day decline in its history, likely as a result of herding error from high-frequency trading bots [21]. In finance, herding is defined as widespread investment guided not by individual market calculations or analysis but rather through mimicry of other investors. These bots, noticing the rapid trading of other bots, joined in chain reaction to drive up stock values until the bubble inevitably burst.

Automatic trading programs remain a vital arm of finance, but there is evidence to suggest that even independent of malfunction they are vulnerable to external manipulation. Similar algorithms scouting online market chatter were misled by an orchestrated bot campaign to invest in a mysterious tech company by the name of Cynk. It was not until the market value had increased 200-fold ($5 billion) that analysts recognized the scheme [3]. The report is anecdotal but doubtless a microcosm of a pervasive crisis in online authenticity: *digital astroturfing*.

Digital astroturfing is described in [2] as “a form of manufactured, deceptive, and strategic top-down activity on the Internet initiated by political actors that mimics bottom-up activity by autonomous individuals”. While the terminology of social bots among researchers remains to some degree colloquial and diverse, the same cannot be said of astroturfing which is partitioned into the following intuitive categories [2]:

I. **Political Actors**: A political group with specific and directed interest. Often researchers may disagree as to whether purely commercial groups can participate in astroturfing campaigns.
II. **Targets**: A recipient of behavior from targets. Researchers distinguish between specific political actors and the public at large as they play dissimilar roles in the execution of online astroturfing.
III. **Goal**: Communication of positive or negative valence (peripheral messaging) on either political actors or policy.

In short, political actors create (or hire) agents in an effort to influence the public with respect to a political actor or policy. Accounts affiliated with an astroturfing project on Twitter may coordinate with one another through joint retweeting of external posts (with which their interests align), retweeting one another’s posts, and posting identical messaging seemingly independently (co-tweeting) [1]. Researchers in [1] also suggest a number of central constraints of the approach:

(a) Unless planned years in advance (a rarity), accounts must be created within a shorter time period in aggregate in comparison to conventional accounts
(b) Accounts must start and stop posting about similar topics as they are informed by (or children of) a central authority, from which they take orders
(c) Attempts to conceal an astroturfing campaign will likely also limit its progress

II – Case Study and Analysis

Among the earliest of examples, the 2012 South Korean election in which National Intelligence (NIS) agents operated an expansive disinformation campaign in support of candidate Geun-hye Park, is uniquely informative of the architecture and distribution of labor within a digital astroturfing campaign. As a result of the public nature of the scandal, court proceedings remain available to the public that identify 1,008 accounts allegedly involved in the affair [10].

Researchers found that campaign-affiliated profiles frequently posted identical content in short time-spans and social network densities for message coordination differed significantly from those of conventional users. Court documents confirm that supervisors incentivized retweets, likes, and posts over convincing account profiles. Ironically, while NIS accounts had considerably more followers than conventional users, they received less mentions (@-tags), and retweets attracted outside of the campaign (40% of total) comprise a minuscule fraction of total retweets from the sample data provided from the election cycle [1]. As consequence, there is little evidence to suggest that the campaign significantly influenced online discussion of the 2012 South Korean election.

The authors of [17] pursue bot identification in an unsupervised manner: their research seeks to determine the impact of social bots on online discussion over the course of the 2016 presidential election. Collecting 20 million tweets (2.8 million users), researchers classified social bot accounts through feature-based machine learning algorithms; of the 2.8 million users, 400,000 are suspected bot accounts (3.8 million tweets). They then compared the two sets of accounts (that being conventional and social bot) along temporal, geographic, partisan, engagement, and sentiment lines:

Temporal: When observing period of high interest (peaks in tweets posted), humans appear to post disproportionately. Suspected bots, conversely, post at more or less the same frequency over all periods, resembling noise.
Geographic: Human profiles tend to map, unsurprisingly, to U.S population densities (see Figure 2). Suspected bots show strong bases in midwestern and southern states, particularly Georgia (see Figure 3).

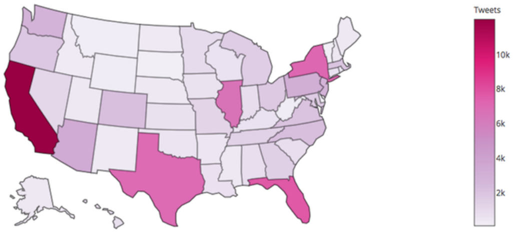


Figure 2: Human Accounts by Claimed Geography [17]

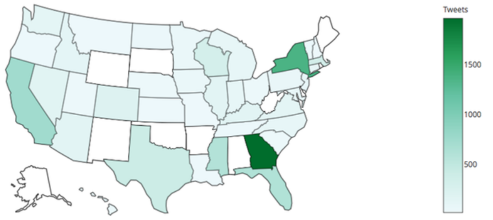
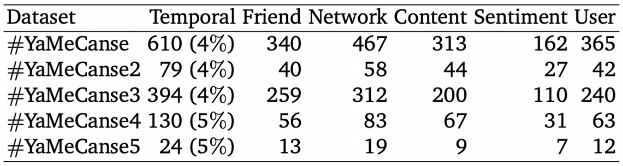


Figure 3: Suspected Bot Accounts by Claimed Geography [17]

Partisan: There exists minimal difference between the sets with respect to partisanship. Of Clinton supporters, 9% are suspected bots which is comparable to Trump's 12% of bots.
Engagement: Human profiles appear to engage significantly more with other humans than with bots, while the latter prefers other bot accounts; however, there is no substantial difference in the amount of retweets between humans and bots. Interestingly, both groups retweet one another at nearly the same rate.
Sentiment: Post from Trump users are notably more likely to express positive sentiment than Clinton supporters; however, artificial Trump accounts are also significantly more positive than their human counterparts. The overwhelming tone of the 2016 dataset is negative, suggesting some level of potential distortion.



Dataset	Temporal	Friend	Network	Content	Sentiment	User
#YaMeCanse	610 (4%)	340	467	313	162	365
#YaMeCanse2	79 (4%)	40	58	44	27	42
#YaMeCanse3	394 (4%)	259	312	200	110	240
#YaMeCanse4	130 (5%)	56	83	67	31	63
#YaMeCanse5	24 (5%)	13	19	9	7	12

Figure 4: Number of accounts scoring above a BotOrNot score of 0.65 according to each feature [18]

Examination of the #YaMeCanse online protest movement within Mexico highlights similar artificial influence. Over the course of the campaign, users claimed that astroturf accounts began gaming the hashtag to seemingly drown out discourse. In response, protesters flocked from #YaMeCanse to #YaMeCanse2, #YaMeCanse3, etc., maintaining significant support until #YaMeCanse25 [18]. Researchers collected a 152,757 tweet database and used a publicly available machine learning service, *BotOrNot,* to classify accounts as authentic or artificial. Results indicate that there is indeed a significant concentration of bots within the set (see Figure 4), likely a lower bound as 10-14% of the set is deleted.

III – Identification Measures

The last of the two case studies are instances of the greater astroturf problem: how may a researcher identify social bot accounts without access to public records identifying key patterns of each agent. Essentially, is it possibly to generalize the identification problem?

Researchers of the South Korea case were able to identify suspect NIS accounts through comparison of confirmed bots with existing accounts in their database. Examination of co-tweet networks yields an additional 662 suspect accounts. Further analysis of temporal and content patterns demonstrates strong similarity to confirmed NIS accounts [1]. While the ground-truth values (court documents) are unavailable to researchers in the vast majority of astroturf

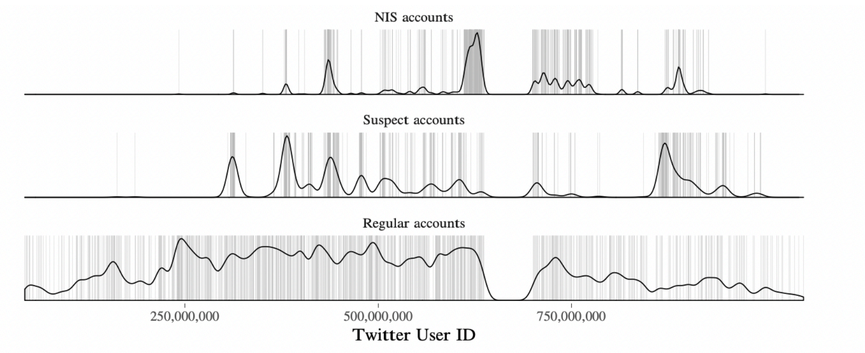


Figure 5: Distribution of Twitter IDs by category [1]

studies, the authors suggest that the findings remain informative to the task: by shifting detection criteria away from bot-identification to a more nuanced inspection of account coordination, future researchers may find greater success [1]. Analysis of social structures have proven effective in detection algorithms such as *SybilRank* [3].

Further academic inspection confirms that a singular approach to identification is unlikely to capture the breadth of a campaign. While behavior is often similar amongst actors, the scale of the larger projects necessarily entail a division of labor that disrupt its hegemony; this would suggest that any one-size-fits-all approach may be prone to miss nuances in the conspiracy [10].

Rather than frame astroturf campaigns through broad network coordination, recall that the majority of campaigns are operated by limited agents. Therefore, we may expect to observe widespread linguistic similarity amongst clusters of posts: essentially, an authorship attribution problem. If provided a set of written works from a particular author, we may generate an n-gram “spectra”, a range of n-gram groupings of the same text. For background, an n-gram is a partition of text separated by every n words. Through inventing a set of some number of n-grams, we conceivably produce an author profile with which we may compare to others. We may determine similarity of the spectra through k-nearest neighbors; however, the required data may be often inaccessible [11].

N value	**N-gram**
2	“I am”, “am not”, “not a”, “a social”, “social bot”
3	“I am not”, “am not a”, “not a social”, “a social bot”
4	“I am not a”, “am not a social”, “not a social bot”
5	“I am not a social”, “am not a social bot”

Figure 6: Example of n-gram spectra for the sentence “I am not a social bot” for n-values 2 to 5

Simpler machine learning measures are commonplace in social bot identification. *BotOrNot* and *BotOMeter* are popular instances of classification-based detection models [18] [19]: provided some set of features (number, likes, bio, number of friends) and training information (known bots and their corresponding features), the model separates authentic and artificial profiles. Humans and bots tend to cluster separately across many behaviors: the latter trend lower in followers, less likely to receive retweets, and half as likely to mention other users [15]. Recent analysis has applied similar models to identify cases of human users masquerading under false identity but with little success [12].

In situations of little available training data, a simpler approach proves fruitful. Crowdsourcing detection describes a process of democratic classification from human agents between human and bot profiles. After constructing an online Turing test of artificial and conventional accounts, researchers discovered that while accuracy drops overtime, choosing the majority evaluation yields a near-zero false negative rate [3].

However, the process is resource intensive, scaling poorly for the average platform unless implemented at an early stage [3].

IV – The Wider Social Context

Interacting in the wider social space, a clever astroturf agent may disrupt online protest, confuse election expectations, or magnify the impact of disinformation [18] [17] [1]. Researchers examining the spread of misleading news note that a small number of accounts were responsible for a large proportion of shares, in some cases shortly after the original posting of the article. Subsequent *BotOrNot* analysis found that these profiles in aggregate had significantly higher bot scores than the broader public. Spreader accounts tend to target users with higher median follower counts and their geographical distribution appears inconsistent with the wider distribution of all posters [19].

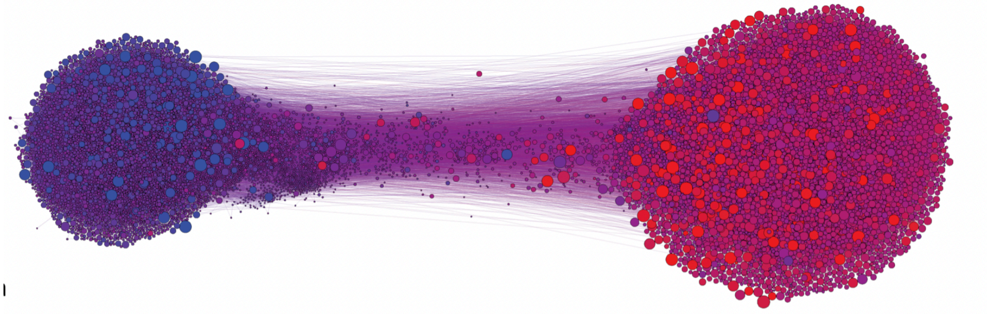


Figure 7: Graph of Twitter partisanship (color) with respect to propensity towards distributing defunct articles (size)

Online platforms remain opaque in their handlings of astroturf campaigns. Transparency is a rarity on such platforms: ad placement and disclosures all suffer from systemic dysfunction. Facebook's ad explanations, provided as a data transparency initiative, remain consistently incomplete and likely inaccurate [20]. Ad disclosures suffer from similar inconsistencies, and it is worth noting that users presented with two advertisements posts, varying in likes, tend to trust the most popular of the two [8] [5] [4] [14]. Automated spread of sensationalist and misleading articles may potentially snowball existing correlations between partisanship and disinformation [21].

V – Further Study

Analysis of the current literature dictates a broad range of implications. Further research is needed to determine the extend of artificial impact on human behavior, and researchers should take care to uncover which behaviors seem most predictive of social bot success. The broader landscape of social media demands scrutiny: to what degree can platforms identify false interactions, and what measures currently exist that can be implemented? Perhaps more specifically, a timely investigation into the implications of artificial interactions on partisan perception is critical to predict future outcomes of online politics.

References

[1] Franziska B. Keller, David Schoch, Sebastian Stier & JungHwan Yang (2020) Political Astroturfing on Twitter: How to Coordinate a Disinformation Campaign, Political Communication, 37:2, 256-280, DOI: 10.1080/10584609.2019.1661888.

[2] Kovic, M., Rauchfleisch, A., Sele, M. and Caspar, C. 2018. Digital astroturfing in politics: Definition, typology, and countermeasures. *Studies in Communication Sciences*. 18, 1 (2018), 69–85. DOI:https://doi.org/10.24434/j.scoms.2018.01.005.

[3] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. Commun. ACM 59, 7 (July 2016), 96–104. DOI:https://doi.org/10.1145/2818717.

[4] Marcia W. DiStaso, Denise Sevick Bortree. 2012. Multi-method analysis of transparency in social media practices: Survey, interviews and content analysis,. Public Relations Review, 38:3. DOI: https://doi.org/10.1016/j.pubrev.2012.01.003.

[5] Carter Ben. The Illusion of Transparency on Social Media. 2009. https://www.searchenginepeople.com/blog/the-illusion-of-transparency-in-social-media.html

[6] Grabowicz, P.A., Romero-Ferrero, F., Lins, T., Benevenuto, F., Gummadi, K., & Polavieja, G.G. (2015). Bayesian Social Influence in the Online Realm. arXiv: Physics and Society.

[7] Deibert, R.J. (2019). The Road to Digital Unfreedom: Three Painful Truths About Social Media. *Journal of Democracy 30*(1), 25-39. doi:10.1353/jod.2019.0002.

[8] Maheshwari Sapna. Endorsed on Instagram by a Kardashian, but Is It Love or Just an Ad? 2016. https://www.nytimes.com/2016/08/30/business/media/instagram-ads-marketing-kardashian.html.

[9] Neyland, D. (2016). Bearing Account-able Witness to the Ethical Algorithmic System. *Science, Technology, & Human Values*, *41*(1), 50–76. https://doi.org/10.1177/0162243915598056.

[10] Yang, JungHwan & Keller, Franziska & Schoch, David & Stier, Sebastian. (2017). How to Manipulate Social Media: Analyzing Political Astroturfing Using Ground Truth Data from South Korea. Proceedings of the Eleventh International AAAI Conference on Web and Social Media.

[11] Peng, J, Detchon, S, Choo, K-KR, Ashman, H. Astroturfing detection in social media: a binary n-gram–based approach. *Concurrency Computat: Pract Exper.* 2017; 29: e4013. https://doi.org/10.1002/cpe.4013.

[12] E. Van Der Walt and J. Eloff, "Using Machine Learning to Detect Fake Identities: Bots vs Humans," in *IEEE Access*, vol. 6, pp. 6540-6549, 2018, doi: 10.1109/ACCESS.2018.2796018.

[13] Chong Oh, Yaman Roumani, Joseph K. Nwankpa, Han-Fen Hu. Beyond likes and tweets: Consumer engagement behavior and movie box office in social media. 2017. Information & Management, 54, 1. DOI: https://doi.org/10.1016/j.im.2016.03.004.

[14] Seo, Y., Kim, J., Choi, Y.K. and Li, X. (2019), "In “likes” we trust: likes, disclosures and firm-serving motives on social media", *European Journal of Marketing*, Vol. 53 No. 10, pp. 2173-2192. https://doi.org/10.1108/EJM-11-2017-0883.

[15] Stieglitz S., Brachten F., Berthelé D., Schlaus M., Venetopoulou C., Veutgen D. (2017) Do Social Bots (Still) Act Different to Humans? – Comparing Metrics of Social Bots with Those of Humans. In: Meiselwitz G. (eds) Social Computing and Social Media. Human Behavior. SCSM 2017. Lecture Notes in Computer Science, vol 10282. Springer, Cham. https://doi.org/10.1007/978-3-319-58559-8_30.

[16] Stieglitz, S., Brachten, F., Ross, B., & Jung, A. (2017). Do Social Bots Dream of Electric Sheep? A Categorisation of Social Media Bot Accounts. ArXiv, abs/1710.04044.

[17] Bessi, Alessandro and Ferrara, Emilio, Social Bots Distort the 2016 US Presidential Election Online Discussion (November 7, 2016). First Monday, Volume 21, Number 11 - 7 November 2016, Available at SSRN: https://ssrn.com/abstract=2982233.

[18] Suárez-Serrato P., Roberts M.E., Davis C., Menczer F. (2016) On the Influence of Social Bots in Online Protests. In: Spiro E., Ahn YY. (eds) Social Informatics. SocInfo 2016. Lecture Notes in Computer Science, vol 10047. Springer, Cham. https://doi.org/10.1007/978-3-319-47874-6_19.

[19] Shao, Chengcheng & Ciampaglia, Giovanni & Varol, Onur & Flammini, Alessandro & Menczer, Filippo. (2017). The spread of fake news by social bots.

[20] Athanasios Andreou, Giridhari Venkatadri, Oana Goga, Krishna Gummadi, Patrick Loiseau, et al.. Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook's Explanations. *NDSS 2018 - Network and Distributed System Security Symposium*, Feb 2018, San Diego, United States. pp.1-15.

[21] Nikolov, D., Flammini, A., & Menczer, F. (2021). Right and left, partisanship predicts (asymmetric) vulnerability to misinformation. *Harvard Kennedy School (HKS) Misinformation Review*. https://doi.org/10.37016/mr-2020-55.