## United States Senate Select Committee on Intelligence

## Testimony of Sean J. Edgett
## Acting General Counsel, Twitter, Inc.

**November 1, 2017**

Chairman Burr, Vice Chairman Warner, and Members of the Committee:

Twitter understands the importance of the Committee's inquiry into Russia's interference in the 2016 election, and we appreciate the opportunity to appear here today.

The events underlying this hearing have been deeply concerning to our company and the broader Twitter community. We are committed to providing a service that fosters and facilitates free and open democratic debate and that promotes positive change in the world. We take seriously reports that the power of our service was misused by a foreign actor for the purpose of influencing the U.S. presidential election and undermining public faith in the democratic process.

Twitter is familiar with problems of spam and automation, including how they can be used to amplify messages. The abuse of those methods by sophisticated foreign actors to attempt state-sponsored manipulation of elections is a new challenge for us—and one that we are determined to meet. Today, we intend to demonstrate the seriousness of our commitment to addressing this new threat, both through the effort that we are devoting to uncovering what happened in 2016 and by taking steps to prevent it from happening again.

We begin by explaining the values that shape Twitter and that we aspire as a community to promote and embody. We then describe our response to reports about the role of automation in the 2016 election and on social media more generally. As we discuss, that response includes the creation of a dedicated team within Twitter to enhance the quality of the information our users see and to block malicious activity whenever and wherever we find it. In addition, we have launched a retrospective analysis of activity on our system that indicates Russian efforts to influence the 2016 election through automation, coordinated activity, and advertising. Although the work of that review continues, we share what we know, today, in the interests of transparency and out of appreciation for the urgency of this matter. We do so recognizing that our findings may be supplemented as we work with Committee staff and other companies, discover more facts, and gain a greater understanding of these events. Indeed, what happened on Twitter is only one part of the story, and the Committee is best positioned to see how the various pieces fit together. We look forward to continued partnership, information sharing, and feedback.

We also detail the steps we are taking to ensure that Twitter remains a safe, open, transparent, and positive platform for our users. Those changes include enhanced safety policies, better tools and resources for detecting and stopping malicious activity, tighter advertising standards, and increased transparency to promote public understanding of all of these areas. Our work on these challenges will continue for as long as malicious actors seek to abuse our system, and will need to evolve to stay ahead of new tactics.

We are resolved to continue this work in coordination with the government and our industry peers. Twitter believes that this hearing is an important step toward furthering our shared understanding of how social media platforms, working hand-in-hand with the public and private sectors, can prevent this type of abuse both generally and, of critical importance, in the context of the electoral process.

## I. Twitter's Values

Twitter was founded upon and remains committed to a core set of values that have guided us as we respond to the new threat that brings us here today.

Among those values are defending and respecting the user's voice—a two-part commitment to freedom of expression and privacy. Twitter has a history of facilitating civic engagement and political freedom, and we intend for Twitter to remain a vital avenue for free expression here and abroad. But we cannot foster free expression without ensuring trust in our platform. We are determined to take the actions necessary to prevent the manipulation of Twitter, and we can and must make sure Twitter is a safe place.

Keeping Twitter safe includes maintaining the quality of information on our platform. Our users look to us for useful, timely, and appropriate information. To preserve that experience, we are always working to ensure that we surface for our users the highest quality and most relevant content first. While Twitter's open and real-time environment is a powerful antidote to the abusive spreading of false information, we do not rest on user interaction alone. We are taking active steps to stop malicious accounts and Tweets from spreading, and we are determined that our strategies will keep ahead of the tactics of bad actors.

Twitter is founded on a commitment to transparency. Since 2012, we have published the Twitter Transparency Report on a semiannual basis, providing the public with key metrics about requests from governments and certain private actors for user information, content removal, copyright violations, and most recently, Terms of Service ("TOS") violations. We are also committed to open communication about how we enforce our TOS and the Twitter Rules, and about how we protect the privacy of our users.

Following through on those commitments takes both resolve and resources. And the fight against malicious activity and abuse goes beyond any single election or event. We work every day to give everyone the power to create and share ideas and information instantly, without barriers.

## II. Background on Twitter's Operation

Understanding the steps we are taking to address election-related abuse of our platform requires an explanation of certain fundamentals of Twitter's operation. We therefore turn now to a description of the way our users interact with our system, how we approach automated content, and the basics of advertising on Twitter.

### A. User Interaction

Twitter has 330 million monthly active users around the world, 67 million of which are located in the United States. Users engage with our platform in a variety of ways. Users choose what content they primarily see by following (and unfollowing) other user accounts. Users generate content on the platform by Tweeting original content, including text, hashtags, photos, GIFs, and videos. They may also reply to Tweets, Retweet content already posted on the platform, and like Tweets and Retweets; the metric we use to describe such activity is "engagement"—the different ways in which users are engaged with the content they are viewing. Users can also exchange messages with users and accounts they follow (or, if their privacy settings permit, with any other user) through direct messaging ("DM").

The volume of activity on our system is enormous: Our users generate thousands of Tweets per second, hundreds of thousands of Tweets per minute, hundreds of millions of Tweets per day, and hundreds of billions of Tweets every year.

Another metric we use is how many times a specific piece of content such as a Tweet is viewed. That metric—which we refer to as "impressions"—does not require any additional engagement by the user; viewing content generates an impression, although there is no guarantee that a user has actually read the Tweet. Impressions are not "unique," so multiple impressions may be created by one account, by a single person using multiple accounts, or by many accounts.

A third important concept is "trends." Trends are words, phrases, or hashtags that may relate to an event or other topic (*e.g.*, #CommitteeHearing). Twitter detects trends through an advanced algorithm that picks up on topics about which activity is growing quickly and thus showing a new or heightened interest among our users. Trends thus do not measure the aggregate popularity of a topic, but rather the velocity of Tweets with related content. The trends that a user sees may depend on a number of factors, including their location and their interests. If a user clicks on a trend, the user can see Tweets that contain that hashtag.

### B. Malicious Automation and Responsive Measures

Automation refers to a process that generates user activity—Tweets, likes, or following behavior—without ongoing human input. Automated activity may be designed to occur on a schedule, or it may be designed to respond to certain signals or events. Accounts that rely on automation are sometimes also referred to as "bots."

Automation is not categorically prohibited on Twitter; in fact, it often serves a useful and important purpose. Automation is essential for certain informational content, particularly when time is of the essence, including for law enforcement or public safety notifications. Examples include Amber Alerts, earthquake and other storm warnings, and notices to "shelter in place" during active emergency situations. Automation is also used to provide customer service for a range of companies. For example, as of April 11, 2017, users are able to Tweet @TwitterSupport to request assistance from Twitter. If a user reports a forgotten password or has a question about our rules, the initial triage of those messages is performed by our automated system—a Twitter-developed program to assist users in troubleshooting account issues.

But automation can also be used for malicious purposes, most notably in generating spam—unwanted content consisting of multiple postings either from the same account or from multiple coordinated accounts.  While "spam" is frequently viewed as having a commercial element since it is a typical vector for spreading advertising, Twitter's Rules take an expansive view of spam because it negatively impacts the user experience.  Examples of spam violations on Twitter include automatically Retweeting content to reach as many users as possible, automatically Tweeting about topics on Twitter in an attempt to manipulate trends, generating multiple Tweets with hashtags unrelated to the topics of those hashtags, repeatedly following and unfollowing accounts to tempt other users to follow reciprocally, tweeting duplicate replies and mentions, and generating large volumes of unsolicited mentions.

Our systems are built to detect automated and spam accounts across their lifecycles, including detection at the account creation and login phase and detection based on unusual activity (*e.g.*, patterns of Tweets, likes, and follows).  Our ability to detect such activity on our platform is bolstered by internal, manual reviews conducted by Twitter employees.  Those efforts are further supplemented by user reports, which we rely on not only to address the content at issue but also to calibrate our detection tools to identify similar content as spam.

Once our systems detect an account as generating automated content or spam, we can take action against that account at either the account level or the Tweet level.  Depending on the mode of detection, we have varying levels of confidence about our determination that an account is violating our rules.  We have a range of options for enforcement, and generally, the higher our confidence that an account is violating our rules, the stricter our enforcement action will be, with immediate suspension as the harshest penalty.  If we are not sufficiently confident to suspend an account on the basis of a given detection technique, we may challenge the account to verify a phone number or to otherwise prove human operation, or we may flag the account for review by Twitter personnel.  Until the user completes the challenge, or until the review by our teams has been completed, the account is temporarily suspended; the user cannot produce new content (or perform actions like Retweets or likes), and the account's contents are hidden from other Twitter users.

We also have the capability to detect suspicious activity at the Tweet level and, if certain criteria are met, to internally tag that Tweet as spam, automated, or otherwise suspicious.  Tweets that have been assigned those designations are hidden from searches, do not count toward generating trends, and generally will not appear in feeds unless a user follows that account.  Typically, users whose Tweets are designated as spam are also put through the challenges described above and are suspended if they cannot pass.

For safety-related TOS violations, we have a number of enforcement options.  For example, we can stop the spread of malicious content by categorizing a Tweet as "restricted pending deletion," which requires a user to delete the Tweet before the user is permitted to continue using the account and engaging with the platform.  So long as the Tweet is restricted—and until the user deletes the Tweet—the Tweet remains inaccessible to and hidden from all Twitter users.  The user is blocked from Tweeting further unless and until he or she deletes the restricted Tweet.  This mechanism is a common enforcement approach to addressing less severe content violations of our TOS outside the spam context; it also promotes education among our users. More serious violations, such as posting child sexual exploitation or promoting terrorism, result in immediate suspension and may prompt interaction with law enforcement.

### C. Advertising Basics

Advertising on Twitter generally takes the form of promoted Tweets, which advertisers purchase to reach new groups of users or spark engagement from their existing followers. Promoted Tweets are clearly labeled as "promoted" when an advertiser pays for their placement on Twitter. In every other respect, promoted Tweets look and act just like regular Tweets and can be Retweeted, replied to, and liked.

Advertisers can post promoted Tweets through a self-service model on the Twitter platform or through account managers, who manage relationships with advertising partners. When purchasing a promoted Tweet, an advertiser can target its audience based on information such as interests, geography, gender, device type, or other specific characteristics. For most campaigns, advertisers pay only when users engage with the promoted Tweet, such as following the advertiser; liking, replying to, or clicking on the Tweet; watching a Tweet's video; or taking some other action.

Because promoted Tweets are presented to our users from accounts they have not yet chosen to follow, Twitter applies to those Tweets a robust set of policies that prohibit, among other things, ads for illegal goods and services, ads making misleading or deceptive claims, ads for drugs or drug paraphernalia, ads containing hate content, sensitive topics, and violence, and ads containing offensive or inflammatory content.

Twitter relies on two methods to prevent prohibited promoted content from appearing on the platform: a proactive method and a reactive method. Proactively, Twitter relies on custom-built algorithms and models for detecting Tweets or accounts that might violate its advertising policies. Reactively, Twitter takes user feedback through a "Report Ad" process, which flags an ad for manual human review. Once our teams have reviewed the content, typically one of three decisions will be made: if the content complies with our policy, we may approve it; if the content/account violates the policy, we may stop the particular Tweet from being promoted to users; or, if Twitter deems the account to be in repeated violation of our policies at the Tweet level, we may revoke an account's advertising privileges (also known as off-boarding the advertiser).

## III. Malicious Automation in the 2016 Election: Real-Time Observations and Response

Although Twitter has been fighting the problem of spam and malicious automation for many years, in the period preceding the 2016 election we observed new ways in which accounts were abusing automation to propagate misinformation on our platform. Among other things, we noticed accounts that Tweeted false information about voting in the 2016 election, automated accounts that Tweeted about trending hashtags, and users who abused their access to the platform we provide developers.

At the time, we understood these to be isolated incidents, rather than manifestations of a larger, coordinated effort at misinformation on our platform. Once we understood the systemic nature of the problem in the aftermath of the election, we launched a dedicated initiative to research and combat that new threat.

### A. Malicious Automation and Misinformation Detected in 2016

We detected examples of automated activity and deliberate misinformation in 2016, including in the run-up to the 2016 election, that in retrospect appear to be signals of the broader automation problem that came into focus after the election had concluded.

On December 2, 2016, for example, we learned of @PatrioticPepe, an account that automatically replied to all Tweets from @realDonaldTrump with spam content. Those automatic replies were enabled through an application that had been created using our Application Programming Interface (“API”). Twitter provides access to the API for developers who want to design Twitter-compatible applications and innovate using Twitter data. Some of the most creative uses of our platform originate with applications built on our API, but we know that a large quantity of automated spam on our platform is also generated and disseminated through such applications. We noticed an upward swing in such activity during the period leading up to the election, and @PatrioticPepe was one such example. On the same day we identified @PatrioticPepe, we suspended the API credentials associated with that user for violation of our automation rules. On average, we take similar actions against violative applications more than 7,000 times per week.

Another example of aberrant activity we identified and addressed during this period involved voter suppression efforts. In particular, Twitter identified, and has since provided to the Committee, examples of Tweets with images in English and Spanish that encouraged Clinton supporters to vote online, vote by phone, or vote by text.

In response to the attempted “vote-by-text” effort and similar voter suppression attempts, Twitter restricted as inaccessible, pending deletion, 918 Tweets from 529 users who proliferated that content. Twitter also permanently suspended 106 accounts that were collectively responsible for 734 “vote-by-text” Tweets. Twitter identified, but did not take action against, an additional 286 Tweets of the relevant content from 239 Twitter accounts, because we determined that those accounts were seeking to refute the “text-to-vote” message and alert other users that the information was false and misleading. Notably, those refuting Retweets generated significantly greater engagement across the platform compared to the Tweets spreading the misinformation—8 times as many impressions, engagement by 10 times as many users, and twice as many replies.

Before the election, we also detected and took action on activity relating to hashtags that have since been reported as manifestations of efforts to interfere with the 2016 election. For example, our automated spam detection systems helped mitigate the impact of automated Tweets promoting the #PodestaEmails hashtag, which originated with Wikileaks’ publication of thousands of emails from the Clinton campaign chairman John Podesta’s Gmail account. The core of the hashtag was propagated by Wikileaks, whose account sent out a series of 118 original Tweets containing variants on the hashtag #PodestaEmails referencing the daily installments of the emails released on the Wikileaks website. In the two months preceding the election, around 57,000 users posted approximately 426,000 unique Tweets containing variations of the

#PodestaEmails hashtag. Approximately one quarter (25%) of those Tweets received internal tags from our automation detection systems that hid them from searches. As described in greater detail below, our systems detected and hid just under half (48%) of the Tweets relating to variants of another notable hashtag, #DNCLeak, which concerned the disclosure of leaked emails from the Democratic National Committee. These steps were part of our general efforts at the time to fight automation and spam on our platform across all areas.

### B. Information Quality Initiative

After the election, we followed with great concern the reports that malicious actors had used automated activity and promoted deliberate falsehoods on social media as part of a coordinated misinformation campaign. Along with other platforms that were focused on the problem, we realized that the instances our automated systems had detected in 2016 were not isolated but instead represented a broader pattern of conduct that we needed to address in a more comprehensive way.

Recognizing that elections continue and that the health and safety of our platform was a top priority, our first task was to prevent similar abuse in the future. We responded by launching an initiative to combat the problem of malicious automation and disinformation going forward. The objective of that effort, called the Information Quality initiative, is to enhance the strategies we use to detect and deny bad automation, improve machine learning to spot spam, and increase the precision of our tools designed to prevent such content from contaminating our platform.

Since the 2016 election, we have made significant improvements to reduce external attempts to manipulate content visibility. These improvements were driven by investments into methods to detect malicious automation through abuse of our API, limit the ability of malicious actors to create new accounts in bulk, detect coordinated malicious activity across clusters of accounts, and better enforce policies against abusive third-party applications.

Our efforts have produced clear results in terms of our ability to detect and block such content. With our current capabilities, we detect and block approximately 450,000 suspicious logins each day that we believe to be generated through automation. In October 2017, our systems identified and challenged an average of 4 million suspicious accounts globally per week, including over three million challenged upon signup, before they ever had an impact on the platform—more than double our rate of detection at this time last year.

We also recognized the need to address more systematically spam generated by third-party applications, and we have invested in the technology and human resources required to do so. Our efforts have been successful. Since June 2017, we have suspended more than 117,000 malicious applications for abusing our API. Those applications are collectively responsible for more than 1.5 billion Tweets posted in 2017.

We have developed new techniques for identifying patterns of activity inconsistent with legitimate use of our platform (such as near-instantaneous replies to Tweets, non-random Tweet timing, and coordinated engagement), and we are currently implementing these detections across our platform. We have improved our phone verification process and introduced new challenges, including reCAPTCHAs (utilizing an advanced risk-analysis engine developed by Google), to give us additional tools to validate that a human is in control of an account. We have enhanced our capabilities to link together accounts that were formed by the same person or that are working in concert. And we are improving how we detect when accounts may have been hacked or compromised.

In the coming year, we plan to build upon our 2017 improvements, specifically including efforts to invest even further in machine-learning capabilities that help us detect and mitigate the effect on users of fake, coordinated, and automated account activity. Our engineers and product specialists continue this work every day, further refining our systems so that we capture and address as much malicious content as possible. We are committed to continuing to invest all necessary resources into making sure that our platform remains safe for our users.

We also actively engage with civil society and journalistic organizations on the issue of misinformation. Enhancing media literacy is critical to ensuring that voters can discern which sources of information have integrity and which may be suspect. We are creating a dedicated media literacy program to demonstrate how Twitter can be an effective tool in media literacy education. Moreover, we engage in collaborations and trainings with NGOs, such as Committee to Protect Journalists, Reporters without Borders, and Reporters Committee for Freedom of the Press. We do so in order to ensure that journalists and journalistic organizations are familiar with how to utilize Twitter effectively and to convey timely information around our policies and practices.

## IV. Retrospective Reviews of Malicious Activity in the 2016 Election

In addition to the forward-looking efforts we launched in the immediate aftermath of the election, we have initiated a focused, retrospective review of malicious Russian activity specifically in connection with last year's presidential election. Those reviews cover the core Twitter product as well as the advertising product. They draw on all parts of the company and involve a significant commitment of resources and time. We are reporting on our progress today and commit to providing updates to the Committee as our work continues.

### A. Malicious Automated and Human-Coordinated Activity

For our review of Twitter's core product, we analyzed election-related activity from the period preceding and including the election (September 1, 2016 to November 15, 2016) in order to identify content that appears to have originated from automated accounts or from human-coordinated activity associated with Russia. We then assessed the results to discern trends, evaluate our existing detection systems, and identify areas for improvement and enhancement of our detection tools.

#### 1. Methodology

We took a broad approach for purposes of our review of what constitutes an election-related Tweet, relying on annotations derived from a variety of information sources, including Twitter handles, hashtags, and Tweets about significant events. For example, Tweets mentioning @HillaryClinton and @realDonaldTrump received an election-related annotation, as did Tweets that included #primaries and #feelthebern. In total, we included more than 189 million Tweets annotated in this way out of the total corpus of more than 16 billion unique Tweets posted during this time period (excluding Retweets).

To ensure that we captured all relevant automated accounts in our review, Twitter analyzed the data not only using the detection tools that existed at the time the activity occurred, but also using newly developed and more robust detection tools that have been implemented since then. We compared the results to determine whether our new detection tools are able to capture automated activity that our 2016 techniques could not. These analyses do not attempt to differentiate between "good" and "bad" automation; they rely on objective, measurable signals, such as the timing of Tweets and engagements, to classify a given action as automated.

We took a similarly expansive approach to defining what qualifies as a Russian-linked account. Because there is no single characteristic that reliably determines geographic origin or affiliation, we relied on a number of criteria, including whether the account was created in Russia, whether the user registered the account with a Russian phone carrier or a Russian email address, whether the user's display name contains Cyrillic characters, whether the user frequently Tweets in Russian, and whether the user has logged in from any Russian IP address, even a single time. We considered an account to be Russian-linked if it had even one of the relevant criteria.

Despite the breadth of our approach, there are technological limits to what we can determine based on the information we can detect regarding a user's origin. In the course of our analysis—and based in part on work conducted by our Information Quality team—we observed that a high concentration of automated engagement and content originated from data centers and users accessing Twitter via Virtual Private Networks ("VPNs") and proxy servers. In fact, nearly 12% of Tweets created during the election originated with accounts that had an indeterminate location. Use of such facilities obscures the actual origin of traffic. Although our conclusions are thus necessarily contingent on the limitations we face, and although we recognize that there may be other methods for analyzing the data, we believe our approach is the most effective way to capture an accurate understanding of activity on our system.

### 2. Analysis and Key Findings

We began our review with a universe of over 16 billion Tweets—the total volume of original Tweets on our platform during the relevant period. Applying the methodology described above, and using detection tools we currently have in place, we identified 36,746 accounts that generated automated, election-related content and had at least one of the characteristics we used to associate an account with Russia.

During the relevant period, those accounts generated approximately 1.4 million automated, election-related Tweets, which collectively received approximately 288 million impressions.

Because of the scale on which Twitter operates, it is important to place those numbers in context:

- The 36,746 automated accounts that we identified as Russian-linked and tweeting election-related content represent approximately one one-hundredth of a percent (0.012%) of the total accounts on Twitter at the time.
- The 1.4 million election-related Tweets that we identified through our retrospective review as generated by Russian-linked, automated accounts constituted less than three quarters of one percent (0.74%) of the overall election-related Tweets on Twitter at the time. *See* Appendix 1.
- Those 1.4 million Tweets received only one-third of a percent (0.33%) of impressions on election-related Tweets. In the aggregate, automated, Russian-linked, election-related Tweets consistently underperformed in terms of impressions relative to their volume on the platform. *See* Appendix 2.

In 2016, we detected and labeled some, but not all, of those Tweets using our then-existing anti-automation tools. Specifically, in real, time we detected and labeled as automated over half of the Tweets (791,000) from approximately half of the accounts (18,064), representing 0.42% of overall election-related Tweets and 0.14% of election-related Tweet impressions.

Thus, based on our analysis of the data, we determined that the number of accounts we could link to Russia and that were Tweeting election-related content was small in comparison to the total number of accounts on our platform during the relevant time period. Similarly, the volume of automated, election-related Tweets that originated from those accounts was small in comparison to the overall volume of election-related activity on our platform. And those Tweets generated significantly fewer impressions as compared to a typical election-related Tweet.

### 3. Level of Engagement

In an effort to better understand the impact of Russian-linked accounts on broader conversations on Twitter, we examined those accounts' volume of engagements with election-related content.

We first reviewed the accounts' engagement with Tweets from @HillaryClinton and @realDonaldTrump. Our data showed that, during the relevant time period, a total of 1,625 @HillaryClinton Tweets were Retweeted approximately 8.3 million times. Of those Retweets, 32,254—or 0.39%—were from Russian-linked automated accounts. Tweets from @HillaryClinton received approximately 18 million likes during this period; 111,326—or 0.62%—were from Russian-linked automated accounts. The volume of engagements with @realDonaldTrump Tweets from Russian-linked automated accounts was higher, but still relatively small. The 851 Tweets from the @realDonaldTrump account during this period were Retweeted more than 11 million times; 416,632—or 3.66%—of those Retweets were from Russian-linked, automated accounts. Those Tweets received approximately 27 million likes across our platform; 480,346—or 1.8%—of those likes came from Russian-linked automated accounts.

We also reviewed engagement between automated or Russia-linked accounts and the @Wikileaks, @DCLeaks_, and @GUCCIFER_2 accounts. The amount of automated engagement with these accounts ranged from 47% to 72% of Retweets and 36% to 63% of likes during this time—substantially higher than the average level of automated engagement, including with other high-profile accounts. The volume of automated engagements from Russian-linked accounts was lower overall. Our data show that, during the relevant time period, a total of 1,010 @Wikileaks tweets were retweeted approximately 5.1 million times. Of these retweets, 155,933—or 2.98%—were from Russian-linked automated accounts. The 27 tweets from @DCLeaks_ during this time period were Retweeted approximately 4,700 times, of which 1.38% were from Russian-linked automated accounts. The 23 tweets from @GUCCIFER_2 during this time period were Retweeted approximately 18,000 times, of which 1.57% were from Russia-linked automated accounts.

We next examined activity surrounding hashtags that have been reported as potentially connected to Russian interference efforts. We noted above that, with respect to two such hashtags—#PodestaEmails and #DNCLeak—our automated systems detected, labeled, and hid a portion of related Tweets at the time they were created. The insights from our retrospective review have allowed us to draw additional conclusions about the activity around those hashtags.

We found that slightly under 4% of Tweets containing #PodestaEmails came from accounts with potential links to Russia, and that those Tweets accounted for less than 20% of impressions within the first seven days of posting. Approximately 75% of impressions on the trending topic were views by U.S.-based users. A significant portion of these impressions, however, are attributable to a handful of high-profile accounts, primarily @Wikileaks. At least one heavily-retweeted Tweet came from another potentially Russia-linked account that showed signs of automation.

With respect to #DNCLeak, approximately 23,000 users posted around 140,000 unique Tweets with that hashtag in the relevant period. Of those Tweets, roughly 2% were from potentially Russian-linked accounts. As noted above, our automated systems at the time detected, labeled, and hid just under half (48%) of all the original Tweets with #DNCLeak. Of the total Tweets with the hashtag, 0.84% were hidden and also originated from accounts that met at least one of the criteria for a Russian-linked account. Those Tweets received 0.21% of overall Tweet impressions. We learned that a small number of Tweets from several large accounts were principally responsible for the propagation of this trend. In fact, two of the ten most-viewed Tweets with #DNCLeak were posted by @Wikileaks, an account with millions of followers.

### 4. Human-Coordinated Russian-Linked Accounts

We separately analyzed the accounts that we have thus far identified through information obtained from third-party sources as linked to the Internet Research Agency ("IRA"). We have so far identified 2,752 such accounts. Those 2,752 accounts include the 201 accounts that we previously identified to the Committee. In responding to the Committee and through our cooperation with its requests, we have since linked the 201 accounts to other efforts to locate IRA-linked accounts from third-party information. We discovered that we had found some of the 201 accounts as early as 2015, and many had already been suspended as part of these previous efforts. Our retrospective work, guided by information provided by investigators and others, has thus allowed us to connect the 201 accounts to broader Russian election-focused efforts, including the full set of accounts that we now believe at this point are associated with the IRA. This is an active area of inquiry, and we will update the Committee as we continue the analysis.

The 2,752 IRA-linked accounts exhibited a range of behaviors, including automation. Of the roughly 131,000 Tweets posted by those accounts during the relevant time period, approximately 9% were election-related, and many of their Tweets—over 47%—were automated.

While automation may have increased the volume of content created by these accounts, IRA-linked accounts exhibited non-automated patterns of activity that attempted more overt forms of broadcasting their message. Some of those accounts represented themselves as news outlets, members of activist organizations, or politically-engaged Americans. We have seen evidence of the accounts actively reaching out to journalists and prominent individuals (without the use of automation) through mentions. Some of the accounts appear to have attempted to organize rallies and demonstrations, and several engaged in abusive behavior and harassment. All 2,752 accounts have been suspended, and we have taken steps to block future registrations related to these accounts.

## B. Advertising Review

In the second component of our retrospective review, we focused on determining whether or how malicious Russian actors may have sought to abuse our platform using advertising.

### 1. Methodology

To evaluate the scope and impact of election-related advertisements, we used a custom-built machine-learning model that we refined over a number of iterations to maximize accuracy. That model was designed to detect all election-related content in the universe of English-language promoted Tweets that appeared on our system in 2016.

Our model yielded 6,493 accounts. We then divided those accounts into three categories of high, medium, and low interest based on a number of factors: the number of promoted Tweets the account had purchased in 2016, the percentage of promoted Tweets from the whole that our model suggested were election-related (a concept known as "election density"), whether the account had Russian-specific characteristics, and whether the account had non-Russian international characteristics.

For the purpose of this review, we deemed an account to be Russian-linked if any of the following criteria were present: (1) the account had a Russian email address, mobile number, credit card, or login IP; (2) Russia was the declared country on the account; or (3) Russian language or Cyrillic characters appeared in the account information or name. (As in the core-product review, here too, we encountered technological challenges associated with VPNs, data centers, and proxy servers that do not allow us to identify location.) We treated as election-related any promoted Tweets that referred to any candidates (directly or indirectly), political parties, notable debate topics, the 2016 election generally, events associated with the election, or any political figures in the United States.

Experienced advertising policy content reviewers then engaged in a manual evaluation of each account to determine whether they had promoted violative content in 2016. While we reviewed every account, the level of review corresponded to the category in which the account belonged. For high-interest accounts (197), we reviewed 100% percent of the account's promoted content, as well as information about the account itself, including location and past advertising activity. For other types of accounts, we adjusted our level of manual review according to the interest category of the account. For the medium interest accounts (1,830), we reviewed approximately three quarters of the promoted content associated with the account, together with the account information. For the low interest accounts (4,466), we reviewed about one quarter of the promoted content, together with other account information. For each Tweet our reviewers examined, the reviewers evaluated its contents, including any attached media, geographical and keyword targeting, and account-level details, such as profile, avatar, and non-promoted Tweets. Reviewers looked at the Russian signals connected to any account, regardless of its interest category.

Finally, we tested our results against accounts we knew to be Russian, such as Russia Today accounts, to ensure that our methodology was sound. As we did with the retrospective review of election-related Tweets, we evaluated the advertising data both using the policies in place at the time and using our new policies that we have since introduced. That permitted us to compare what we would have detected and stopped promoting during the relevant time period had the more recent improvements been in place.

### 2. Analysis and Key Findings

We identified nine accounts that had at least one of the criteria for a Russian-linked account and promoted election-related content Tweets that, based on our manual review, violated existing or recently implemented ads policies, such as those prohibiting inflammatory or low-quality content.

Two of those accounts were @RT_COM and @RT_America. Those two accounts represented the vast majority of the promoted Tweets, spend and impressions for the suspect group identified in our review. Together, the two accounts spent $516,900 in advertising in 2016, with $234,600 of that amount devoted to ads that were served to users in the U.S. During that period, the two accounts promoted 1,912 Tweets and generated approximately 192 million impressions across all ad campaigns, with approximately 53.5 million representing impressions generated by U.S.-based users.

On Thursday, October 26, 2017, Twitter announced that it would no longer accept advertisements from RT and will donate the $1.9 million that RT had spent globally on advertising on Twitter to academic research into elections and civil engagement.

The remaining seven accounts that our review identified represented small, apparently unconnected actors. Those accounts spent a combined total of $2,282 on advertising through Twitter in 2016, with $1,184 of this amount spent on ads that were served to users in the U.S. Our available impressions data indicates that in 2016, those accounts ran 404 promoted Tweets and generated a total of 2.29 million impressions across all ad campaigns. Approximately

222,000 of those impressions were generated by U.S.-based users. We have since off-boarded these advertisers.

## V. Post-Election Improvements and Next Steps

While Russian, election-related malicious activity on our platform appears to have been small in comparison to overall activity, we find any such activity unacceptable. Our review has prompted us to commit ourselves to further enhancing our policies and to tightening our systems to make them as safe as possible. Over the coming months, we will be focusing on a series of improvements both to our user safety rules and our advertising policies that we believe will advance the progress we have already made this year.

### A. Enhancements to User Safety and Prevention of Abuse

In 2017, Twitter prioritized work to promote safety and fight abuse across much of the platform. Our engineering, product, policy, and user operations teams worked with urgency to make important and overdue changes designed to shift the burden of reporting online abuse away from the victim and to enable Twitter proactively to identify and act on such content.

As a result of that focus, we have:

- Improved Twitter's detection of new accounts created by users who have been permanently banned;

- Introduced safer search, which is activated by default and limits potentially sensitive and abusive content from search results;

- Limited the visibility and reach of abusive and low-quality Tweets;

- Provided additional user controls both to limit notifications from accounts without verified email or phone numbers and/or profile photos and to allow more options to block and mute; and

- Launched new forms of enforcement to interrupt abuse while it is happening.

While we have made progress on many of our goals, our CEO recently acknowledged that much work remains and that we recognize the need for greater openness about the work we are doing. We are therefore increasing our efforts on safety. Consistent with our commitment to transparency—and to offer full visibility to the Committee, the public, and the Twitter community—on October 19, 2017, we published a calendar of our immediate plans. That calendar identifies dates for upcoming changes to the Twitter Rules that we plan to make in the next three months. These changes will enhance our ability to remove non-consensual nudity, glorification of acts of violence, use of hate symbols in account profiles, and various changes to user-reported Twitter Rules violations. *See* https://blog.twitter.com/official/en_us/topics/company/2017/safetycalendar.html. We plan to offer periodic, real-time updates about our progress.

We are implementing these safety measures alongside the enhanced techniques and tools that the Information Quality initiative has generated for stopping malicious automated content. As described above, we have recently made enhancements to our enforcement mechanisms for detecting automated suspicious activity and have more improvements planned for the coming weeks and months. One of our key initiatives has been to shorten the amount of time that suspicious accounts remain visible on our platform while pending verification—from 35 days to two weeks—with unverified accounts being suspended after that time. While these suspicious accounts cannot Tweet while they are pending verification, we want to further reduce their visibility. We will also introduce new and escalating enforcement mechanisms for suspicious logins, Tweets, and engagements, leveraging our improved detection methods from the past year. Such changes are not meant to be definitive solutions, but they will further limit the reach of malicious actors on the platform and ensure that users have less exposure to harmful or malicious content.

These new threats to our system require us to continually reevaluate how to counter them. As the role of social media in foreign disinformation campaigns comes into focus, it has become clearer that attempts to abuse technology and manipulate public discourse on social media and the Internet through automation and otherwise will not be limited to one election—or indeed to elections at all. We will provide updates on our progress to Congress and to the American people in real time.

### B. Enhancements to Advertising Policy

Last week, we announced a new policy to increase transparency regarding advertising on Twitter. We will soon launch an industry-leading transparency center that will provide the public with more detail than ever before about social media and online advertisers. The enhancements include the ability to see what advertisements are currently running on Twitter, how long the advertisements have been running, and all creative pieces associated with an advertising campaign.

Users will also have greater insight into and control over their experience with advertising on Twitter. Individual users will be able to see all advertisements that have been targeted to them, and all advertisements that the user is eligible to see based on a campaign's targeting. We will also make it possible for users to provide negative feedback regarding an advertisement, whether or not the user has been targeted by the campaign.

Our new policy also changes how Twitter treats electioneering advertisements, or advertisements that clearly identify a candidate or party associated with a candidate for any elected office. Electioneering advertisers will be required to identify themselves to Twitter, and they will be subject to stricter requirements for targeting and harsher penalties for violations of our policies. Any campaign that an electioneering advertiser runs will be clearly marked on the platform to allow users to easily identify it. In addition to the information provided about all advertisements on Twitter, this disclosure will include current and historical spending by an electioneering advertiser, the identity of the organization funding the campaign, and targeting demographics used by the advertiser, such as age, gender, or geographic location.

We recognize that not all political advertising is electioneering advertising. While there is not yet a clear industry definition for issue-based advertisements, we will work with our industry peers and with policymakers to clearly define them and develop policies to treat them similarly to electioneering advertisements.

*          *          *

We have heard the concerns about Twitter's role in Russian efforts to disrupt the 2016 election and about our commitment to addressing this issue. Twitter believes that any activity of that kind—regardless of magnitude—is intolerable, and we agree that we must do better to prevent it. We hope that our appearance today and the description of the work we have undertaken demonstrates our commitment to working with you, our industry partners, and other stakeholders to ensure that the experience of 2016 never happens again.

Indeed, cooperation to combat this challenge is essential. We cannot defeat this novel, shared threat alone. As with most technology-based threats, the best approach is to share information and ideas to increase our collective knowledge. Working with the broader community, we will continue to test, to learn, to share, and to improve, so that our product remains effective and safe.

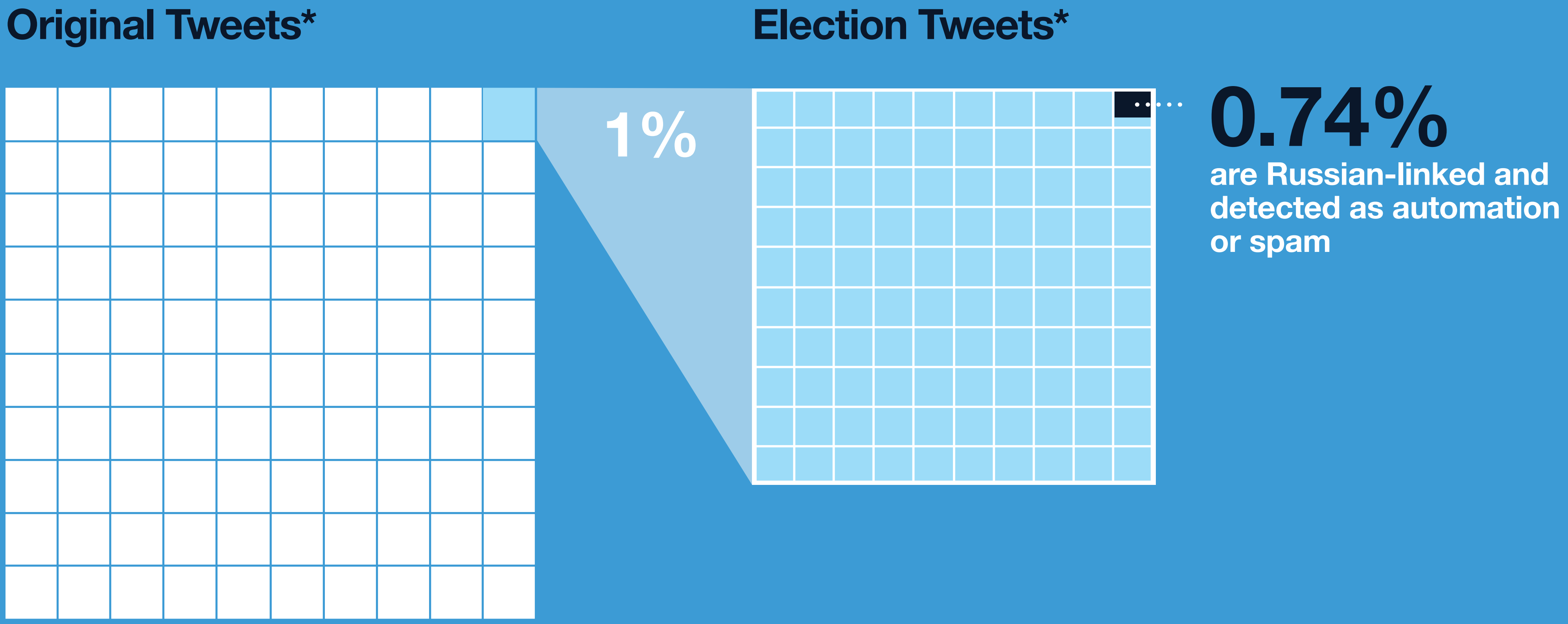
We look forward to answering your questions and working with you in the coming months.

# APPENDIX 1

# Original Tweets

## From September 1 to November 15, 2016

Russian Automated Activity Represented a Small Fraction of Overall Election-Related Tweets



*1) Original Tweets are all Tweets excluding Retweets. 2) A Tweet is considered related to the election when it mentions people and topics related to the 2016 US election, Hillary Clinton, or Donald Trump.

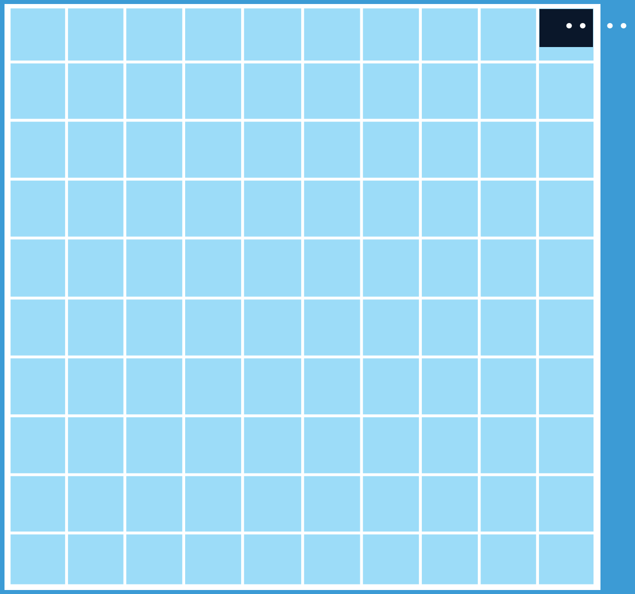
# APPENDIX 2

# Impressions
## on Election-Related Tweets*

### From September 1 to November 15, 2016

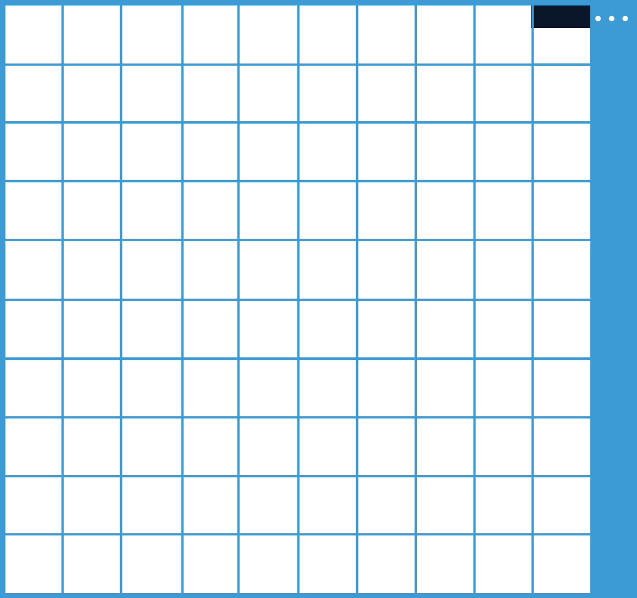
Our Efforts to Disrupt this Activity Had Significant Impact

### Election Tweets*



**0.74%** are from Russian-linked automated accounts

### Election Impressions*



**0.33%** of impressions are on Tweets from Russian-linked automated accounts

*1) Tweets impressions include impressions of original Tweets and Retweets. 2) A Tweet is considered related to the election when it mentions people and topics related to the 2016 US election, Hillary Clinton, or Donald Trump.

NATIONAL
SECURITY
ARCHIVE

National Security Archive,

Suite 701, Gelman Library, The George Washington University,

2130 H Street, NW, Washington, D.C., 20037,

Phone: 202/994-7000, Fax: 202/994-7005, nsarchiv@gwu.edu