When Bad Actors Adhere to Group Norms

Extended Abstract

Leo G. Stewart
Human Centered Design & Engineering
University of Washington
lgs17@uw.edu

Ahmer Arif
Human Centered Design & Engineering
University of Washington
ahmer@uw.edu

Kate Starbird
Human Centered Design & Engineering
University of Washington
kstarbi@uw.edu

ABSTRACT

The term, "bad actor" can suggest an individual that is easily identifiable by their offensive or antisocial behavior—perhaps as suggested by the growing focus on online harassment. In this paper, we examine a set of accounts that do not fit this image. These accounts do not necessarily engage in vulgarity or abuse, but rather purposeful, targeted, and systematic manipulation. Consequently, we take the position that a useful definition of bad actors has to consider not just behaviors, but intent. We also argue that social media companies need to move beyond sanctioning "bad actors" to helping users understand our vulnerabilities within these information ecosystems.

CCS CONCEPTS

• **Human-centered computing → Social media**; **Empirical studies in collaborative and social computing**;

KEYWORDS

Social media, Black Lives Matter, Twitter, trolling

ACM Reference Format:
Leo G. Stewart, Ahmer Arif, and Kate Starbird. 2018. When Bad Actors Adhere to Group Norms: Extended Abstract. In *Proceedings of ACM Woodstock conference (CHI'18).* ACM, New York, NY, USA, Article 4, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 GOVERNMENT-AFFILIATED "BAD ACTORS"

In November, 2017, Twitter released a list of 2,752 accounts known to be affiliated with the Internet Research Agency (IRA-RU), an entity accused of operating Russian accounts that use a variety of automated and non-automated strategies to influence online discourse [5, 6, 8]. Recognizing some of the Twitter handles from our prior work, we cross-referenced this list with a dataset examining frames in #BlackLivesMatter and #BlueLivesMatter discourse in the context of police-related shootings [9]. This dataset contained 248,719 tweets from 160,217 accounts collected over a period of 9 months in 2016 with tweets matching shooting-related keywords such as "shooting" or "gun man" and at least one of the phrases "blacklivesmatter", "bluelivesmatter", or "alllivesmatter". Within this collection, we identified 96 IRA-RU accounts. To understand how these accounts contributed to the informational ecosystem,we constructed a directed retweet graph, setting a threshold of a retweet degree of at least two (meaning that an account retweeted or was retweeted at least twice) with the goal of revealing more established information channels and mitigating the effect of viral tweets. To see close-knit communities in the graph, we used the highest-level community assignments generated by the Infomap optimization of the map equation [2, 7]. Our final step was to use the Force Atlas 2 algorithm in Gephi [1] to visualize the network graph and communities. Most immediately, the resulting graph (shown in Figure 1) and communities echoed our prior work by manifesting an extremely polarized information space divided into two distinct communities (purple and green). To contrast the two communities, we applied the community categorization methods from our prior work, using hashtag frequencies in the accounts' aggregated Twitter bios to reveal the accounts' affiliations and supplementing this data with frequently retweeted and followed accounts by each community. As shown in Table 1, we see that the purple community expresses alignment with hashtags like #blacklivesmatter, #imwithher, and #blm and thus categorize this community as broadly politically left-leaning. Similarly, turning to the green community, we see that hashtags related to Donald Trump's presidential campaign and gun rights are prevalent (#trump2016, #maga, #2a) and categorize this community broadly politically right-leaning.

Having established the context of the information space, we next locate the RU-IRA accounts in the graph.

Of the 96 accounts identified in the broader dataset, 29 accounts were present in the subset shown in the retweet graph, with 22 troll accounts in the left-leaning community and 7 troll accounts in the right-leaning community. These accounts demonstrated a wide range of influence in this heavily curated dataset—@BleepThePolice was retweeted 702 times by 614 distinct accounts on our graph while six troll accounts were not retweeted at all. Table 2 shows the top 5 most-retweeted accounts across both clusters.

Figure 2 shows that the IRA-RU accounts are located far from the center of the graph and their retweets spread through but not across communities. This suggests that there are two distinct groups of actors - tied to the same agency, and perhaps even colocated - that systematically participated in both sides of this polarized conversation. Qualitative analysis of the content shared by these accounts is the subject of future work, but we have observed that the IRA-RU accounts took advantage of the divided conversation by producing and amplifying content that's aligned with each audience's preferences. For example, accounts in the right-leaning cluster shared

CHI'18, April 2018, Montréal, Canada

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

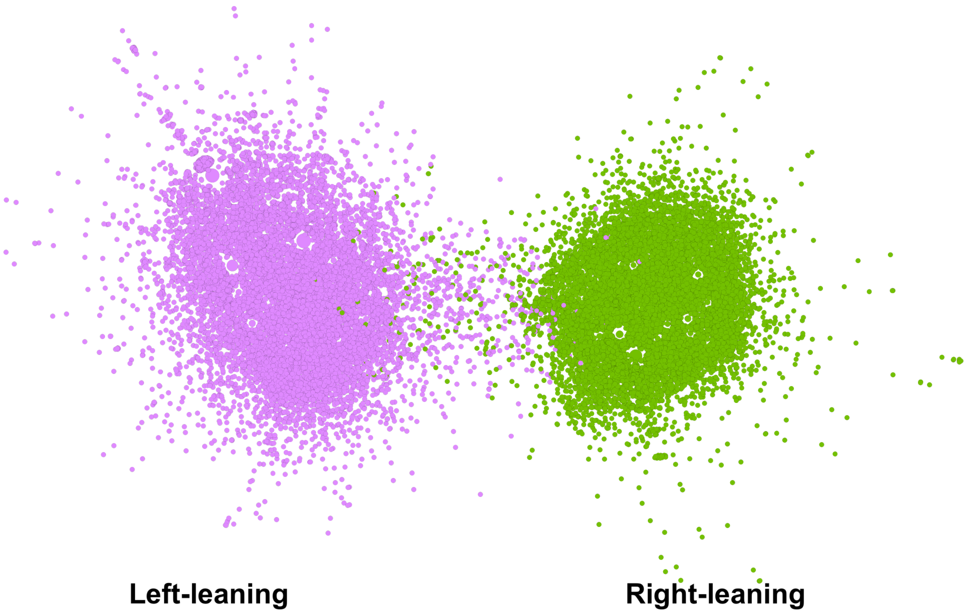

Figure 1: The retweet graph shows two distinct clusters.

Table 1: Classification of Clusters

Cluster	Top 10 hashtags in account descriptions	Size	Top 10 most retweeted	Top 10 accounts by follower count
Purple	blacklivesmatter (8.529%), imwithher (1.442%), blm (1.105%), uniteblue (1.039%), feelthebern (1.021%), allblacklivesmatter (0.721%), bernieorbust (0.599%), neverhillary (0.571%), nevertrump (0.571%), freepalestine (0.524%)	10681	trueblacknews (3773), YaraShahidi (2108), ShaunKing (1553), ShaunPJohn (1214), BleepThePolice (692), Crystal1Johnson (573), DrJillStein (524), meakoopa (409), kharyp (387), tattedpoc (307)	YouTube, ABC, ELLEmagazine, RollingStone, USATODAY, YourAnonNews, RickeySmiley, globeandmail, ntvkenya, BigBoi
Green	trump2016 (6.615%), maga (6.099%), 2a (5.237%), tcot (2.787%), trump (2.776%), neverhillary (2.524%), makeamericagreatagain (2.461%), nra (2.229%), trumptrain (1.998%), bluelivesmatter (1.872%)	9509	PrisonPlanet (4945), Cernovich (1704), LindaSuhler (1034), MarkDice (789), DrMartyFox (758), _Makada_- (591), andieiamwhoiam (510), LodiSilverado (500), BlkMan4Trump (458), JaredWyand (447)	Newsweek, Independent, michellemalkin, AppSame, VOANews, theblaze, RealAlexJones, BraveLad, AnthonyCumia, NY1

memes about #BlackLivesMatter activists celebrating the death of police officers after a Baton Rouge shooting, whilst accounts on the left shared content about police killing an elderly African American man whilst shouting racial epithets.

This points to a vulnerability in audience-driven information systems. On Twitter, where narratives and frames can be crowd-constructed [3], RU-IRA accounts and other bad actors can access the framing discourse by constructing personas that blend in with

Table 2: Top-5 RU-IRA Accounts across both clusters by Retweets

Handle	Tweet Count	Total retweets on graph	# Accounts who retweeted	% Retweets by Left	% Retweets by Right	Retweet rank[1]
BleepThePolice	18	702	614	86.2	0.427	10
Crystal1Johnson	14	585	462	76.9	0.855	12
BlackNewsOutlet	2	63	57	85.7	3.17	35
SouthLoneStar	2	235	232	0.851	94.9	130
gloed_up	15	53	53	100	0	157

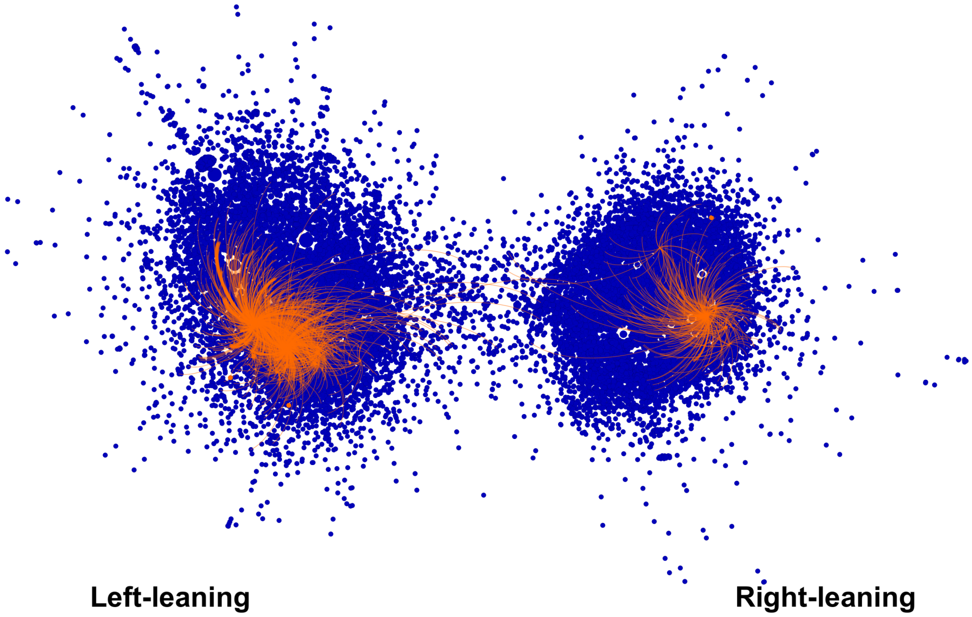

Figure 2: Retweets of RU-IRA trolls suggest polarization.

the crowd. Furthermore, these personas can embody extreme caricatures of each side with the goal of building loyalty toward one side and animosity toward the other. We could say that what makes this set of accounts "bad actors" is that they were seemingly trying to promote bad behaviors in others.

2 BAD ACTS VS. BAD ACTORS

The basis by which we might judge the RU-IRA accounts as "bad" actors is worth clarifying. By some narrow definitions that cast bad actors as unruly or contentious individuals [4], these accounts were conceivably "good actors" since they were able to blend in with the communities they targeted (Twitter's own CEO was retweeting one of them). And because these accounts were not unruly in this context, they were unlikely to be associated with significantly more user reports of negative behavior. This highlights a flaw in approaches that try to sort out "good" and "bad" actors by judging their actions using a rules-driven logic that pays less attention to the actors themselves (e.g. no user reports equals a good actor). Such rules can be gamed by sophisticated actors, and the algorithms implementing these rules often lack the interpretive flexibility that is necessary to adapt to changing circumstances.

The fact is that two actors might take similar actions for very different reasons, and from the standpoint of many ethical frameworks (e.g. Confucian, Buddhist, Aristotelian ethics), we would be perfectly justified in feeling that one actor was good, and the other bad [10]. That is because these frameworks invite us to consider not just actions, but the actors themselves in terms of things like their socio-emotional capacities and intentions. It is by considering intent for instance, that we can begin to discern the differences between "good" users that are using #Blacklivesmatter to seek social justice, and the RU-IRA accounts—bad actors that aren't likely to be reported for online harassment, but carry an intent to promote division and disagreement via the hashtag.

There are no quick fixes here. In practice, prioritizing bad actors over bad actions means that we have to become more willing to deal with the real messiness of human activity. Things like intent can be difficult to establish but there are reasons why it is an integral part of our legal systems. Social media companies need to focus less on sanctioning behaviors and more on simply talking with and understanding different types of bad actors to establish a more nuanced ethical, legal and policy framework.

In light of strategic infiltration by government-affiliated disinformation actors, such as the RU-IRA accounts, this feels more and more necessary. The crowdsourced nature of social media discourse by design makes that discourse open to whoever might be in the crowd, including varieties of "real users," disinformation actors, trolls, bots, and everything in between. As social media platforms increasingly serve as access points to political conversations, a platform that is compromised by bad actors also compromises the democratic processes facilitated by that platform—for example, how events are interpreted within a broader politicized context.

Social media companies also need to move beyond sanctions like suspending accounts to consider topics like education and 'recovery' for certain types of bad actors and the people who interacted with them. We require a greater understanding of how audience-driven systems can promote user agency in understanding the information they consume and how greater protections might coexist with the benefits of crowd-driven online communities, such as access to discourse and anonymity. Users must be equipped with the tools to understand how their information landscape—trending hashtags, viral tweets, and influential accounts—has been shaped and the awareness that elements of this landscape may have been manufactured with the intent to manipulate or mislead, even in spaces perceived as particularly safe or genuine. To accomplish this, social media companies need to reach out to users, perhaps even on a personal level, about how they interacted with these propaganda accounts.

3 ABOUT THE AUTHORS

Leo G. Stewart is a prospective PhD student who hopes to study framing and narrativity on social media. He uses Facebook, Snapchat, and Instagram to keep up with his friends and family.

Ahmer Arif is a PhD candidate who studies rumoring and disinformation on social media. He is currently running a research project that's examining a disinformation campaign that targets the White Helmets (a humanitarian group) in Syria. He is also running a project around designing activities to help students in higher education learn about disinformation and emotional reasoning in our online spaces. Ahmer's primary form of participation in online communities takes the form of being a member of an internet forum since 1998—that currently has about 190,000 registered members. In his time with this community, he has lurked, commented and moderated (policing the political debate & discussion section of the forum).

Kate Starbird is an Assistant Professor of Human Centered Design & Engineering at the University of Washington. Her research focuses on the use of social media during crisis events—and more recently online rumors, misinformation, and disinformation in that context. As a graduate student at the University of Colorado, Kate and her colleagues conducted action research via Twitter, performing as digital volunteers during numerous crisis events, beginning with the 2010 Haiti earthquake. More recently, Kate engages with friends and colleagues on Facebook and with a broader community of crisis responders, researchers, and others on Twitter.

ACKNOWLEDGMENTS

This research is a collaboration between the emCOMP lab and DataLab at the University of Washington and was supported by National Science Foundation Grant 1420255 and Office of Naval Research Grant N00014-17-1-2980. We also wish to thank the UW SoMe Lab for providing infrastructure support.

REFERENCES

[1] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: An Open Source Software for Exploring and Manipulating Networks. (2009). http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154

[2] Daniel Edler and Martin Rosvall. [n. d.]. The MapEquation software package. ([n. d.]). http://www.mapequation.org

[3] Sharon Meraz and Zizi Papacharissi. 2013. Networked Gatekeeping and Networked Framing on #Egypt. *The International Journal of Press/Politics* 18, 2 (April 2013), 138–166. https://doi.org/10.1177/1940161212474472

[4] Merriam-Webster. 2018. bad actor. (2018).

[5] United States House of Representatives Permanent Select Committee on Intelligence. 2017. (Nov. 2017). https://democrats-intelligence.house.gov/uploadedfiles/exhibit_b.pdf

[6] United States House of Representatives Permanent Select Committee on Intelligence. 2017. Testimony of Sean J. Edgett. (Nov. 2017). https://intelligence.house.gov/uploadedfiles/prepared_testimony_of_sean_j._edgett_from_twitter.pdf

[7] Martin Rosvall, Daniel Axelsson, and Carl T. Bergstrom. 2009. The map equation. *The European Physical Journal Special Topics* 178, 1 (2009), 13–23.

[8] Scott Shane and Vindu Goel. 2017. Fake Russian Facebook Accounts Bought $100,000 in Political Ads. (Sept. 2017). https://www.nytimes.com/2017/09/06/technology/facebook-russian-political-ads.html

[9] Leo G. Stewart, Ahmer Arif, A. Conrad Nied, Emma S. Spiro, and Kate Starbird. 2017. Drawing the Lines of Contention: Networked Frame Contests Within #BlackLivesMatter Discourse. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 96 (Dec. 2017), 23 pages. https://doi.org/10.1145/3134920

[10] Shannon Vallor. 2016. *Technology and the virtues: A philosophical guide to a future worth wanting.* Oxford University Press.