Turning the Virtual Tables: Government Strategies for Addressing Online Opposition with an Application to Russia

Sergey Sanovich*

Denis Stukal

Joshua A. Tucker

New York University†

October 7, 2016

Abstract

Web and social media – once a great hope for online activists from autocratic countries to obtain a platform where they could counter state propaganda in traditional media – are no longer safe heavens for alternative opinions as governments are increasingly as assertive and comfortable there as activists. We introduce a novel classification of strategies employed by autocrats to combat online opposition generally, and opposition on social media in particular. Our classification distinguishes both online from offline responses and exerting control from engaging in opinion formation. For each of the three options – offline action, technical restrictions on access to content, and online engagement – we provide a detailed account for the evolution of Russian government strategy since 2000. In addition, for online engagement option we construct the tools for detecting such activity on Twitter and test them on a large dataset of politically relevant Twitter data from Russia, gathered over a year and a half. We make preliminary conclusions about the prevalence of "bots" in the Russian Twittersphere.

*Corresponding author: sanovich@nyu.edu.

†All three authors participated in the research design, editing, and revision of this paper. Sanovich wrote the first draft of what is now Sections 1-4 and Appendixes A-C, and Stukal wrote the first draft of Sections 5 and 6, as well as the code for the computational analysis. Stukal recruited and Sanovich trained coders, whose coding is presented in Section 5. We are extremely grateful to Duncan Penfold-Brown, who has collaborated with us on the larger project of which this paper is a part. The data was collected by the NYU Social Media and Politlcal Participation (SMaPP) laboratory (`https://wp.nyu.edu/smapp/`), of which Tucker is one of the co-Directors along with Richard Bonneau, John T. Jost, and Jonathan Nagler. The SMaPP lab is supported by the INSPIRE program of the National Science Foundation (Award SES-1248077), the New York University Global Institute for Advanced Study, the Moore-Sloan Data Science Environment, and Dean Thomas Carew's Research Investment Fund at New York University.

1 Introduction

On December 3rd, 2014, the Russian news website *Meduza.io* reported that the 100th mirror of another Russian news website, *Grani.ru*, had been banned by the Russian Federal Service for Supervision of Communications, Information Technology and Mass Media (*Roskomnadzor*). *Grani.ru* was a popular news website with extensive coverage of opposition activity and alternative opinions. It was blocked in the Spring of 2014, at the height of Russian-Ukranian conflict, using a technical system developed by *Roskomnadzor* to block content deemed as extremist, as allowed under Russian law. *Meduza.io* itself is a new Russian media, established in the neighboring Baltic state of Latvia by Galina Timchenko, the former editor-in-chief of the most popular Russian news website *Lenta.ru*, who was dismissed by *Lenta.ru's* owner over the coverage of the Russian-Ukrainian conflict and moved to Latvia's capital, Riga, along with most of *Lenta.ru's* editorial staff. Around the same time, one of the most popular political blogs in Russia, which belonged to Russian opposition leader Alexey Navalny, was also permanently banned on the "LiveJournal" platform, and in early 2015 authorities began to crack down on its mirrors too.

While one might expect this type of response from regimes like Putin's one, its response to unfriendly activity online (and on social networks in particular) has not always been through bans and legal action. As late as 2010, a report of the *Internet in Russian Society* program at the Berkman Center for Internet and Society at Harvard University noted that "the political blogosphere appears to remain a free and open space for Russians of all political stripes to discuss politics, criticize or support government, fight corrupt practices and officials, and to mobilize others around political and social causes" (Etling et al. 2010). Moreover, as recently as 2009 the newly elected Russian president Dmitry Medvedev opened his own blog at LiveJournal and subsequently established a presence on Twitter and Facebook, as did many of his aides. Accordingly, the pro-government youth movements, which were created to confront possible "colored revolutions" on the streets of Moscow, were charged with the duty of competing with oppositional voices in the cyberspace and promoting government-friendly content (Kelly et al. 2012). In some cases they had even engaged directly with leading oppositional bloggers on the pressing issues of the day. In more recent years we have also witnessed a widespread proliferation of pro-government bots in the Russian Twittersphere as well as the notorious "troll factory" in St. Petersburg documented in the pages of the *New York Times*[1].

Why were the changes in policy so quick and dramatic? What is the menu of options the government can choose from to respond to emerging online challenges? Might a different country (or leader) have responded differently? The goal of this article is to (1) argue that these are indeed important questions for political science research to address; (2) introduce an organizational framework for doing so, and (3) provide a proof of concept that we can employ digital forensic techniques to both identify and analyze a particular form of online response to opposition by governments: the use of automated accounts on Twitter known as "bots". Accordingly, we begin by presenting a new classification system for different forms of government response to online opposition. In particular, we focus on how users of online media experience government attempts to address online opposition. We suggest there are essentially three types of options for governments: offline responses, which include legal action and market regulation in addition to more traditional forms of (physical) intimidation; online infrastructure responses, which rely on digital tools to filter the information available to end users; and direct online engagement with users that aims to shape online conversations, usually through content generation; in all cases we provide empirical examples of how governments have utilized these strategies. As an illustration of the utility of this framework we provide a detailed case study of the evolution of Internet policies in Russia during Putin's first term in office (2000-2008), the Medvedev's interregnum (2008-2012), and the period of time since Putin's return to the Kremlin after 2012. In particular, we investigate why the government almost completely ignored the Internet when it was actively imposing its will on traditional media and why this policy changed after Putin left the Kremlin in 2008. We also look at why,

[1] Adrian Chen, "The Agency," The *New York Times*, June 2, 2015,http://www.nytimes.com/2015/06/07/magazine/the-agency.html.

during Medvedev's presidency, online engagement rather than imposing heavy restrictions was chosen as a primary strategy and why this choice was reversed when Putin and Medvedev switched offices in 2012.

In the second half of the paper, we turn to quantitative methods to demonstrate the feasibility of actually trying to identify attempts by governments to utilize the third strategy of *online engagement*. More specifically, we apply digital forensic techniques to identify "bots", or automated (machine controlled) accounts, in the Russian political Twittersphere[2]. We introduce a new, exhaustive framework for classifying Twitter accounts as official accounts, bots, cyborgs, human accounts, or spam, plus sub-categories for bots and humans. We demonstrate five different techniques for identifying bots, which prove remarkably adept at finding bots in a collection of politically-relevant Twitter data. These techniques also helped us to locate bots with highly ideologically charged content.

Although this empirical work is largely intended to function as a proof of concept analysis, our initial hand-coding of a small number of accounts identified by our bot-detection methods suggests one interesting finding that warrants further investigation. For the accounts we can identify as politically oriented bots, only slightly more than half of them appear to have a pro-government orientation; we also find evidence of both pro-opposition and pro-Ukrainian bot activity. This suggests that simply assuming all political bots are pro-government – and therefore part of a government strategy for online engagement – in an analysis of this type would be a mistake.

2 Literature

Various forms of government reaction to online activities have been the subject of intensive research in recent years. This literature covers everything from legal regulation of the Internet

[2] We draw upon a collection of 28 million tweets collected using key-word filtering over a year and a half period (November 25, 2013 - July 13, 2015) during a particularly tumultuous period in recent Russian history including the Sochi Olympics, the Euromaidan uprising in Ukraine and the subsequent annexation of Crimea and Russian involvement in the war in Eastern Ukraine.

in world's most advanced democracies (Giacomello 2008; Travis 2013) to online censorship tools used by different types of autocratic governments around the world (Ko, Lee, and Jang 2009; MacKinnon 2011; King, Pan, and Roberts 2013; Nabi 2013) to establishing an active presence of governments on social media platforms (Barash and Kelly 2012; Pearce 2014). Freedom House even produces an annual international index of Internet freedom (Freedom House 2011, 2012, 2013, 2014). However, few studies attempt to provide a framework for the systematic analysis of government behavior on the web that would allow us to analyze why particular tools are chosen under any given circumstances, or even what makes up the choice set in the first place.

Notable exceptions are Deibert et al. (2011), Morozov (2011) and Roberts (2014). Deibert et al. (2011) provide a historical framework, which traces the evolution of state-web relationships from initial laissez-faire non-involvement (until 2000) to attempts to either deny altogether (2000-2005) or carefully control access to cyberspace (2005-2010) all the way to the current (after 2010) stage of active contestation between state and corporate censors on the one hand, and cyber-activists on the other. This framework is meaningful to describe global trends, but is not closely followed in that order by each particular country. As we shall see, the Russian government turned to access denial only after it tried and failed both (in Deibert's terms) control and contestation. In addition, while they provide many relevant examples of the ways government tried to deny or control access online, they do not provide any systematic classification of the types of action the state may take. Roberts (2014) does provide such a classification, but distinguishes between the effects (fear to speak or to listen and either friction or flooding as impediments for access) rather than the tools employed. Indeed the same tools could lead to both flooding and fear, and different kinds of tools – online and offline – could be used to increase friction.

Morozov (2011) does distinguish between technological and what he calls "sociopolitical" means of controlling online activity, the latter combining technology with online and offline actions by humans.[3]

Morozov hypothesizes that if "liberation technologies", such as those promoted by Diamond (2010), were to succeed, embattled governments could turn to potentially more violent methods such as smearing campaigns, criminal prosecutions or even physical attacks on bloggers: "as technological methods lose efficacy, sociopolitical methods could simply overtake them: an authoritarian government might find it harder to censor blogs, but still rather easy to jail bloggers" (Morozov 2011, 63).

While this option is certainly not hypothetical, and Morozov's classification is useful for studying the dangers and promises of "liberation technologies" (what makes sociopolitical response different is exactly being beyond the reach of these technologies), it does not fully distinguish between government actions that restrict, or otherwise structure, online media environments and those where the government actively engages in shaping the formation of opinions online. This distinction is important for at least three reasons. First, while in *censoring* online media the government could build on strategies from long before the Internet was created, online *propaganda* in distributed networks is fundamentally different from a top-down broadcasting of the party line through hierarchical monopolies of traditional media. Second, this part of the government response is experienced differently by users: not as an outcome (e.g., an inaccessible web page), but as a point of interaction with the state (e.g., a paid pro-government troll replying to your tweet). Last but not least, the study of government online activities will increasingly focus on social media, which is simultaneously the most abundant and versatile data source and the key point of contestation between the government and civil society (Deibert et al. 2011; Etling, Roberts, and Faris 2014; Pearce 2014; Lange 2014; Gunitsky 2015; Munger et al. 2015). Since social media data capture exactly the moment of interaction between the user and the state, it is important to understand the place

[3] Technological responses include Internet-filtering, which spans from targeted bans of particular webs-sites and keywords to larger national-level schemes to block entire segments of the Internet (China) or – in the extreme – any outside Internet access outside the country (North Korea). Sociopolitical responses range even more widely from distributed denial-of-service (DDoS) attacks to employing both automated bots and paid trolls to destroy online communities' social capital to physical attacks on bloggers.

of government action leading to this interaction in the wider menu of options available to the government.

Therefore, we propose a new classification of government options for responding to independent online activity. In addition to differentiating between "**offline**" and "**online**" tools, it distinguishes between online tools aimed merely at restricting the flow of information and those entailing active engagement with the users on behalf of the government. While the former operates largely – although not exclusively – through exerting control over Internet infrastructure, the latter typically involves some content generation. Since users experience each of these possible government actions differently, our classification, effectively, dissects government options *from the Internet user's point of view.* In the next section we discuss each option in detail, providing examples and identifying the key resources needed to employ each of these options. This classification will then inform our analysis of the strategy pursued by the Russian government in Section 4.

3 A classification system for government responses to online opposition

In this section, we introduce our tripartite system for classifying government responses to online opposition. We begin with *offline response*, which primarily refers to changing country's legal Internet regulations, but also includes attempts to change ownership structure of online media and intimidate particular users. The second category encompasses various ways to technically *restrict access* to online content, from firewalls to DDoS attacks to sophisticated online censorship systems. The final category also involves online activity, but instead of focusing on restricting access to content, this tactic involves creating content to *engage* with users online.

3.1 Offline response

The first set of options at any government's disposal is based on digital age implications of traditional governing advantages: nodality ("network centrality"), organizational capacity, legal authority to enforce the law, a monopoly on the legitimate use of violence and the right to regulate human activity and, last but not least, the ability to expend large financial resources through taxation (see a detailed discussion in Ackland 2013, Chapter 8). The actions facilitated by these advantages could have a huge impact online, but take place offline; thus end-users either observe the consequences online after-the-fact (and, if necessary, adjust their behavior) or encounter these actions in person, but offline. The latter case, of course, applies to legal prosecution and violence against (usually, social media) users. In more favorable circumstances, users just observe the outcomes of government actions when commenting functionality is turned off by their favorite news web-sites after readers' comments become legally designated as media content, with all related legal obligations and risks on part of media outlets hosting them.

Another option is to require popular bloggers to register with the government, making each individual blogger responsible for her own content, on par with actual commercial media outlets, as it has recently been done in Russia[4]. As Ackland notes, such a legal designation – as well as other forms of regulating user-generated content – could be attributed by the government to either genuine or fabricated popular demand stemming from concerns over public safety or morality (143, 145-146).

Finally, the government could attempt to change the landscape of the digital media market and alter the choice of online platforms available to users. Relying on their authority to regulate commerce, autocrats around the world designate certain companies and industries, including telecommunications, as "strategic", upon which they start to enforce various restrictions, such as banning foreign ownership and/or investments, appointing state representative

[4] Neil Macfarquhar, "Russia Quietly Tightens Reins on Web With 'Bloggers Law,'" *The New York Times*, May 6, 2014, http://www.nytimes.com/2014/05/07/world/europe/russia-quietly-tightens-reins-on-web-with-bloggers-law.html. See also Appendix B.3.

to the board, etc. For example, in late 2013 the publicly-owned – but heretofore relatively independent editorially (especially, in its popular social media operations) – major Russian news agency *RIA Novosti* was stripped of its leadership, restructured, rename, and put under the leadership of a fervent regime supporter[5]. Nevertheless, in order to ensure the complete control of this already loyal news outlet, in early 2014 it was included in the list of "strategic enterprises", along with the second largest Russian news agency, *ITAR-TASS*[6].

In 2014 Russia changed its media law to ban (beginning January 1, 2016) foreign ownership (defined as more than 20%) of any media operating in Russia. In addition, any foreigner or Russian citizen with another citizenship was banned from serving as the founder or editor of any media[7]. While this law primarily affected print media, many print publications maintain a significant presence on the web and are widely shared through social media[8].

If control over digital media is challenging or costly to legislate or order, especially in the case of private companies, then governments can use other means, in particular purchasing power and extra-legal pressure, to assume control over important Internet platforms. The so-called "Russian Google", Yandex, sold a "golden share" to state-owned *Sberbank* in 2009, allegedly after negotiations with Dmitry Medvedev and multiple proposals to designate companies such as Yandex as "strategic", which would have forced them to re-register in Russia[9] and severely

[5] Sergej Sumlenny, "Bad News: What Does the Closure of RIA Novosti Mean for Media in Russia?," *Calvert Journal*, December 12, 2013, http://calvertjournal.com/comment/show/1837/RIA-novosti-putin-russian-media-kiselyov.

[6] Gabrielle Tetrault-Farber, "RIA Novosti Begins Cutting 1/3 of Staff," *The Moscow Times*, March 12, 2014, http://www.themoscowtimes.com/news/article/ria-novosti-begins-cutting-13-of-staff/495980.html.

[7] Natalia Gulyaeva, Maria Baeva, Oxana Balayan, and Maria Sedykh, "Russia Tightens Foreign Ownership Restrictions in Media," *Hogan Lovells Global Media and Communications Watch*, October 20, 2014, http://www.hlmediacomms.com/2014/10/20/law-restricting-foreign-ownership-in-media-business-in-russia/; Michael Birnbaum, "Russia's Putin Signs Law Extending Kremlin's Grip over Media," *The Washington Post*, October 15, 2014, http://www.washingtonpost.com/world/europe/russias-putin-signs-law-extending-kremlins-grip-over-media/2014/10/15/6d9e8b2c-546b-11e4-809b-8cc0a295c773_story.html.

[8] The Moscow Times, "15 Global Firms Hit by Russia's Law Limiting Foreign Ownership of Media," September 28, 2014, http://www.themoscowtimes.com/article/507968.html.

[9] Yandex is incorporated in the Netherlands as Yandex N.V. – a fact that in 2014 was publicly condemned by Vladimir Putin at his meeting with People's Front for Russia. See Christopher Brennan, "Putin Says CIA Created the Internet, Cites Foreign Influence at Yandex," *The Moscow Times*, April 24, 2014, http://www.themoscowtimes.com/news/article/putin-says-cia-created-the-internet-cites-foreign-influence-at-yandex/498903.html.

diminish their appeal to international capital markets[10]. In 2014 Yandex founder Arkady Volozh resigned as the company's Russian CEO[11]. A similar attempt was made in the case of Vkontakte, known as the "Russian Facebook", which resulted in the hostile takeover of the company by business groups loyal to the Russian government[12] and founder and former owner and CEO Pavel Durov (and his team) fleeing the country[13].

Of course, instead of attempting to take control over existing public and private media and communications platforms, the government could try to increase its influence through artificially generated competition. In several countries, including Russia and Turkey, governments reportedly allocated generous funds to the creation of "national" search engines, social networks or email services.[14] A Russian national search engine has been discussed since at least 2008, when it was mentioned by then President Dmitry Medvedev after the so-called 5-days war with Georgia. Turkey began a similar project in 2009 (Morozov 2011). The results of government investments in creating artificial competition are less than impressive, though. In late 2014, the Turkish project was still in development, with a scheduled launch in 2016[15]. In Russia, *Rostelecom* – the largest telecommunications company in the country, with the government as its majority shareholder – released a beta-version of the search engine "Sputnik" in May of 2014, but it has not been able to acquire a noticeable market share as of yet. In the Russian case, though, loyalty and successful hostile takeovers of existing platforms eliminated the need to build new ones from scratch.

[10] Nikolay Grishin, "Yandexed Everything," *Kommersant - Trade Secret*, March 12, 2012, http://www.kommersant.ru/doc/2065978.

[11] He kept the executive position in the international operations, though. See Nadia Beard, "Founder and CEO of Yandex, Arkady Volozh, Resigns," *Calvert Journal*, August 26, 2014, http://calvertjournal.com/news/show/3035/founder-of-yandex-resigns-amid-controversy-arkady-volozh.

[12] Nickolay Kononov, "The Kremlin's Social Media Takeover," *The New York Times*, March 10, 2014, http://www.nytimes.com/2014/03/11/opinion/the-kremlins-social-media-takeover.html; Joshua Yaffa, "Is Pavel Durov, Russia's Zuckerberg, a Kremlin Target?," *Bloomberg Businessweek*, August 1, 2013, http://www.businessweek.com/articles/2013-08-01/is-pavel-durov-russias-zuckerberg-a-kremlin-target.

[13] Ingrid Lunden, "Durov, Out For Good From VK.com, Plans A Mobile Social Network Outside Russia," *Techcrunch*, April 22, 2014, http://techcrunch.com/2014/04/22/durov-out-for-good-from-vk-com-plans-a-mobile-social-network-outside-russia/.

[14] In the Russian case these pleas were especially suspicious given that Russia already had a "national" search engine, social network and email service, all created without any government aid. Yandex and Vkontakte have larger market shares in Russia than Google and Facebook, respectively, an impressive result in the absence of any protectionist measures against foreign competitors.

[15] http://english.yenisafak.com/news/tubitak-to-develop-national-search-engine-2024960

3.2 Online response: restrictions

Offline means of controlling online activity are popular among autocrats around the world, not the least because they usually require zero IT competence or resources. However, rapid growth in Internet penetration rates and the emergence of the Internet as a principal source of information for increasing numbers of people creates challenges even for autocrats who are able to successfully employ offline tools of control. To begin with, information can be produced and distributed by foreign citizens and entities that are out of reach for the autocrat's security apparatus. Second, some local activists and/or journalists can use their digital proficiency to distribute information anonymously and therefore avoid offline prosecution. Finally, for various reasons autocrats could simply prefer putting flows of information under their control rather than going after its producers. If, for example, an autocrat wants to avoid taking responsibility for the government's actions, a DDoS attack on a popular oppositional blog can be blamed on "unidentified" hackers, while most types of offline response require at least some involvement of the state apparatus.[16] Thus, governments facing a serious threat from the opposition or an insurgency may try to acquire tools to exert control over Internet infrastructure and deploy such tools either routinely or in times of political instability.

Of course, there always exists the option to completely monopolize the telecommunication infrastructure inside the country and cut any connections with international networks. North Korea did just that: it maintains *Kwangmyong*, a national intranet, and a national mobile phone service *Koryolink*. The North Korean authorities have complete control over the information accessible through *Kwangmyong* and tightly monitor personal communications through *Koryolink*. Communications with the outside world through both channels are prohibited (except for the ruling elite and foreign tourists). However, such a system imposes a heavy toll on the national economy.

[16] Even if the attack on a blogger is carried out by private citizens, and the government claims to have no relation with them, there is a strong expectation and laws on the books that require criminal investigation in the case of violence or even threat of violence. DDoS attacks on the other hand, and cybercrime in general, remain weakly (if at all) legally regulated in most countries, and usually neither the law nor public opinion demands any action on the part of the government in the case of a cyberattack.

A step removed from this extreme approach – albeit still with non-trivial costs – is the highly sophisticated Chinese "Great Firewall", probably the best example of blocking sensitive information without fatally hurting either government communications or commercial activity[17]. Country experts anticipate that even North Korea will eventually take the Chinese and Cuban path of establishing such a "Mosquito-Net" model of Internet access, which facilitates the "use [of] the Internet as a propaganda machine in addition to taking advantage of it economically, [...] while keeping out information deemed threatening by the regime" (Chen, Ko, and Lee 2010; see also Ko, Lee, and Jang 2009).

Targeted Internet-censorship is well documented by King, Pan, and Roberts (2013), who describe the immense Chinese system of monitoring and censoring of UGC across the country's dispersed social media platforms. They estimate that around 13% of all social media posts get censored, and find that most of the censorship happens within 24-hours from initial posting, and claim that in spite of obvious technical difficulties the level of censorship increases disproportionately during periods of especially voluminous discussions on social media after both significant political and non-political events.

The ability of autocratic governments to filter Internet communications (including the access to social media) is, primarily, a function of three factors: control over the critical infrastructure; planning ahead and implementing a long-term, comprehensive, but not overly costly solution; and financial and human resources as well as the technical expertise necessary to build filtering tools. The first component is rarely an advantage of autocratic regimes: most of the world's Internet infrastructure (ISPs, search and social media platforms, transaction services, etc.) is located in advanced democracies and therefore out of reach for autocrats seeking to control them. Consequently, the latter two factors – i.e., preparedness and money – determine ultimate success. According to Howard and Hussain (2013, 71-72) "sophisticated long-term investments in managing information infrastructure" made by countries such as Iran, Saudi Arabia and Bahrain have made these countries much less vulnerable to growing

[17] The Economist, "The Art of Concealment," April 4, 2013, http://www.economist.com/news/special-report/21574631-chinese-screening-online-material-abroad-becoming-ever-more-sophisticated.

discontent than Egypt, Libya, and Yemen, who have had to rely on ad hoc solutions. The key risk of the latter approach is that an abrupt shutdown of Internet access can end up hurting the government's ability to communicate with its own allies while simultaneously provoking further unrest by interfering with the non-political uses of Internet.

Two primary technological options for regimes are filtering/blocking of particular websites or segments of the web and DDoS attacks. The former has the advantage of being permanent and customizable. China, for example, blocks only certain platforms and content (by keywords), while North Korea famously maintains its local web segment in complete isolation from the outside web. Both policies, though, share a common disadvantage of this approach: high transparency for local users and susceptibility to documentation by outsiders (including other governments, human rights organizations, etc.) (Morozov 2011).

DDoS attacks, on the other hand, are usually hardly traceable, relatively cheap, can be deployed during particularly sensitive political events such as elections or protests and can be more easily outsourced to loyal but independent groups, such as the Syrian Electronic Army[18]. On the other hand, their ability to break up online communications is limited in time and web space, i.e., a small set of web-sites at best. Moreover, the most popular platforms, such as Google and Twitter, are highly protected from DDoS attacks.

The most distinctive (and important for our discussion) feature of Internet filtering is that it is observed by users as an end result and does not create any kind of interaction with the state (or its representatives) in course of user's *activity* online. If Twitter is blocked in your country (permanently, as in China, North Korea, Iran and several other countries, or temporarily, as in Venezuela and Turkey in 2014), you either cannot get there, or you can use one of the available tools (anonimizers, proxy-servers, etc.) to restore your access. In either case, users observe government actions as end results. The impact of such actions is either in successful breaking of inter-personal communication or access to websites, or the lack thereof.

[18] Helmi Noman, "The Emergence of Open and Organized Pro-Government Cyber Attacks in the Middle East: The Case of the Syrian Electronic Army," *The Information Warfare Monitor*, May 30, 2011, http://www.infowar-monitor.net/2011/05/7349/.

In other words, the government cannot possibly shape the conversation through these means. To achieve this latter goal government has to directly engage with users online.

3.3 Online response: engagement

Establishing a government presence on the web and using it to promote the government's agenda constitutes the third and final option at a government's disposal. This type of government response actually takes place online and users encounter it in course of their online – particularly social media – activity. Mainly, it includes the government creating content, either through artificial intelligence or real human effort. However, hacking and publishing bloggers' personal communications (such as emails, instant messages, etc.) could allow the government to expose and implicate the opposition and shape the conversation that way. A typical example of the latter is the case of Russian blogger and hacker *torquemada_hell*, a Russian-speaking person allegedly living in Germany. In 2010-11 he successfully hacked the email accounts of multiple Russian opposition politicians and released potentially damning information to the public[19].

Still the most obvious – and increasingly popular – tool employed by governments to alter political conversations on the social media is using either "bots" or real people to advocate pro-government positions, turn conversation meaningless or prohibitively divisive, or distract users from sensitive political issues altogether.

Bots could perform two key functions: either clutter conversations with "digital dust" or alter search results, Internet rankings, top lists and other automated tools for sorting, sharing, discovering and consuming online content. As such, bots could be used to support real people. For instance, a ranking of the most popular Russian blog posts, maintained by Yandex, was closed in 2009, inundated by bots promoting mostly pro-government posts[20].

[19] Alexey Sidorenko, "Russia: Analysis of Hacker Attacks On Bloggers," *Global Voices*, June 20, 2010, http://globalvoicesonline.org/2010/06/20/russia-analysis-of-hacker-attacks-on-bloggers/.

[20] Alexey Sidorenko, "Russia: Major Search Engine Closes Its Blog Rating," *Global Voices*, November 6, 2009, http://globalvoicesonline.org/2009/11/06/russia-major-search-engine-closes-its-blog-rating/. The more pressing concern for Yandex, though, was government outrage in the cases when, instead, anti-government posts got traction in the ratings, see Alexandra Odynova, "Yandex to Close List That Annoyed

The possible functions of humans acting on the government side are much more diverse. It is useful, therefore, to provide a basic classification of pro-government users. This classification does not look into users' honesty, consciousness and convictions. Instead, it is based on formal or informal ties with the government (or the lack of thereof).

To begin with, the government could hire students or other low-paid workers to submit rather simple messages, which would nevertheless pass the human intelligence tests integrated in many modern social media platforms. Their messages could be either identical or contain the same message worded differently. One particular example of this type of bloggers are the so-called Chinese 50-centers[21]. Russian pro-government youth movements, such as *Nashi* and *Young Guard of United Russia* were often accused of running a similar network of 11-rublers[22]. Leaks released by the Russian arm of Anonymous in 2012 indicated that Nashi paid hundreds of thousand of dollars in fees for comments, statuses, Facebook likes, Youtube dislikes, etc.[23].

Cheap bloggers paid per comment are not the only group of friendly users that could be put on the government payroll. Bribing prominent and trusted bloggers, celebrities or journalists – although potentially much more expensive – could turn out to be a better investment in terms of persuading the public. The same leaks as noted in the previous paragraph revealed that along with paying small fees to thousands of low-skilled bloggers, Nashi also put aside tens of thousand of dollars to be paid to a small group of popular and heretofore considered independent bloggers for highly sophisticated positive publicity for the Russian leadership[24].

State — News," *The Moscow Times*, November 6, 2009, http://www.themoscowtimes.com/news/article/yandex-to-close-list-that-annoyed-state/388969.html.

[21] Sarah Cook, "China's Growing Army of Paid Internet Commentators," *Freedom At Issue Blog*, October 11, 2011, http://www.freedomhouse.org/blog/china%E2%80%99s-growing-army-paid-internet-commentators.

[22] Anton Nossik, "11 Rubles and 80 Kopecks per Comment," *Echo of Moscow*, September 10, 2013, http://www.echo.msk.ru/blog/nossik/1154616-echo/.

[23] Miriam Elder, "Hacked Emails Allege Russian Youth Group Nashi Paying Bloggers," *The Guardian*, February 7, 2012, http://www.theguardian.com/world/2012/feb/07/hacked-emails-nashi-putin-bloggers; Miriam Elder, "Polishing Putin: Hacked Emails Suggest Dirty Tricks by Russian Youth Group," *The Guardian*, February 7, 2012, http://www.theguardian.com/world/2012/feb/07/putin-hacked-emails-russian-nashi.

[24] Miriam Elder, "Emails Give Insight into Kremlin Youth Group's Priorities, Means and Concerns," *The Guardian*, February 7, 2012, http://www.theguardian.com/world/2012/feb/07/nashi-emails-insight-kremlin-groups-priorities.

The next group consists of government supporters whose social media activity is not paid *per se*, but is facilitated through participation in various political projects or actual employment by the government. These sets of bloggers range from members of various youth political movements to the MPs from the ruling (or affiliated) parties to relatively prominent politicians (ministers, party leaders) who are encouraged to take on the challenge of representing the "government's point of view" in an often hostile social media environment.

Finally, the government could also try to mobilize genuine supporters with no ties – formal or informal – to the government or ruling party. Obviously, this group could include various types of people, but two groups deserve special attention. First, if famous people – particularly, celebrities and journalists – volunteer to support the government agenda, it could help the autocrat both directly and indirectly through endowing the ideas already promoted by the armies of bots and paid bloggers with the weight of fame, reputation and personal independence. Second, people in the opposition could occasionally be legitimately attracted by government policies and join the ranks of pro-government bloggers. Prominent example from recent Russian history is the brief, but notable, excitement of previously largely oppositional nationalists about the Russian annexation of Crimea and support for separatists in Eastern Ukraine[25]. Obviously, this latter category rarely would end switching sides permanently. However, given their relative immunity to direct bribing and indirect manipulation, fluctuations in their support of the government could provide a useful baseline for studying other groups.

In the next section, we illustrate the usefulness of this taxonomy by discerning the evolution of Russian government policy regarding the Internet over the past decade and a half, or, put

[25] Tom Balmforth, "From The Fringes Toward Mainstream: Russian Nationalist Broadsheet Basks In Ukraine Conflict," *Radio Free Europe/Radio Liberty*, August 17, 2014, http://www.rferl.org/content/feature/26534846.html; Natalia Yudina, "Beware the Rise of the Russian Ultra-Right," *The Moscow Times*, September 11, 2014, http://www.themoscowtimes.com/opinion/article/beware-the-rise-of-the-russian-ultra-right/506876.html; Paul Goble, "Ukrainian Events Have Deeply Split Russian Nationalists," *The Interpreter*, July 20, 2014, http://www.interpretermag.com/ukrainian-events-have-deeply-split-russian-nationalists/; Ivan Nechepurenko, "How Nationalism Came to Dominate Russia's Political Mainstream," *The Moscow Times*, August 3, 2014, http://www.themoscowtimes.com/news/article/how-nationalism-came-to-dominate-russia-s-political-mainstream/504495.html.

another way, Russia in the age of Putin. This evolution proves to be not the usually discussed linear increase in the censorship efforts towards and especially during Putin's third term, but a complicated process of choosing the optimal strategy, which directly reflects both the political struggles inside the regime and the distinctive challenges associated with the implementation of each of the three options we identify. In the Section 5 of the paper, we turn to the feasibility of studying the least understood of the three response options, online engagement.

4 Russian government online: a constantly evolving strategy[26]

Russian government activities online gained serious international attention when they were redirected towards aiding Russian offensive in Ukraine in the wake of the Euromaidan Revolution in 2014. The resourcefulness and inventiveness of these actions as well as their reach (that went far beyond Russian borders and even Russian-speaking world) were all the more surprising for Western observers and policy makers since until then Russian authorities were not considered to be particularly artful in their digital operations, even for domestic purposes.

Indeed, compared with Chinese decades-long massive effort to block access to content deemed dangerous for domestic consumption, Vladimir Putin's government appeared to many as not only lacking the expertise, but even the interest in doing propaganda and counter-propaganda online. U.S. defense analysts, who discovered a 2013 article in an obscure military-industrial magazine by the Chief of the Russian General Staff, General of the Army Valery Gerasimov about "ambiguous warfare", went as far as suggesting that online media tools deployed by Kremlin to brainwash the Ukrainian population and whitewash Russian actions in the West were part of the elaborate military strategy clandestinely developed by the Russian military planners.

While it is undeniable that digital assets were strategically employed by Kremlin to achieve its foreign policy and military goals in Ukraine, Syria and elsewhere, their origins are much

[26] While this section provides an analytical summary of the Russian experience, the online Appendix B adds the necessary details on many specific cases of government interaction with online media, including evidence from the primary sources.

more likely to be civilian than military. And while it is true that early in his tenure Putin decided for the time being to avoid almost any interference in the online media, it was not done out of ignorance or lack of attention. Kremlin's approach to online media was from the very moment Putin came to power guided by a deliberate strategy. This strategy evolved under changing circumstances and through a long trial-and-error process, which was much more complicated than simple linear increase in the level of censorship and amount of propaganda.

Vladimir Putin is famously old-fashioned when it comes to digital tools: he rarely uses computer and never had any personal online presence, which in 2016 makes him an exception among heads of state. However, policy-wise he showed both interest in new technologies and awareness of potential government strategies regarding them. Even before he became acting President, in late 1999 he convened the leaders of nascent Russian IT industry and online media and made a clear commitment to protect their freedom and avoid Chinese-style filtering. While it is unclear whether he was concerned with the Russian image abroad or with offsetting the damage this image suffered after he took near total control over the traditional media, his choice was politically expedient, and not without a precedent in Russian history.

At only 2% Internet penetration in 2002 (and 16% at the end of Putin's second term in 2008), online media were a medium for personal communication more than of mass persuasion and as such were hardly an asset of any political significance. Following an old Soviet tradition, Putin avoided direct interference with personal communication channels[27]. This resulted in an emergence of thriving and competitive internet industry, whose leading companies – Yandex and Vkontakte – have won competition over Google and Facebook, respectively, and did it without the aid of any protectionist measures, a rare achievement for any country. Years ahead of most Western countries, Russian news media created from scratch online overtook websites of traditional media in popularity and began doing their own original reporting (instead of relying on existing offline news agencies and outlets). Meanwhile, Russian public created a vibrant blogosphere large enough to completely overtake the major blog platform of the time,

[27] However, again in line with the Soviet blueprint, an elaborate system of digital surveillance called *SORM*) was set up (Zasursky 2004, 181-183)

LiveJournal, which was eventually purchased by a Russian company.

As Internet penetration continued to rise steadily in the late 2000s and most traditional media became completely sanitized of any alternative opinion, online news media, most of which were rather critical of the regime, became more and more influential. However, the government first saw this as an opportunity rather than a threat.

To a large extent it was the result of a change in the government[28]. In 2008, freshly installed into Kremlin, Dmitry Medvedev and his team were looking for ways to build their own support base, sorely needed both to implement their modernization agenda and to get a fighting chance to stay in power for the second term. Medvedev published his modernization manifesto “Go, Russia” in online-only liberal newspaper *Gazeta.ru*. Recognizing the differences with traditional Russian media, where alternative opinions could be bought out or shut down, Medvedev and his team made a serious attempt to engage Russian online public in a genuine discussion of the country’s way forward. Both he and his aids created presence on multiple blog platforms, which earned for Medvedev a nickname “Blogger-in-Chief”. Both the tone and (to a certain extent) the message they have put forward there were different from the one intended to a wider audience of Russian TV and press.

Most crucially, they sought, received and responded to critical feedback from the audience, the practice unheard of for years in the traditional media, but required to get any attention in the vibrant Russian blogosphere at the time. Pro-government youth movements were mobilized to spread Medvedev’s message to every corner of the Russian segment of the Internet. While their activities were not without controversy (more due to corruption and incompetence than ideological zeal[29]), even they had to engage in genuine discussion with bloggers critical of the government, thus facilitating the public debate on important issues. While the

[28] That it became one of Medvedev’s government’s defining policies suggests how limited was the change (see on limits of Medevedev’s modernization Jonson and White 2012)

[29] Miriam Elder, “Emails Give Insight into Kremlin Youth Group’s Priorities, Means and Concerns,” *The Guardian*, February 7, 2012, https://www.theguardian.com/world/2012/feb/07/nashi-emails-insight-kremlin-groups-priorities.

government did occasionally use DDoS attacks, particularly in relation to the 2008 Russo-Georgian war, the Russian internet remained remarkably free (in a growing contrast with the traditional media) and the government activities there were primarily targeted to mobilize genuine support based on the compelling message and (limited) interaction with the public.

This engagement came to the abrupt end in the wake of the 2011 – 2012 Russian popular protests, which coincided with Putin's return to Kremlin[30]. Putin openly expressed disdain to the "ungrateful" protesters, who benefited from the economic growth he presided over, but turned away from him. This logic guided his conclusions regarding the strategy of engagement with active, but not necessarily loyal segments of the Russian public online. As protesters at Bolotnaya square were to a large extent the target audience of Medvedev's efforts, these people were deemed the lost cause and the strategy of engaging them was declared a failure. However, given the role of media, and social media in particular, in coordinating and sustaining the largest and the longest wave of protests in Russia in two decades, it was also impossible for Putin to go back to "disengagement strategy" he used during his first two terms in office.

Instead, Putin began to actively employ both offline means of controlling media production and online means of controlling access to it. The first included pressuring media moguls into either replacing the editorial stuff of online media they own (*Lenta.ru*, *Gazeta.ru* and *RBK* are the most prominent among dozens of examples) or into selling them to more loyal owners (Russian Forbes and *Vkontakte*). The government also adopted laws making online media liable for the content of comments posted by their readers, thus requiring these websites either to actively police user-generated content, or shut commenting tools down altogether. In addition, various laws were adopted to prosecute individual bloggers for alleged extremism and other content deemed inappropriate. Since 2012 these laws are applied increasingly promiscuously, punishing with large fines and real prison terms not only original authors

[30] According to most observers, protest, triggered by the alleged major irregularities in vote count during Duma elections, did not just coincide with Putin's return to Kremlin, but were largely caused by the perceived undemocratic nature of the deal between Putin and Medvedev, which was announced just a few weeks before elections at the ruling party convention, and was kept secret until last minute even from the convention delegates (Sakwa 2014, 111-134).

of the messages, but also those who reposted them. Finally, many prominent bloggers and journalists of online media faced threats and assaults, including life-threatening, which are never investigated.

Online tools of controlling access to content include creation of the Russian Internet Blacklist, maintained by the dedicated government agency, *Roskomnadzor*. Blacklisting initially required a court order, but later was also allowed on a simple request from the Office of the Prosecutor General. While theoretically it is supposed to be easy to exit the blacklist (after removing the content deemed unlawful), after several prominent opposition news websites and opposition leader's blogs were blocked in March 2014, in the midst of the Russian-Ukrainian conflict, they were not informed what content they have to remove to exit the Blacklist[31]. Government refused to respond to their requests even after they sued for an answer, and they remain blacklisted to this day[32].

Still, neither offline, nor online tools allowed the government to shut down hostile activity online completely. While a long period of unrestricted development of domestic alternatives diminished the market share of Facebook and Google in Russia (which makes government's job easier, as local platforms are easier to coerce into compliance), Facebook and Google are still used by millions of Russians on a daily basis. And when Vkontakte, immediately after getting a request, removed the event page of pro-opposition rally, Facebook (after some uncertain moves) refused to comply[33]. Journalists fired by the pressured owners could move abroad and set up a news media there (as *Meduza.io* did). Therefore, the space for engagement strategy remains, but instead of playing the leading role, it supports offline and online restrictions. Rather than trying to engage in dialog or persuade, the government simply attempts to hummer down the official message, artificially increase the indicators of its take-up (propel

[31] Human Rights Watch, "Russia: Halt Orders to Block Online Media," March 23, 2014, https://www.hrw.org/news/2014/03/23/russia-halt-orders-block-online-media.

[32] Global Freedom of Expression, "Grani.ru v. Office of Prosecutor General," *Columbia University*, September 2, 2014, https://globalfreedomofexpression.columbia.edu/cases/grani-ru-vs-office-of-prosecutor-general/.

[33] Sergei Guriev, "Facebook Faces Down Putin," *Project Syndicate*, January 9, 2015, http://www.project-syndicate.org/commentary/facebook-versus-putin-by-sergei-guriev-2015-01.

the politicians into the lists of top bloggers and their messages into the list of top posts), while simultaneously cluttering the communication channels used by the opposition. This created a huge market (at which Russian government spares no resources) for various troll and bot factories, which produce pro-government content in volumes, caring about the quantity much more than quality and persuasion capacity. This content requires a new set of tools to study it properly. In the next section we describe in detail how we are beginning to build these tools.

5 Online engagement: preliminary analysis

5.1 Introduction

Online engagement is a complex phenomenon, ranging from completely automated bots producing large volumes of gibberish in order to flood popular communication platforms to high-profile paid bloggers with independent reputations, who send nuanced, targeted messages to different groups of the public. While of course it would be useful to study all forms of this activity, for the sake of space constraints in this manuscript we limit ourselves to studying bots.[34]

There are three main reasons for this choice. First, bots produce by far the largest volume of content, and without the tools to remove it, studying the human-generated content would be almost impossible. Second, the only practical way to identify bots is by using automated algorithms; starting the empirical part of our research in this manner has the advantage of therefore creating an objective and replicable approach that can be employed in future analysis. There is, however, another and less methodologically inspired reason to start with bots, which is that they are both important and interesting objects to study. While social media provide citizens and politicians with new and powerful tools for expressing their political beliefs and preferences, affecting the political agenda, mobilizing supporters, and organizing political actions, they also bring the challenge of differentiating between real political com-

[34] We do, however, intend to greatly expand this focus in future research.

munication on the one hand, and interaction with computer programs that imitate human activity on the other.

It is important to note from the outset that we seek to identify bots in the most strict – and technical – sense of the word. **Bots** are accounts that are, of course, set up and maintained by humans (usually in bulk), but they set them up to be filled with content by computer programs which *automatically* lift this content from pre-defined set of sources, such as news feeds, other twitter accounts or search queries. Hence, a troll who is hired to post her own tweets about a certain topic, would be a **human** rather than a bot in our classification. Similarly, a real human being who does not post her own tweets, but only retweets accounts she follows, would also be classified as human rather than a bot. If we believe that an account alternates between periods of human and automated control, we put it in the intermediate category: **cyborgs**. We discuss this classification in more detail in Section 5.4.

Bot detection is a relatively new topic in political science emerging from the burgeoning research in political communication on social media platforms. The scant literature that exists on the subject mostly borrows methods developed in computer science to detect email spam. Bot detection in social media is a different (but closely related) task that can be achieved with a wider range of techniques that make use of both textual and non-textual account information (Chu et al. 2012). Nevertheless, bot detection in social media is still regarded as a challenging task within the computer science community, making Boshmaf et al. (2011) claim that 80% of bots are undetectable. Here, we combine some of the existing automated bot-detection techniques with domain specific knowledge and human coding to radically improve bot-detection capabilities.

5.2 Data

We used the Twitter's Streaming API to collect a large dataset of tweets that contain specific keywords related to Russian politics. We created a list of politically relevant keywords and hashtags (including major politicians' names, events and slogans[35]) that represent the entire political spectrum, including Putin and United Russia, loyal and radical opposition, Russian nationalists and others. This allowed us to collect a dataset with more than 14 million tweets posted by approximately 1.3 million Twitter users who list Russian as their account language, between November 25, 2013 and July 13, 2015[36].

The Twitter API returns both tweet-level information (i.e. text, date and time of the tweet, its language, etc.) and metadata (various characteristics of the account sending the tweet including the author's ID and screen name, the number of followers and friends and official account language). There is a large variation in the number of tweets from different users in our collection, ranging from 1 to almost **97,000**. Around 37% of all tweets in our collection are retweets. Among the rest 63% we were able to identify about 6.7 million of repeating tweets. Most of them are short and repeat just once, which might be a pure coincidence (or language artifact), whereas others repeat numerous times with a maximum of **24,558** times. Such extreme cases are arguably a good indication of a particular type of bots presence in our collection. Bots repeating strings of text, however, are by no means the only kind of bots, and we introduce multiple detection methods to capture different types of bots.

5.3 Detection methods

Our approach to detecting bots and cyborgs focused on three account characteristics: the *entropy* of inter-tweeting time intervals, the *followers/friends ratio* of accounts (which produced two different methods of finding bots), and the presence of *identical tweets* in our collection (this has also produced two methods). Below we justify and describe each method in detail; we verify whether the accounts they recover are indeed bots in Section 5.4.

[35] Politicians' names include Putin, Medvedev, Navalny, Khodorkovsky, Udaltsov, etc. The events feature Sochi Olympics, opposition rallies at Bolotnaya Square in Moscow, "Direct Line with Vladimir Putin", etc. We also searched for slogans like "Party of thieves and crooks", "Sobyanin is our mayor", "Stop feeding the Caucasus", etc. The full list of keywords and hashtags is provided in the Appendix A.

[36] We have data on 483 days between those dates, with the exception of two short periods: October 1 – 30, 2014 and November 5, 2014 – January 30, 2015. We stopped collecting tweets between those days for technological reasons. Our collection includes additional 14 million tweets from 3.3 million users who do not list Russian as their account language. Nevertheless, most of their tweets are in Russian and we will incorporate this data in the subsequent research.

Entropy of inter-tweeting time intervals. Our first technique is predicated on the idea that bots show a much higher regularity in their activity on Twitter than human beings. This is also true for cyborgs, at least to the extent they rely on automated sourcing of content. In the most simple case, for instance, bots may be programmed to send tweets every k seconds. On the contrary, humans' tweeting activity is much more sporadic. These differences in predictability may be captured by entropy which is a measure of uncertainty popular in computer science and information theory. In order to compute entropy, we created a list of all accounts that have at least three tweets in our collection.[37] Then, for each of these accounts, we computed the length of time intervals between consecutive tweets, and used those time intervals to compute an average entropy as follows:

$$Av.\ Entropy_i = \frac{1}{T_i} \sum_{t=1}^{T_i} p_t^{(i)} \times log_2(p_t^{(i)}),$$

where p_t is the probability of interval t for account i; T_i is the total number of time intervals for account i. The higher the value, the more unpredictable an account is. We expect that accounts with low entropy are either bots or cyborgs.

Although entropy proved to be the most informative bot-detection technique in Chu et al. (2012), we did not restrict ourselves to this measure due to a special type of data we analyze. Instead of having all the tweets from a given set of accounts, we have all the tweets mentioning predefined keywords. Thus, we are likely to miss most non-political tweets from most of the accounts in our collection. For this reason, the entropy measure might be noisy in our case.[38]

Followers/friends ratio. Our second method of detection, however, does not depend on the twitting activity, and thus can not be affected by whether or not we have all of user's tweets in the dataset. Instead, we rely on the idea that bots should have fewer followers than normal human users. Indeed, most humans would probably refuse to follow a bot that does

[37] Since we are computing the entropy of *inter*-tweeting time intervals, we need at least two intervals to compute a meaningful entropy value, or, in other words, three tweets.

[38] Still, if we are looking for bots designed to produce *political content*, this might not be too much of a problem: for bots only tweeting about politics, we get most of their tweets, depending on the extent they use our chosen keywords and hashtags.

not show signs of a normal human online activity. At the same time, bots would tend to follow lots of users (in Twitter parlance, have many friends) in the hope that some of them will accidentally follow them back. Thus, we expect that some bots will tend to have a very small followers/friends ratio defined simply as:

$$ratio_i = \frac{|\{followers_i\}|}{|\{friends_i\}|},$$

where $|\{followers_i\}|$ denotes the number of accounts that follow account i, and $|\{friends_i\}|$ stands for the number of accounts that are followed by account i.

Another interesting case includes accounts that have no friends (i.e. do not follow anybody). Obviously, the followers/friends ratio is undefined for these accounts due to division by zero, and we code them separately.

Identical tweets. Our two final bot-detection techniques involve identifying accounts sending identical tweets. There are two subtypes of identical tweets an account could send out: *intra-* and *inter-account* identical tweets. The first subtype refers to the case when an account is repeatedly sending the same tweet. We doubt that a human being would engage in such an activity, whereas a pre-programmed primitive bot could easily do that. The second subtype refers to the situation when a group of accounts are sending out identical tweets. This strategy can be employed by bots to maximize the spread of specific information over the network.

5.4 Empirical assessment of bot detection methods

We use all the methods outlined above to identify bot accounts for the subsequent verification using human coding. Doing so required setting a set of thresholds that determine how many accounts each method recovers. Similar to most cases of threshold selection, this is not a straightforward exercise since no theory has so far been developed to justify the choice. Given the lack of theory-driven guidance, we followed an empirical approach with two main considerations in mind. On the one hand, thresholds should be sufficiently loose to produce enough accounts for a meaningful verification. On the other hand, thresholds should be sufficiently stringent to keep hand-coding of the recovered accounts feasible. Thus, these thresholds should be considered simply a first cut in developing bot-detection methods: we check whether we can reliably identify bots if we set the thresholds at their most extreme levels keeping the sample size for coding reasonably large but feasible for hand-coding (roughly 100 accounts per method). In future work we intend to use sophisticated machine learning tools to empirically identify both the levels that distinguish bots from humans and those account characteristics that are most informative for bot detection.

For the *entropy* measure, we restricted our attention to those accounts that produced at least 300 tweets throughout the period under study. As there were more than $5,500$ such accounts, we selected 100 accounts with the lowest entropy values.

A similar threshold (300+ tweets in our collection) was applied to $36,500$ accounts with no friends, yielding 99 accounts.

To set the threshold for the *followers/friends ratio*, in the right panel of Figure 1 we plotted its distribution for all accounts in our dataset that have the ratio below 1 (i.e. follow more people than they are followed by). There are more than $900,000$ accounts like that, or about 70%. Among them, a disproportionally large number of accounts have the ratio around 0 (when nobody follows them) and 1 (when they are followed by the same number of accounts that they follow themselves). We expect the bots to have a very low ratio and have chosen 0.01 as our threshold, thus selecting accounts that follow (at least) 100 times more accounts than they are followed by. Together with a 50+ tweets in the collection activity requirement, it yields 135 accounts for verification.

Lastly, $43,000$ accounts in our collection tweeted the same text several times and another $600,000$ tweeted text that some other accounts in our collection tweeted (distributions shown in Figure 2). We expect those who did it in large quantities to be bots. For the former category we have set the threshold at 50 repetitions, yielding 90 accounts to verify. For the

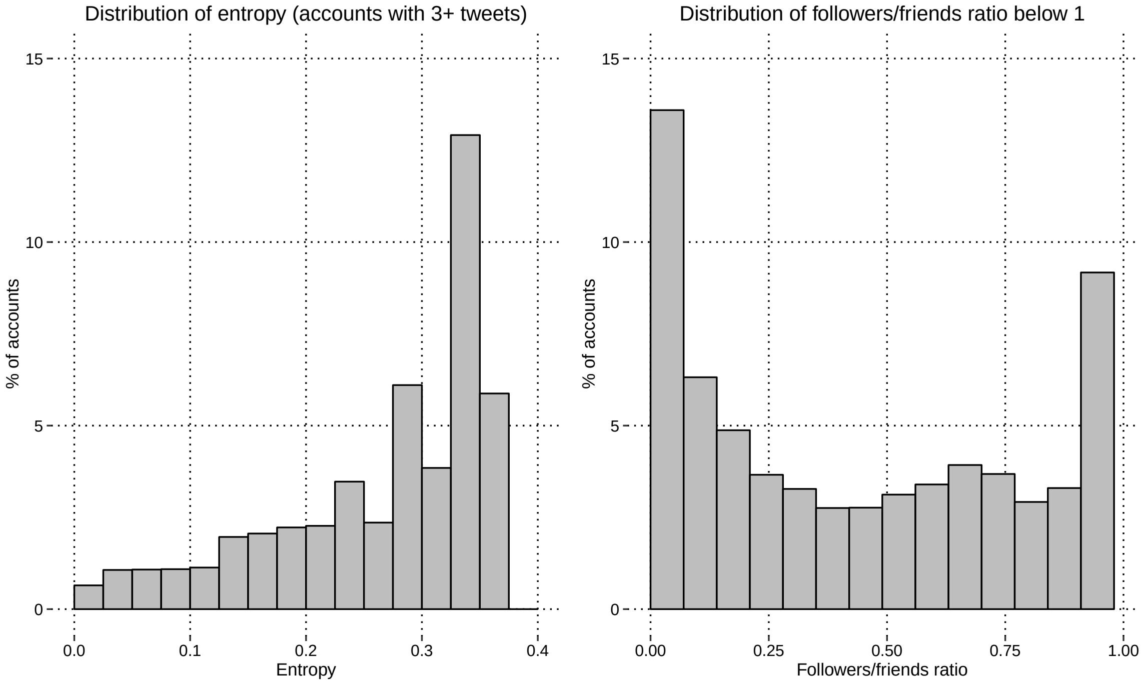

Figure 1: Distribution of Twitter accounts in the data set by the entropy of inter-tweeting time intervals (left panel) and followers/friends ratio (right panel)

latter category we set the threshold at $1,000$ repetitions, which leaves us with 102 accounts to verify.

Thus, we end up with five sets of "suspicious" Twitter accounts (a total of 526) that we identified using different bot-detection techniques. Cross-table 1 shows intersections of these sets. Net of 14 duplicates, we proceed with the analysis of 512 accounts. The low number of accounts these sets have in common suggests that we managed to identify different kinds of bots that may be used for different purposes, or at the very least created using different techniques.

In order to assess the reliability of our bot-detection algorithms, we enlisted 20 coders (native Russians, undergraduate students in Political Science, and familiar with Twitter) and tasked them with classifying 512 accounts into five categories: in addition to humans, bots, and cyborgs, described above, we have two miscellaneous ones, spam and official accounts. Spam includes accounts that feature no meaningful content, and consist mostly of gibberish

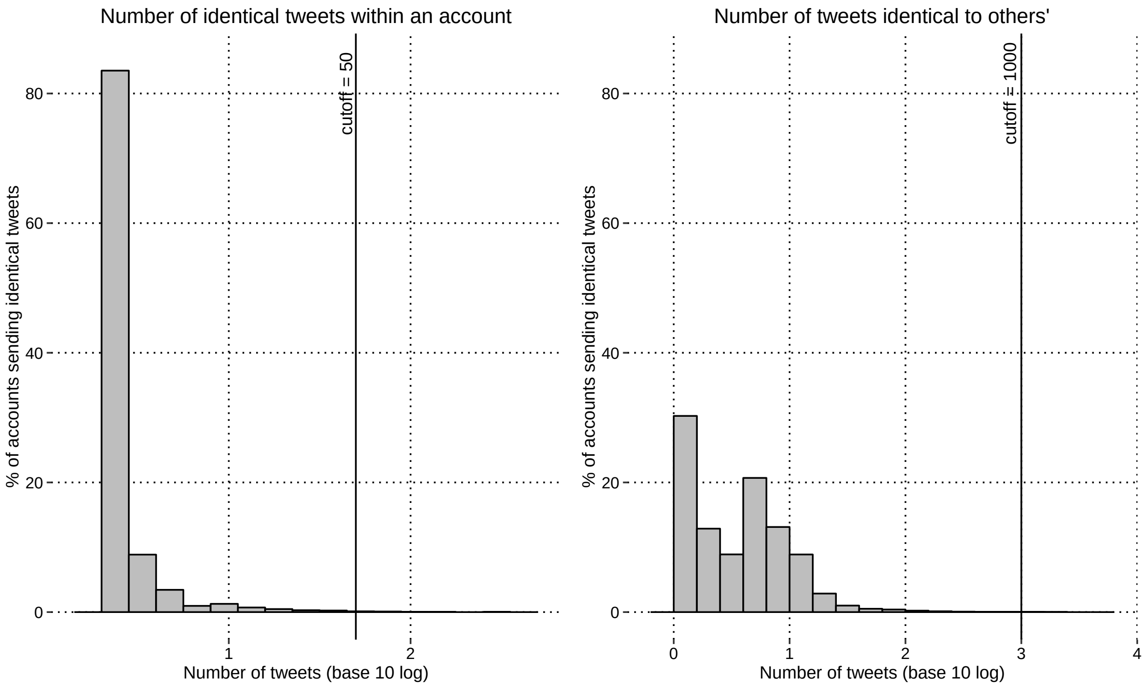

Figure 2: Distribution of accounts sending repeated tweets

Table 1: Intersections of sets with suspicious accounts

	No friends	Low ratio	Entropy	Repeat themselves	Repeat others
No friends	**99**	0^a	4	2	1
Low ratio		**135**	0	0	0
Entropy			**100**	0	5
Repeat itself				**90**	2
Repeat others					**102**

Note: Entries are numbers of Twitter accounts that are common to a pair of sets.
[a] "Low ratio" and "no friends" sets are mutually exclusive by definition, as ratio of followers to friends is undefined for accounts with no friends.

or consumer advertisement. Official accounts refer to the accounts run by organizations (such as media and government bodies) whose tweeting patterns are expected to be different from personal accounts. Some of these five categories were further split into subcategories. This was primarily done to reduce noise in human coding (more narrow categories are easier to define precisely for coders), and they were reaggregated at the analysis stage. With respect to bots, however, we present some interesting findings about their subtypes below. For details on coding schema, see Appendix C.

Coders have received detailed coding instructions in Russian (including, in schematic tree-form), and were provided with a list of 20 Twitter accounts (different from 512 accounts of interest) to code as an exercise. Next, each of them went through a 90-minutes Skype session with one of the co-authors, to ensure their clear understanding of the coding schema. We then randomly split coders into 4 groups of 5 people, and randomly assigned different Twitter accounts of interest to different groups. All coders were instructed to work independently and did not know the names of other coders in their groups. Thus, every account was classified independently by 5 coders.

It is important to note that coders did not work with live Twitter accounts. Instead, we used tweets and other data from our collection to re-create Twitter accounts as they looked like at the time our collection was running. Obviously, there is no chance to replicate accounts exactly (for example, we do not include non-political tweets because they did not contain any of our keywords or hashtags and therefore were ignored by our collection filters). However, in addition to featuring up to 100 tweets from the account[39], we made use of account metadata to reproduce the number of followers, friends, tweets, as well as account description, geo-location, user and background pictures, date of account creation, and other information typically available on a Twitter web-page.

In order to insure the reliability of coding, after our 5-coders groups finished coding all 512 accounts, we examined the inter-coder agreement in each group for each account and imposed a stringent requirement for an account to be considered reliably classified: it had to be put into a particular category (bot/cyborg/human/official account/spam) by at least 4 out of 5 coders. This is a rather stringent approach that guarantees the inter-coder reliability is at least as high as 80%. Stringency of this requirement notwithstanding, more than 80% of accounts pass this requirement.

Table 2 shows the results of the verification. Of the accounts classified reliably, 77% belong to bots, and further 1% are cyborgs who share more characteristics with bots than with

[39] If an account had more than 100 tweets in our collection, we used the most recent 100 tweets.

humans. Reliably identified humans constitute only 1.4% of our dataset. These results reveal an outstanding precision of the five bot-detection methods we have used.

Table 2: Results of suspicious accounts verification

	No friends	Low ratio	Entropy	Repeat themselves	Repeat others	**Totals**
Bots	82	77	83	72	93	**394**
Cyborgs	1	0	3	1	0	**5**
Humans	1	0	2	4	1	**7**
Official accounts	2	0	0	0	0	**2**
Spam	0	7	1	1	0	**9**
Unclear	13	51	11	12	8	**95**
Totals	**99**	**135**	**100**	**90**	**102**	**512**

Note: Entries are frequencies. Row sums do not equal row totals because some methods produced overlapping sets of accounts (see Table 1).

The share of bots is probably even higher than above calculations indicate. Only 25 of 95 unclear accounts have been claimed to belong to a human being by at least a single coder. Typically, the discrepancies in the coding that generate unclear accounts are due to attributing accounts to spam or cyborgs by some coders and to bots by others. Hence, even those accounts that belong to the category of unclear are much more likely to be bots than humans. Still, from here forward, we restrict our analysis to 394 accounts that were reliably classified as bots. We will begin our analysis with examining different types of bots we have uncovered. In the next subsection, we will describe our preliminary findings regarding bots' political orientations.

As already mentioned, our coders coded accounts by subtypes, which includes seven subtypes of bots. Examining the distribution of bots across subtypes (last column of Table 3), reveals that more than 90% belong to just three subtypes: news headlines with and without links and (the very distant third) accounts that consist entirely of retweets from other accounts. This would not be a surprise given how easy it is to setup accounts like this: one does not need to upload pictures of videos or combine information of different kind as accounts with diverse content do.

Several interesting patterns emerge from breaking types of bots by their source of detection (first five columns in Table 3). Even though differences in effort required to maintain accounts of different type are unlikely to be large, creators seems to be aware of them and take them into account when they program network behavior of accounts and their way to engage other twitter users. For example, 38% of accounts that do not follow any other accounts (first column) feature tweets with links to the news story; for accounts that do follow other accounts (second column) this share jumps to 74%. While the former simply aim to promote specific news in search engine rankings, or rely on hashtags to enter Twitter own popular searches, the latter hope that at least few people would follow them back and then click on the news link. Accounts repeating others are similar to accounts with no friends: 73% of them tweet news headlines with no link to the story. Accounts that repeat themselves are different, they more often feature diverse content (quotes from famous people, beauty and character advice, etc.). It is this content that tends to be repeated to attract followers, who are then exposed to fresh news stories, often with a link to the website. Low entropy method catches well a different kind of bots: those who retweet from other accounts. There could be two explanations for this. It's possible that tweeting in this case could not be triggered by the news headline appearing in the news agency feed, and bot owner has to specify the frequency of posting himself. Alternatively, accounts who retweet everything from many other accounts simply get to tweet more often than a typical news agency comes up with a new story. Further analysis of network behavior and content choices of different types of bots is warranted to understand the mechanisms and effects of their engagement with human users.

5.5 Political orientations in a subsample of verified political bots

In addition to coding each account as bot, human, cyborg, spam or official account, for bots in particular we asked our coders to determine their political orientation. Given a rather small sample size, this analysis remains preliminary, but interesting patterns we uncover suggest that our bot-detection methods are promising as a tool for quantitative studies of propaganda.

Table 3: Verified bots by type and method of identification

	No friends	Low ratio	Entropy	Repeat themselves	Repeat others	**Total**
Retweets only	3	8	26	3	3	9
Videos only	2	0	0	0	0	< 1
Pictures only	2	0	0	0	1	1
Text only:						
– News headlines only:						
News headlines with links	38	74	40	41	15	40
News headlines without links	48	15	26	36	73	42
– Other text	3	2	2	5	1	3
Diverse content	3	2	7	15	7	6

Note: Entries are column percentages (may not sum up to 100 due to rounding).

We distinguish between three different political orientations: *pro-Kremlin*, *pro-opposition* and *pro-Ukrainian*. This choice was dictated both by the our research framework and by the activity patterns in Russian segment of Internet during the time our collection was running. As anecdotal evidence of Russian governments activity in social media continues to mount, our principle interest was in studying accounts that post content friendly to the Russian government. However, Russian governments, as well as governments around the world, often claim that they are victims, nor perpetrators of bot attacks, and blame opposition leaders for running anti-government botnets. Explicitly looking for both pro- and anti-government content in the same dataset provides a valuable opportunity to verify both claims on the level playing field. However, since the timing of our data collection coincided with the political crisis in Ukraine and further Russian involvement there, we have to distinguish between anti-government content that mostly has to do with Russian domestic politics, and content spread by (again, according to anecdotal evidence) Ukrainian bots. Thus, the accounts that our coders find to spread content unfriendly to the Russian government is split into two categories: pro-opposition and pro-Ukrainian.

Given that our dataset – and hence artificial account profiles coders worked with – by construction includes only political tweets, one could try to discern political orientation of the accounts by closely reading news headlines or pictures they publish, accounts they retweet, etc. This was not our goal, however. We were interested only in accounts whose political identification is strikingly clear and unambiguously manifest in most of their tweets to even a casual reader. To this end, we deliberately defined the *neutral* category as broadly as possible, so that those accounts that end up classified as non-neutral have indeed a very strong political bias. Furthermore, in the coding schema provided to coders accounts that do not neatly fit either of our "camps" (for example, praise Kremlin for its foreign ventures, but loathe its for economic policy) were also left in the neutral category.

Still given the relatively difficult nature of the task, we had to adopt stringent inter-coder reliability requirements to ensure high-quality of classification. To this end we imposed the following rule: if even just 2 of 5 coders put the account into two different partisan categories (for instance, pro-Ukrainian and pro-opposition), we considered this account as coded unreliably ("Unclear" category in Table 4). If there was no disagreement between coders regarding partisanship of the account, and coders only differed if it belongs to a particular group or is neutral, we opted for neutral if three or more coders categorized it as such; only otherwise we have put it in the partisan category. Table 4 shows the resulting classification, broken up by the method used to uncover bots.

Table 4: Ideological distribution of verified bots

	No friends	Low ratio	Entropy	Repeat themselves	Repeat others	**Totals**
Pro-Kremlin	13	16	8	8	8	**11**
Pro-opposition	1	1	11	1	2	**4**
Pro-Ukrainian	4	3	12	3	3	**5**
Neutral	65	55	41	51	76	**58**
Unclear	17	26	28	36	11	**23**

Note: Entries are column percentages (may not sum up to 100 due to rounding).

Table 4 highlights interesting findings. First, more than a half of bots are neutral, meaning that they do not carry any explicit political message. This does not mean they are not setup with political purposes. As we mentioned above, many bots (particularly among those who have no friends and post the same text with many other bots) feature primarily news headlines to promote them in search rankings. These headlines usually come from large media agencies that produce enough routine factual news (who said what, went where and signed which memorandum) to appear neutral according to our deliberately broad notion of neutrality. This applies even to the state-owned media. It does not mean, however, that if their stories make it to the top, and readers will go to their website, they will find a politically neutral and objective media.

Second, taken together pro-opposition and pro-Ukrainian bots (9%) are almost as common as pro-Kremlin ones (11%). This result may seem unexpected given the mass media's clamor about Kremlin's social media propaganda campaigns. At the same time, this result might also imply that Kremlin prefers more sophisticated and expensive online propaganda techniques like paid trolls, whereas the Russian opposition and pro-Ukrainian users may so far lack resources to employ those techniques at a large scale[40].

How different are bots with different ideological orientation? We explore this issue from two different perspectives by looking at their content and the similarity of their dynamics of tweeting activity. Figure 3 presents bar plots illustrating the popularity of 15 most common hashtags within every orientation group, including neutral bots. One can see from the graph that all types of bots discussed political developments in Ukraine and include a variety of hashtags about Euromaidan protests, and further tragic events in Odessa and Eastern Ukraine. Despite these similarities, there are also important differences across orientation groups both in the popularity of different hashtags and their contents. For instance, the Organization for Security and Co-operation in Europe (OSCE) that plays a peace-keeping role in Eastern Ukraine features twice among pro-Ukrainian bots, whereas the U.S. and NATO appear among common hashtags for pro-Kremlin bots. Hashtag in support of a Ukrainian officer Nadezhda Savchekno, who was captured and put on trial in Russia (allegedly in relation to the death of Russian journalists who covered the conflict in the Eastern Ukraine), is among the most popular among pro-opposition, but not pro-Ukrainian bots.

[40] We also cannot rule out the possibility that at least some of the anti-Kremlin bots, particularly pro-Ukrainian bots with the most vicious content, are created as a provocation against the Ukrainian cause. Methods presented here are not suited to test this hypothesis, but the analysis of network structure that we plan to undertake could shed light on this matter.

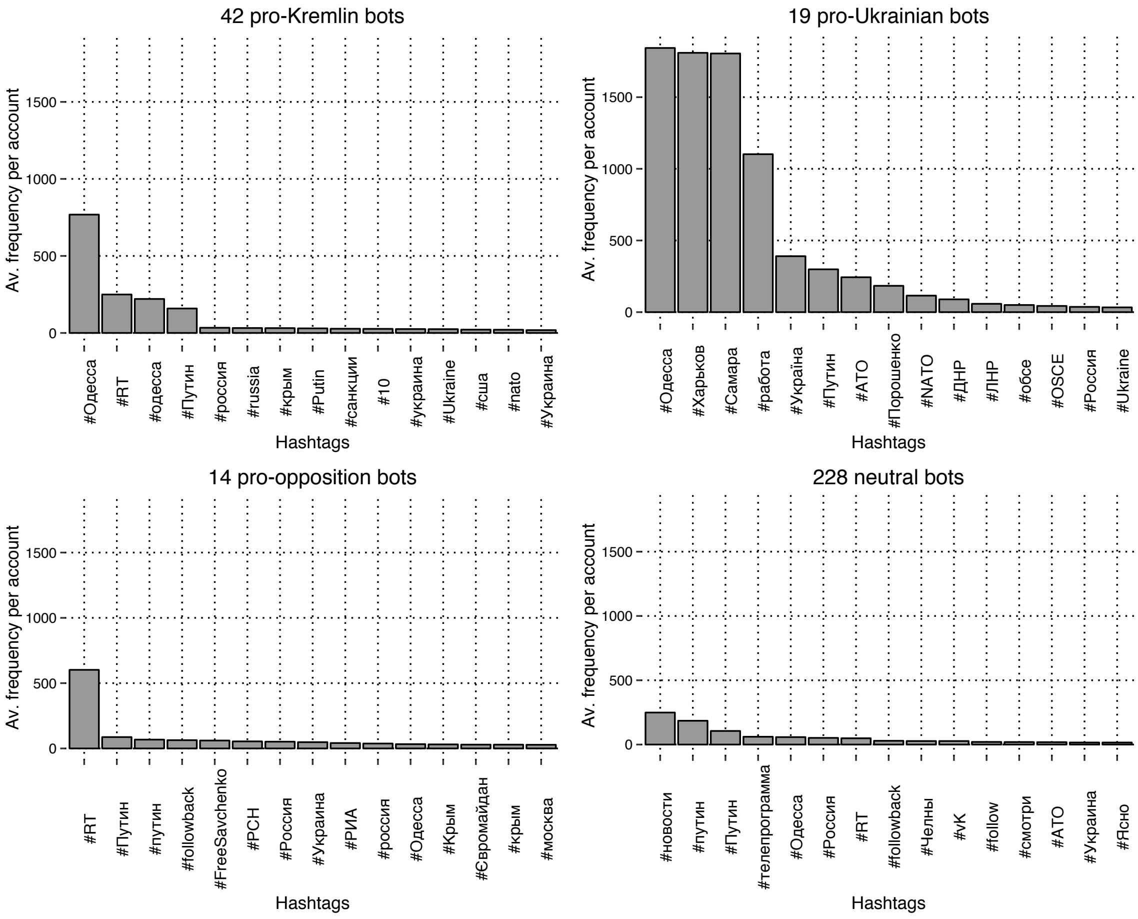

Figure 3: Most Common Hashtags

Overall, pro-opposition and pro-Ukrainian bots are closest in terms of their used hashtags, whereas neutral bots are furthest apart from the rest, as Jaccard similarities in Table 5 show. Jaccard similarity is a popular measure of similarity between two sets that shows how many elements the two sets have in common in comparison to the total number of elements in the sets[41]. Table 5 presents Jaccard similarities for sets of 50 most common hashtags in different groups of bots by political orientation.

Another perspective at the similarity of bots with different political orientation is to look at when and how much they tweet. Figure 4 presents their total political activity on Twitter

[41] Technically, the Jaccard similarity coefficient (J(A,B)) between sets A and B is $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$, where $|A \cap B|$ is the number of elements that A and B have in common, and $|A \cup B|$ is the total number of elements in the two sets.

Table 5: Jaccard similarities of 50 most popular hashtags

	pro-Kremlin	pro-opposition	pro-Ukrainian	Neutral
pro-Kremlin	1.00	0.27	0.30	0.25
pro-opposition		1.00	0.39	0.19
pro-Ukrainian			1.00	0.23
Neutral				1.00

Note: Entries are Jaccard similarities between sets of 50 most popular hashtags in every group of bots. Jaccard similarities range from 0 to 1, where 0 refers to a pair of sets that have no common element, whereas 1 refers to a pair of sets that are identical.

per day over time. One can see that pro-opposition activity was much higher than pro-Kremlin one in the very beginning of 2014, when Russian involvement in Ukraine was already apparent for anybody interested (like Russian opposition), but not yet advertised by the Russian government itself (Kremlin even actively tried to conceal it at that stage). Pro-Ukrainian activity was virtually non-existent: Ukraine was no more ready for the cyber-war than for on the ground one. With the annexation of Crimea, the conflict escalation and Russian involvement becoming more assertive, pro-Kremlin bot activity quickly caught up with the pro-opposition by the late Spring of 2014. By summer pro-Ukrainian bot activity was in full swing too.

All three remained on roughly the same levels through the end of 2014, but then began to diverge again. Pro-opposition activity was gradually dwindling, this trend however was interrupted when late at night on February 27th, 2015 one of the most prominent leaders of the Russian opposition, Boris Nemtsov, was shot dead in front of the Kremlin. His tragic death generated one of the largest waves of identical tweets in our collection that involved more than $1,800$ users we have data on. Meanwhile, pro-Kremlin bots' activity remained relatively constant and did not match the marked increase in activity by pro-Ukrainian bots in the beginning of 2015, when the war in Eastern Ukraine escalated again. This disparity in cyber-war effort appears to mirror the changes on the ground, as Kremlin was no longer interested in the conflict and instead of taking advantage of escalation to stage an offensive,

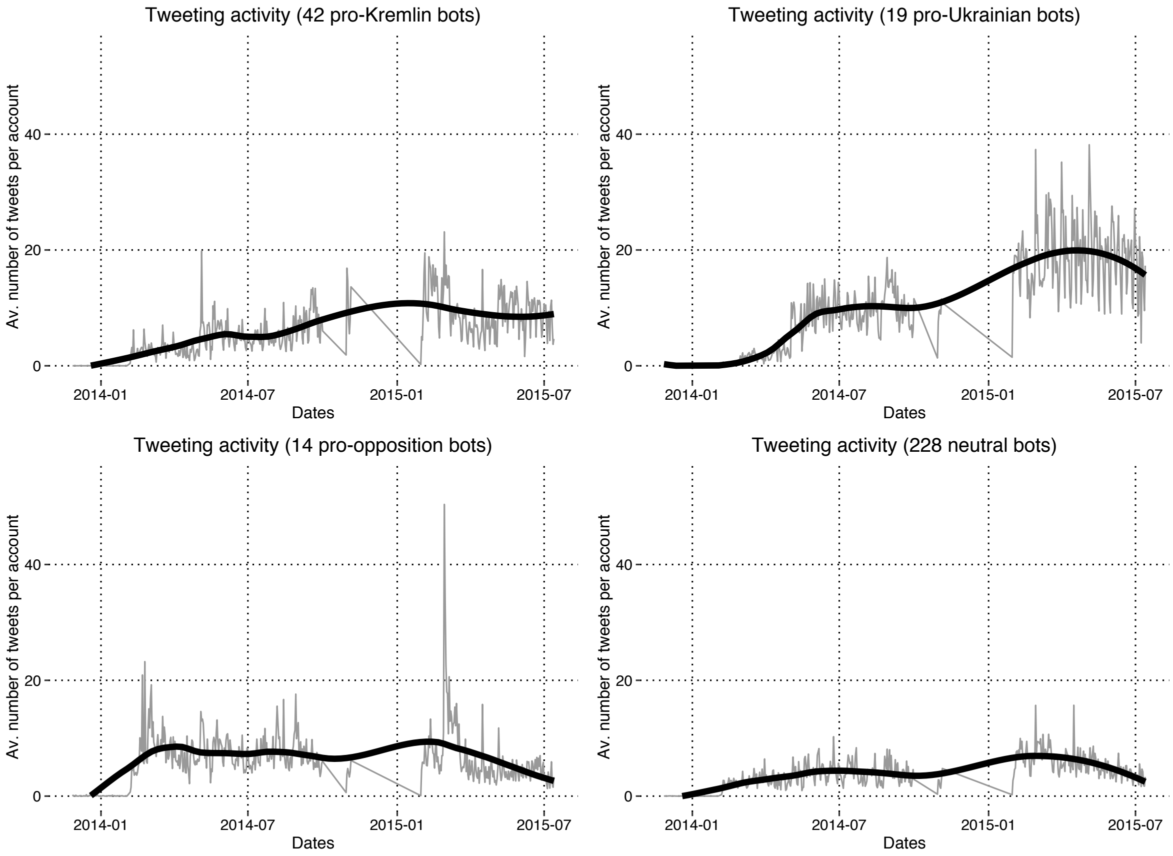

Figure 4: Political Bots' Tweeting Activity over Time

put just enough resources to coerce Ukrainians into the Minsk agreement, which essentially preserved the status-quo and frozen the conflict on the Transnistria model.

6 Machine learning tools for bot detection

6.1 Methodology

The results presented in Section 5 demonstrate high precision of our bot-detection techniques, but, obviously, cannot estimate their recall. Neither can they prove that metrics we selected for detection and thresholds we set for them were optimal. Last but not least, evaluation of our techniques using just hand-coding entails a serious limitation. While it allows to answer questions about a (rather small) sample of accounts (which gives us some interesting insights about different types of bots employed for political purposes), expanding to the entire dataset to study the tangible impact of bot activity on the Twittersphere as a whole remains unfeasible. In this section, we begin to address these limitations by employing powerful tools of supervised machine learning. Using the insights from Section 5 about the behavior of bots, we construct two ensemble classifiers: based on textual and non-textual features of Twitter accounts. After training them on the labeled set that includes cases coded for Section 5, we apply both classifiers to a large set of unlabeled Twitter accounts.

Our first classifier makes use of textual data for classification purposes, identifying bots by comparing word frequencies in tweets from the test set to word frequencies in tweets from the training set. The second classifier relies on various account metrics and metadata that we have collected for each tweet. In the future, we plan to follow the growing literature on multi-view learning (Xu, Tao, and Xu 2013) and treat these two types of features as different views of the same data, thus building a multi-view learning classifier. For now, however, we present separate results for each.

To build our **text-based** classifier we have extracted all unigrams (i.e. words) and bigrams (word pairs) that appear in tweets from at least 20 users in our training set (this gave us a total of $83,000$ features) and stored information on how many times each of them appeared in tweets from every training user.

Our **non-textual** classifier incorporates many of the account characteristics that we suspect are indicative of bots or humans. They include those discussed in detail in Section 5, but also many others:

- number of tweets in the collection: bots are obviously able to produce many more tweets than humans;
- percentage of tweets that feature a hyperlink (URL): bots that are programmed to refer to specific websites would tend to have a higher share of tweets with hyperlinks;
- average and maximum number of hyperlinks per tweet: some types of bots might tend

to include lots of hyperlinks in tweets;

- retweets as a share of all tweets from the account: some bots are purposefully programmed to do retweets;
- average number of retweeted accounts per retweet: if a bot is programmed to retweet a particular account or set of accounts, it will score lower on this indicator than a human account that retweets tweets from different accounts;
- an indicator variable for using the default Twitter profile image: bots might be less inclined to replace the automatically provided eggs (of various colors) with an original picture;
- an indicator variable for a change in using the default Twitter profile image: since all the data is stored at tweet-level, we can trace changes in some account characteristics over time; here we record information about whether we see any change in the previous indicator over time;
- indicators for using the default Twitter profile theme and background, as well as for change in this characteristic over time;
- an indicator for an empty user description field: humans are expected to fill that field more often than bots;
- the length of the description in the user description field: humans are expected to be more elaborate in filling out that field;
- an indicator for a change in the user description field: we record only changes in the length of that description, expecting that humans might be more prone to change that field than bots;
- an indicator for having digits other than years of birth in the screen name: we noticed that bots tend to have digits in their names, while this is not the case for humans, except for the digits that might correspond to the year of birth;

- an indicator for filling in location in the user profile: we expect humans to do it more often than bots;

- an indicator for having more than one word in the name field: humans tend to have both the first and the last name, but we expect this to be less common for bots;

- the number of tweets a given user favored;

- an indicator for enabling geotagging and changes in that parameter over time;

- share of directed tweets (tweets mentioning other users using @ symbol): bots that are programmed to respond to tweets of a particular type (for example, if they contain certain keywords) or from a particular user would probably tend to have a larger share of directed tweets;

- percentage of tweets with a hashtag: bots might be pre-programmed to either use hasthags in every tweet, or never use them; intermediate situations are also possible and would indicate more sophisticated bots;

- average number of hashtags per tweet: although we do not have any specific prior expectations about this feature, one can imagine bots and humans having different strategies of using hashtags in their tweets;

- average followers-to-friends ratio: since our collection stores tweet-level (rather than account-level) data, we know an account's number of followers and friends at every point in time when a tweet was stored in our collection. Thus, if an account produced more than one tweet in our collection, it can be characterized by an array of followers-to-friends ratios, instead of a single number. To compress this array to a scalar, we take averages over time;

- dummy for accounts with undefined followers-to-friends ratio (i.e. having no friends): as we have mentioned above, some bot accounts have no friends, so the followers-to-friends ration is undefined;

- entropy of platform usage: users can write tweets using different platforms (Twitter website, Twitter apps for Android, iPhone, etc.; third-party applications), and we expect that humans will be much less predictable in their choice of platforms than bots;

- number of times a given platform was used: there are $18,559$ platforms users in our dataset used to write tweets; we chose 15 most popular ones; these 15 platforms account for more than 50% of tweets written by bots and humans in the training set.

To train these classifiers we use data we have already coded at the previous stage. These accounts, however, are by design almost all bots, so to give our algorithms enough information about human behavior on Twitter, we augment the training data set with additional accounts that we draw from our collection at random and got coded following the same coding schema. Trained classifiers then assign unlabeled cases in our test set to two classes: bots or humans[42].

To validate our results we take a random sample of machine-classified accounts and get them coded by humans. Validation not only estimates the quality (both precision and recall) of our classifiers, but also allows us to expand the training set further in case validation results are unsatisfactory. Thus, this procedure could be iterative and allows for continuous improvement of the machine classification accuracy. Currently, the training set consists of 1041 accounts including 858 bots and 183 humans[43] [44].

In order to allow our classifiers to have as much flexibility as possible, we use non-parametric ensemble tree classifiers. On the one hand, this popular family of methods in machine learning goes beyond linearity built in many other techniques (including logistic regression or discriminant analysis). On the other, it builds upon lower-level classifiers and helps improve their

[42] For simplicity, we dropped all other categories – such as cyborgs and spam – and focused on the most technically straightforward task of binary, rather than multi-class, classification. Obviously, we removed all accounts that are neither bots nor humans from the training set.

[43] Relatively small share of human accounts in the training dataset reflects that even among the randomly chosen accounts bots presence is overwhelming.

[44] These accounts were coded following our schema outlined in Appendix C, but for speed it was done by only one coder (one of the coauthors of this paper). In the future versions, these accounts would be recoded by multiple coders following exactly the same reliability standards as described in Section 5.4.

individual performance (Hastie, Tibshirani, and Friedman 2009). In particular, we use xgboost package in R (Chen, He, and Benesty 2015) to run gradient boosted classification trees. This gradient boosting machine implements greedy function approximation developed and popularized by Friedman (2001).

Following Chen (2014), we define the minimized objective function as follows:

$$Obj^{(t)} = \sum_{i=1}^{n} L\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \sum_{k=1}^{t} \Omega(f_k),$$

where n is the number of observations in the training set; t is the current level of the ensemble, ranging from 1 to T, the maximum number of classifiers in the ensemble; x_i is a vector of predicting features for observation i; y_i is the indicator variable for being a bot for observation i; $\hat{y}_i^{(t-1)}$ is the predicted label for the ith observation on the previous level of the ensemble; $f_t(x_i)$ is the regression tree applied at the current iteration t to the feature vector x_i; $L(y_i, \hat{y}^{(t)})$ is the logistic loss function; and $\Omega(f_k)$ is the penalty function used to prevent kth regression tree from overfitting. More specifically,

$$\left(y_i, \hat{y}_i^{(t)}\right) = \sum_{i=1}^{n} y_i ln(1 + e^{-\hat{y}_i}) + (1 - y_i) ln(1 + e^{\hat{y}_i}),$$

$$\Omega(f_k) = \gamma k + \frac{1}{2}\lambda \sum_{j=1}^{k} \omega_j^2,$$

where k is indexes regression trees and thus refers to the number of leaves; ω_j is the leaf mean value for the jth leaf; γ and λ are penalty parameters.

We choose all parameters using 10-fold cross-validation, i.e. we randomly split the training data into 10 chunks, train classifiers on 9 chunks and predict the label for the remaining observations. We select the set of parameters with the smallest mean prediction error. For the textual classifier it was $\alpha = 0.2$, $\lambda = 1$, $\eta = 0.4$, $depth = 4$ and number of rounds $= 3$. For the non-textual classifier the best performance was achieved with $\alpha = 1$, $\lambda = 1.2$, $\eta = 0.55$, *depth* = 4 and number of rounds = 8.

Finally, we apply the trained classifiers with the parameters we chose with cross-validation to the unlabeled test set of 141,654 accounts[45], thus producing a classification of accounts into bots and humans.

6.2 Results

The textual classifier recovered 90.2% of bots (128,000 accounts) and 9.8% of humans (14,000) in our test set. The strikingly large share of predicted bots partly explains our previous suspicion that bots have an extremely large presence in the Russian political Twitter. In order to characterize the performance of the textual classifier, we randomly sampled 100 account ids from the classified ones and verified them using human coding. Table 6 shows the results of the verification.

Table 6: Text Classification Accuracy

		True Label		
		Humans	Bots	Total
Prediction	Humans	38 (69%)	17 (31%)	55
	Bots	17 (33%)	35 (67%)	52
	Total	55	52	107

Note: Obs are accounts. 100 accounts sampled by id. Some ids correspond to different user names.

The overall quality of classification is moderately high, with $Pr(human|\widehat{human}) = 0.69$ and $Pr(bot|\widehat{bot}) = 0.67$. Although these values show that we are doing better than a random guess, there is still a big potential for improvement.

[45] The size of the test set was limited due to computational restrictions; in the future we will expand it to the entire dataset described in Section 5.2.

The non-textual classifier confirms the finding of a large presence of bots, but provides slightly more optimistic numbers in terms of human activity in Russian political Twitter. It classified 78.2% of accounts as bots ($111,000$ accounts) and 21.8% of accounts ($31,000$) as humans. Validation shows significantly higher performance characteristics for the non-textual classifier, as compared with the textual one. Table 7 summarizes the results for the same random sample of 100 account ids.

Table 7: Non-Textual Classification Accuracy

		True Label		
		Humans	Bots	Total
Prediction	Humans	41 (80%)	10 (20%)	51
	Bots	14 (25%)	42 (75%)	56
	Total	55	52	107

Note: Obs are accounts. 100 accounts sampled by id. Some ids correspond to different user names.

As one can see, both $Pr(human|\widehat{human}) = 0.80$ and $Pr(bot|\widehat{bot}) = 0.75$ are higher for the non-textual classification results than for the textual classification ones.

Interestingly, the two classifiers show considerable variability in their predictions, as shown in Table 8 with 18% of observations assigned to different classes.

These results show that although various account-level characteristics might be a better predictor of bots than the text alone, there is still substantial difference in text produced by bots and humans. They also indicate that there is potential for classification improvement if we achieve greater consistency among classifiers, which is a machine learning task we are currently working on using different multi-view learning approaches.

Table 8: Cross-Table for Predicted Classes

		Text	
		Humans	Bots
Non-Text	Humans	9,588	21,284
		(6.8%)	(15%)
	Bots	4,301	106,481
		(3%)	(75.2%)

Note: Obs are accounts. N = 141,654.

References

Ackland, Robert. 2013. *Web Social Science: Concepts, Data and Tools for Social Scientists in the Digital Age.* SAGE Publications. [8].

Agora. 2011. *Internet Freedom in Russia 2011.* Kazan, Russia: Association of Human Rights Organizations "Agora". http://openinform.ru/fs/j_photos/openinform_353.pdf. [69].

———. 2012. *Internet Freedom in Russia 2012.* Kazan, Russia: Association of Human Rights Organizations "Agora". http://www.hro.org/node/15685. [74].

———. 2013. *Internet Freedom in Russia 2013.* Kazan, Russia: Association of Human Rights Organizations "Agora". http://eliberator.ru/files/Internet_2013.pdf. [74].

Alexanyan, Karina, Vladimir Barash, Bruce Etling, Robert Faris, Urs Gasser, John Kelly, John G. Palfrey, and Hal Roberts. 2012. "Exploring Russian Cyberspace: Digitally-Mediated Collective Action and the Networked Public Sphere." SSRN Scholarly Paper. Accessed April 30, 2014. http://papers.ssrn.com/abstract=2014998. [62, 64, 75].

Banks, Arthur S., and Kenneth A. Wilson. 2013. *Cross-National Time-Series Data Archive.* Jerusalem, Israel.: Databanks International. [61].

Barash, Vladimir, and John Kelly. 2012. "Salience vs. Commitment: Dynamics of Political Hashtags in Russian Twitter." SSRN Scholarly Paper. Accessed April 30, 2014. http://papers.ssrn.com/abstract=2034506. [5, 68].

Becker, Jonathan. 2004. "Lessons from Russia A Neo-Authoritarian Media System." *European Journal of Communication* 19 (2): 139–163. [62].

Bershidsky, Leonid. 2014. "Google's Retreat From Moscow." BLOOMBERGVIEW.com. December 12. Accessed January 12, 2015. http://www.bloombergview.com/articles/2014-12-12/googles-retreat-from-moscow. [76].

Black, J. L. 2014. *The Russian Presidency of Dimitri Medvedev, 2008-2012: The Next Step Forward Or Merely a Time Out?* Routledge. [66].

Black, J. L., and Michael Johns. 2013. *Russia after 2012: From Putin to Medvedev to Putin – Continuity, Change, or Revolution?* Routledge. [66].

Boshmaf, Yazan, Ildar Muslukhov, Konstantin Beznosov, and Matei Ripeanu. 2011. "The Socialbot Network: When Bots Socialize for Fame and Money." In *Proceedings of the 27th Annual Computer Security Applications Conference,* 93–102. ACSAC '11. New York, NY, USA: ACM. doi:10.1145/2076732.2076746. [23].

Burrett, Tina. 2008. "The end of independent television? Elite conflict and the reconstructing the Russian television landscape." In *The Post-Soviet Russian Media: Conflicting Signals,* edited by Birgit Beumers, Stephen Hutchings, and Natalia Rulyova, 71–86. Routledge. [62].

———. 2010. *Television and Presidential Power in Putin's Russia.* Routledge. [60, 62].

Chen, Cheng, Kyungmin Ko, and Ji-Yong Lee. 2010. "North Korea's Internet strategy and its political implications." *The Pacific Review* 23 (5): 649–670. [12].

Chen, Tianqi. 2014. "Introduction to Boosted Trees." Accessed March 12, 2016. https://homes.cs.washington.edu/~tqchen/pdf/BoostedTree.pdf. [43].

Chen, Tianqi, Tong He, and Michael Benesty. 2015. "CRAN - Package xgboost." Accessed March 12, 2016. https://cran.r-project.org/web/packages/xgboost/. [43].

Chu, Zi, S. Gianvecchio, Haining Wang, and S. Jajodia. 2012. "Detecting Automation of Twitter Accounts: Are You a Human, Bot, or Cyborg?" *IEEE Transactions on Dependable and Secure Computing* 9 (6): 811–824. [23, 25].

Clover, Charles. 2011. “Internet subverts Russian TV’s message.” The Financial Times. December 1. Accessed May 26, 2014. http://www.nytimes.com/2014/03/11/opinion/the-kremlins-social-media-takeover.html. [64].

Deibert, Ronald, John Palfrey, Rafal Rohozinski, and Jonathan Zittrain, eds. 2011. *Access Contested: Security, Identity, and Resistance in Asian Cyberspace.* Information revolution and global politics. MIT Press. [5, 6].

Denisova, Irina, Markus Eller, and Ekaterina Zhuravskaya. 2010. “What Do Russians Think About Transition?” *Economics of Transition* 18 (2): 249–280. [60].

Diamond, Larry. 2010. “Liberation Technology.” ¡p¿Volume 21, Number 3, July 2010¡/p¿, *Journal of Democracy* 21 (3): 69–83. [6].

Dunn, John A. 2008. “Where did it all go wrong? Russian television in the Putin era.” In *The Post-Soviet Russian Media: Conflicting Signals,* edited by Birgit Beumers, Stephen Hutchings, and Natalia Rulyova, 42–55. Routledge. [62].

Enikolopov, Ruben, Vasily Korovkin, Maria Petrova, Konstantin Sonin, and Alexei Zakharov. 2013. “Field experiment estimate of electoral fraud in Russian parliamentary elections.” *Proceedings of the National Academy of Sciences* 110 (2): 448–452. [71].

Enikolopov, Ruben, Maria Petrova, and Ekaterina Zhuravskaya. 2011. “Media and Political Persuasion: Evidence from Russia.” *American Economic Review* 101 (7): 3253–3285. [60].

Etling, Bruce, Karina Alexanyan, John Kelly, Robert Faris, John G. Palfrey, and Urs Gasser. 2010. “Public Discourse in the Russian Blogosphere: Mapping RuNet Politics and Mobilization.” SSRN Scholarly Paper. Accessed April 29, 2014. http://papers.ssrn.com/abstract=1698344. [2, 62, 67, 68].

Etling, Bruce, Hal Roberts, and Robert Faris. 2014. "Blogs as an Alternative Public Sphere: The Role of Blogs, Mainstream Media, and TV in Russia's Media Ecology." SSRN Scholarly Paper. Accessed April 29, 2014. http://papers.ssrn.com/abstract=2430786. [6, 68].

Freedom House. 2011. *Freedom on the Net 2011.* Washington, DC: Freedom House. http://www.freedomhouse.org/report/freedom-net/freedom-net-2011. [5, 69].

———. 2012. *Freedom on the Net 2012.* Washington, DC: Freedom House. http://www.freedomhouse.org/report/freedom-net/freedom-net-2012. [5, 72].

———. 2013. *Freedom on the Net 2013.* Washington, DC: Freedom House. http://www.freedomhouse.org/report/freedom-net/freedom-net-2013. [5, 72].

———. 2014. *Freedom on the Net 2014.* Washington, DC: Freedom House. https://freedomhouse.org/report/freedom-net/freedom-net-2014. [5].

Friedman, Jerome. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *The Annals of Statistics* 29 (5): 1189–1232. [43].

Gaidar, Egor. 2003. *The Economics of Transition.* MIT Press. [60].

Gehlbach, Scott. 2010. "Reflections on Putin and the Media." *Post-Soviet Affairs* 26 (1): 77–87. [62].

Gel'man, Vladimir, Dmitry Travin, and Otar Marganiya. 2014. *Reexamining Economic and Political Reforms in Russia, 1985–2000: Generations, Ideas, and Changes.* Lexington Books. [60].

Gessen, Masha. 2012. *The Man Without a Face: The Unlikely Rise of Vladimir Putin.* Penguin. [60, 62].

Giacomello, Giampiero. 2008. *National Governments and Control of the Internet: A Digital Challenge.* Routledge. [5].

Greenall, Robert. 2012. "LiveJournal: Russia's unlikely internet giant." BBC News. February 29. Accessed May 26, 2014. http://www.bbc.co.uk/news/magazine-17177053. [63].

Gunitsky, Seva. 2015. "Corrupting the Cyber-Commons: Social Media as a Tool of Autocratic Stability." *Perspectives on Politics* 13 (1). [6].

Hale, Henry. 2006. "Democracy or autocracy on the march? The colored revolutions as normal dynamics of patronal presidentialism." *Communist and Post-Communist Studies,* DEMOCRATIC REVOLUTIONS IN POST-COMMUNIST STATES, 39 (3): 305–329. [67].

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning.* Springer Series in Statistics. New York, NY: Springer New York. [43].

Howard, Philip N., and Muzammil M. Hussain. 2013. *Democracy's Fourth Wave?: Digital Media and the Arab Spring.* Oxford University Press. [12].

Human Rights Watch. 2014. "Russia: Halt Orders to Block Online Media — Human Rights Watch." March 24. Accessed May 27, 2014. http://www.hrw.org/news/2014/03/23/russia-halt-orders-block-online-media. [72].

Jonson, Lena, and Stephen White. 2012. *Waiting For Reform Under Putin and Medvedev.* Palgrave Macmillan. [19].

Judah, Ben. 2013. *Fragile Empire: How Russia Fell In and Out of Love with Vladimir Putin.* Yale University Press. [60, 71].

Kelly, John, Vladimir Barash, Karina Alexanyan, Bruce Etling, Robert Faris, Urs Gasser, and John G. Palfrey. 2012. "Mapping Russian Twitter." SSRN Scholarly Paper. Accessed April 29, 2014. http://papers.ssrn.com/abstract=2028158. [3, 68].

King, Gary, Jennifer Pan, and Margaret Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* 107 (2): 326–343. [5, 12].

Ko, Kyungmin, Heejin Lee, and Seungkwon Jang. 2009. "The Internet Dilemma and Control Policy: Political and Economic Implications of the Internet in North Korea." *Korean Journal of Defense Analysis* 21 (3): 279–295. [5, 12].

Kobak, Dmitry, Sergey Shpilkin, and Maxim S. Pshenichnikov. 2012. "Statistical anomalies in 2011-2012 Russian elections revealed by 2D correlation analysis." *arXiv:1205.0741 [physics, stat].* arXiv: 1205.0741. [71].

Koltsova, Olessia. 2006. *News Media and Power in Russia.* Routledge. [60].

Komaromi, Ann. 2004. "The Material Existence of Soviet Samizdat." *Slavic Review* 63 (3): 597. [61].

Kononov, Nickolay. 2014. "The Kremlin's Social Media Takeover." The New York Times. March 10. Accessed May 17, 2014. http://www.nytimes.com/2014/03/11/opinion/the-kremlins-social-media-takeover.html. [76].

Lange, Sarah. 2014. "The End of Social Media Revolutions." *The Fletcher Forum of World Affairs* 38 (1): 47–68. [6].

Lipman, Maria. 2009. "Media Manipulation and Political Control in Russia." Chatham House. Accessed May 16, 2014. https://www.chathamhouse.org/sites/default/files/public/Research/Russia%20and%20Eurasia/300109lipman.pdf. [60, 62].

Lunden, Ingrid. 2014. “Durov, Out For Good From VK.com, Plans A Mobile Social Network Outside Russia.” Techcrunch. April 22. Accessed May 17, 2014. http://techcrunch.com/2014/04/22/durov-out-for-good-from-vk-com-plans-a-mobile-social-network-outside-russia/. [76].

———. 2015. “Intel Shuts Down Russian Developer Forums To Comply With Russia’s ’Blogger Law’.” TECHCRUNCH. January 5. Accessed January 12, 2015. http://techcrunch.com/2015/01/05/intel-shuts-down-russian-developer-forums-to-comply-with-russias-blogger-law/. [75].

MacKinnon, Rebecca. 2011. “China’s ”Networked Authoritarianism”.” *Journal of Democracy* 22 (2): 32–46. [5].

Mills, Elinor. 2009. “Twitter, Facebook attack targeted one user.” CNET. August 6. Accessed May 27, 2014. http://www.cnet.com/news/twitter-facebook-attack-targeted-one-user/. [69].

Morozov, Evgeny. 2011. “Whither Internet Control?” *Journal of Democracy* 22 (2): 62–74. [5, 6, 10, 13].

Munger, Kevin, Richard Bonneau, John Jost, Jonathan Nagler, and Joshua Tucker. 2015. “Elites Tweet to get Feet off Streets: Measuring Elite Reaction to Protest Using Social Media.” New York University. http://kmunger.github.io/pdfs/Munger_post_APSR.pdf. [6].

Nabi, Zubair. 2013. “The Anatomy of Web Censorship in Pakistan.” *CoRR* abs/1307.1144. [5].

Nossik, Anton. 2014. “I Helped Build Russia’s Internet. Now Putin Wants to Destroy It.” New Republic. May 15. Accessed May 26, 2014. http://www.newrepublic.com/article/117771/putins-internet-crackdown-russias-first-blogger-reacts. [63, 65].

Oates, Sarah, and Tetyana Lokot. 2013. "Twilight of the Gods?: How the Internet Challenged Russian Television News Frames in the Winter Protests of 2011-12." SSRN Scholarly Paper. Accessed May 26, 2014. http://papers.ssrn.com/abstract=2286727. [65].

Pearce, Katy. 2014. "Two Can Play at that Game: Social Media Opportunities in Azerbaijan for Government and Opposition." *Demokratizatsiya: The Journal of Post-Soviet Democratization* 22 (1): 39–66. [5, 6].

Przeworski, Adam, Michael E. Alvarez, Jose Antonio Cheibub, and Fernando Limongi. 2000. *Democracy and Development: Political Institutions and Well-Being in the World, 1950-1990.* Cambridge University Press. [61].

Razumovskaya, Olga. 2011. "Russian Social Network: FSB Asked It To Block Kremlin Protesters." The Wall Street Journal. December 8. Accessed January 12, 2015. http://blogs.wsj.com/emergingeurope/2011/12/08/russian-social-network-fsb-asked-it-to-block-kremlin-protesters/. [76].

Remnick, David. 2011. "Putin's Television." The New Yorker Blogs. December 9. Accessed May 26, 2014. http://www.newyorker.com/online/blogs/newsdesk/2011/12/putins-television.html. [64].

Ries, Brian. 2014. "Founder of 'Russia's Facebook' Says Government Demanded Ukraine Protestors' Data." Mashable. April 16. Accessed January 12, 2015. http://mashable.com/2014/04/16/vkontakte-founder-fsb-euromaidan/. [76].

Roberts, Hal, and Bruce Etling. 2011. "Coordinated DDoS Attack During Russian Duma Elections." Internet & Democracy Blog. December 8. Accessed May 27, 2014. http://blogs.law.harvard.edu/idblog/2011/12/08/coordinated-ddos-attack-during-russian-duma-elections/. [69].

Roberts, Margaret. 2014. “Fear, Friction, and Flooding: Methods of Online Information Control.” Doctoral dissertation, Harvard University, June 6. Accessed June 25, 2016. https://dash.harvard.edu/handle/1/12274299. [5].

Rothrock, Kevin. 2015. “Meet Russia’s 369 Kremlin-Registered Bloggers.” Global Voices. January 8. Accessed January 12, 2015. http://globalvoicesonline.org/2015/01/08/meet-russias-369-kremlin-registered-bloggers/. [75].

Sakwa, Richard. 2014. *Putin Redux: Power and Contradiction in Contemporary Russia.* Routledge. [20, 65, 71, 77].

Shleifer, Andrei, and Daniel Treisman. 2001. *Without a Map: Political Tactics and Economic Reform in Russia.* MIT Press. [60].

Simons, Dr Greg. 2013. *Mass Media and Modern Warfare: Reporting on the Russian War on Terrorism.* Ashgate Publishing, Ltd. [70].

Sivkova, Alena. 2014. “”We don’t see much risk in blocking Twitter in Russia”.” Izvestia. May 16. Accessed January 9, 2015. http://izvestia.ru/news/570863. [73].

Smyth, Regina, and Irina Soboleva. 2014. “Looking beyond the economy: Pussy Riot and the Kremlin’s voting coalition.” *Post-Soviet Affairs* 30 (4): 257–275. [77].

Snyder, Timothy. 2014. “Fascism, Russia, and Ukraine.” The New York Review of Books. March 20. Accessed May 27, 2014. http://www.nybooks.com/articles/archives/2014/mar/20/fascism-russia-and-ukraine/. [77].

Soldatov, Andrei, and Irina Borogan. 2013. “Russia’s Surveillance State.” *World Policy Journal* 30 (3): 23–30. [62].

Travis, Hannibal, ed. 2013. *Cyberspace Law: Censorship and Regulation of the Internet.* Routledge. [5].

Tregubova, Yelena. 2003. *The Tales of a Kremlin Digger.* Moscow: Ad Marginem. [60, 62].

West, Darrell. 2010. "President Dmitry Medvedev: Russia's Blogger-in-Chief." The Brookings Institution. April 14. Accessed January 3, 2015. http://www.brookings.edu/research/opinions/2010/04/14-medvedev-west. [66].

Womack, Helen. 2014. "Making waves: Russian radio station is last bastion of free media." *Index on Censorship* 43 (3): 39–41. [62].

Xu, Chang, Dacheng Tao, and Chao Xu. 2013. "A Survey on Multi-view Learning." *arXiv:1304.5634 [cs].* arXiv: 1304.5634. [39].

Yagodin, Dmitry. 2012. "Blog Medvedev: Aiming for Public Consent." *Europe-Asia Studies* 64 (8): 1415–1434. [66].

Yudina, Natalia. 2012. "RuNet, hate crime and soft targets: how Russia enforces its anti-extremism law." Open Democracy. October 30. Accessed May 27, 2014. http://www.opendemocracy.net/od-russia/natalia-yudina/runet-hate-crime-and-soft-targets-how-russia-enforces-its-anti-extremism-la. [71].

Zasursky, Ivan. 2004. *Media and Power in Post-Soviet Russia.* M.E. Sharpe. [18, 60].

Appendix A Keywords and hashtags for collecting Twitter Data

- медведев
- духовныескрепы
- путинвор
- путинах
- пжив
- Стратегия31
- триумфальная
- болотная
- оппозиция
- горожанепротив
- сурковскаяпропаганда
- навальный
- занавального
- команданавального
- судвкирове
- PussyRiot
- толоконникова
- народныйсход
- сднемпобеды
- #sochi2014
- #sochi
- #putinsgames
- #сочи - Sochi
- #витишко -
- #считаемвместе
- #sochifail
- #sochi2014problems
- голодовка
- МинутаНеМолчания
- жалкий
- путин
- спасибопутинузаэто
- прямаялиния
- партияжуликовиворов
- едро
- 6мая

- собяниннашмэр
- маршмиллионов
- зачестныевыборы
- болотноедело
- 6may
- свободуполитзаключенным
- свободуузникам6мая
- росузник
- одинзавсех
- всезаодного
- рассерженные
- честныевыборы
- удальцов
- высурковскаяпропаганда
- 37годвернулся
- ДМП
- привет37год
- кровавыйрежим
- кировлес
- ПуссиРайот
- толокно
- бирюлево
- хватиткормитькавказ
- хватитвинитькавказ
- русскиймарш
- Сочи2014
- #олимпиада
- #олимпийскаязачистка
- #sochiproblems
- Одесса
- #Nemtsov
- #Немцов
- немцов
- савченко
- #FreeSavchenko
- #Putinkiller
- майдаун
- майданер
- майданутый
- #ПутинУмер
- #МинутаНеМолчания
- Su24
- Су-24

- Су24
- #самолет
- #RussianJet
- #ExpelTurkeyFromNATO
- #Russianplane
- #Erdogan
- #Latakia

Appendix B Russian government online: a long way from Putin to Putin

B.1 2000-2008: Internet vs. traditional media in Putin's Russia

Russian reforms of 1990s got mixed assessment from both outside observers (Shleifer and Treisman 2001) and reformers themselves (Gaidar 2003), and their reception among Russians remains ambiguous at best (Denisova, Eller, and Zhuravskaya 2010). However, a wide consensus holds that if anything worked during the painful transition from Communism, it was media freedom (Zasursky 2004). Diverse, influential and competitive news outlets emerged almost immediately and by the end of 1990s several powerful media conglomerates were operating alongside a large network of independent federal and regional media, usually free of any government control (Lipman 2009). This is proved by the enormous role media played in political fights throughout 1990s (Koltsova 2006) and, above all, during so-called war for Yeltsin's succession (Gel'man, Travin, and Marganiya 2014, 104-108), when high-powered media was mobilized by both Putin and his opponents. Role of the media in Putin's rise to power is well documented by rigorous quantitative studies (Enikolopov, Petrova, and Zhuravskaya 2011), Putin's biographers (Gessen 2012, Chapter 2), Western observers (Judah 2013, Chapter 2) and Russian political memorialists alike (Tregubova 2003, Chapter 10). This experience allowed Putin to fully appreciate the power of the media to change public opinion and reverse political fortunes. Putin's media policy in the next 15 years demonstrates that he took this lesson extremely seriously and worked tirelessly to put media under his control (Lipman 2009; Burrett 2010). However, as we shall see, this policy was not universally applied to all types of media. To the contrary, online media enjoyed the *laissez-faire* regime that was in many respects on par with most advanced democracies. To uncover the reason, why Putin had drawn such sharp line between traditional and online media, we will start by examining the Soviet experience of media control. Seemingly inconsistent strategy Putin adopted would become much less surprising if discussed in light of the strategy once devised by Putin's (former KGB lieutenant colonel) employers at the Central Committee and Lubyanka. Next, we will discuss main features of Putin's dual strategy and assess the changes in political and media environment, which ultimately rendered it unworkable.

Soviet Union, as a relatively long-lived 20th century dictatorship[46], survived several waves of technological advancement. Already in 1917 Bolsheviks famously recognized the role of modern technologies and communications by seizing (along with the Winter Palace, railway stations, bridges and army headquarters) Petrograd's Telegraph and Telephone Exchanges. This recognition entailed two different strategies – for mass and personal communications – which were implemented fairly persistently throughout Soviet history. The government maintained the complete monopoly over the mass communication and fiercely prosecuted those who tried to challenge this monopoly. Most famous examples include jamming of Western radio broadcasters (Radio Liberty, Voice of America, BBC, etc.) and strict regulations over using printers, plotters and photocopier after they were installed at various Soviet administrative departments and institutions[47] (Komaromi 2004).

Personal communications was a different matter. Instead of monopolizing their usage, government allowed Soviet citizens to use them promiscuously and then used it to identify and prosecute those who were disloyal. Phones, for example, were spreading rapidly in the USSR, approaching roughly 37 million, or about 13 per 100 inhabitants in 1990 (Banks and Wilson 2013). It was still six times as small as the U. S. at the time, but the difference was due to technological and economic reasons, not political restrictions. However, government used every opportunity to spy over its citizens using wiretapping. Under Stalin serious efforts were put in the research on speaker identification, which was famously depicted by Aleksandr Solzhenitsyn in the autobiographical novel *In the First Circle*. Later the elaborate system of surveillance and spread of personal, but publicly registered home phones eliminated the need in voice identification. Similarly, typewriters were allowed for personal use, however their printout had to be handed to the *First Department* (local KGB affiliates at any Soviet

[46] According to Przeworski et al. (2000) average dictatorship which was overthrown between 1950 and 1990 lasted 27.4 years and average dictatorship still in course in 1990 was 26.2 years old.

[47] The last among these regulations, which was not enforced anymore, was struck down by the Russian Supreme Court only in 2009.

enterprize or institution) and could be cross-verified to identify the exact typewriter used in printing "inappropriate" materials.

After assuming power in 1999, Pitin gradually implemented similar strategy of complete monopolization of mass media and liberal policy on personal communications. While the latter was virtually left free from interference (but not from surveillance[48]), national television networks were returned to government ownership and Soviet-style management with weekly instructions delivered at Kremlin to the news executives (Gehlbach 2010). Press and radio remained more diverse, with some pro-government and some relatively independent outlets competing with each other (Lipman 2009; Womack 2014). The process of putting traditional media under government control in the early 2000s included such colorful episodes as imprisonment of media mogul and oligarch Vladimir Gusinsky (in order to be released he signed a secret protocol with Russian Minister of Communications Mikhail Lesin and handed over his media assets to state natural gas monopoly Gazprom) and stunning reversal of fortunes for the architect of both Yeltsin's electoral victory in 1996 and Putin's one in 2000 Boris Berezovsky (who lost control of main Russian TV channel and went to exile already by 2001) (Becker 2004). This and other episodes and their consequences for the media landscape are well documented in literature (see academic stydies in Burrett 2010; Dunn 2008; Burrett 2008; for more informal discussion see Gessen 2012, Chapter 7; and an illuminating personal memoir by Tregubova 2003, Chapters 11-13). Putin's approach to the internet, on other hand, was rarely assessed. Observers simply noted that "the Russian blogosphere is a space that appears to be largely free of government control" (Etling et al. 2010, 33); "the absence of Internet filtering is notable. Based on tests run through the OpenNet Initiative, we continue to find no evidence of significant technical filtering of the Russian Internet" (Alexanyan et al. 2012, 10-11), etc. A recent account by one of the leading Russian internet news producers Anton Nossik suggests that it was no accident. Instead, already in 1999, still prime-minister Vladimir Putin had a clear preference for non-interference in the internet space:

[48] Sophisticated "deep pocket" web surveillance system known as SORM (-2,-3) was installed by the Russian government no later than in the late 1990s and has been being updated constantly ever since (see Soldatov and Borogan 2013).

> ... in December 1999, three days before he became acting president of Russia, Vladimir Putin [...] summoned all the heads of Russia's nascent Internet industry for a meeting [...]. In his brief but passionate speech that day, Putin made special mention of Chinese and Vietnamese models of Internet regulation, stating that he viewed them as unacceptable. "Whenever we'll have to choose between excessive regulation and protection of online freedom, we'll definitely opt for freedom," he concluded [...]. (Nossik 2014)

Under the auspices of such a benevolent government policy, Russian online media flourished, creating a vibrant sector of the economy and a reliable source of information for millions of Russians. Russia is one of the few countries where Google is not the most popular search engine and Facebook is not the most popular social network. Remarkably, both occurred without restrictions on American competitors. Unlike Baidu and Weibo, Yandex, Odnoklassniki and Vkontakte won virtually fair competition with their American counterparts[49]. Successful development of local services did not mean that foreign ones were not actively used by Russian bloggers and readers. LiveJournal, the most popular Russian social network in 2001-2011, while being originally American and predominantly English-speaking, developed Russian community so large that it was eventually overtaken by a Russian media holding and became dominated by Russian users (Greenall 2012). And as of April, 2014 Facebook has 24 million users from Russia [50] and Twitter has more than 8 million[51], which makes Russian one of 10 most popular languages on Twitter[52].

Again, in stark contrast with most other countries, Russian most popular news web-sites do not represent traditional media such as newspapers, radio and TV broadcasters (Nossik 2014). Instead, *Gazeta.Ru*, *Lenta.Ru*, *NewsRu.com*, *Polit.ru* and alike were built from scratch and became major news outlets in their own right (i.e. their staff does original reporting,

[49] Still, Vkontakte (but not Yandex or Odnoklassniki) had significantly benefited from the lax enforcement of the property rights. However, this doesn't make comparison with China less impressive, given that China is also famous for wide-spread piracy.

[50] http://ria.ru/technology/20121004/766127348.html

[51] http://digit.ru/internet/20131031/407481403.html

[52] http://bits.blogs.nytimes.com/2014/03/09/the-languages-of-twitter-users/

often as an eyewitness, rather than just digitalize other's content).

As a result, Russia developed a strong, powerful and independent internet media sphere, which was a remarkable achievement for any non-democratic country, but especially for one, where traditional media are so tightly controlled. As Alexanyan et al. (2012) note "Russia is unusual in the degree of freedom found online compared to offline media and political restrictions". Such imbalance, however, proved to be unsustainable. In the late 2000s internet media increasingly supplemented and eventually supplanted TV as the main news source at least for the educated Russians (Clover 2011). One of the leading Russian TV anchors Leonid Parfenov, who has been banned from air since 2004, aptly summarized this process in 2010 speech, which went viral on YouTube (Remnick 2011):

> These evergreen tricks are known to everyone who has witnessed the Central Television of the USSR. Reports are replaced by protocol shootings like "Meeting at the Kremlin"; reporter's intonations support the officials in the picture; broadcasting models are implemented to show "the leader receiving a minister or a governor", "the leader campaigns among the masses ", "the leader holding a summit with his foreign colleague", etc. These are not news; this is old record that repeats the already established patterns of broadcasting. Even a news hook isn't a must. In the emasculated media environment any small fry will pass for a big shot just because of getting some airtime.
>
> [...]
>
> It hurts twice as much to speak about television journalism, given the evident success of the large-scale TV shows and Russian school of television series. Russian TV is getting more and more sophisticated in exciting, fascinating, entertaining and amusing people, but it hardly could be called civic social and political institution. I am convinced it is the reason for the dramatic decline in TV viewership among the most active part of the population. People of our type say: "Why bother turning on the box? It's not intended for us".

However, as Nossik (2014) notes, this dual strategy – tight control of the traditional media and almost complete nonintervention in the web – was devised when Russian internet penetration was almost negligible. Even three years after Putin came to power, in 2002, Russia has 2.1 million people (2% of the adult population) who used internet daily [53]. By 2008 this share increased to 14 million (16% of the adult population), and by 2013 to 52.2 million people (46% of the adult population)[54]. Needless to say, the quality of access changed dramatically after wide access to broadband connection replaced slow dialup. This circumstances diminished the value of monopoly in TV broadcasting and strong influence in other traditional media which Kremlin enjoyed (Oates and Lokot 2013) and simultaneously made the online communities sufficiently large and well-structured to become politically significant. Dual nature social media, which is simultaneously mass and personal communication, presented a particular challenge for the government.

These changes coincided with the constitutionally required transition of power from Putin (who served two consecutive terms) to Medvedev in 2008. While Putin was appointed Prime Minister of Russia immediately after elections and Medevdev was widely considered as a weak leader, who never freed himself from Putin's oversight, Medvedev had his own agenda and probably nowhere else it was more visible than in his approach towards information technologies and internet in particular.

B.2 2008-2012: "Blogger-in-Chief" and his followers

Dmitry Medvedev's approach towards internet was integral part of his general agenda. Laid out in an article "Russia, Forward!", which was published in liberal (and online-only) newspaper *Gazeta.ru*, his *modernization* plan aimed to preserve the basic parameters of the political system built by Putin, but make it more efficient and friendlier towards businesses and citizens (Sakwa 2014, Chapters 3-5). This included, for example, establishing Moscow as an international financial center, police reform, boosting higher education international compet-

[53] http://bd.fom.ru/report/map/projects/internet/internet1133/vesna2011

[54] http://fom.ru/SMI-i-internet/11417

itiveness and creation of a functioning e-government (see and overview of Medvedev's reforms in Black 2014; Black and Johns 2013). Medvedev signature project was Skolkovo, a publicly funded, but semi-independently managed high-tech incubator near Moscow. Obviously, the success of these projects was dependent on creative class in major population centers, and IT professionals in particular. Thus establishing a communications channels with these people, who were largely ignored by the blatant Soviet-style TV propaganda, was the first order of business for Medvedev. And unlike in many other areas, he did not hesitate to break with Putin legacy, and put the traditionally solemn and unquestioned presidential speech in the caustic domain of the social networks.

Less then a year after assuming office, in early 2009 Medvedev started a video blog which quickly moved to LiveJournal – then Russian main social network and blogging platform. In 2010 he visited Silicon Valley, met Steve Jobs and opened a twitter account at Twitter headquarters in San Francisco. Notably, his account began to follow (in addition to foreign heads of states and Russian officials) several bloggers known for their criticism of the government and newsfeed from radio station *Echo of Moscow* – perhaps the most critical of government among major media outlets in Russia. Finally, in 2011 he opened his Facebook page, which he occasionally used to communicate with its readers on the matters not covered or ill-covered by the official media (such as 2011 protests) using a different, more frank tone. In all social networks he build a large readership, which is typical for heads of states, but still notable since the environment was completely different from the general media environment Medvedev was used to: here he could not get his message through simply by eliminating competition and controlling the platform and the agenda (Yagodin 2012). In addition, in a rare occasion in 2011 he visited small private TV channel *Rain*, which at the time was mainly accessible online. As a result, Medvedev got permanently associated with blogging and social networks, and even called both in Russia and abroad "Blogger-in-Chief" (see for example West 2010), which simultaneously gave him credit for being up-to-date with the internet age and suggested that his rhetoric translates in little action.

Medvedev was not embarking on social media platforms alone. While it still remained an exception for high-level public officials at the time, several of his aids established significant presence on the social media. In particular, his close aid and economics adviser Arkady Dvorkovich maintains one of the most popular Russian twitter accounts with close to half a million followers; he also has a Facebook page, as does Medvedev's press-secretary Natalya Timakova (who as a former journalist is Facebook friend of many prominent liberal reporters). However, probably even more important was the establishment of large-scale and permanent operation to push pro-government agenda on the web and in social media in particular. Following the long-standing Russian tradition, government action came late, but quickly. Pro-Kremlin youth movements, created to combat color revolution on Moscow streets and squares (Hale 2006) were partially repurposed to push pro-government agenda online. Its leaders (in case of *Nashi* they were called *commissars*) became active bloggers, but they never relied on the persuasion capacity of their messages. Instead they gradually created a network of online support. Until then Russian government presence on social media was very limited. A report by the Berkman Center for Internet and Society at Harvard University, which was published in late 2010 and covered Russian blogosphere – concentrated in LiveJournal at the time – in May 2009 – September 2010, found that "pro-government bloggers are not especially prominent and do not constitute their own cluster" (Etling et al. 2010, 3). Moreover, those affiliated with the government "are not central nodes in any of the political or social clusters [...] investigated" (33).

A network of support started with artificial intelligence rather than human effort. Networks of bots got frequently employed first to flood opposition blogs with meaningless or assaultive content. Later they began to push alternative, pro-government messages to top charts and help pro-government bloggers to attract new followers. A report by the Berkman Center noted that "there is a concentration of bloggers affiliated with pro-government youth groups among the Instrumental bloggers [i.e. bots]" (3). However, real bloggers soon followed. In less than a year – which also witnessed the transition of the discussion core of the Russian blogosphere from LiveJournal to Twitter – pro-government bloggers emerged as a distinct, and indeed, one of the largest clusters on Russian political Twitter (Kelly et al. 2012, 11). This result holds even after filtering out bots and other instrumental accounts, which remained numerous in the pro-government segment.

Continuous monitoring of the Russian blogosphere, undertaken by "Internet in Russian society" program at the Berkman Center for Internet and Society at Harvard University in 2010 – 2014 reveals several distinctive characteristics of pro-government segment in Russian social networks, as compared both to oppositional and "uncommitted" users. First, due to the general weakness and high fragmentation of the Russian opposition, "many active Russian bloggers [...] engage on political topics without 'choosing a team'. [...] most Russian bloggers prefer to declare an independent intellectual posture, and eschew group affiliations" (Etling et al. 2010, 19). In contrast, pro-government bloggers tend to declare their political preferences and affiliation. Moreover, usage of predominantly pro-government hashtags in Twitter was highly concentrated among pro-government users, at least compared to predominately oppositional hashtags, which were more widely used in different clusters. Finally, while pro-government users demonstrate high commitment in terms of the number of hashtag mentions (after the first one), they usually did it in a short time period, producing sharply peaked distribution of hashtag popularity (Barash and Kelly 2012).

As blogosphere remain the most ideologically diverse media environment in Russia, pro-government users experience pressures absent in other media. A comparative study of Russian blogosphere and TV in the year before the Duma elections of 2011 reveals that such competitive environment forces pro-government bloggers to engage with their adversaries in cases when TV and even newspapers could largely ignore oppositional activity. Etling, Roberts, and Faris (2014) give an example of the oppositional youth retreat in the outskirts of Moscow, which was intended to countervail large government-sponsored youth camp "Seliger". Largely overlooked by the traditional media, it became the subject of the heated discussion between leading oppositional and pro-government bloggers on Twitter.

Online response to hostile (or perceived as such) internet activity through direct engagement with users remained the "weapon of choice" during Medvedev presidency, but certainly it wasn't the only one. Both offline response and attempts to go through the online infrastructure did take place, but the latter were relatively rare and quite limited in their scope and the former was not a part of any systematic internet policy, and as such could not (and wasn't intended to) change the digital media landscape.

Up until the end of Medvedev's presidential term the only type of internet infrastructure infringement known in Russia were relatively brief (lasting up to several days) DDoS attacks on particular web resources (Agora 2011; Freedom House 2011). The first major attack was launched on August 6, 2009 – the first anniversary of the Russo-Georgia 5-days war. The target was pro-Georgian blogger *cyxymu*. The attack was strong enough to significantly disrupt Facebook and completely shut down Twitter and LiveJournal (Mills 2009). The series of smaller attacks on various LiveJournal blogs and independent media culminated on the weekend of the Russian Duma elections of 2011, when two dozens of the most prominent independent media (including *The New Times*, *Kommersant*, *Echo of Moscow*, *Novaya Gazeta*, *Slon*, etc.), blogs (including the entire LiveJournal platform) and, most crucially, election monitors' coordinating portals (including the largest one, GOLOS) were shut down for hours (Roberts and Etling 2011). Later many of the very same resources were attacked during oppositional rallies after elections and in the early 2012.

Importantly, DDoS attacks, unlike filtering (and offline response), could be used not only be the government, but also by the opposition. In early 2012 Russian branch of international cyber activist group *Anonymous* blocked web sites of the Russian government, Kremlin and several major state media, such as *Vesti* and *RIA Novosti*[55]. These attacks, however, did last only several hours (compared to several days in the case of LiveJournal), and, obviously, could not impede state response to demonstrations.

[55] http://habrahabr.ru/post/143501/; http://lenta.ru/news/2012/05/10/attack/

Finally, offline response by Russian government to unfriendly internet activity was not yet separated from the general anti-opposition activity and was not legally or organizationally institutionalized. Market regulation and government entrepreneurship was still targeted at traditional media: for example, in 2011 newspaper *Moskovskiye Novosti* was relaunched by state news agency *RIA Novosti*. As it was widely assumed, the project was aimed to provide moderate competition to privately owned (and quite critical) *Vedomosti*, simultaneously being more friendly to Medvedev than most state media, loyal to Putin[56]. Later that year Medvedev announced the establishment of the Public Television of Russia, which faced no private competition, but shared the second goal with *Moskovskiye Novosti*[57].

Violence and legal action against bloggers were relatively rare and mostly took place in the North Caucuses. Legal restrictions, if any, were imposed under the auspices of the general anti-terrorist laws and orders, mostly having to do with combating Chechen and Dagestani insurgencies. While anti-terrorist rational was often abused for the sake of winning over political enemies in the respective republics, these cases were rarely consequential at the federal level (Simons 2013). In few cases outside the Caucuses prosecutions were largely a regional matter or the result of local security apparatus initiatives rather than implementation of any national strategy. Prominent cases from that time included blogger Savva Terentyev from Komi Republic, who in 2008 was convicted of defamation of the "social group 'law enforcement personnel'" and sentenced to one year of imprisonment with a probation period of one year after an anti-police comment at Liverjournal. Another prominent case took place in 2009 in the Republic of Tatarstan, where a former government official turned opposition blogger posted false rumor that the governor of the republic has died. He was convicted of libel and defamation of the "social group 'government officials'" and sentenced for 2 years in

[56] A. Morozov. 2011. "'Moskovskie Novosti' Space". http://os.colta.ru/media/projects/18065/details/21397/; O. Barykova, and M. Zotova. 2011. "Ushlo li vremya 'Moskovskih Novostey'?". BBC Russian. http://www.bbc.co.uk/russian/russia/2011/04/110331_moscow_news_scandal.shtml; Meyer, Henry. 2011. "Putin Revives Gorbachev Glasnost Paper to Widen Election Appeal." Bloomberg. http://www.bloomberg.com/news/2011-03-30/putin-revives-gorbachev-glasnost-paper-to-widen-election-appeal.html

[57] As mentioned, execution of government projects in media suffers from general government inefficiency and the TV channel went on air only in 2013, long after Medvedev switched offices with Putin.

prison (Yudina 2012).

Institutionalization of offline response, as well as means of control over the online infrastructure happened only after Dmitry Medvedev handed his office back to Vladimir Putin in 2012. However, the process was so quick that already by 2014 the relative importance of different types of government response was reversed: sheer force of offline response and establishment of a comprehensive system of internet filtering rendered the online engagement with users, created by Medvedev, almost irrelevant.

B.3 2012 – : Cracking down and giving up

Compared to transition from Putin to Medvedev in 2008, the reverse transition in 2012 was much less smooth. Announced on Septermber 24, 2011 and immediately nicknamed as "castling", it was met with resentment by both Medvedev supporters and those in opposition to both Medvedev and Putin (Judah 2013). This resentment has transformed in large-scale street protests after parliamentary elections in December, 2011, which were widely considered as rigged[58]. As we mentioned above, close relationship between Putin and Medvedev (culminated in "castling"), did not mean that Medvedev lacked his own agenda. In this case too his response was a program of moderate, but significant political reforms, announced in the Address to the Federal Assembly (Russian equivalent of the State of the Union) in late December of 2011, three weeks after Duma elections, and just after major protests had started. This program included, most importantly, reinstatement of popular elections of Russian governors and elections of MPs in districts (switching back from pure proportional to mixed electoral system) (Sakwa 2014, 129-132). These reforms, however, were either striped of any substance (like change in party registration rules) or explicitly reversed (like decriminalization of libel) (Chapters 7-8). Protest activity, on the other hand, was severely restricted after on May 6, 2012 (one day before Putin's inauguration) an opposition rally was dispersed by force (hundreds of people were arrested and several dozens of them were subsequently prosecuted for

[58] Post-election analysis revealed that suspicion was well-grounded (Kobak, Shpilkin, and Pshenichnikov 2012; Enikolopov et al. 2013).

inciting riots and assaulting police).

It is in this context, when the freedom of Russian internet from filtering came to an end (Freedom House 2012, 2013). Already in July of 2012, despite vocal protests, including Russian Wikipedia temporary voluntary shut down, Russian State Duma adopted (and Vladimir Putin signed into law) so-called Internet Restriction Bill (Federal law of Russian Federation no. 139-FZ), which created a continuously updated Russian Internet Blacklist[59]. The list, maintained by the Russian Federal Service for Supervision of Communications, Information Technology and Mass Media (Roskomnadzor), contains domain names which any Russian ISP has to permanently block on the grounds of containing pornography, copyright infringement or "extremist content". Initially, items were to be included in the list per a court order and only if the hosting website fails to remove the content in 24 hours after receiving the notification. However, in December of 2013 new amendments to the Law on Information, Information Technology, and Information Protection provided the Office of the Prosecutor General with the authority to block websites without any court order. Moreover, the procedure was changed, so the web page were to be blocked first, and allowed to be accessible again only after it removes the content deemed as "calling for mass disorders, extremist activity, and participation in mass public events, which fail to follow appropriate regulations"[60] (Human Rights Watch 2014).

However, when at the height of Russian-Ukrainian conflict in March of 2014 several oppositional news web sites were blocked, even these loose rules were not followed. On March 13, 2014 *Grani.ru*, *Kasparov.ru* and *EJ.ru*, as well as popular opposition politician (in 2013 he ran for Moscow mayor and came second) Alexey Navalny's LiveJournal blog, were blocked by all ISPs per government order. Since then several suits were brought to courts demanding the reason for the blocking. Journalists and Alexey Navalny asked authorities to identify specific materials on these websites that triggered the blocking, so that the materials could be removed and access reestablished. Throughout 2014 authorities repeatedly denied that

[59] http://eais.rkn.gov.ru/

[60] Text of the law: http://www.rg.ru/2013/12/30/extrem-site-dok.html.

they are under any obligation to provide such information and courts repeatedly dismissed the cases. Early in 2015, all three websites and Navalny's LiveJournal page remain completely blocked in Russia.

Still, Russian government incomplete control over the online infrastructure significantly impedes its ability to crack down on opposition activity simply by blocking web pages. The greatest thereat, of course, are largest social media platforms – Facebook and Twitter. First, unlike most other web resources, Facebook's and Twitter's individual pages (say, a particular post or user profile) could not be blocked by the filtering software currently available to the Russian authorities (Sivkova 2014). Blocking the entire platforms, on the other hand, is still considered undesirable: it would further hurt Putin's regime reputation abroad and simultaneously hurt and potentially antagonize a large number of politically indifferent (and regime-friendly) users in Russia. In a rare event, a public official, Roskomnadzor deputy head Maxim Ksenzov, who speculated over such possibility, was publicly disproved by Prime Minister Dmitry Medvedev in a Facebook post[61] and later was formally reprimanded. Of course, instead of blocking them, Russian government could ask them to police themselves and remove access to certain pages at least for users inside Russia. However, unlike Vkontakte, foreign social networks could easily ignore such orders. For example, in December of 2014 authorities requested Facebook and Vkontakte to block access to pages, allowing supporters of Alexey Navalny to register for a rally protesting his looming criminal conviction and receive updates about the place and time of the event. Vkontakte blocked the page and all subsequent attempts to create a copy, posting a warning that "This page is blocked upon receiving a Roskomndazor notification of restricting access to information, which contains calls to participate in mass public events, which fail to follow appropriate regulations, as per the request of the Office of the Prosecutor General of Russia."[62]. Facebook also blocked access to a similar page inside Russia[63], but after a huge outcry in Western media, refused to block

[61] https://www.facebook.com/Dmitry.Medvedev/posts/10152047885946851

[62] https://vk.com/blank.php?rkn=32274605; It should be noted that according to those "appropriate regulations" authorities could not be notified about the upcoming rally earlier than 15 days in advance. The page was blocked 26 days before the event it announced was scheduled to take place.

[63] https://www.facebook.com/events/417200101767938

any other pages. Moreover, some Russian media outlets, which were afraid to report the scheduling of the event itself, covered the Roskomnadzor order and social networks response. As a result, more people learned about the event and the new event page opened on Facebook attracted even more people[64].

Given that second page attracted more than 33 thousands people, who stated that they are “going to the rally” (plus almost 6 thousands, who stated that they are “likely going”), it’s not surprising that the authorities resorted to offline response: they simply changed the data of the return proceedings to two weeks earlier. The new date was the day before the largest Russian holiday (The New Year’s Eve) and Navalny was informed less than 24 hours in advance. While the third event also attracted considerable number of supporters, combination of suddenness, cold weather and pre-holidays preparation likely reduced the turnout.

Offline response was certainly not limited to ad hoc solutions just described. Instead, government complete control over law enforcement apparatus and law making was actively used to augment its limited ability to censor social media platforms. Criminalization of online activity was first implemented trough targeted amendments to existing criminal law, but was soon institutionalized in a dedicated law. Using media to spread information deemed extremist was always an aggravating circumstance in Russian criminal law. Laws missing such provision were sooner or later corrected: for instance, when in 2011 punishment for Article 280 of the Criminal Code was severed, using mass media for “extremism propaganda” became an aggravating circumstance. However, when just two years later, in 2013, a new extremism crime appeared in the Criminal Code (Article 280.1, Public Appeals to the Violation of the Territorial Integrity of the Russian Federation), using “mass media, including telecommunication networks (including ‘Internet’)” was added as the aggravating circumstance (Agora 2012, 2013).

In May, 2014 Vladimir Putin signed into law a requirement for any blogger with the daily readership in excess of 3000 people to register with the government and reveal her true identity

[64] https://www.facebook.com/events/406603402849183

and email address [65]. In addition, bloggers will be held accountable for failure to verify the information they "spread", have to keep archives of their postings and follow laws which regulate news production during electoral campaigns. However, institutionalized regulations expectedly are much less effective then targeted actions: in half a year after the law came into force just 369 people got registered with Roskomnadzor (Rothrock 2015) and the only known real consequence is the shut down of Intel's forum for developers – hardly a platform of political significance, which was closed by Intel voluntarily out of precaution (Lunden 2015). Among the reasons is unclear definition of "readership": Roskomnadzor guidelines on the subject[66] call to use rigorous "page views" count (rather than hits, number of friends or followers or any other metric), but not all platforms generate such statistics, and it is especially hard to do in social networks.

Using loyal business groups to restructure the online media market proved much more reliable tool to ensure that at least Russian major platforms are under control. Hitherto mostly focused on traditional media (TV and press), power brokers in the Presidential Administration, Ministry of Communications and largest media conglomerates have been increasingly preoccupied with online news outlets and platforms. The methods they used were not much different. Two cases are particularly revealing. In 2014 billionaire Alexander Mamut fired the editor-in-chief of the most popular Russian online news portal *Lenta.ru*, allegedly on the grounds of insufficiently "pro-Russian" coverage of the Ukrainian revolution of 2013-2014. Complete lock out of the entire editorial staff was strikingly similar to the one at the NTV channel in 2001 and countless others since then. However – and here comes the difference between TV and website – this fired team of journalists was able to relaunch their media. The insurance of their independence and security from outside pressure was physical relocation of most of the editorial staff to neighboring Baltic country of Latvia and opening website in *.io* domain zone, which belongs to British Indian Ocean Territory and is administered by a UK

[65] According to the survey reported by Alexanyan et al. (2012), even without any legal requirement Russian bloggers rarely conceal their identity. They do use pseudonyms (following internet tradition), but usually alongside, not instead of their real names. This is particularly true for politically-engaged bloggers.

[66] http://rkn.gov.ru/docs/prikaz_Roskomnadzora_ot_09.07.2014_N_99.pdf

company. New media name *Meduza* (Russian for jellyfish) matches the geographical location of its domain.

The hostile takeover of Vkontakte in 2013-2014 by Kremlin-affiliated businessmen also followed the approach which earlier successfully secured the loyalty of various media outlets (such as *Izvestia* and *Kommersant*): involuntary ownership transfer, usually, compensated at the market rate (dependent on the cooperation of the former owner). This transfer usually came after former owners and/or managers refuse to cooperate in politically sensitive matters for too long. Vkontakte received requests from FSB similar to the ones described (and followed) in case of pro-Navalny rally in late 2014 for years. Specifically, requests to remove pro-Navalny groups came first in the wake of large-scale protests after Duma elections in 2011 (Razumovskaya 2011), but Vkontakte owner and CEO, libertarian internet-guru Pavel Durov refused to comply. However, when in early 2014 Vkontakte was served with request to disclosure personal data of administrators of Euromaidan-related pages in Vkontakte [67], government did not take no for an answer. Durov had to sell what was left of his share, resigned and left the country (Ries 2014; Kononov 2014; Lunden 2014).

The lesson of Vkontakte was taken seriously not only by Russian media managers and owners, who wanted to keep their positions and businesses. Foreign companies which wanted to be able to refuse involuntary cooperation with Russian government had to assess if they had any vulnerable assets in Russia. For instance, Facebook ability to change its response and refuse to block any more groups in late 2014 episode of pro-Navalny rally was secured by company's retreat from Russia. Its development office was closed and the entire engineering team was invited (and accepted the offer) to move to Google offices in Europe and elsewhere. While reasons for this move were not disclosed, observers assumed company concerns with the potential access to Google code to the Russian government and even coercive methods to get this access through pressure on individual engineers (Bershidsky 2014).

[67] https://vk.com/wall1_45621.

What in this context happened with the online response? Did response offline and through the online infrastructure supplanted any serious engagement with users on behalf of the government? Yes, but it was not the only factor. Another key change was in the target audience of government online effort. If Medvedev was trying to build a coalition around values of modernization and reformist policy agenda, "Putin Redux" entailed rather dramatic change in regime ideology, not just compared to Medvedev years, but also compared to Putin's first two terms (Sakwa 2014). Reorientation towards conservative, even traditionalist values in domestic policy was paired with expansionist, revanchist foreign policy (Smyth and Soboleva 2014; Snyder 2014). This change had implications of truly historical scale, but one of less known consequences was reorientation of online propaganda machine from winning over neutral or even already oppositionally-inclined users towards protecting wider public, those receiving most of news news from TV but starting to use internet for entertainment or consumption, from the dangerous influence of the opposition voices. Typical example from the same episode with pro-Navalny rally in late 2014 was Youtube videos with prominent Navalny supporters, who were showing on air the web address of the supposedly pro-Navalny website with information about the rally. In reality they were fooled and the address was leading to page full of anti-Navalny videos. These videos and apparent endorsement of them by famous artists and journalist were then promoted on social media.

Such provocations obviously could not build a reputation that Medvedev, or his economic advisor Arkady Dvorkovich, or Perm governor in 2004 – 2012 Oleg Chirkunov were seeking to build online. Their goal is not to engage, agitate and invite for discussion; it is to disorganize, discourage and mute opposition. And this goal is much better served by filtering technologies and targeted prosecution of influential bloggers. Extensive online debates between oppositional politicians and pro-Putin "Nashi" youth movement, which occasionally happened before 2012, are no longer possible. With the gradual, but persistent political retreat of Medvedev's team, and government officials of liberal and pro-Western inclination in general (which in many cases includes leaving public service or even the country for good), the government presence online would not vanish. However the government officials and their speakers, both paid and volunteers, would speak with themselves and their most loyal supporters. Government would be online, but it would not be responding to anybody online, much less waiting for response from anyone.

Appendix C Classifying Twitter Accounts

For all accounts, the first order of business was to check if it features some meaningful content or it exclusively contains advertisement and/or gibberish. These accounts mostly feature various kinds of **spam** links (sales coupons, lottery, etc), pictures of consumer products and instant earnings ads. Sometimes, they switch to feeds of spam links in other languages.

If an account features meaningful content, coders first checked whether it is an *official* account for an organization. As institutional accounts are qualitatively different from accounts that individual people run (for example, they have larger content volume), we put them in a separate category. Both public and private account holders belong to this category. In our collection, these accounts usually belong to the news media. Note that we put only official accounts[68] in this category: if a bot features the same news feed as a news agency, it would be classified as a bot, not as an official account.

If an account belongs neither to spam nor to official institutional accounts, the goal was to classify it most reliably as human or bot. We proceeded by developing a comprehensive and empirically motivated classification scheme that would allow assigning unambiguously each account of interest to a category.

First, we checked how diverse is the account content. Some accounts feature one and only type of content, thus indicating bot activity. It could be just few hundred or millions of tweets, but all of them are identical in their form. For example, this could be easily identifiable **news headlines** ("President to visit China Tuesday"), posted one after another without tweets of any other kind. Alternatively, it could be only **retweets** from other Tweeter accounts, or just **pictures**, or just **videos**. In this case, the substantial content of the tweets does not matter: pictures could be from tennis events, videos from theater performances in Saint-Petersburg and news headlines about agriculture in the U. S. What matters here is that these are not tweets written by a human being, but posted by a robot, likely querying the web or feeding

[68] But not necessarily verified as such by Twitter.

content from a pre-prepared RSS feed. The only exception to this rule are retweets: they can feature different content, produced by both bots and humans. Here the decisive factor is the absence of content actually posted by the user in question, as well as consistency and volume of retweets: few humans maintain Tweeter accounts featuring exclusively retweets from others, but even those who do could rarely demonstrate the dedication to the task exhibited by bots.

News headlines are further distinguished into those **containing a link** to a news website (it could be both the real source of news, like a major newspaper, and makeshift website that simply republishes content taken from elsewhere) and those **without any links**. The last group that belongs to this category are accounts posting *text* that does not come from the news headlines, but clearly is not generated by the account holder either. For instance, feeds entirely consisting of famous quotes from historical figures belong to this category.

The most challenging classification task arises in cases when the account features **diverse** content. First group in this category are accounts that may contain retweets, pictures, videos and text, but clearly do not contain anything that is not available elsewhere on the web and thus is not personally produced by the account holder. These could be pictures from a news agency, alternating with retweets, alternating with links to news stories or unidentified quotes from other blogs and twitter feeds. If in doubt about the latter, coders were instructed to google a chunk of text from a randomly chosen tweet to check if it was available elsewhere on the Internet.

All of the account types we mentioned above usually exhibit a level of content consistency so high that one can spot bot activity even without paying close attention to the content of the tweets: multiple links that lead to the same news website; pictures of the identical size posted one after another and never featuring a person unfamiliar to the general public, etc. Naturally, the first encounter with human content comes when the consistent pattern is broken. For instance, take a Twitter feed consisting of many retweets featuring similar content, such as pictures of Ukrainian soldiers allegedly killing civilians in the most brutal way possible. But occasionally something different is popping up: a reply to another user, wishing her a nice morning, or lamenting about bad weather. Googling this content confirms that it could not be found, at least easily, elsewhere. If such kind of content is sparsely dispersed between many similar tweets clearly lifted from elsewhere, we are likely dealing with a bot manually maintained by (a team of) supervisors, who occasionally tweet by simply writing a real reply to another user (who could be both a real person and another bot). We classify this account as **cyborg**. We believe that this tactics is used to avoid Twitter spam filters as well as to increase the so-called account reputation (a computed characteristic of a Twitter account that might increase its visibility inside and outside the Twitter network).

Importantly, we clearly distinguish between trolls and cyborgs. Trolls are human beings dedicated (because of their persuasion, for money or out of fun) to spreading a particular kind of message online. We classify them, together with perfectly "normal" individual users into **personal accounts**. A different kind of human account is a feed featuring links to user's online presence elsewhere. For example, one might want to post links to her Instagram pictures or Facebook posts. We call such accounts **translations**. Finally, sometimes a small group of people maintain a community, for example, with local news or dedicated to a particular artist. It is a rare case on Twitter, but for such cases we keep the **community** category.

Therefore, the coding schema for our verification effort is as follows (terminal categories are in bold):

1. **Official institutional accounts**
2. Bots
    - (a) Single-content
        - i. **Retweets**
        - ii. **Pictures**
        - iii. **Videos**
        - iv. Text
            - A. News headlines
                - **Text with links**
                - **Text without links**

        B. **Other text**

    (b) **Diverse content**

3. **Cyborgs**

4. Humans

    (a) ***Translations***

    (b) **Personal accounts**

    (c) **Communities**

5. **Spam**