How Pro-Government "Trolls" Influence Online Conversations in Russia

Anton Sobolev (UCLA)*

The latest version of this paper is available here:

www.asobolev.com/files/Anton-Sobolev-Trolls.pdf

I explore the behavior and impact of several hundred "trolls" — paid supporters of Vladimir Putin's regime in Russia who were allegedly employed in late 2014 and early 2015 to leave pro-government comments on the popular social media platform *LiveJournal*. First, I devise a classification method of the possible objectives that would motivate governments to employ Internet trolls, the strategies trolls use to achieve these objectives, and these strategies' observable implications. Second, combining text analysis with modern approaches in causal inference, I develop a method to measure the natural evolution of online discussions so as to estimate the causal effect of troll interventions. Using a modified regression discontinuity approach and a set of partially testable assumptions about the timing of such interventions, I discover that Russian troll activity was more successful in diverting online discussions away from politically charged topics than in promoting a pro-government agenda. Moreover, while trolls succeeded in diverting discussions away from purely political topics, their interference apparently had no effect when the topic under discussion was the national economy. Those social media users who were discussing poor economic growth, unemployment, or price inflation seemed not to be responsive to troll interventions.

*Anton Sobolev: PhD Candidate, Department of Political Science, UCLA

1. Introduction

The problem of political control is one of the most important issues faced by authoritarian leaders, and social media have the unbridled potential to empower anti-regime movements. Using online blogs and forums, citizens can access information unavailable in state controlled newspapers or on TV, thereby learning more about the competence and popularity of the regime. They can also find like-minded individuals and coordinate amongst themselves on the time and place of protest activities. To combat such dangers, these governments introduce various forms of media control. These include exerting pressure on owners of social media platforms, banning websites, censoring content, and employing paid commentators to interfere with online conversations that espouse pro-government views and challenge the narrative of the political opposition.

Novel tools of collecting and analyzing textual data have allowed scholars of authoritarian regimes to look closely at how political control can be organized within social media. King, Pan and Roberts (2013, 2014) demonstrate that the Chinese government is more likely to censor posts related to citizens' coordination of protest activity than those criticizing the government. By creating accounts on numerous social media sites and randomly submitting different texts to these accounts, researchers demonstrate that even posts written in opposition to the ongoing protests have a good chance of be censored. Nevertheless, censorship, while a popular tool of oppression, is not the only option for political leaders: Munger et al. (2015) find that, in Venezuela, the loyalist members of the parliament actively tweeted non-political messages to shift the public agenda by reducing the share of dissidents tweeting about the impending protest events of 2014. Keller et al. (2017) report that during the South Korean 2012 presidential race, the National Intelligence Service actively used accounts on Twitter to wage a campaign in favor of the eventual winner, Park Geun-hye. Moreover, they identify three different groups of accounts that targeted specific social media audiences. Sanovich, Stukal and Tucker (2017) founded that around 60% of Twitter accounts tweeting about politics in Russia were merely automated software bots. Miller (2017) investigates how regional administrations in China used 'Big Data' systems to monitor "public opinion emergencies" as well as astroturfed to alter the public perception of the authorities. King, Pan and Roberts (2016) study pro-government commentators in the Chinese blogosphere and find that those bloggers spent time celebrating different aspects of Chinese social life while not necessarily engaging in a political debate.

In the last decade, modern information technologies have changed the world of politics as we knew it, erasing the clear distinction between domestic and foreign spheres. Today, policy battles and elections are fought not just through traditional lobbying, party activities, and TV ads, but by means of covert interventions by murky actors, who may be located anywhere and funded by almost anyone. These new behaviors are important for both democracies and non-democracies. Their impact is hard to assess. For instance, the debate continues as to whether or not hackers and Internet trolls affected voting in the 2016 U.S. election.[1] To date, researchers have focused on developing tools to identify paid online actors, their target groups, and the scale of their Internet presence. The researched described in this paper takes the next logical step, addressing the question whether or not users of social media pay attention to posts by paid agents. Can such agents successfully engage users with pro-government rhetoric? Can they divert them from criticizing political leaders? This paper is an attempt to shed some light on these questions within an observational setting using recently leaked information on what has been described as an attempt by the Russian government to create "an army of well-paid trolls" in order to "wreak havoc all around the Internet".[2]

In early 2015, journalists of the Russian independent newspaper *Novaya Gazeta* leaked the account names of 700 users that had allegedly been employed as paid "trolls". These trolls had published blog posts and participated in discussions on the popular Russian social media platform *LiveJournal* (*LJ*). As paid actors were trying to maximize their reach, i.e., the number of people who saw their posts, they had kept their accounts t, including their lists of friends and the communities to which they belonged, opened and had not deleted their posts and comments. Employing the leaked list of troll accounts, I collected two datasets: one containing almost a half a million troll posts and the other comprised of eighty thousand discussions infiltrated by these trolls.

[1] See recent reports from The Guardian (https://goo.gl/FkFKLp), CNN (https://goo.gl/Qgm6uX), and GQ Magazine (https://goo.gl/UQoGaK).

[2] For more information, see the famous piece in New York Times by Adrian Chen (https://goo.gl/HcHsts)

The major goal of the paper is to develop a method to test whether troll participation in an ongoing conversation on a social medium had changed the direction of the conversation. To identify the effect of trolls interventions on the direction of online conversations, one had to be able to trace the evolution of such discussions. To do so, I took the following steps . First, I identified troll comments within a discussion. Second, I pooled all non-troll comments into thirty-minute slices before and after the time of the first comment made by a troll. Third, for each thirty-minute slice, I employed Latent Dirichlet Allocation algorithm (Blei, Ng and Jordan, 2003) to estimate the mixture of topics. Finally, I determined the topic that dominated the discussion before the troll's intervention occurred. The propensity of a thirty-minute slice to cover this topic is used to trace the evolution of the discussion both before and after trolls intervene.

A simple before-and-after comparison to identify a change in the topic of conversations, however, might fail to identify the causal effect of the troll interference since the trolls might have chosen to enter only the conversations that were already trending in the desired direction. To remedy this problem, I focus on estimating whether or not an appearance of trolls in a discussion constituted a disruption in topics discussed by the non-troll users within a narrow time frame. To estimate the local effect of the trolls' intervention on the evolution of the online conversations, I fit a flexible model to comments appearing before the first troll intervention and the same flexible model to comments appearing after the troll intervention. This approach allows me to take into account each discussion's topical trend. Mechanically, this estimation is similar to a regression discontinuity, with the time of the appearance of the first troll comment acting as a cut-off. Put simply, I estimate the change in the prominence of topics while also taking into account the natural evolution of the discussion. A partially testable identification assumption suggests that, in a narrow time frame, the time of the appearance of the first troll comment could effectively be assumed to be random. Under this assumption, the set of comments appearing before a troll intervention constitutes a contrafactual allowing the local average treatment effect to be calculated.

Paid commentators can have different objectives. One of them is to promote the trend and skew the influence of the ranking of online newsfeeds by commenting and "*liking*" posts that are supportive of the government. Such behavior results in moving these posts to the top of the social media front page and thus increases their visibility. Another approach is to attack the participants of the conversations or authors of posts who criticize the regime. In this manner, they attempt to distract users from discussing anti-government topics, to promote a pro-government agenda, to stop the discussion itself, and to project the strength or the popularity of the incumbent. Still another approach is to imitate anti-government extremism to provide legal grounds for banning posts and accounts of anti-government activists. In this paper, I focus on the first two goals: the diversion of discussions from politically charged topics and the promotion of pro-government agenda.

My research yielded no evidence of the promotion effect, but did suggest large and statistically significant diversion effect. Upon checking for heterogeneity, I found that this latter effect to be driven not by discussions of Russia's then-current economic crisis in Russia, but only by the political discussions that primarily referencedPutin's political regime and his foreign policy. *LJ* users were found to be easily distracted if they were discussing political opinions about Russia's involvement in Ukraine's political crisis, but they paid little or no attention to troll comments while discussing the poor performance of the national economy, the volatile Ruble exchange rate, and rising prices. Thus, my findings indicated that economic grievances are much more resilient to governmental tactics of distraction than ideological opposition to the regime.

Several factors can undermine the validity of these results. *First*, the proposed approach can confuse the effect of troll interventions with that of new participants joining the discussion. To address this concern, I conducted a set of placebo-tests where randomly chosen participants of targeted online conversations were treated as trolls. *Second*, the author of a post could have deleted comments. While an owner of an account on *LJ* can try to selectively delete comments by trolls, the share of conversations in the data that contained any deleted comments was negligible (comprising approximately 3% of the data). These discussions were therefore not taken into account for hypotheses tests. *Third*, the leaked list of troll accounts could have been incomplete. In this case, some trolls could have been treated as ordinary users, and pooling their comments with the others could have generated a false positive effect. To deal with this issue, I assumed that the overall number of troll accounts was most likely negligible relative to the overall community of forty million users (with almost three million accounts in Russia). A large sample of Cyrillic *LJ* accounts was selected randomly and their owners treated as non-trolls. All posts published by these accounts were collected and then combined with posts published by accounts on the troll list, and a set of classification models was trained to predict whether a given account was likely to belong to a troll. Next, I randomly selected 650 non-troll participants of those conversations targeted by trolls, collected their posts, and applied the trained models to calculate their propensity to be *de-facto* trolls. A negligible number of ordinary participants in the targeted conversations exhibited a feasible propensity to be trolls, thus lending credibility to the claim that that the leaked list of troll accounts was exhaustive.

The research described in this paper attempted to make four contributions. *First*, it proposes a framework for analyzing political engaging in social media. Traditional studies of the political role of social media have tended to primarily concern devote the effects of political messages. However, the exposure to such messages was found to have a statistically significant but negligible effect (Bond et al., 2012; Jones et al., 2017). A potential explanation for this discrepancy is that social media users easily identify that these messages originate within someone's political campaign and discount their value. At the same time, political actors can target users in more sophisticated ways, including intensely engaging them through online conversations. This paper describes an approach to analyze political targeting that can occur through multiple mechanisms, including political socialization and learning. An important distinction of targeting through conversations is that paid commentators hide their pro-government affiliation from regular users, thus reducing the ability of users to attribute received messages to specific political forces. *Second*, this paper proposes a method for estimating the effect of troll interventions on politically charged online discussions. In contrast to standard matching techniques, this method allows the evolution of discussion to be controlled for and thus could prove helpful in alleviating selection bias in cases where trolls can choose to target a discussion after observing the direction of its movement. The proposed method can be combined with existing approaches in causal inference with text data (Egami et al., 2018). *Third*, this paper intends to add to the existing literature on the problem of authoritarian control. Previous studies (King, Pan and Roberts, 2013, 2014; Gunitsky, 2015; Munger et al., 2015; King, Pan and Roberts, 2016; Keller et al., 2017; Miller, 2017; Sanovich, Stukal and Tucker, 2017) have established that to deter political dissidents, authoritarian governments attempt to deter political dissidents by preventing online discussions by censoring or creating informational noise. This research establishes that a particular type of such interventions – the injection of paid pro-government commentators into online political conversations — might in fact be effective, but that the effectiveness of this technique is limited. *Fourth*, it investigates the difference in behavioral patterns of trolls and regular social media users and presents an algorithm to identify trolls by the observed online behavior.

The focus of this paper is limited in scope. While it analyzes the effects of trolls' interventions on the behavior of participants in social media conversations, it does not consider the potential effects of such interventions on the larger audience of readers of these conversations and on the overall social media agenda.

The remainder of the paper is organized as follows. The next section considers political astroturfing in the context of the strategies employed in information control and hypothesizes as to the possible tactics and goals that the governments intend to achieve by using paid social media commentators. The third section provides the background information for the study, evaluating the role of online activism in Russia and government attempts to limit it. Section four describes the data collection methods and the measurements employed in the study. The fifth section describes the research design and states key identification assumptions. The sixth section presents the study's results. The seventh section addresses threats to result validity. The final section draws conclusions and discusses limitations of the study.

2. Political astroturfing as a tool of information control

2.1. *Political effects of social media and authoritarian response*

Scholars see political astroturfing or the masked engagement in political conversations, as a tool for information control by authoritarian regimes. The role of social media in political discussion has because indispensable because their use has dramatically reduced the costs of communication and helped citizens who support opposition to such regimes in two key ways (see, fig. 1). First, enhanced informational exchange helps citizens learn about the successes and failures of public policies and so evaluate government competence. It also helps citizens to obtain more accurate information about overall public satisfaction with the regime. Third, social media provide improved dissident coordination. By discussing the failures of government policies, civic activists can develop a political agenda or choose a leader who can efficiently compete with the current incumbent. Finally, social media simplify the organization of protestors' collective action (Tufekci and Wilson, 2012). While most observers agree that protesters actively employ social media for political purposes, establishing a causal effect of social media on protest participation has been difficult. Nevertheless, using an instrumental variable approach, Enikolopov, Makarin and Petrova (2015) demonstrate that the increase in social media penetration across Russia's cities significantly increased both the probability of a protest and the number of protesters during the 2011-12 electoral protests.

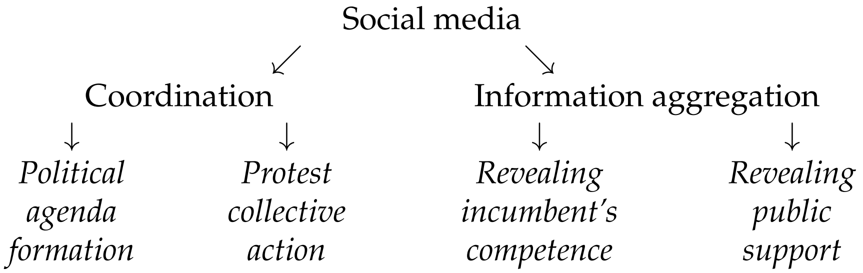

Social media
Coordination
Information aggregation
Political agenda formation
Protest collective action
Revealing incumbent's competence
Revealing public support

Figure 1: Political effects of social media

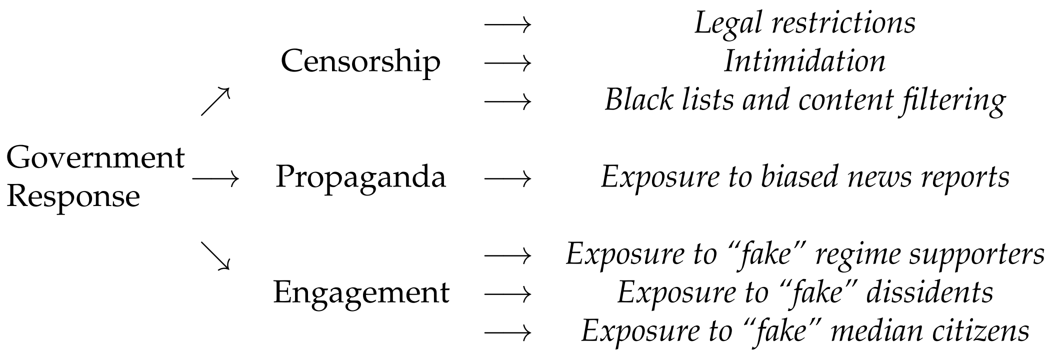

Government Response
Censorship
Propaganda
Engagement
Legal restrictions
Intimidation
Black lists and content filtering
Exposure to biased news reports
Exposure to "fake" regime supporters
Exposure to "fake" dissidents
Exposure to "fake" median citizens

Figure 2: Government responses to political threats of social media

Incumbents in Russia, China, and other authoritarian regimes can employ three options to mitigate the political consequences of social media development: censorship, propaganda, and engagement (see Figure 2). *A traditional tool of information control, the goal* *of censorship is to restrict flow of information, and governments can achieve this by employing different means. The traditional tools whereby censorship is enacted include legal restrictions on traditional media / social media platforms (including banning foreign / private ownership) and prosecution and intimidation of journalists, activists, and regular users.* Online tools of censorship consist primarily of including the websites into "black lists" while blocking user access to all members of such lists, and content filtering (a set of restrictions to prevent web-aggregators and search engine services from indexing information contained in blocked web sites). Content filtering often takes into account the existence of "blacklists of topics", politically sensitive topics about which news-aggregators and online media are not allowed to publish news. Some scholars believe that propaganda and engagement represent the same phenomenon, and paid commentators can be used for both engagement and propaganda purposes. However, there is an important distinction between them. Propaganda sources do not hide their affiliation with the state or the incumbent political party whereas paid commentators typically pose as regular social media users. A famous example of a contemporary propaganda channel is Russia Today (RT), a state-owned company that broadcasts Russian propaganda abroad. In order to deflect attention from its editorial policy, which espouses specific political lines, Editor-in-Chief Margarita Simonyan has declared all media outlets to, in fact, be channels of propaganda. [3] The goal of *propaganda* is to maximize the exposure of citizens to biased news reports so as to prevent political learning. In case of *political astroturfing*, commentators are employed to engage with political activists and regular social media users. *Political astroturfing* is probably the most flexible means of information control. Paid commentators can pretend to be people having differing political views and goals from those of extreme pro-government supporters to "undecided citizens" to extreme dissidents who see terrorism as an acceptable means of political struggle. In contrast to censorship and propaganda, political astroturfing allows targeting of specific groups and chooses tactics to maximize the probability of successfully achieving the goal. In the following sections, I discuss the goals of hiring paid commentators, targets of their engagement in online conversations, potential communication styles, and tactics.

[3] See: Simonyan's Interview to NBC news

2.2. *Political astroturfing: a classification of goals, targets, and tactics*

Most of the scholars consider the promotion of pro-government political agenda to be a major goal of political engagement (see Gunitsky, 2015; Sanovich, Stukal and Tucker, 2017 for a review). As Sanovich, Stukal and Tucker (2017) write: *"establishing a government presence on the web and using it to promote the government's agenda constitutes ... the final option at government's disposal.* "An exhaustive literature review has shown King, Pan and Roberts (2016) to be the only study in comparative politics that explicitly considers other potential goals of paid commentators, including criticism and cheerleading. I build the on results from Gunitsky (2015), King, Pan and Roberts (2016), Sanovich, Stukal and Tucker (2017) to develop a classification of trolls' goals, targets, and tactics.

Different levels. Paid commentators can try to target macro- and micro-level goals. At the macro-level, trolls can try to shape the overall public narrative by affecting news trends and tops of newsfeed. They achieve this approach through massive reposts, comments, and "likes" of the post having the desired content. At the micro-level, trolls can target two separate groups of users: the authors of posts and the participants of social media conversations. Under constant attack, the former can either stop posting to their blog, or change the content of their posts.

Different goals. In this paper, I focus on the micro-level goals of paid commentators. More specifically, I consider the potential effects of troll interventions on participants in conversations, not on the authors of posts. Trolls can attempt to achieve five goals by engaging with participants of social media conversations: to project strength, to project popularity, to imitate anti-government extremism, to promote pro-government agenda, and to distract opposition activists. Examples, as well as, the details of the corresponding tactics, are provided in Table 1.

In this paper, I focus on the last two goals: promotion of a pro-government agenda and distraction of opposition activists. *Promotion* and *diversion* are different. *Promotion* implies that, regardless of the initial topics of conversation, the trolls engage platform users in a discussion of a pro-government topic (for example, increases in international respect for the Russian army, the assertiveness of Russia's foreign policy, or how divided and weak is the political opposition to President Vladimir Putin). Measuring a promotion

Goal	Hypothetical Example	Tactics	
Project Strength	*Users:* “Should suit Putin for bribes in public procurement?” *Troll 1:* “We see all of you and know where you are.”	communication style	*show that you are the army of paid trolls, show the capacity of the state to monitor and locate online activities of dissidents*
(Magaloni and Wallace 2008, Roberts and Stewart, 2015)	*Troll 2:* “If you try to find another place to discuss, we will locate you in seconds.”	target group	*civic activists*
		desired reaction of target group	*textual response is not important*
Project Popularity	*Troll 1:* “Putin is the only hope”	communication style	*pretend to be ordinary users / regime supporters*
(Egorov and Sonin, 2014)	*Troll 2:* “Who if not Putin?”	target group	*civic activists, top bloggers*
	Troll 3: “86% of Russians support Putin?” *Users:* “...”	desired reaction of target group	*textual response is not important / conversation stops*
Imitate Anti-government Extremism	*Users:* “We need to protest against corruption” *Troll 1:* “Let’s bribe the military and assault the Kremlin”	communication style	*pretend to be dissidents, conduct illegal online actions, like extremist messages*
	Troll 2: “Let’s also kill [an ethnic minority group], because all of us hate them ”	target group	*civic activists*
		desired reaction of target group	*textual response is not important; ban the blog, prosecute the author and participants*
Promote Pro-government Agenda	Troll: “Russia’s reunion with Crimea is good”	communication style	*pretend to be ordinary users, polite*
(Geddes and Zaller, 1989, Roberts and Stewart 2015)	User: “Well... yeah”	target group	*ordinary users*
		desired reaction of target group	*engage in discussion of pro-government topic*
Distract Opposition Activists	“‘*Users:* “We cannot buy olives because of the Crimea sanctions!”	communication style	*pretend to be ordinary users, polite*
	Troll: “But in the next couple of years, there will be massive olive production in Crimea”	target group	*civic activists, top bloggers*
	Users: “Is it that fast?”	desired reaction of target group	*redirect discussion from anti-government topics*

Table 1: Goals for using paid pro-government commentators with respect to participants of online conversations.

effect involves looking at how prominently a pro-government topic would emerge after trolls appear in a conversation. *Diversion* is a different activity. Even if trolls are unable to shift the topic of the conversation into something beneficial for the government, they might be able to shift people's attention away from criticizing the government. Thus, the diversion effect shows itself as a decrease in the prominence of some critical topic after the appearance of one or more trolls in the discussion. One of the popular tactics of diversion is *whataboutism*: if people in a conversation criticize Russia's government (for example, for supporting rebels in Eastern Ukraine), trolls would appear and ask, "What about the US...?" (for example, "What about US interference in the domestic affairs of other countries?"). The topic then naturally shifts away from discussion of the Russian government. Below, I provide some examples of *diversion* and *promotion* at work.

- *Diversion:*
  - A conversation about corruption in the Russian government shifts toward a discussion of the IQ-levels of the participants in the current talk.
- *Promotion:*
  - A conversation about corruption in the Russian government shifts toward a discussion about corruption in the opposition
  - A conversation about Russia's alleged support for the insurgencies in Eastern Ukraine shifts toward a discussion of the legitimacy of US military involvement in Middle Eastern countries.

Distinguishing between diversion and promotion implies the two major hypotheses of this study:

Diversion hypothesis: *the propensity of an online conversation to cover an anti-government topic decreases after trolls intervene.*

Promotion hypothesis: *the propensity of an online conversation to cover a topic that benefits pro-government propaganda increases after trolls intervene.*

For testing these hypotheses, the population of interest would be all comments in political conversations that are critical of the government and that are parts of discussions infiltrated by pro-government trolls. The *Diversion Hypothesis* implies that the commentators who participate in the conversation right after the appearance of trolls are less likely to follow the initial topic (i.e., the one critical towards the government) and are more likely to follow some other topic. Thus, the appearance of pro-government trolls creates a discontinuity in topic structure. The *Promotion Hypothesis* implies that a topic to which the conversation is diverted by trolls is more likely to be among the topics that we designate as favoring pro-government discourse.

To obtain insights in these hypotheses, one can look at a particular political conversation on LJ. On August 7, 2014, Orthodox cleric Deacon Andrei Kuraev, who is an author of several books on Orthodox Christianity, published a short post titled "Fasting Will Be Less Pleasant" in which he mildly criticized the Russian government for imposing a ban on almost all food products produced in the European Union. His particular concern was olives, which, according to him, provide Orthodox Russians with enjoyment in the austere time of the Great Lent preceding Easter. He finished the post by stating that he intended to buy sufficient olives to last through this time while they were still available. His post sparked a lively discussion about the impact of food ban on the diets of Russian churchgoers that continued until a user *glycmamroga* joined the conversation. *Glycmamroga*, a user-account that had appeared on the *Novaya Gazeta* troll list, argued that the olive problem would be solved in a couple of years because Crimea (annexed from the Ukraine) provides a perfect place to grow olives. If this troll's intervention did serve to influence the topic of this conversation, after–troll-comments by regular users would be expected to respond positively to the troll's comment. The *diversion* mechanism implies that such comments would shift from the discussion of the negative effects of sanctions toward less sensitive topics (such as the general problem of olive cultivation). The *promotion* mechanism would imply that after-troll-comments would would shift the conversation toward discussion of positive aspects of the Crimea annexation.

3. Background: Political regime, social media and information control in Russia

3.1. *Russia's political regime, civic activism and social media*

This paper explores the strategies of an authoritarian government to influence online conversation in a specific context: an alleged attempt by the Russian government to employ paid commentators to inject themselves into discussions on the popular social media platform LJ. This section discusses this case in more detail. Vladimir Putins' political regime in Russia is categorized as a personalist autocracy (Geddes, Wright and Frantz, 2014). In 2014, the experts of the Polity IV project gave Russia a score of 4, placing Russia into the same category as Venezuela, Zimbabwe, Nigeria, and the Ukraine (Marshall and Jaggers, 2002). Freedom House puts Russia into the '*Not Free*' category. Russia's civil society has been traditionally perceived as weak and disorganized. It is commonly believed that communist rule as well as centuries-long monarchy have hampered the formation of social trust and in Russia. This, in turn, has caused Russian politicians and especially those in the the executive branch of government to be unaccountable to civic groups while the political opposition remains unstructured and weak. In addition to the country's history, scholars find the Putin regime's policies designed to curb international funding and suppress independent activists responsible for the lack of a strong civil society in Russia (McFaul and Treyger 2004). Russia's geography with enormous but sparsely populated territories also constitutes a challenge for forming nationwide groups of any kind (Turovsky, 2005). In addition, state attempts to control the media are also viewed as preventing citizens from converting private grievances into public ones (Pfaff and Kim 2003, Oates 2006, Mickiewicz 2008, Greene 2014).

This situation has changed after 2010. With a broad introduction of cellular network, the Internet, and especially social media ordinary citizens significantly increased their capacity for social coordination. In 2016, more than three-quarters of Russian households had a computer, and almost 70% of the population was logging on to the Internet at least once a month. As of 2013, social media had attracted 35 millions of Russian Internet users (Treisman, 2018).

Armed with these new tools of social coordination, dissidents challenged the leadership of Vladimir Putin in 2011 and early 2012 with an online-coordinated protest movement. Several hundred thousand people took to the streets in major cities to express their dissatisfaction with alleged manipulation of the parliamentary elections. The government responded by offering some policy concessions to the pro-democracy movement but also stepped up repressions by arresting some protesters and passing laws that increased the punishment for unsanctioned protest activity. According to many observers, social media played an important role in the protest mobilization. Activists, including the future leader of the Russian political opposition, actively encouraged citizens to take to the street via their online blogs.Smyth, Sobolev and Soboleva (2013) point out that belonging to ”at least one online network” was one of the strongest predictors of individual participation in protests. Using a plausibly exogenous variation with respect to penetration of the major online social network, VK.com, Enikolopov, Makarin and Petrova (2015) found that social media penetration increased both the probability of protest onset and the size of the protest in Russia. In line with these results, Bodrunova and Litvinenko (2013) show that social media played not only the organizational but also a ‘cultivational' role in fomenting protests by mediating the public discourse that emerged during the electoral campaign. Kolsova and Shcherbak (2014) established a statistical relationship between the increase in the weekly pre-election ratings of the opposition parties and the intensity of political activity in the blogosphere.

3.2. *State response to social media activism*

Because the effect of social media on the political and economic life of Russia has the potential to be nontrivial, the regime has attempted to employ strategies that would interfere with citizens' co-ordination and dissemination of knowledge through social media. Indeed, there is substantial evidence that such interference exists.

At least since 2008, the Russian government has been trying to identify and target opposition activists online. Borogan and Soldatov (2017) suggest that the youth league *Nashi* was created by deputy head of presidential administration Vladislav Surkov as part of a campaign to prevent the "Orange revolution” in Russia. In 2013, investigative reporters of independent outlet *Novaya Gazeta* found evidence that *Nashi* had been hiring people to comment on social networks. (Specifically, the article reported that employees of that project were required to write around 100 comments per day.) While the government never confirmed these allegations, they were later corroborated by leaked email exchanges between operatives of this pro-regime movement and their contacts in the presidential administration. Most importantly for this project, in March 2015, *Novaya Gazeta* published a list of account names of people who had been tasked with leaving comments on the blog platform *LiveJournal*.[4] A follow-up investigation by the *New York Times* showed the existence of a huge industry of paid commentators in Russia and indicated that Russian trolls may not only be engaged in fighting political opposition in Russia but also may be organizing sabotage against other countries. Among the most popular ones was promotion of fake news about a serious explosion at a processing plant in Louisiana.[5]

The fact that paid commentators appear on *LiveJournal* is not surprising. *LJ* is one of the most popular blogging platforms in Russia, leading in both content production and number of discussions concerning current affairs in 2010 (Etling et al., 2010). Historically, *LJ* has been the most commonly used social media platform of dissidents of the regime. The website has around 40 million registered users with 50% of its traffic generated by Russian users. Although its popularity has been declining since 2014, it is still one of the most popular websites in Russia, ranked 15 by the web traffic aggregator Alexa.com. Originally developed and maintained by US programmer Brad Fitzpatrick, *LiveJournal* is now owned by the Russian company SUP Fabric, which is controlled by Alexander Mamut and Alisher Usmanov, both entrepreneurs with ties to the Kremlin.

4. Data

4.1. Data collection

Immediately following the publication of the list of 700 paid commentators on *LiveJournal*, I identified the links to the comments attributable to each of these accounts. At that time, the Russian search engine *Yandex* allowed comments to be searched by user-ID for any

[4] See: http://www.novayagazeta.ru/inquests/67574.html

[5] See: http://www.nytimes.com/2015/06/07/magazine/the-agency.html

Author	**Time**	**Post Title**
Post		
User 1	Time	Comment
User 2	Time	Comment
Troll 1	Time	Comment
User 2	Time	Comment
Troll 2	Time	Comment
User 1	Time	Comment
User 3	Time	Comment

Table 2: Structure of Collected Data on Posts

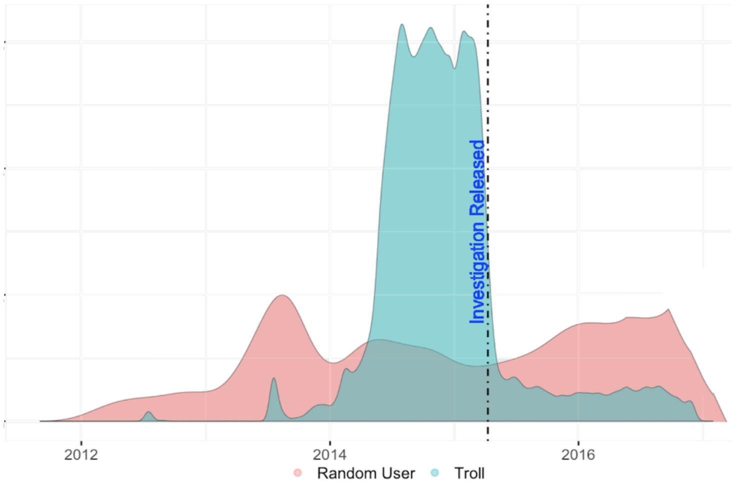


Figure 3: Time of Postings by Trolls and Random LiveJournal Users

social media, including *LiveJournal*. Its search range was limited to the last 1000 comments made by a user, and thus only a fraction of the posts in which these paid trolls intervened was accessible. After collecting the set of comments made by these trolls, I identified posts that appeared to have been under attack by trolls. and then collected all posts that involved at least one comment by a troll along with all comments relevant to those posts, yielding a corpus of around 180,000 posts and seven million associated comments .

For each post the following features are available: *text of the post, date, day, and time of posting*, *author's name* and his *suggestive type* (*troll* or *non-troll*). The same features are available for comments to posts. I treat all comments to a particular post as an online conversation.

It is worth mentioning two things. First, the very next day after the list was released, most of the accounts on the list stopped any activity (see Figure 1). Second, *Yandex* suspended its comment search functionality shortly thereafter.

The collected data consist of posts and discussions from 2014 and early 2015. In Russia, this was a period of political conflict with Ukraine, economic stagnation, declining oil prices, rising food and consumer goods prices, and intensive government propaganda. Most importantly for mass economic expectations, Russia's currency – the ruble – was depreciated by half, contributing further to rising prices and imposing a severe financial strain on people whose mortgages and consumer loans were denominated in US dollars.

Figure 3 represents the timeline of the troll interventions captured in the data I collected. Because the scraping procedure was limited to the last thousand comments made by each troll, no conclusion can be made about the dynamics of these interventions. However, it is easy to see a set of spikes on the timeline – days or even hours when more than 30-40 posts were attacked at the same time.

4.2. Post classification and processing of conversations

4.2.1 Automated data classification with Latent Dirichlet Allocation. Several parts of this study rely on automatic text classification using latent Dirichlet allocation (LDA), a generative statistical model that allows sets of texts to be described by their propensity to clusters (topics) (Blei, Ng and Jordan, 2003). LDA assumes that each text is composed of a mixture of topics and that the intensity of usage of specific words reflects the propensity of the text to cover a specific topic.

For example, imagine that all online conversations discuss the recent Russia-Ukraine conflict (specifically, the problem of control over the Crimea) and consist of only two terms: "Reunion" and "Annexation". Conversations that mainly consist of the word "Reunion" are probably organized by the supporters of President Putin, whereas those that primarily use the word "Annexation" are initiated by the Russian dissidents. Figure 4 depicts this example.

First, the LDA algorithm tries to identify clusters within these texts. Conversations that mostly use the word "Reunion" are classified as pro-government ones. Opposing conversations are classified as "anti-government" ones. To identify the propensity of a particular conversation to cover a specific topic, LDA algorithm conducts two steps. After first calculating the central values of the two clusters, it then calculates the relative distances of a specific conversation from the center of each cluster. These relative distances

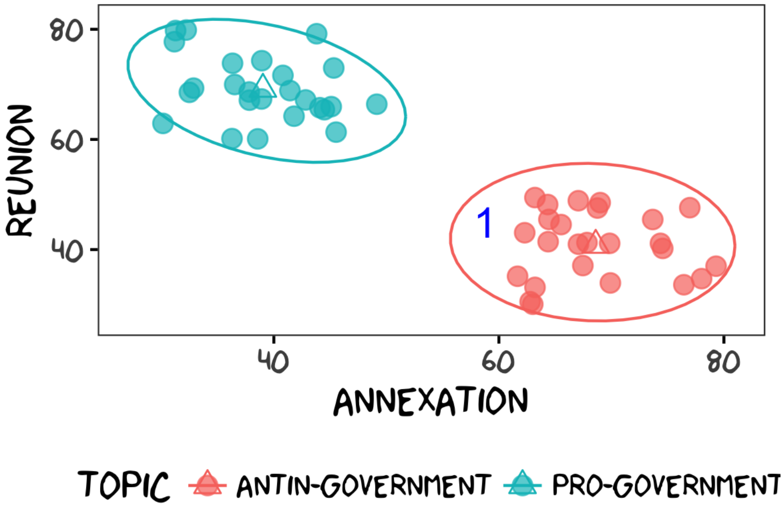


Figure 4: A hypothetical example of LDA classification

represent the propensities of a given conversation to each of the two topics. An important property of the LDA algorithm is that it requires the researcher to pre-specify the number of topics to estimate.

4.2.2 Processing online conversations. This section describes the processing method I employed for the collected online conversations. First, I analyzed the posts that provoked the online discussions and attracted the attention of the trolls. Almost 45 percent of the these posts were written by the trolls themselves. Another 7 percent were generated by the automatic media outlets' robots, which basically post links to the media outlet. Thus, around 80 thousand posts were written by non-troll users. I applied the LDA model to classify these posts by estimating a mixture of ten topics for each of the posts. Number of topics selected ranging from 8 to 10 did not change any results. Increasing in the number of topics to more than 10 returns produced duplication of topics. The choice of seven topics or fewer returned topics consisting primarily of various sparse terms. Next, the dominant topics (i.e., those with the highest propensities) were identified for each post . Eight out of ten estimated topics referenced non-political content, and the other two described the economic crisis in Russia as well as Russia's recent conflicts with Ukraine, Europe, and the United States (around eight thousand and twelve thousand posts, respectively).

I analyzed the conversations that were provoked by each of these posts and identified the time of the first troll comment for each of the twenty thousand posts. I then removed

	Author	Time		Post Title
	Post			
	User 1	Time	Comment	Pooled Comments $t = -2$
	User 2	Time	Comment	
	User 1	Time	Comment	Pooled Comments $t = -1$
	User 2	Time	Comment	
drop ←	Troll 1	Time	Comment	$t = 0$
	User 2	Time	Comment	Pooled Comments $t = +1$
drop ←	Troll 2	Time	Comment	
	User 1	Time	Comment	
	User 3	Time	Comment	

Figure 5: Processing online conversations

all troll comments from the conversation and pooled the rest of the comments into 30-minute slices centered on the time of the first troll comment. iThus, for each post, all comments occurring within 30 minutes after the first troll comments were combined to form a new text. This operation was repeated for all comments in the five-hour range following the first troll comment (an average *LJ* talk continues for 17-22 hours). Thus, for most of the conversations, I generate twenty slices (ten before and ten after the troll intervention). Figure 5 provides an example of the implementation of this algorithm with ordered 30-minute slices of conversations as units in this analysis.

4.3. *Measurement*

4.3.1 How to track evolution of online conversations. In this section, I develop a simple approach to estimate the evolution of an online conversation. The underlying idea is simple and straightforward: estimating changes in a conversation's mixture of topics in each of the subsequent time slices permits the evolution of the conversation to be traced.[6]

Recall the updated example from the previous section (see Figure 6). Each observation represents a thirty-minute slice of a conversation. The distance of this conversation from the "centers" of the anti-government and pro-government topics would change if

[6] An alternative approach suggests using a *Dynamic LDA:* a method that establishes initial distribution of topics in the first time slice of each conversation and to track their evolution in subsequent slices. While being a reasonable alternative to my method, *Dynamic LDA* suffers from a specific problem: if a topic emerges at the late stages of conversations, the method has a risk of not catching the topic of interest at all by assign important words to pre-existing topics. Thus, while *Dynamic LDA* can be to perform well in testing *diversion* hypothesis, the researcher can fail to use it for "promotion hypothesis" tests. Later, the results of *Dynamic LDA* will be reported in the Appendix.

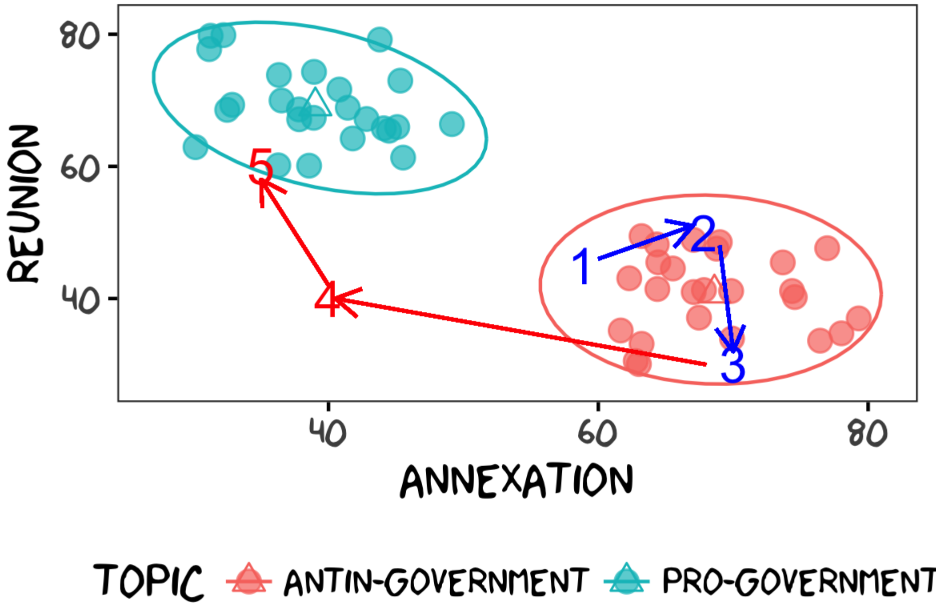


Figure 6: An example of tracing the evolution of a conversation that consists of two words: “Reunion” and “Annexation”.

participants were to begin using the terms “Reunion” and “Annexation” more or less frequently in the following slice, respectively. Since conversations consist of multiple words, in my actual analysis “centers” of topics are defined in multi-dimensional space with each dimension representing the frequency of a specific word in the slice.

4.3.2 Outcomes of interest. In the constructed dataset, I employed LDA to estimate a mixture of topics and their corresponding propensities (separately, for political and economic conversation) for every time slice for each conversation. First, for each conversation, I identify a topic that was dominant before one or more trolls joined the conversation (separately for political and economic conversations). I used the estimated propensity of a conversation’s slice to cover this topic to test the *diversion* hypothesis. Noteworthy is that all topics that were dominant before a troll intervention appeared to be anti-government (see the first row of Table 3). Next, I estimated the propensity of each time slice to cover the appropriate anti-government topic. Interesting to note is that both an anti-government and a pro-government topic constitute from sixty-five to eighty-five percent within the topic mixture.

	Economics	**Politics**
Anti-Government topic	“ruble” + “price” +”oil”+ “USD” + “exchange rate” + “Economics” + “crisis” + “Putin”	“war” + “Ukraine” + “military” + “Donbas” + “Donetzk” + “Boeing”
Pro-government topic	“good” + “salary” + “employed” + “better” + “income” + “can afford”	“Ukraine” + “USA” + “plot” + “Crimea” + “great” + “peace”

Table 3: Anti-government and pro-government topics in online conversations. *Note:* other topics are dropped from this table. The full table will be presented in the Appendix to this paper.

Two dependent variables are used in my analysis:

- *Propensity of a slice of a conversation to cover the anti-government topic,*
- *Propensity of a slice of a conversation to cover a pro-government topics.*

5. Research design and identification strategy

The focus of my research was assessing whether the appearance of one or more trolls in a discussion constituted a disruption in the topics being discussed by non-troll users. To estimate the local effect of troll interventions on online conversations, I fit a flexible model to the data representing the conversation before the appearance of the first troll in that conversation, and then I fit the same flexible model to the data representing the conversation after the troll intervention. This approach allowed me to take into account the existing topical trend of each discussion. Mechanically, this estimation is similar to the regression discontinuity, where the time of the appearance of the first troll is treated as a cut-off and the order of a slice of the conversation is used as a forcing variable. I calculated standard errors for clusters on the conversation-level.

My estimand of interest was the local average treatment effect, i.e. an immediate change in the evolution of an anti-government topic after a troll joins the conversation. A key assumption allowing this identification is that, within a narrow time frame, the time at which trolls begin to intervene in an online conversation is effectively random, as assumption with some evidentiary support. For example, no systematic patterns are evident in the timing of the troll attacks. The relative order of the first troll comment is

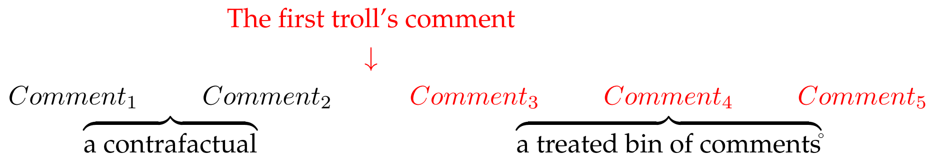


Figure 7: A plausible contrafactual for "treated" slices under the *narrow time frame* assumption

almost uniformly distributed across the timespan of the conversation. Moreover, this time apparently did not depend on the initial topic, the number of pre-existing comments or participants, or the previous course of the conversation. If this assumption holds, within a narrow time frame, the set of comments appearing before the troll intervention constituted a contrafactual (see 7).

Although the proposed identification assumption may not be applicable to online conversations in general, given the specific operational conditions of this Troll Factory, it is most likely valid? valid. These conditions possibly include the following: *LJ* trolls are required to post numerous comments on numerous posts per day; they are required to attack posts including a specific type of content; they need to manually read a large number of post abstracts via the *LJ* search engine in order to identify appropriate posts to target; and they have fixed working shifts. The documents leaked concerning "these particular trolls" suggest that all these conditions were met.

6. Results[7]

Figures 8 and 9 depict the main results of the regression discontinuity analysis. As can be seen, trolls appear to have been more successful in diverting discussions from politically charged topics than in promoting a pro-government agenda. When a discussion considered politics, troll intervention reduced the propensity of the conversation to cover an anti-government topic by fifteen percentage points. As shown in the figures, the intervention also switched the trend of the conversation from positive to flat and stable throughout the conversation. The effect of an intervention in promoting a pro-government agenda appears to be statistically significant but negligible. Troll interventions increase the propensity of a conversation to cover a pro-government topic by about one percent. Trolls were

[7] Supplementary materials are available in the online appendix of this paper

successful in diverting discussions from purely political topics but had no effect on discussions on the national economy. Discussions on poor economic growth, unemployment, and/or price inflation seemed not to have been responsive to troll interventions.

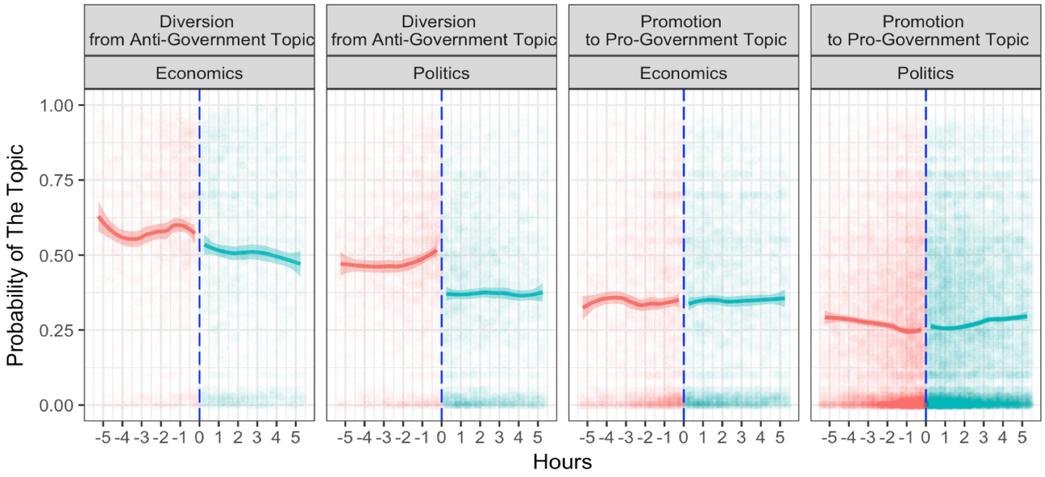
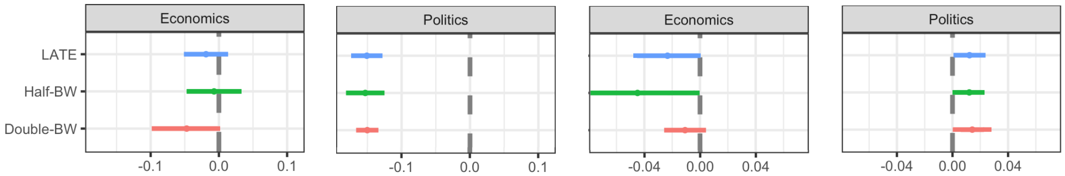


Figure 8: Troll interventions in online conversations



(a) Diversion from economic discussions

(b) Diversion from political discussions

(c) Promotion in economic discussions

(d) Promotion in political discussions

Figure 9: Effects of trolls' interventions on online conversations

7. Robustness and threats to validity

7.1. *Effect of a random user*

A part of the estimated effect of the entry of a troll, a new poster, into a conversation was due to the fact that new participants introduce their own lexicon into the conversation, which evokes a response from other participants. To account for that effect, I replicated the analysis for all comments posted before the first troll comment. Next, I assigned troll status to a random non-troll participant and then analyzed the effect of this poster on the propensity of a conversation to cover an anti-government or a pro-government topic. On average, the resulting analysis suggested that the entry of a new, non-troll participant into a conversation does not affect its evolution .

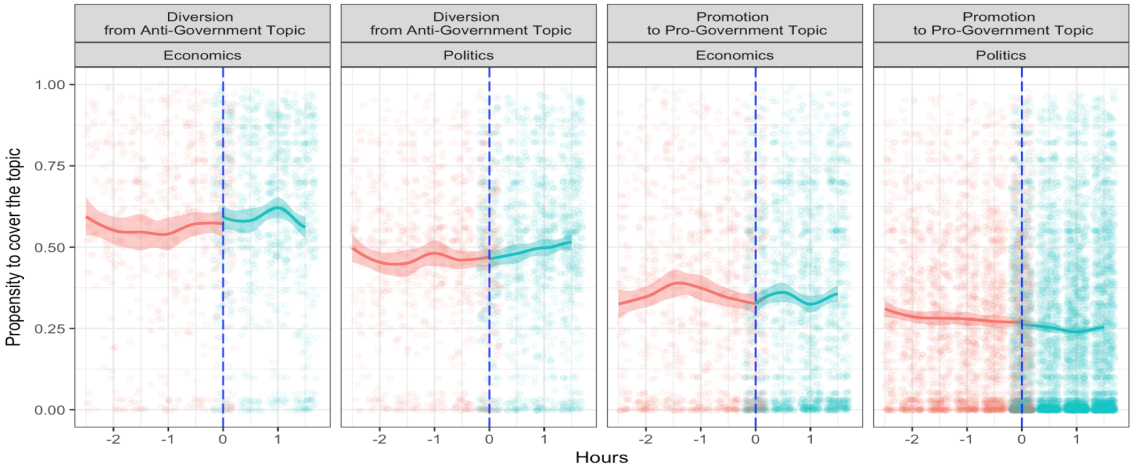


Figure 10: A random user's intervention effect on the evolution of conversations targeted by trolls

7.2. The problem of unobserved trolls

The list of troll accounts found in the leaked documents could have been incomplete, meaning that actual trolls, those who did not appear in the *Novaya Gazeta* list, were treated as non-troll participants in the analysis and that their comments could therefore have been used to measure the propensity of different conversation parts to include anti-government or pro-government topics. This fact could have generated systematic measurement error and so biased the study results . I relied on three strands of evidence to address this problem.

7.2.1 Evidence from journalist investigations. Media investigations suggest that the published list was exhaustive. For example, Lyudmila Savchuk, a former troll who helped to leak the documents to the press, pointed out that these trolls were organized into groups and worked in twelve-hour shifts with every other day as a day off.[8] The leaked list of trolls was divided into four shifts named for the shift's supervisor. [9] If, each shift worked for twelve hours every other day, the activities of all four shifts fully covered each hour of the week with no overlap. Table 4 displays a possible working schedule for the troll shifts. If a group worked on Monday, in the next week it would work on Tuesday. In Russian companies, this schedule is typical for employees who work in twelve-hour shifts. The post data reflects this pattern. On average, the trolls on the the list published

[8] Read Ludmila Savchuk's interview to Sguchenka.Com

[9] See, the list of trolls at NovayaGazeta.ru

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Group 1	Group 3	Group 1	Group 3	Group 1	Group 3	Group 1
Group 2	Group 4	Group 2	Group 4	Group 2	Group 4	Group 2

Table 4: Potential working schedule of troll shifts

approximately the same number of comments during each day and night of the week.

7.2.2 Randomness of unobserved interventions hypothesis. Another possible threat to study validity is that the journalist's account could have been incorrect, meaning that unlisted trolls could have been active on *LJ* at the time covered by the data. However, there is no reason to believe that these unidentified trolls should have commented only *after* the first comment of a troll whose account was included in the list. If no systematic difference between the known and the unknown trolls' accounts can be observed, the comments of the latter should have approximately the same likelihood to appear *before* as well as *after* the first comment written by the known troll. In this case, the resulting estimated local average treatment effect should remain unbiased. However, no tools exist to verify whether possibly unidentified trolls followed a different logic when determining the point at which to join a conversation. For this reason, I developed a third way to address possible implications of the incomplete list problem.

7.2.3 Identification of similarities in behavior of trolls and other participants in targeted conversations. In this section, I discuss similarities in the behavioral patterns of trolls and non-troll participants in the conversations they target. I conducted this comparison in four steps. First, I randomly sampled *LJ* accounts, collected their associated posts, and combined them with posts written by known trolls. Second, I extracted a set of features from these posts. Third, based on the extracted features, I trained a set of classification models to distinguish between the randomly sampled *LJ* accounts and accounts belonging to the leaked list of trolls. Finally, I sampled a group of participants in the targeted conversations who did not appear on the troll list. After collecting their posts and performing the feature extraction, I applied the most accurate trained model to predict the propensity of the sample participants of the targeted the conversations to be trolls.

Step 1: sample random LJ accounts. I initially assumed that the true share of troll accounts in the total population of *LJ* accounts was negligible, an assumption corroborated by the fact that most investigations report the total number of troll accounts on a specific platform (e.g., *Facebook*, *Twitter*, *VK.com*, or *LJ*) as limited to several hundreds.[10] At the same time, *LJ* has around 40 million registered users with 50 percent of its traffic generated by Russian users. In order to conduct my comparison, I assumed that an account randomly drawn from the population of all Cyrillic *LJ* accounts did not belong to a troll.

Under this assumption, I randomly sampled 900 Cyrillic *LJ* accounts and, for each of these , I collected all posts written by their owners from early 2014 to early 2015 (a period when 96 percent of troll posts were written). I ended up with close to fifty thousand posts, which I combined with the contents of a dataset containing troll posts made over the same period (more than 380,000 posts in total).

Step 2: extract features from posts. I extracted features from the posts described above as follows. First, on the corpus of the texts of posts, I fit a topic model with LDA. Second, I calculated the *term frequency–inverse document frequency* (TF-IDF), a numerical statistic that reflects the importance of a particular word to a specific post. As the vocabulary of words used in all of the posts was extremely large and the resulting matrix of TF-IDF components was very sparse, I performed *truncated single value decomposition* to reduce the TF-IDF matrix to fifty features. In addition, I introduced the following time-dependent features: length of post, day of the week and hour of posting (one, seven, and twenty-four features, respectively). Next, I aggregated post features to user-level by calculating the mean values of LDA topic probabilities, mean values of TF-IDF components, the mean length of each post by computing the relative share of posts written by a specific user on each day of the week and on each hour of the day , ending with 102 user-level features.

Step 3: train classification models. I then used a suite of machine learning techniques to classify post authors as random users or trolls. To do so, I first verified that the extracted features could indeed help in classification. Figure 11 depicts the results of performing a principal components analysis on the space of the first two principal components. Here, the green dots represent trolls and red dots random users . As can be easily seen, most trolls are located far away from random users. The troll group has a much smaller variance

[10] See, these reports for a review.

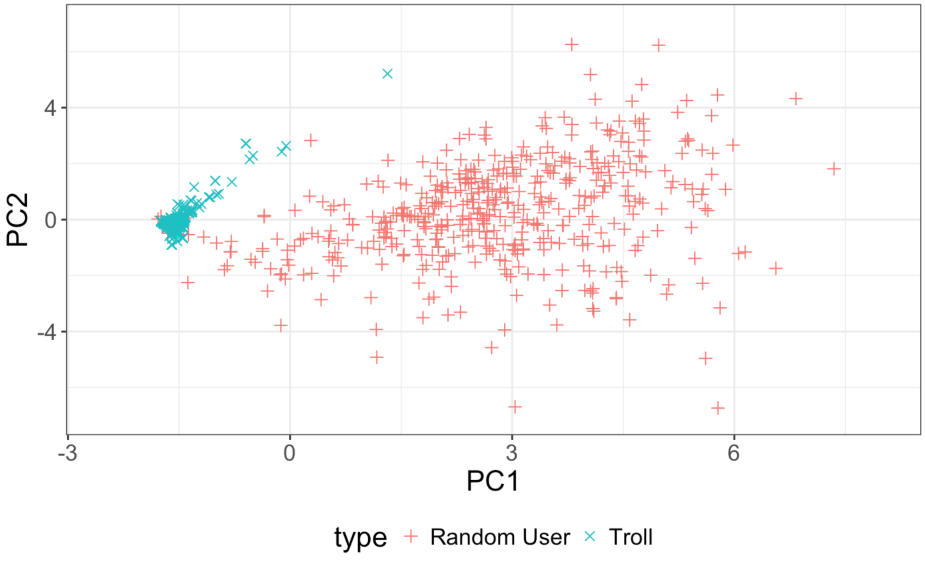


Figure 11: Trolls VS random *LJ* users: principal component analysis

than that of the random users group, a reasonable outcome if (as journalist accounts have suggested) trolls tend to employ the same terms , use the same message templates, and follow a regular time-schedule. While random *LJ* users differ in these particulars, most trolls exhibit very similar behavioral patterns.

I conducted the classification analysis using *regression, linear support vector machine, Gaussian support vector machine, Gaussian naive Bayes, multinomial naive Bayes model, random forest, gradient boosted tree,* and *deep neural network* to identify how trolls differ from random *LJ* users.[11] Then, I randomly split the data into training set and test sets and, on the training dataset, performed a *grid search* (Hsu et al., 2003) over the hyperparameter space of each model with five-fold cross-validation. Lastly, I applied trained models to the test data to evaluate each model's performance. Table 5 displays the the statistics measuring the performance of the various classification models. With a 96% precision and a 92% accuracy, *the random forest* appears to be the most efficient classification model.[12]

Apart from the accuracy of prediction, the *random forest* model identified the most important features distinguishing trolls from random users. With respect to word usage, trolls more frequently used such terms as "*USA*", "*America*", and "*Obama*" whereas random users were much more likely to use the words "*Sanctions*", "*Crimea*", and "*Putin*". Moreover, trolls and non-trolls differed greatly in the timing of their posts. Random users

[11] The choice of models was driven by the popularity of these models for classification tasks

[12] Note that, because model performance was evaluated by applying the models to the test dataset, overfitting should not be an issue.

Model	Recall	Precision	F1-score	Accuracy	ROC
Logistic Regression	0.924	0.746	0.825	0.870	0.884
Support Vector Machine (*linear*)	0.793	0.646	0.712	0.787	0.789
Support Vector Machine (*gaussian*)	0.750	0.873	0.807	0.881	0.848
Naive Bayes *(gaussian)*	0.924	0.545	0.685	0.718	0.770
Naive Bayes *(multinomial)*	0.880	0.587	0.704	0.755	0.786
Random Forest	**0.804**	**0.961**	**0.876**	**0.924**	**0.894**
Gradient Boosted Tree	0.815	0.938	0.872	0.921	0.894
Deep Neural Network	0.848	0.857	0.852	0.903	0.889

Table 5: Performance of classification models

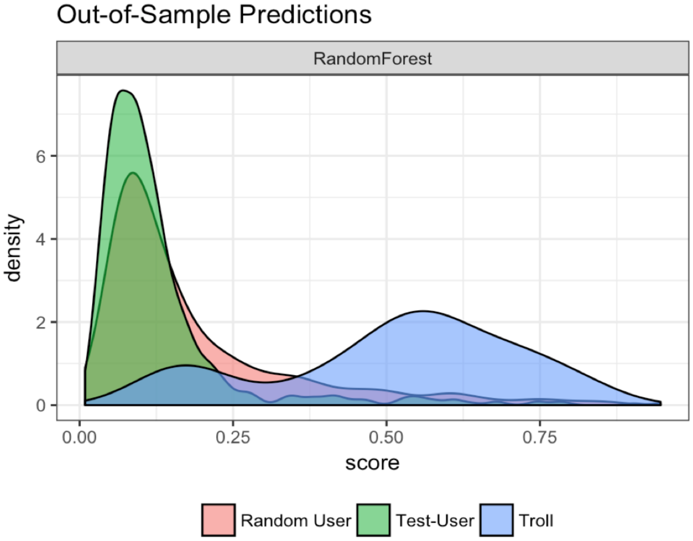


Figure 12: Propensity of trolls, random users, and participants of targeted conversations to be a troll

almost never publish posts between 2 am and 12 pm.

Step 4: apply the most accurate model to employ propensity scores to predict whether a poster is a troll or a random user. To test whether participants of targeted conversations exhibited behaviors such as that of the trolls on the *Novaya Gazeta* list, I sampled 650 participants of those conversations that were not mentioned in the list and then collected their posts and calculated the scores representing their corresponding features. Next, I applied the trained *random forest* model to calculate the propensity of those participants to behave like trolls, and Figure 12.displays the distribution of the propensities of trolls, random users, and participants in targeted conservations.

Figure 12 shows that both randomly sampled users and the randomly sampled participants of targeted conversations differ greatly from trolls. The calculated propensity scores for most of the accounts in these two groups are extremely low. Moreover, the results of the analysis show that, in fact, a randomly sampled *LJ* user has an even higher propensity to be a troll than the participants of targeted conversations. One possible explanation is that trolls target specific types of conversations, ones in which participants are very likely to be critical of the Vladimir Putin regime than an average user of a social medium platform. As a result, they would tend to use "important non-troll words" more frequently than would random *LJ* users. The study results also show that a small fraction of known troll accounts looked very much like accounts of regular users. One explanation is that paid trolls use their real *LJ* accounts to publish both personal and working posts. According to the findings of my analysis, less than three percent of the participants in targeted conversations had a propensity to be a troll higher that 50 percent, while less than half a percent had a propensity exceeding 60%. This evidence lends credibility to the hypothesis that the list of troll accounts published by *Novaya Gazeta* was, in fact, exhaustive.

8. Discussion and conclusions

The research described in this paper yielded three major results. *First*, it proposed a framework for analyzing the effects of political engagement on social media. This framework allows analysis of online political targeting such as occurs through multiple mechanisms, including political socialization and learning. This framework takes into account the fact that paid commentators hide their pro-government affiliation, thus reducing the ability of users to attribute received messages to specific political forces. *Second*, the paper proposes a method for estimating the effect of troll interventions on politically charged online discussions under a set of assumptions. These assumptions may not be applicable to online conversations in general but can be plausible given the specific operational conditions of Russian trolls such as those studied in this research. *Third*, it adds to the the existing literature on the problem of authoritarian control. Previous studies have established that to deter political dissidents, authoritarian governments try to prevent online discussions by censoring or creating informational noise. This research has established that a particular type of such interventions – the injection of paid pro-government commentators into online political conversations — might in fact be effective but that this effectiveness is limited to political discourse. Trolls appear to be successful in diverting the discussions from politically charged topics. When a conversation considers politics, troll intervention reduces the probability of an anti-government topic by fifteen percentage points and changes the the trend of the evolution of this conversation. The effect on promotion of a pro-government agenda thus appears to be negligible. While trolls are successful in diverting discussions from purely political topics, their interventions have no effect if the users discuss problems involving the national economy.

The focus of this paper has been limited in scope. *First*, it has considered only two potential effects of troll interventions in online conversations: the diversion of discussions from politically charged topics and the promotion of a pro-government agenda. *Second,* while the paper has analyzed the effects of troll interventions on behavior of participants in social media conversations, it does not consider the potential effects of such interventions on the broader audience of readers who eventually read these conversations and on the social media agenda. *Third*, while this paper has identified the effects of troll interventions on the evolution of online conversations, it has not provided evidence that they can change the preferences or offline political behavior of users. Further research will be required to explore these possibilities.

References

Blei, David M, Andrew Y Ng and Michael I Jordan. 2003. "Latent dirichlet allocation." *the Journal of machine Learning research* 3:993–1022. 1, 4.2.1

Bond, Robert M, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle and James H Fowler. 2012. "A 61-million-person experiment in social influence and political mobilization." *Nature* 489(7415):295. 1

Egami, Naoki, Christian J Fong, Justin Grimmer, Margaret E Roberts and Brandon M Stewart. 2018. "How to make causal inferences using texts." *arXiv preprint arXiv:1802.02163* . 1

Enikolopov, Ruben, Alexey Makarin and Maria Petrova. 2015. "Social Media and Protest Participation: Evidence from Russia." *Available at SSRN 2696236* . 2.1, 3.1

Geddes, Barbara, Joseph Wright and Erica Frantz. 2014. "Autocratic breakdown and regime transitions: A new data set." *Perspectives on Politics* 12(02):313–331. 3.1

Gunitsky, Seva. 2015. "Corrupting the Cyber-Commons: Social Media as a Tool of Autocratic Stability." *Perspectives on Politics* 13(01):42–54. 1, 2.2

Hsu, Chih-Wei, Chih-Chung Chang, Chih-Jen Lin et al. 2003. "A practical guide to support vector classification.". 7.2.3

Jones, Jason J, Robert M Bond, Eytan Bakshy, Dean Eckles and James H Fowler. 2017. "Social influence and political mobilization: Further evidence from a randomized experiment in the 2012 US presidential election." *PloS one* 12(4):e0173851. 1

Keller, Franziska B, David Schoch, Sebastian Stier and JungHwan Yang. 2017. How to Manipulate Social Media: Analyzing Political Astroturfing Using Ground Truth Data from South Korea. In *ICWSM*. pp. 564–567. 1

King, Gary, Jennifer Pan and Margaret E Roberts. 2013. "How censorship in China allows government criticism but silences collective expression." *American Political Science Review* 107(02):326–343. 1

King, Gary, Jennifer Pan and Margaret E Roberts. 2014. "Reverse-engineering censorship in China: Randomized experimentation and participant observation." *Science* 345(6199):1251722. 1

King, Gary, Jennifer Pan and Margaret E Roberts. 2016. "How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, not Engaged Argument." *Copy at http://j. mp/1Txxiz1 Download Citation BibTex Tagged XML Download Paper* 2. 1, 2.2

Marshall, Monty G and Keith Jaggers. 2002. "Polity IV project: Political regime characteristics and transitions, 1800-2002.". 3.1

Miller, Blake. 2017. Surveillance-Driven Authoritarian Learning from "Public Opinion Emergencies" in China. 1

Munger, Kevin, Rich Bonneau, John T Jost, Jonathan Nagler and Joshua Tucker. 2015. "Elites Tweet to get Feet off the Streets : Measuring Elite Reaction to Protest Using Social Media." pp. 1–31. 1

Sanovich, Sergey, Denis Stukal and Joshua Tucker. 2017. Turning the Virtual Tables: Government Strategies for Addressing Online Opposition with an Application to Russia. In *Annual Conference of the International Society of New Institutional Economics*. 1, 2.2

Smyth, Regina, Anton Sobolev and Irina Soboleva. 2013. Patterns of Discontent: Identifying the Participant Core in Russian Post-Election Protest. Upsala University. 3.1

Tufekci, Zeynep and Christopher Wilson. 2012. "Social media and the decision to participate in political protest: Observations from Tahrir Square." *Journal of communication* 62(2):363–379. 2.1