The Harvard Kennedy School Misinformation Review*
January 2020, Volume 1, Issue 1
© The Author(s) 2020
Reprints and permissions: misinforeview@hks.harvard.edu
DOI: https://doi.org/10.37016/mr-2020-003
Website: misinforeview.hks.harvard.edu

Russian Twitter disinformation campaigns reach across the American political spectrum

Evidence from an analysis of Twitter data reveals that Russian social media trolls exploited racial and political identities to infiltrate distinct groups of authentic users, playing on their group identities. The groups affected spanned the ideological spectrum, suggesting the importance of coordinated counter-responses from diverse coalitions of users.

Authors: Deen Freelon (1), Tetyana Lokot (2)
Affiliations: (1) Hussman School of Journalism and Media, University of North Carolina at Chapel Hill; (2) School of Communications, Dublin City University

How to cite: Freelon, Deen; Lokot, Tetyana (2020). Russian Twitter disinformation campaigns reach across the American political spectrum, *The Harvard Kennedy School (HKS) Misinformation Review*, Volume 1, Issue 1
Received: Nov. 11, 2019 Accepted: Dec. 16, 2019 Published: Jan. 14, 2020

RESEARCH QUESTION

- What authentic audiences did the Internet Research Agency (IRA) interact with, and with what messages?
- To what extent did these audiences share the ideological orientation of the IRA accounts to which they replied?
- Are IRA strategies different for different communities?
- What strategies might be the most effective to counter IRA activities?

ESSAY SUMMARY

- The IRA is a private company sponsored by the Russian government, which distributes Kremlin-friendly disinformation on social media under false identities (see DiResta et al., 2018; Howard, Ganesh, Liotsiou, Kelly, & Francois, 2018).
- The IRA engaged with several distinct communities of authentic users—primarily conservatives, progressives, and Black people—which exhibited only minimal overlap on Twitter.
- Authentic users primarily engaged with IRA accounts that shared their own ideological and/or racial identities.
- Racist stereotyping, racial grievances, the scapegoating of political opponents, and outright false statements were four of the most common appeals found among the most replied-to IRA tweets.

* A publication of the Shorenstein Center on Media, Politics and Public Policy at Harvard University's John F. Kennedy School of Government.

- We conducted a network analysis of 2,057,747 authentic replies to IRA tweets over nine years, generated ideology ratings for a random sample of authentic users, and qualitatively analyzed some of the most replied-to IRA tweets.
- State-sponsored disinformation agents have demonstrated success in infiltrating distinct online communities. Political content attracts far more engagement than non-political content and appears crafted to exploit intergroup distrust and enmity.
- Collaboration between different political groups and communities might be successful in detecting IRA campaigns more effectively.

IMPLICATIONS (Why does this matter? And to whom?)

This study's results support two broad conclusions and two practical implications regarding State-supported social media disinformation in general and the IRA's efforts in particular:

Politically active communities present substantial vulnerabilities that disinformation agents can exploit. By far, the IRA accounts and content that attracted the most attention were explicitly political in nature. In contrast, the organization was less successful in engaging users with its hashtag games, health appeals, and general-interest news headlines. This indicates that politically engaged users should be especially mindful of attempts by foreign governments and others to co-opt their social media activities for surreptitious disinformation purposes.

Our results make it clear that group identity lies at the core of the IRA's attack strategy. Political audiences were addressed as liberals, conservatives, and Black people to provoke anger against oppositional outgroups. Each group was paired with a specific set of opponents: the IRA presented conservatives with outrages committed by liberals, immigrants, CNN, George Soros, and others; liberals witnessed the travesties of the Trump administration, Republicans in general, and evangelical Christians; and Black users were confronted with an endless cavalcade of racism, often perpetrated by white police officers. There was very little policy-related or even horse-race campaign content to be found—most tweets were devoted to vilifying political and social adversaries. Other tweets supported the core group identity in affirmative ways, such as conservative tweets celebrating law enforcement and the military and Black posts spotlighting Black history and achievements. Individuals who identify as members of targeted groups ran a disproportionate risk of exposure to IRA disinformation over the study time period. The 2020 election may put them in a similar position (Linvill & Warren, 2019).

Facts, inflammatory opinions, and outright falsehoods are all components of a successful disinformation playbook. In their typology of IRA Twitter accounts, Linvill and Warren (2018) separate political users on the left and right from so-called "Fearmongers" whose main purpose is to spread fabricated news stories. Our results reveal that political IRA users also trafficked in falsehoods alongside factual content and extreme opinions. While previous research has noted this tendency (Howard et al., 2018), we find false content among the ranks of the most widely-discussed tweets, especially on the right. The IRA's strategy of posting about nonexistent events can only be successful if authentic users engage with and spread such content at scale. This opens the possibility that it may have had some degree of political impact.

This study's findings imply that 1) combating State-sponsored disinformation requires cross-ideological engagement, and 2) to protect and empower their users, social media platforms need to do more than simply delete disinformation messages upon detection. The relative sizes of the communities we detected suggest that neither side of the political aisle is immune to foreign disinformation. The ideological breadth of the threat presents opportunities for anti-propaganda collaborations across lines of political difference. We have already seen bipartisan efforts to this effect in the US Congress (e.g., US Senate Select Committee on Intelligence, 2017), but civil society could do more. Because disinformation messages targeted at one group are unlikely to be seen by others, members of different targeted groups could coordinate to identify and expose suspicious behaviors, perhaps by using private messaging tools. While they may not agree on the issues, they should at least be able to identify foreign meddling in domestic elections as a common threat.

Social media platforms could also do more to empower their users against foreign manipulation. Both Facebook and Twitter's current policies require that "coordinated inauthentic behavior" be removed immediately upon detection. But this practice robs users of opportunities to understand and recognize attempts at manipulation in context. Platforms could balance user disinformation education with the understandable desire to stop such messages from spreading by:

- Labeling disinformation messages as such,
- Providing links to supporting evidence for the labels,
- Showing statistics on how far the account's messages had spread before detection,
- and disabling the share and reply functions for such messages.

Policy changes such as these might help users understand how politically polarizing and hostile messages are marshaled as nonpartisan weapons of information warfare, and perhaps even discourage them from circulating their own such messages. Extensive user testing should be conducted before implementing such measures to ensure that they do not backfire by inviting users to believe disinformation content.

FINDINGS

We find that the IRA engaged with several distinct communities of users on Twitter. We used a network analysis technique called community detection to determine the sizes of the most popular IRA accounts' respective audiences. A "community" is defined as a group of authentic (non-IRA) users that mostly reply to the same popular IRA accounts and only rarely to other accounts. We discovered ten distinct communities, all featuring varying degrees of overlap with one another: four devoted to right-wing politics, one left-wing and generally anti-Trump, one focused on Black American issues, one focused on false news outlets that mostly discussed real news, one devoted to hashtag games (i.e., hashtags that pose challenges for users to answer cleverly, such as #ReasonsIAintInARelationship and #3wordsBetterThanILoveYou), one about health and diet issues, and one Russian-language community.

Figure 1 is a network visualization of the ten communities. The size of each circle is proportional to the number of members within each community—the larger the circle, the more populous the community. The lines connecting each circle indicate how many replies crossed community boundaries, with thicker lines corresponding to more replies. We found that 47.5% of unique users across all communities were placed in one of the four right-wing communities, compared to 16.5% in the Left community, 15.2% in the Russian-language community, 14% in the Black community, 2.3% in the News community, 2.2% in the Hashtag Gamer community, and 2.2% in the Health and Diet community.

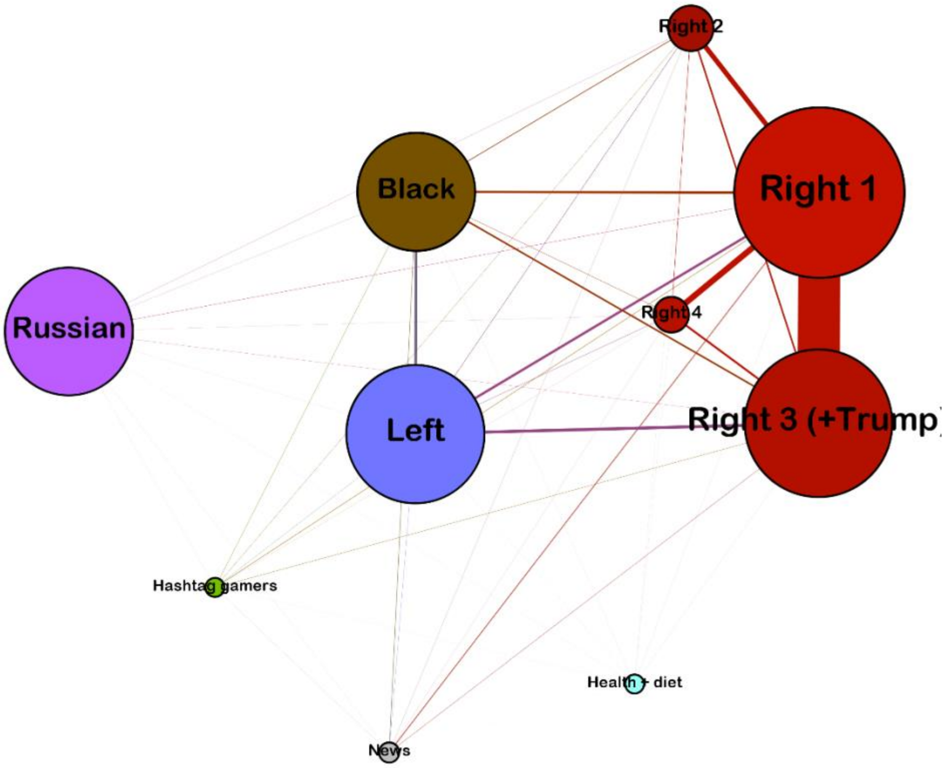

Figure 1. *Sociogram of IRA network communities.*

Each community is anchored by one or more leaders that are responsible for a plurality or majority of replies. Table 1 displays the top three of each community's most prominent leaders and the proportion of replies for which each is responsible, out of all replies pointing to community members. Twenty-three of the 30 leader accounts are affiliated with the IRA. The remaining seven authentic accounts (highlighted in italics) all appear in one of the Right communities. Some authentic accounts were included in these communities simply because someone mentioned them, while others actively participated in conversations in which IRA accounts were involved.

Community	**Leader(s)**	**Percent replies to community**
Right 1	@pamela_moore13	27.1%
	@usa_gunslinger	3.5%
	@potus	1.9%
Right 2	@jenn_abrams	83.4%
	@youtube	0.6%
	@vine	0.2%
Right 3	@ten_gop	85.5%
	@realdonaldtrump	5.9%
	@anncoulter	0.6%

Right 4	@southlonestar	85.5%
	@jk_rowling	0.8%
	@kthopkins	0.8%
Left	@wokeluisa	45.1%
	@kanijjackson	21.5%
	@jemishaaazzz	7.0%
Black	@crystal1johnson	33.3%
	@blacktolive	8.6%
	@blacknewsoutlet	5.3%
News	@chicagodailynew	9.6%
	@dailylosangeles	8.1%
	@seattle_post	4.7%
Hashtag gamers	@giselleevns	18.6%
	@danageezus	9.7%
	@chrixmorgan	5.1%
Health & diet	@exquote	11.7%
	@funddiet	7.5%
	@finddiet	4.6%
Russian	@kadirovrussia	16.7%
	@lavrovmuesli	11.7%
	@margosavazh	8.0%

Table 1. *IRA community leaders.*

Most communities exhibited only minimal overlap with one another, except for the conservative communities. The other communities that engaged with the IRA's messages had -in general- minimal direct contact with one another. For each community, Figure 2 shows the proportion of replies in which both accounts reside within the community out of all those in which at least one account resides within the community. The right-wing communities proved the most outward-facing, with three of the four sharing most of their replies with other communities. (The Right 3 community included @realDonaldTrump, which understandably attracted substantial amounts of attention from other communities.)

In contrast, nearly 70% of the replies in which at least one participant was classified as Left or Black remained internal. This means that these communities mainly interacted within themselves, overlapping minimally with their neighbors. The Russian-language community was by far the most insular, most likely due to the language difference.

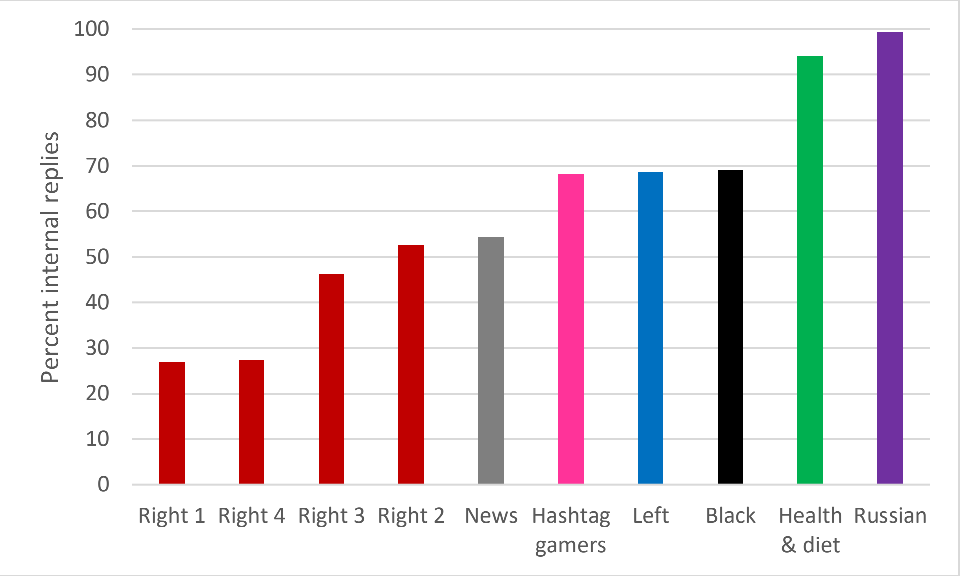

Figure 2. *Percent internal replies for IRA network communities.*

Our network analysis technique allows us to determine whether authentic users mostly engaged with IRA accounts sharing similar identities, or whether they mostly replied to accounts of vastly different identities. Our findings strongly favor the former conclusion; in other words, most of the authentic users shared the political ideologies of the IRA accounts to which they replied. Figure 3 depicts average ideology scores for random samples of 500 unique, authentic users replying to members of each community. The ideology scores are on a unidimensional scale in which lower negative values indicate more liberal ideologies, and higher positive values indicate more conservative ideologies. (See the Methods section for details on how we calculated these scores.) The ideology averages for all four Right communities are right of center, while those for Left and Black are left of center. Health & Diet and Hashtag Gamers also have left of center averages.

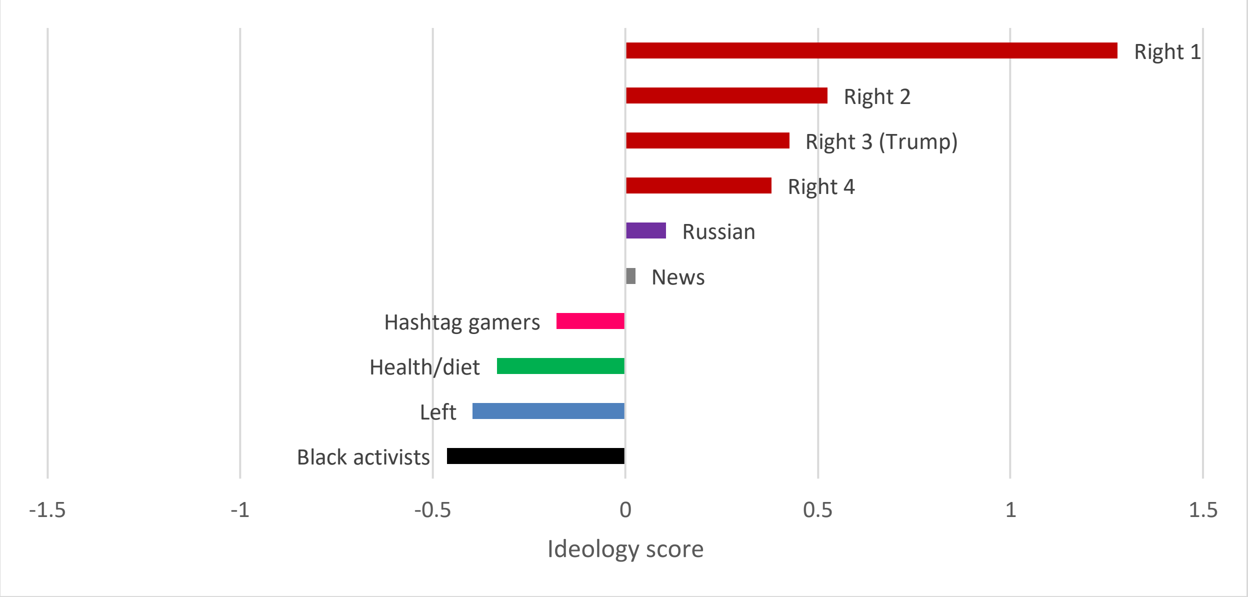

Figure 3. *Mean ideology scores for IRA network communities.*

Political IRA accounts were more effective at eliciting reactions than apolitical ones, with identity-specific appeals—particularly racial (i.e., invoking race) and racist (i.e., expressing racial animus) ones—frequently appearing among the former's most replied-to tweets. IRA accounts tailored their messages to exploit prejudices held by community members against disfavored outgroups. Common targets among conservative-presenting IRA accounts' top tweets included Democrats, Liberals, Antifa, Muslims, immigrants/refugees, George Soros, the Black Lives Matter movement, and CNN. Some tweets presented inflammatory interpretations of undisputed facts, e.g., "VIDEO: Biker Revs Engine, Drives Through Anti-Trump Activists Laying in the Street for a 'Die-In' RT if you'd buy the biker a beer!" (@pamela_moore13). But others presented false stories as factual, for example, "This is big! Hillary Clinton covered up child trafficking investigation at the State Department." (@ten_gop). Thinly veiled racism was common, often manifesting in the form of such outgroup-directed pejoratives as "creeping sharia," "BLM domestic terrorists," "Muslim no-go zones," and "illegals," among others. All four conservative-presenting communities used similar tactics, differing from one another mainly in terms of size.

The Left and Black communities also relied heavily on racial appeals, although from an opposite political stance from the Right communities. Topically, there was some overlap between the two, with nearly all the Black community's tweets directly addressing race, while this was the case with only some of the Left community's. The Left community's non-racial tweets typically targeted Trump, his political allies, and evangelical Christians, e.g., "RT if you want Mueller to arrest Trump on live TV during the State of the Uniom [sic] address #SOTU" (@wokeluisa) and "Michael Flynn (convicted felon) gets a standing ovation at a republican fundraising event. Andrew McCabe (defended America from terrorist threats post 9/11) gets fired without a pension. This is a shining example of what the republican party has become." (@kanijjackson). The IRA leaders of the Black community posted two main types of tweets: first, denunciations of racism, e.g., "Ohio cop shatter [sic] windshield of police cruiser with handcuffed black man's face. Stop police brutality!" (@blk_voice); and second, apolitical celebrations of Black achievement, e.g. "8th grader, Kory Terrell is the Texas Spelling Bee champion! Show him some love. These things go unnoticed!" (@crystal1johnson).

The remaining communities deviated sharply from these patterns. While many of the headlines that emerged from the News community focused on controversial topics such as guns and immigration, their tone was reserved and journalistic, in sharp contrast to the more political communities. The hashtag gamers engaged in a mishmash of political, apolitical, and vulgar jokes in reaction to hashtag prompts such as #RenameMillionWomenMarch, #IKnewWeWereDoomed, #3WordsBetterThanILoveYou, and #ThingsNotToMicrowave. The accounts devoted to health and diet issues remained unswervingly on-topic, avoiding politics altogether.

Our community detection method collated nearly all the Russian-language accounts into a single community. These accounts posted a combination of many of the types of content documented above, including divisive political opinions, jokes, news headlines, and historical facts. Many of the tweets voiced opposition to the government of Ukraine, a common IRA position noted in prior research (Hjorth & Adler-Nissen, 2019). One major difference between the Russian-language IRA accounts and their Anglophone counterparts is that some of the former parodied or impersonated real people, including Ramzan Kadyrov (head of the Chechen Republic) and Sergei Lavrov (Russian foreign affairs minister), whereas none of the latter did so.

METHODS

We collected our data between October 17 and 19, 2018, using a Twitter data collection program called Twint. We searched for all tweets that replied to any screen name on Twitter's list of 3,814 confirmed IRA accounts. A complete list of the screen names we used can be found here: https://intelligence.house.gov/uploadedfiles/ira_handles_june_2018.pdf. This process yielded 2,057,747 tweets posted between May 2009 and October 2018. Fewer than 1% of these tweets appeared prior to 2014, and almost half (46%) appeared in 2017. Given that we collected our data directly from Twitter in real time, we have a high degree of confidence that the true number of authentic replies is no lower than this. However, it may be higher, as tweet deletions and account suspensions almost certainly removed access to at least some replies.

We used a network community detection algorithm called the Louvain method (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008) to generate our communities. We chose to retain the ten largest detected communities based on the insight that larger communities are generally more important than smaller ones. 92% of all unique users in our dataset were classified into one of these ten communities. We labeled the communities based on a qualitative reading of the highest-ranking community members by reply count and the content of their tweets. The network visualization in Figure 1 was created with the network analysis program Gephi.

To generate mean ideology scores for each community, we used an algorithm that infers political ideology based on whom Twitter users follow (Barberá, 2015). Briefly, it uses a list of "elite" users whose ideologies are known to estimate the political ideologies of any user who follows at least one of them. The algorithm assumes that liberals will tend to follow more liberals, and conservatives will follow more conservatives. It produces a unidimensional score in which negative values indicate liberal ideology, positive values indicate conservative ideology, and zero indicates a balanced or moderate ideology. Because collecting followers for the algorithm to analyze is time-consuming, for each community, we randomly sampled 500 authentic users who replied to an IRA member. Our initial sample was thus 5,000 (500 users x 10 communities), but we removed 1,314 users (26.3%) because they did not follow any elites. The ideology scores of the remaining 3,686 users (73.7%) were used to compute each community's mean ideology scores.

BIBLIOGRAPHY

Barberá, P. (2015). Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data. *Political Analysis*, 23(1), 76–91. https://doi.org/10.1093/pan/mpu011

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

DiResta, R., Shaffer, K., Ruppel, B., Sullivan, D., Matney, R., Fox, R., … Johnson, B. (2018). *The Tactics & Tropes of the Internet Research Agency*. Retrieved from New Knowledge website: https://cdn2.hubspot.net/hubfs/4326998/ira-report-rebrand_FinalJ14.pdf

Hjorth, F., & Adler-Nissen, R. (2019). Ideological Asymmetry in the Reach of Pro-Russian Digital Disinformation to United States Audiences. *Journal of Communication*, 69(2), 168–192. https://doi.org/10.1093/joc/jqz006

Howard, P. N., Ganesh, B., Liotsiou, D., Kelly, J., & Francois, C. (2018). The IRA, Social Media and Political Polarization in the United States, 2012-2018. Retrieved from University of Oxford website: https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2018/12/IRA-Report-2018.pdf

Linvill, D., & Warren, P. (2019, November 25). That Uplifting Tweet You Just Shared? A Russian Troll Sent It. Retrieved December 2, 2019, from Rolling Stone website: https://www.rollingstone.com/politics/politics-features/russia-troll-2020-election-interference-twitter-916482/

Linvill, D., & Warren, P. L. (2018). Troll Factories: The Internet Research Agency and State-Sponsored Agenda Building. Retrieved from http://pwarren.people.clemson.edu/Linvill_Warren_TrollFactory.pdf

US Senate Select Committee on Intelligence. (2017). DISINFORMATION: A PRIMER IN RUSSIAN ACTIVE MEASURES AND INFLUENCE CAMPAIGNS (Panel I) (No. 115–40). Retrieved from U.S. Government Publishing Office website: https://www.govinfo.gov/content/pkg/CHRG-115shrg25362/html/CHRG-115shrg25362.htm

Funding

No external or internal funding was used in this research.

Competing interests

None of the authors has any conflicts of interest.

Ethics

Institutional review for this project was unnecessary as it analyzes only public social media posts and the only posts reproduced here are available in public data archives.

Copyright

Data Availability

The IRA tweet data used in this study is available by request from Twitter (recipients are bound by a non-disclosure agreement not to share it). Tweet IDs for the reply data will be shared upon publication.