Update: Russian Interference in 2016 US Election, Bots, & Misinformation

By Twitter PublicPolicy
Thursday, 28 September 2017

Earlier this summer we outlined some of our work to combat bots and networks of manipulation on Twitter. Since then, we have received a number of questions about how malicious bots and misinformation networks on Twitter may have been used in the context of the 2016 U.S. Presidential elections.

Twitter deeply respects the integrity of the election process, which is a cornerstone for all democracies. We will continue to strengthen Twitter against attempted manipulation, including malicious automated accounts and spam, as well as other activities that violate our Terms of Service.

Twitter is in dialogue with congressional committees with respect to investigations into Russian interference in the 2016 U.S. election. Today, Twitter Vice President for Public Policy Colin Crowell met with staff from Senate Select Committee on Intelligence and House Permanent Select Committee on Intelligence to discuss these issues. This is an ongoing process and we will continue to collaborate with investigators.

Due to the nature of these inquiries, we may not always be able to publicly share what we discuss with investigators. And there will always be tools or methods we cannot talk about, because doing so would only help bad actors circumvent them. But we know there is a huge appetite for more transparency into how Twitter fights bots and manipulative networks. That’s why we’ll do our best to keep you informed of both our findings on these specific issues and, more broadly, our efforts to fight bots, spam, and malicious information networks on Twitter.

It’s important to note our work to fight both malicious bots and misinformation goes beyond any one specific election, event, or time period. We’ve spent years working to identify and remove spammy or malicious accounts and applications on Twitter. And we continue to improve our internal systems to detect and prevent new forms of spam and malicious automation, in real time, while also expanding our efforts to educate the public on how to identify and use quality content on Twitter.

Today we are sharing a lot of information that is meant for a wide range of audiences. There are many ways to approach these issues, whether you’re a concerned U.S. voter, a journalist, a developer, etc. We are sharing as much as we can at this stage because of our commitment to be as transparent as possible.

With hundreds of millions of Tweets globally every day, scaling these efforts continues to be a challenge. We will continue to look into these matters on an ongoing basis, and we fully anticipate having more to share as we look into further requests for information. Here are some initial findings that we would like to share publicly at this stage:

**Our Internal Inquiries**

Of the roughly 450 accounts that Facebook recently shared as a part of their review, we concluded that 22 had corresponding accounts on Twitter. All of those identified accounts had already been or immediately were suspended from Twitter for breaking our rules, most for violating our prohibitions against spam.

In addition, from those accounts we found an additional 179 related or linked accounts, and took action on the ones we found in violation of our rules. Neither the original accounts shared by Facebook, nor the additional related accounts we identified, were registered as advertisers on Twitter. However, we continue to investigate these issues, and will take action on anything that violates our Terms of Service.
Russia Today

The US intelligence community released a report in January, 2017, highlighting the role that RT (Russia Today), which has strong links to the Russian government, allegedly played in seeking to interfere in the 2016 U.S. election and undermine trust in American democracy. RT has accounts on Twitter and tweets regularly. The open nature of the Twitter platform means this activity was public.

Today we proactively shared with committee staff a round-up of ads that three RT accounts (@RT_com, @RT_America, and @ActualidadRT) targeted to the U.S. market in 2016. As of our meetings today we believe this is the complete list from these three accounts within that time frame, but we are continuing to review our internal data and will report back to the committees as we have more to share.

Based on our findings thus far, RT spent $274,100 in U.S. ads in 2016. In that year, the @RT_com, @RT_America, and @ActualidadRT accounts promoted 1,823 Tweets that definitely or potentially targeted the U.S. market. These campaigns were directed at followers of mainstream media and primarily promoted RT Tweets regarding news stories.

Election Vote Issues

We are concerned about violations of our Terms of Service and U.S. law with respect to interference in the exercise of voting rights. When we become aware of such activity we take appropriate and timely action. During the 2016 election, we removed Tweets that were attempting to suppress or otherwise interfere with the exercise of voting rights, including the right to have a vote counted, by circulating intentionally misleading information.

For instance: when we were alerted to Hillary Clinton “text-to-vote” examples, we proactively tweeted reminders that one cannot vote via text, examined the content reported to us, used our proprietary tools to search for linked accounts that violated our rules, and, after careful review, took action on thousands of Tweets and accounts. We have not found accounts associated with this activity to have obvious Russian origin but some of the accounts appear to have been automated. We have shared examples of the content of these removed Tweets with congressional investigators.

Political Advertising Disclosure

We note recent calls for increased public disclosure with respect to political advertisements on social media, including Twitter. Twitter supports making political advertising more transparent to our users and the public.

Internally, we already have stricter policies for advertising campaigns on Twitter than we do for organic content. We also have existing specific policies and review mechanisms for campaign ads, but will examine them with an eye to improving them. We welcome the opportunity to work with the FEC and leaders in Congress to review and strengthen guidelines for political advertising on social media.

Automated Traffic & Spam

Every online platform has to deal with spam, and there is no silver bullet. For example, the Internet Society estimated in October 2015 that up to 85 percent of all global email is spam -- and that’s after decades of every email platform in the world tackling this challenge. Obviously email is very different from Tweets, but it’s important to understand the scale of what we are dealing with, and that this is a global issue for all platforms.

Russia and other post-Soviet states have been a primary source of automated and spammy content on Twitter for many years. Content that violates our rules with respect to automated accounts and spam can have a highly negative effect on user experience, and we have long taken substantial action to stem that flow.

As patterns of malicious activity evolve, we’re adapting to meet them head-on. On average, our automated
systems catch more than 3.2 million suspicious accounts globally per week — more than double the amount we detected this time last year. As our detection of automated accounts and content has improved, we’re better able to catch malicious accounts when they log into Twitter or first start to create spam. We have not and do not include spam accounts that we have identified in the active user numbers that we report to shareholders. As we have said, we estimate that false or spam accounts represent less than 5% of our MAUs.

These are just some of our tools:

- **The most effective way to fight suspicious bots is stopping them before they start.** To do this, we’ve built systems to identify suspicious attempts to log in to Twitter, including signs that a login may be automated or scripted. These techniques now help us catch about 450,000 suspicious logins per day. Importantly, much of this defensive work is done through machine learning and automated processes on our back end, and we have been able to significantly improve our automatic spam and bot-detection tools, resulting in a 64% year-over-year increase in suspicious logins we’re able to detect.

- **We’re investing in systems to stop bad content at its source** if its point of origin corresponds with a known bad actor. However, the use of proxy servers, VPNs, and other forms of authentication, especially outside of the United States, may obscure the true origin of traffic on Twitter. We are working on better identifying the true origins of traffic and blocking activity from suspicious sources.

- **We’re also improving how we detect and cluster accounts that were created by a single entity or a single suspicious source.** We used these techniques to stop more than 5.7 million spammy follows from a single source just last week (9/21/2017).

- **Detecting non-human activity patterns:** Using signals like the frequency and timing of Tweets and engagements, we’ve built models that can detect whether an activity on Twitter is likely automated. We’re expanding how we use these signals to restrict how users see suspicious accounts.

- **Compromised account detection:** To stop bad actors from exploiting otherwise healthy accounts to spread malicious content, we’re investing in new ways to identify potentially compromised accounts. For example, we’re building systems to detect when login activity is inconsistent with a user’s typical behavior, to help get compromised accounts back under the control of their owners.

- **Checking suspicious content:** Accounts and content detected by our systems are subject to a number of enforcement actions and limitations including: being placed in a “read only” mode pending authentication, having the reach of Tweets limited based on suspicious origin or low quality content, the removal of associated content, and account suspension.

- **Third-party apps:** We’re also continuing to invest in proactively identifying and taking action against applications that violate our developer policies -- including bots and automated apps. While some bots can provide a vital public utility in times of crisis and natural disaster, we’re committed to combatting the minority of apps that create spam and abuse via our API. Since June 2017, we’ve suspended more than 117,000 malicious applications for abusing our API, collectively responsible for more than 1.5 billion low-quality Tweets this year.

- **Preventing false positives:** Any automated system for detecting spam or bots has a chance of false positives, and it’s our goal to bring this as low as possible. That’s why we typically give users caught by our spam detections an opportunity to verify that they’re legitimate before we suspend them from the platform. During this window, accounts may still appear on Twitter and via our public API, even though they are not able to create new Tweets and engagements. We also limit the visibility of these accounts and their content in both Search and Trends.

- **Improving phone verification:** When we detect suspicious activity from an account, we may require that user to verify their phone number to regain access to Twitter. But, as spammers have adapted their techniques, we’ve found that not all phone numbers are equally trustworthy. We’ve improved our phone reputation system to identify suspicious carriers and numbers and prevent their repeated use to pass
verification challenges.

A note about third-party research:

Studies of the impact of bots and automation on Twitter necessarily and systematically under-represent our enforcement actions because these defensive actions are not visible via our API, and because they take place shortly after content is created and delivered via our streaming API. Furthermore, researchers using an API often overlook the substantial in-product features that prioritize the most relevant content. Based on user interests and choices, we limit the visibility of low-quality content using tools such as Quality Filter and Safe Search -- both of which are on by default for all of Twitter’s users and active for more than 97% of users.

Human-Directed Accounts

While we have made important progress addressing spam and other forms of malicious automation on Twitter, we’ve identified new and emerging challenges dealing with non-automated content -- i.e., human-directed accounts instead of bots -- that coordinate their activities to spread information.

When large numbers of human-directed accounts act in coordinated fashion, it can have an effect similar to that of spam. That doesn’t mean that these accounts aren’t violating our Terms of Service -- in some cases, they are -- but we have to approach this problem differently than we do bots and other forms of automated content. It’s much trickier to identify non-automated coordination, and the risks of inadvertently silencing legitimate activity are much higher. We are continuing to work on this, and are committed to getting better at it.

Gaming Trending Topics

Attempting to game trending topics is a practice as old as Trends on Twitter themselves, and over the years we’ve invested heavily in thwarting spam and other automated attempts to manipulate Trends. We take active measures to protect against trend gaming, such as excluding automated Tweets and users from our calculations of a Trend.

Importantly, as spammers change their tactics we actively modify our technological tools to address such situations. Since June 2017, we’ve been able to detect an average of 130,000 accounts per day who are attempting to manipulate Trends -- and have taken steps to prevent that impact. This is an area where saying more about those steps would only help bad actors, but we will keep looking for new ways to illustrate with examples the important progress we’ve made on this front.

Electoral Outreach

We engage with national election commissions regularly and consistently bolster our security and agent review coverage during key moments of election cycles around the world. We will continue to do this, and expand our outreach so that we’ve got strong, clear escalation processes in place for all potentialities during major elections and global political events.

Supporting Media Literacy and Accurate Emergency Information

In our increasingly polarized public sphere, we believe developing critical media skills is more important than ever. That’s why we are creating a dedicated media literacy program to demonstrate how Twitter can be an indispensable educational tool. We’re also sponsoring, contributing to, and hosting a Media Literacy Week in some of our key markets.

From the Zika Virus to Hurricane Harvey, Twitter’s real-time nature is vital in emergency scenarios and can save lives. We want to ensure factual information is elevated across our platform in times of crisis. As part of our flagship philanthropic initiative, we continue to fund promoted campaigns, called Ads for Good, that promote disaster relief efforts and elevate vital communication from reliable sources during emergencies.
Finally, whether it’s our formal partnerships with Bloomberg News, or our journalistic NGO training and outreach with the likes of Reporters without Borders, the Committee to Protect Journalists, and the Reporters Committee for Freedom of the Press, we have redoubled our engagement on critical questions of modern journalism. We will continue to work with reporters and media organizations to ensure that Twitter’s real-time capacity for dispelling untruths is built into the approach of newsrooms and established media outlets worldwide.

**Next Steps**

Over the coming weeks and months, we’ll be rolling out several changes to the actions we take when we detect spammy or suspicious activity, including introducing new and escalating enforcements for suspicious logins, Tweets, and engagements, and shortening the amount of time suspicious accounts remain visible on Twitter while pending confirmation. These are not meant to be definitive solutions. We’ve been fighting against these issues for years, and as long as there are people trying to manipulate Twitter, we will be working hard to stop them.

Twitter is where the world goes to see what’s happening. We are a platform founded on a commitment to transparency and we take that legacy and responsibility seriously. We will continue to work with official inquiries into these issues, and to share updates publicly as we are able.