ROBOTROLLING
2018. ISSUE 1

ROBOTROLLING

PREPARED AND PUBLISHED BY THE

**NATO STRATEGIC COMMUNICATIONS
CENTRE OF EXCELLENCE**

Executive Summary

Generational change in malicious activity on social media seems to be at hand. Primitive bots indiscriminately promoting links to news sites are on the decline. They are being replaced by coordinated accounts that target conversations centred upon individual media outlets or members of different elites.

In recent months on Twitter, the volume of automated content about NATO activity in the Baltics and Poland has declined at an increasingly rapid pace. The number of bot-tweets dropped by 15 percentage points for Russian and 20 percentage points for English.

We infer that this reduction is best explained by changes introduced by the platform. Our findings are verified by drawing on thirty times more data than for previous Robotrolling issues. For the first time we include messages from VKontakte as a control.

We see a marked rise in organised trolling activity conducted by humans using fake accounts compared to early 2017. As of January 2018, 50% of all Russian-language messages are directed at other Twitter users.

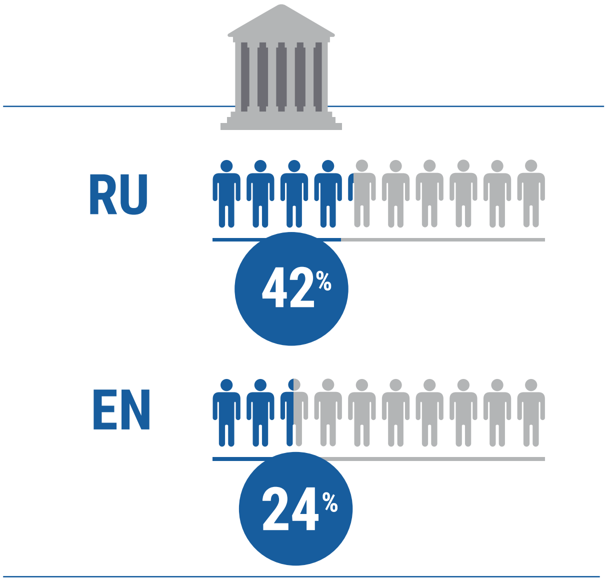
As social media companies intervene to clean up automation, they should take care that changes they introduce may enable new forms of manipulation. Russian-language bot activity is in decline in absolute terms, but Twitter in Russian remains more polluted than Twitter in English. ■

The Big Picture

This quarter, human-controlled trolling through fake profiles has occupied space vacated by automated activity. We attribute this in part to Twitter‘s efforts to cut down on misuse of the platform.

In the period 1 November 2017 – 31 January 2018, Russian-language bots created roughly 55% of all Russian-language messages about NATO in the Baltic States and Poland. Overall, 42% of accounts active in Russian were predominantly automated. In comparison, bots make up 24% of accounts tweeting in English. These bots created 31% of all English-language messages this quarter. Compared to the previous issue of Robotrolling, we observe a sharp drop in bot activity this quarter for both Russian and English language spaces. According to our latest estimates, the number of bot-tweets dropped by 15 percentage points for Russian and 20 percentage points for English.

This issue of Robotrolling analyses Twitter-mentions of NATO together with one or more of the host countries Estonia, Latvia, Lithuania, and Poland. The total number of posts considered is 5 500, of which 30% are in Russian. The number of active users is 3 400. In this issue we introduce two new data sets to help us understand bot dynamics. The first of these is a collection of 81 000 Russian-language Twitter posts mentioning NATO; the second is a collection of 73 000 posts mentioning NATO on the Russian social network VKontakte (VK). In both cases, observations are for the period 1 September – 31 January. Whereas on Twitter we see a reduction in Russian-language posts about NATO, no such decline is visible for VK. ■



Country Overview

Bot activity in English peaked in early November, prompted by Lithuanian Defence Minister Karoblis' assertion that NATO would soon agree to extend air defence in the Baltic region. By contrast, Russian-language bot activity peaked in response to a statement by General Terras that the Russian exercises Zapad-2017 had simulated full-scale war against NATO. The promoted links pointed principally to outlets representing the imaginary state of Novorossia and the Russian nationalist community such as NovorosInform, Infopolk, and Ruvesna. These sites pushed a divisive narrative, positioning Estonia as a third-party seeking to exacerbate NATO-Russia tensions: 'Estonia accuses Russia of preparing war against NATO.'

Figure 2 shows that most Russian-language bot content was about Estonia. In a first for the Robotrolling series, the chart shows that Russian-language messages about NATO in Poland and Latvia were more likely to come from human-controlled accounts than from automated accounts. Figure 3 shows that, compared to the previous quarter, the reduced levels of bot activity were especially visible for Latvia and Poland. In contrast, Lithuania saw the lowest reduction in bot activity.

Estonia

This quarter Estonia regained its position as the prime target of Russian-language bot activity, with 63% of the 620 mentions being made by bots. The main events commented upon were the Cyber Coalition 2017 Exercise conducted in Tartu, and news reports that the Zapad 2017 exercises had been aggressive in nature and more extensive than initially thought. Compared to previous periods, the proportion of bot mentions of NATO activity in Estonia has declined considerably. Nonetheless, automated activity still outweighs human activity in the Russian language space.

Latvia

In the current analysis window there have been few news stories about the NATO presence in Latvia, and the volume of Twitter posts has been muted for both English and Russian. Of the 260 mentions, 49% came from bot accounts. The main incidents drawing commentary related to investments in the Adaži military base, speculation by an opposition politician that NATO might stage a provocation on the Russian border, and calls by the Ministry of Defence for increased NATO spending on regional sea and air defence systems.

Lithuania

In November, English-language bots shared content about Lithuanian calls for increased NATO air defence in the Baltics. Russian bots, on the other hand, mentioned Lithuania primarily within reports about NATO troop rotations, namely contingents arriving from Croatia and France. A few bots circulated the assertion by Russian right-wing nationalist politician Vladimir Zhirinovsky that Lithuania would 'be destroyed' in the event of a conflict between Russian and NATO. The proportion of bot activity was comparatively high, with 62% of all Twitter posts mentioning NATO and Lithuania.

Poland

Poland has for some time been the country receiving the lowest level of Russian-language bot attention. This trend continues in the current quarter, with the proportion of bot messages dropping to 41%. The volume of Twitter posts is also lower than in previous quarters. In November, President Andrzej Duda's expressions of support for Georgia joining NATO drew bot attention, as did reports that Poland was prepared to host a new NATO logistics command. In December, video material hosted by Sputnik and RT purporting to show a US military vehicle 'stuck in the mud' was heavily shared. ■

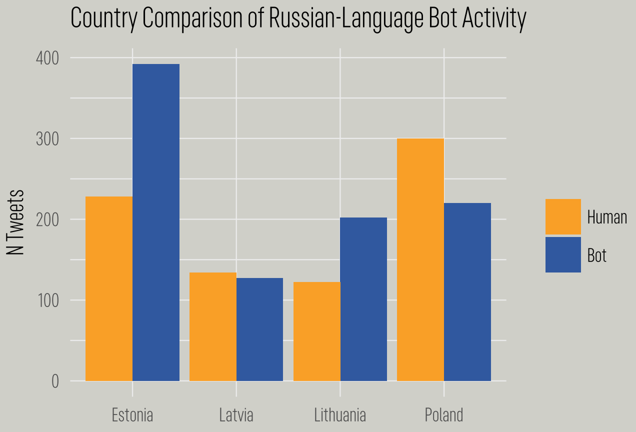


Figure 2: Distribution of Russian-language posts mentioning NATO and Estonia, Latvia, Lithuania, or Poland.

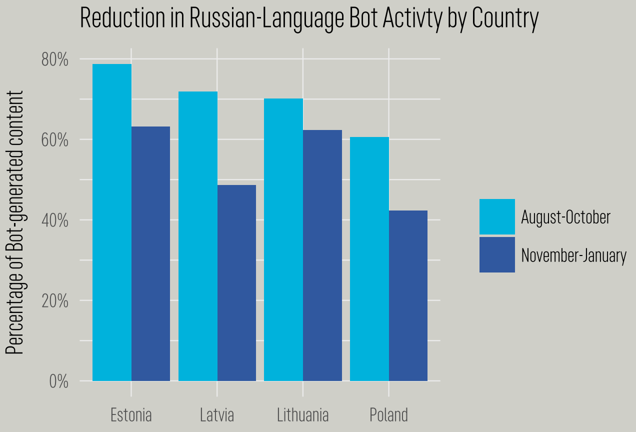


Figure 3: Comparison of the proportion of bot-generated Russian-language content per country for the periods August – October and November – January.

Themes

The main theme in our data this quarter is the double-digit percentage point reduction of automated Twitter posts, both for Russian and for English. Are these improvements attributable to changes implemented by Twitter? Allegations that Russian-backed campaigns exploited Western social media platforms have dominated the media conversation for the past year. Social media companies have been pressured to clarify the extent and impact of the manipulation, and what steps they intend to take in response. As recently as 31 January, Twitter wrote to 1.4 million users who had interacted with one or more of the 3 800 accounts identified as part of a Russian propaganda campaign.

It is also possible that the observed decline is a result of reduced interest in NATO's activity on its eastern flank. In spring 2017 the main theme in bot activity was clear: bots promoting material about the NATO presence in the Baltics and Poland were talking about the news. Reports on the NATO presence in the Baltics and Poland resulted in spikes of hundreds of bot-promoted messages. Since June 2017, such spikes have been absent, despite the media scrutiny of the Russian war-games Zapad 2017.

To help determine which explanation is more plausible, we looked at all messages mentioning NATO on the Russian social network VKontakte. VK is a good reference point, as this platform is unlikely to be under the same pressure to tackle malicious activity as its American counterparts. To test whether the trends hold for a larger dataset, we included all Russian-language Twitter messages mentioning NATO.

Our results are summarised in Figure 4, which shows the dynamic for all posts mentioning NATO in Russian on VK and Twitter. The total volume of NATO-related activity on VK has been stable at roughly 1 300 posts per week (we have not yet separated humans and bots on VK). Human activity on Twitter is similarly constant at about 1 750 weekly mentions. Bot activity, however, exhibits a clear downward trend. It is striking that for eight weeks in a row in December and January, human activity outweighed bot activity.

These data confirm some of our previous findings: in 2017, bot activity regularly accounted for the bulk of the Russian-language Twitter conversation about NATO. Additionally, since June 2017 there has been a steady and generalizable decline in Russian bot activity about NATO on Twitter. No such reduction is visible for VK, suggesting the reduction is better explained by changing dynamics on the Twitter platform than by the news cycle.

The decline in bot activity suggests that Twitter is beginning to tackle the bot-problem also in non-English-language spaces. This is an important development, as clearing out primitive but extensive spam makes it easier to identify and counter more sophisticated disinformation efforts. ■

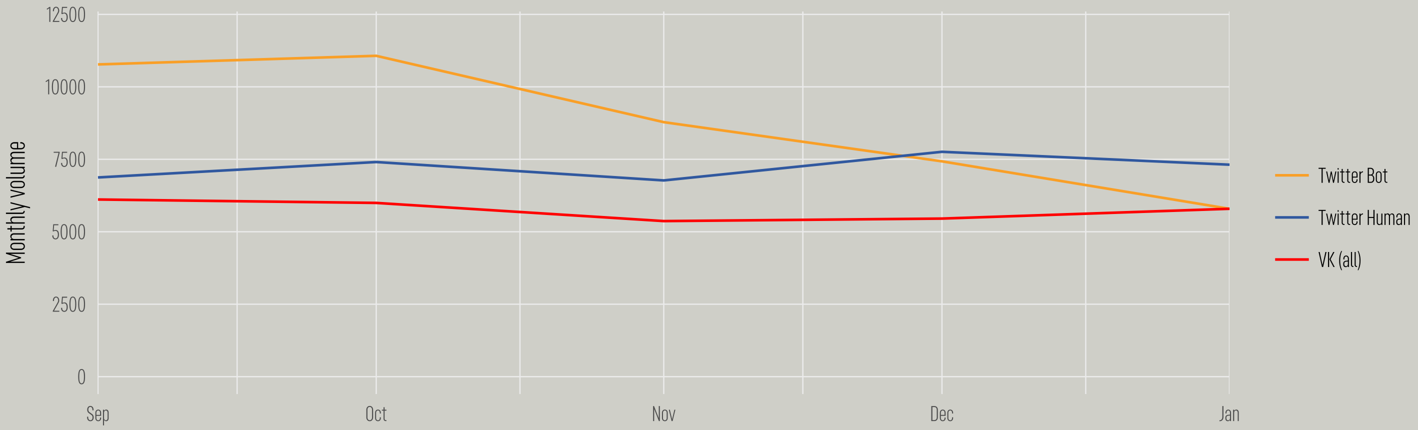


Figure 4:Timeline of all Russian-language mentions of NATO on Twitter and VKontakte for the period September 2017 – January 2018.

Robo-topics

The topics targeted by bots and the dynamics of robotic activity are evolving. Our impression is that there now are fewer synchronised bots pushing high volumes of news-related spam.

The network diagrams in Figure 5 visualize the changing dynamics within the conversation about NATO in the Baltics and Poland. The diagram to the left contains data for March 2017; the diagram to the right for November 2017 to January 2018. The shaded areas contain users active in English; the unshaded space, those active in Russian. The nodes correspond to users; the edges connecting them represent unusually similar follower/following patterns. Bots are coloured in blue, humans in yellow. Node size is scaled to show the proportion of a user's tweets that mention one or more other users. The exact same selection criteria were applied when creating the diagrams.

The figure shows an increase in unusually similar human-controlled accounts, indicating that although coordinated bot activity is in decline, organised troll activity is on the rise. It also demonstrates that bots active in English exhibit greater levels of synchronisation in the current quarter; in March 2017 virtually all anomalous accounts were Russian language bots.

1) The larger, more tightly connected clusters represent substantial groupings of fake accounts or bot networks. A number of the largest, crudest bot networks have disappeared since March.

2) However, bot networks are still in evidence, with a number of smaller, new groupings emerging.

3) One bot group has grown in size since March.

4) The most visible change this quarter is the emergence of large yellow nodes in the centre of the constellation. These represent groups of fake, human-controlled accounts that predominantly send messages directly at other Twitter users. Within the media spaces studied, this is a new form of possibly intimidatory virtual manipulation.

While Russian-language bot activity is in decline in absolute terms, Figure 5 suggests that the Russian-language space continues to be more polluted than the English-language equivalent. Thus, while Twitter's efforts should be commended, our impression remains that manipulation in foreign language spaces is both more primitive and more extensive than would be tolerated for English. ■

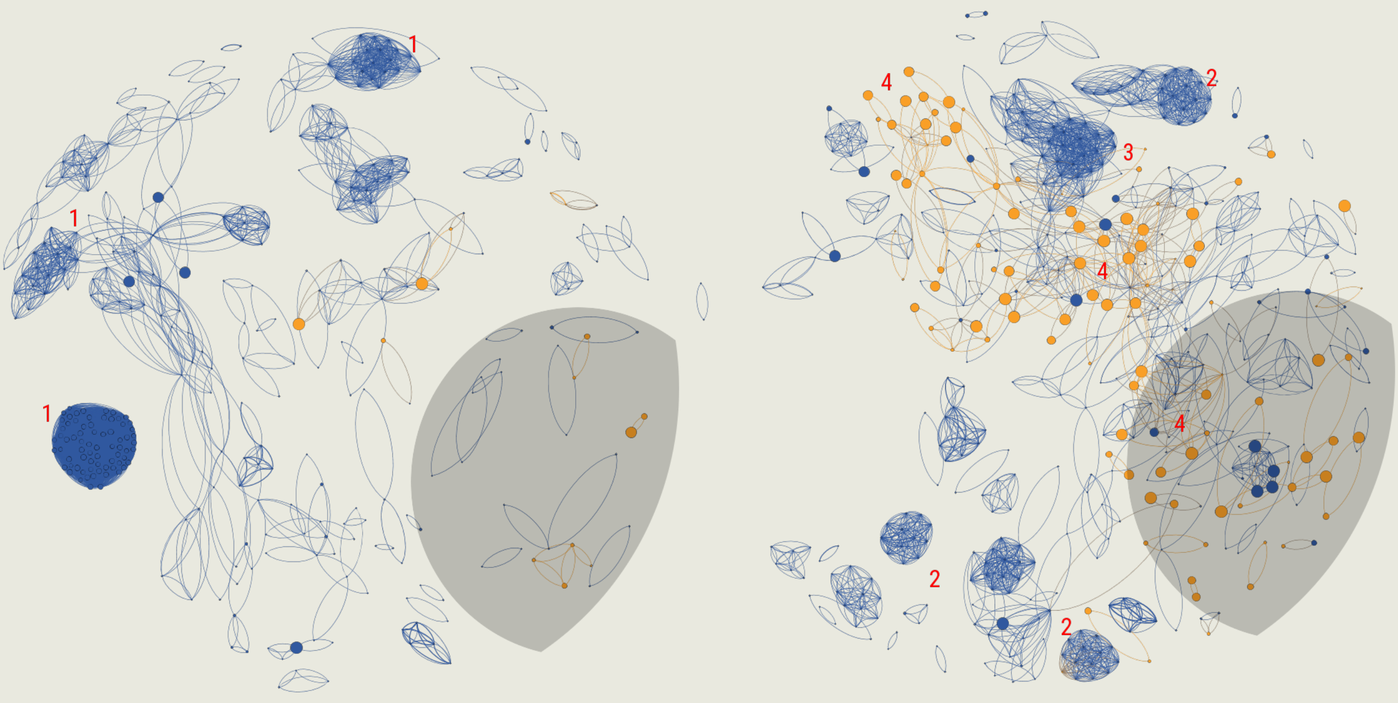


Figure 5: The evolution of bot activity. March 2017 (left) vs. November-January (right)

In Depth: Targeted Manipulation

The quantitative evidence suggests that tactics used to manipulate public opinion through social media are changing. Russian Channel 1's (Pervii Kanal) political talk show Vremya Pokazhet (Time Will Show) offers a practical example of innovation in this space. The show has a record of manufacturing public opinion: from the hand-picked audience's choreographed boos to the 'invited' Western experts whose voices are drowned out by the expert panel, the show is an exercise in carefully controlled political entertainment. On 20 June 2017 the show took the outrage online, as the show's host for the first time encouraged viewers to get involved and tweet their comments. Over the weeks and months that followed, messages from irate Twitter-users scrolled across the screen in accompaniment to the loud on-screen action. Many of the accounts seemed fake, prompting Russian independent media such as TV-Rain and Vedomosti to speculate that Channel 1 had hired a bot-net to artificially promote its show.

Frequently these 'viewers' referenced the NATO presence in the Baltics and Poland, and consequently they ended up in our dataset. The accounts in question are indeed low-quality fakes, featuring randomly generated screen-names (e.g. @1UgO9ZVwodOjiVj). The accounts initially performed no other action than tweeting messages 'at' the show. No retweets, no favourites, no links. According to our analysis, the majority of these accounts are fake, human-operated accounts. Some of them are represented by large yellow nodes in Figure 5.

As Figure 5 demonstrates, activity of this type is not merely more visible due to the disappearance of bots; it is growing in absolute terms, both for English and for Russian online media spaces. Most of the content we observe is directed at media outlets, political scientists, and politicians.

The messages directed at Vremya Pokazhet are the most visible example of fake accounts being used to simulate popular outrage about a subject. As the volume of automated content has declined, the importance of synchronised and targeted messaging from human-controlled accounts has increased. According to our data about NATO in the Baltics and Poland, in March 2017 only 15% of all Twitter messages were 'mentions' directed at other user accounts; in January, more than 50% of all Russian messages and more than 30% of English messages were directed at other users.

Changes introduced by Twitter in September 2017 have helped drive this increase: now it is possible to 'tag' up to 50 users in a Twitter post. And posts that begin with an 'at' mention are now visible on the front page of a user's timeline. As social media companies intervene to clean up automation, they should remain vigilant and mindful that changes introduced to their platforms may enable and incentivise other forms of manipulation. In sum, news links promoted indiscriminately are on the decline, whereas synchronised interventions targeting media outlets and elite figures are on the rise.

It is important to bear in mind that Robotrolling is based on a sample of Twitter-data about military activity in the Baltics and Poland. This sample is not representative of Twitter or content on other social media platforms. Future issues of this product will continue adding more representative data samples, will expand to consider other social networks popular in the region, and will employ more nuanced account classification. See our online FAQ for details on methodology. ■

Prepared by Dr. Rolf Fredheim and published by

NATO STRATEGIC COMMUNICATIONS
CENTRE OF EXCELLENCE

The NATO StratCom Centre of Excellence, based in Latvia, is a Multinational, Cross-sector Organization which provides Comprehensive analyses, Advice and Practical Support to the Alliance and Allied Nations.

www.stratcomcoe.org | @stratcomcoe | info@stratcomcoe.org