UNIVERSITY OF CALIFORNIA

Los Angeles

Dictators in the Spotlight:
What They Do When
They Cannot Do Business as Usual

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Political Science

by

Anton Sobolev

2019

ABSTRACT OF THE DISSERTATION

Dictators in the Spotlight:
What They Do When
They Cannot Do Business as Usual

by

Anton Sobolev
Doctor of Philosophy in Political Science
University of California, Los Angeles, 2019
Professor Daniel Simon Treisman, Chair

This dissertation explores the strategies that modern authoritarian leaders use to survive in office. Unlike many 20th century dictators, today's autocrats must operate "in the spotlight" — new media and information technology enable the political opposition and the public to observe their actions. This greater observability limits the effectiveness of government repression, sometimes forcing the authorities to shift to other tools of political control. I study two of these alternative tools: the staging of pro-government rallies to create an image of invincibility and the recruitment of armies of paid supporters to shape the narrative on the Internet and disrupt online conversation.

To explore these strategies, I focus on the case of Vladimir Putin's regime in Russia. I argue that, faced with a wave of anti-government protests, an autocrat such as Putin can discourage further demonstrations by organizing pro-government rallies that — perhaps surprisingly — convey credible information to regime opponents about the dictator's popularity. Moreover, this discouragement effect will be stronger — under certain conditions — if the autocrat allows some media freedom. I test this theory using data I collected on which Russian cities had access to broadcasts of the independent radio station, "Echo of Moscow."

Combining matching techniques with a difference-in-differences design, I compare protest dynamics in the cities that received broadcasts and in those that did not.

To better understand the second strategy, I explore the behavior and impact of several hundred "trolls" — paid supporters of the regime who are allegedly employed to leave pro-government comments on social media platforms. Using probabilistic topic modeling, I develop a method to estimate the causal effect of troll interventions in online discussions. I find that trolls are able to successfully divert online discussions from politically charged topics, but are ineffective in promoting a pro-government agenda. In a separate chapter, I develop a methodology for the study of such Internet actors. Specifically, I devise a set of classification models to detect paid "political commentators."

The dissertation of Anton Sobolev is approved.

Barbara Geddes

Chad J. Hazlett

Jeniffer Pan

Daniel N. Posner

Daniel Simon Treisman, Committee Chair

University of California, Los Angeles

2019

To my mother

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This dissertation would not be possible without stellar council of my advisor Dan Treisman. I would like to express my special thanks of gratitude to members of my dissertation committee, Barbara Geddes, Chad Hazlett, Jeniffer Pan, and Daniel Posner.

I have benefited enormously from talking to many UCLA faculty members. Specifically, I thank Graeme Blair, Keith Chen, Darin Christensen, Miriam Golden, Mark Handcock, Erin Hartman, Jeff Lewis, Leslie Johns, Susanne Lohmann, Zachary Steinert-Threlkeld.

I thank my fellow UCLA students for their help. Specifically, I am grateful to Maxim Ananyev, Ashley Blum, Natalia Lamberova, Eoghan McGreevy-Stafford, Eugenia Nazrullaeva, Francesca Parente, Aaron Rudkin, Feng Yang, Tamar Zeilberger.

I also appreciate help of faculty members from other institutions who commented informally on the various projects that coalesced in this dissertation. Specifically, I thank Daron Acemoglu, Tim Frye, John Reuter, Molly Roberts, Arturas Rozenas, Konstantin Sonin, Milan Svolik, David Szakonyi, Joshua Tucker, Andrei Yakovlev for their helpful comments.

I thank UCLA graduate advisor Joseph Brown for helping me navigate the program.

Financial support for the various parts of this research has been provided by Columbia University Harriman Institute, George Mason University Institute of Humane Studies, Columbia University Harriman Institute, and UCLA Center for Russian and European Studies.

VITA

2014–2019	Teaching Assistant at UCLA. Taught sections on Russian Politics, Research Design, Statistical Methods, Game Theory
2013–2019	Ph.D. student in the UCLA Political Science Department
2010–2013	Lecturer at Higher School of Economics, Moscow
2008–2013	Research Fellow at Higher School of Economics, Moscow
2009–2011	MA student at Higher School of Economics, Moscow
2005–2009	BA student at Higher School of Economics, Moscow

PUBLICATIONS

Dagaev, Dmitry, Natalia Lamberova, and Anton Sobolev. "Stability of Revolutionary Governments in The Face of Mass Protest." European Journal of Political Economy, 2019, 60: pp. 2–20.

Sobolev, Anton, and Alexei Zakharov "Civic and Political Activism in Russia." The New Autocracy: Information, Politics, and Policy in Putin's Russia, edited by Daniel Treisman, Brookings Institution Press, Washington, D.C., 2018, pp. 249–276.

Lazarev, Egor, Anton Sobolev, Irina Soboleva, and Boris Sokolov. "Trial by Fire: a Natural Disaster's Impact on Support for the Authorities in Rural Russia." World Politics, 2014, 66(4): pp. 641–668.

Yakovlev, Andrei, Anton Sobolev, and Anton Kazun. “Means of Production VS Means of Coercion: Can Russian Business Limit the Violence of Predatory State?,” Post-Soviet Affairs, 2014, 30(1): pp. 171–194.

Remington, Thomas, Anton Sobolev, Irina Soboleva, and Mark Urnov. “Social and Economic Policy Trade-Offs in the Russian Regions: Evidence from Four Case Studies,” Europe-Asia Studies, 2013, 65(10): pp. 1855-1876.

Smyth, Regina, Anton Sobolev, and Irina Soboleva. “Well-Organized Play: Symbolic Politics and the Effect of the Pro-Putin Rallies,” Europe-Asia Studies, 2013, 60(2): pp. 24-39.

CHAPTER 1

Introduction

1.1 Focus of the dissertation

This dissertation is comprised of three essays in political science, focused on a repertoire of tools authoritarian leaders employ to maintain political control, and the response they elicit from citizens. Its focus is on government hiring of regular citizens to engage with its political opposition both online and offline. Offline, these paid supporters are hired to hold pro-leader rallies to enhance an image of invincibility for the leader. Online, they act as "Internet trolls," shaping the online narrative and thereby disrupting politically sensitive discussions. Three essays share common empirical methodologies and intellectual themes. Each of them attempts to measure politically important but difficult-to-observe patterns and determinants of behavior of political agents – participants of pro-government rallies in Essay 1, and paid online pro-government commentators in Essays 2 and 3.

In the last decade, modern information technologies have changed the world of politics as we knew it, erasing any 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, debate continues in the United States over whether or not hackers and Internet trolls affected voting in the 2016

U.S. election.

I focus on the case of Vladimir Putin's regime in Russia. During the unexpected wave of mass protests that broke out seven years ago, the Russian authorities used much less repression than could be expected. Rather than relying on extensive threats, violence, and censorship, they quickly learned how to repurpose the same communication technologies that regime opponents were employing for staging protests. At the hands of authoritarian leaders, communication technologies became a medium of projecting an image of regime invincibility so as to dissuade the discontented from taking to the streets. Recent evidence from other countries suggests that, in doing so, Russian government was not an exception to the rule but rather representative of a new trend. That is why I use Russia as a laboratory to explore the effects of these tools.

1.2 Brief overview of arguments and evidence

My research offers several arguments regarding a repertoire of tools authoritarian leaders employ to maintain political control. Over the last decade, autocratic governments have displayed significantly less violence than those in existence 30-40 years ago. Guriev and Treisman (2015*b*) report that the share of authoritarian leaders whose regimes committed more than 10 political killings a year fell from nearly 60 percent for rulers who entered office in the 1980s to less than 30 percent among those who entered office in the 2010s. Data on political prisoners and torture exhibit a similar downward trend. What explains this shift away from violence? Existing explanations suggest that autocrats shift from violent to non-violent strategies because the latter are cheaper and more effective and have more predictable outcomes. My dissertation research challenges this conventional wisdom.

I propose a new approach to understanding of reduction of political repression. Modern autocrats must operate under conditions of increased transparency —

new information technologies enable the political opposition and the public to observe their actions more closely than ever before. I argue that this greater transparency makes autocratic "business as usual" overwhelmingly expensive and limits the effectiveness of government repression, forcing authorities to employ other tools of political control. In response to increased transparency of the digital era, authoritarian leaders tend to use techniques control that are hard to observe. While staging pro-government rallies and employing paid online commentators can be less efficient than repression, censorship, or propaganda, they also has smaller risks to backfire.

In the first essay, I argue that autocrats can sap the momentum of protest waves by staging large pro-government rallies. This phenomenon is puzzling at first glance because dissidents should recognize that the dictator can pay or intimidate citizens to participate in such pro-regime rallies. However, if mobilizing citizens to march in support of the dictator is sufficiently costly, extremely unpopular dictators will find it more cost-effective to spend their resources in other ways. Upon observing a pro-government rally, citizens can therefore be expected to infer that the dictator is not extremely unpopular and so may revise their estimates of the dictator's popularity upward. However, for citizens who do not directly observe the pro-government rally to make such inferences, they must believe the reports concerning it that they receive. At this point, the role of independent media comes into play: such media can provide credible accounts of such events and hence influence the public's assessment of the dictator's level of support.

To test this theory, I employ matching techniques and a difference-in-differences design to compare cities of Russia that received broadcasts of the independent radio station *Echo of Moscow* with other cities that did not. Studying the relations between the actions of the media, dissidents, and government was challenging because these relations tend to be highly endogenous. However, unique features of the political and media landscape in Russia made it possible to identify a causal relationship between media freedom, pro-government rallies, and protest. For instance, I determined that the occurrence of a massive pro-government rally in Moscow in 2012 discouraged potential protesters significantly more in regions exposed to *Echo of Moscow* broadcasts than in other regions.

The second essay explores how an autocrat such as Russia's Vladimir Putin uses new communication technologies to maintain political control online. It explores the behavior of several hundred Internet trolls. These trolls had published blog posts and participated in discussions on the popular Russian social media platform LiveJournal during 2014-15. As trolls were trying to maximize their audience, they kept their information (posts, comments, lists of friends and communities) open. Using a list of trolls accounts, published by investigative journalists, I collected two datasets: a complete body of text of almost a half a million trolls' posts and eighty thousand discussion threads infiltrated by those trolls. I start my analysis by comparing the behavior of trolls to the behavior of the representative sample of LiveJournal users. In order to do so, I collect additional dataset of the posts of random users of this social media. I combine obtained data with posts written by trolls and apply a set of feature-extraction techniques. Based on the extracted features, I train a set of classification models to distinguish between the randomly sampled LJ accounts and accounts on the leaked list of trolls. I find that while trolls were required to mask themselves as regular users, their behavioral patterns were sharply distinct from those of ordinary citizens. This step serves as evidence of the credibility of the documents leaked by investigative journalists. It contributes to the body of literature that focuses on developing tools to identify paid online actors, their target groups, and the scale of their Internet presence.

The third essay takes the next logical step. It explores the impact of Internet trolls who left pro-government comments on online political discussion. Can such agents successfully engage users with pro-government rhetoric? Can they divert them from criticizing political leaders? To address these questions, I devised a classification of the possible objectives of governments that employ Internet trolls, the strategies trolls use to achieve them, and the observable implications of these strategies. Combining text analysis with existing approaches in causal inference, I have developed 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 discovered that Russian troll activity was more successful in diverting online discussions away from politically charged topics than in promoting a pro-government agenda. At the same time, troll interference apparently had no effect when conversation covered the state of the national economy, with poor economic growth, unemployment, or price inflation under discussion.

CHAPTER 2

Can independent media help non-democratic governments suppress collective action?

2.1 Introduction

Dictators tend to restrict the media for a number of reasons. First, when that dictator's government performs poorly, the presence of independent media enables citizens to infer the quality of government and adopt new political attitudes (Enikolopov, Petrova and Zhuravskaya, 2011; Miner, 2012; Stein, 2012). Second, free media also enables dissidents to coordinate protest activity: for instance, newsreels can serve as focal points that encourage protesters to take to the street (Lohmann, 1994; Miner, 2012; Hassanpour, 2014).

However, not all dictators restrict media completely, and in fact, substantial variation in the degree of media freedom exists even among harsh authoritarian regimes (Egorov and Sonin, 2014). Given that autocratic incumbents do not easily accept criticism of their actions, why do some of them allow some media freedom? Studies of authoritarian politics provide several possible explanations. First, independent media outlets help authorities deal with "the dictator's dilemma" (Wintrobe, 1998), as they allow the gathering of information on public grievances and on the performance of local officials (King, Pan and Roberts, 2013; Lorentzen, 2013; Egorov and Sonin, 2014). Second, informational transparency increases the government's credibility to investors, thus promoting economic growth (Hollyer, Rosendorff and Vreeland, 2014*a*,*b*). Recent theoretical studies suggest that an autocrat can allow some level of media freedom during times of good economic performance, but may be better off increasing censorship when the economy stagnates and public discontent is likely to turn into protest (Edmond, 2013; Guriev and Treisman, 2015*a*). Indeed, VonDoepp and Young (2012) find that, in Africa, media harassment increases if governments are faced with protests and coup plots, while Stein (2012) shows that censored media convinced Brazilians to support the country's military regime throughout its existence from 1964 through 1985.

At the same time, a set of historical cases (e.g., the collapse of socialist regimes in Europe) suggests that revolutions can occur even under full censorship of media and dissemination of aggressive pro-government propaganda (Lohmann, 1994). In contrast, recent events in Venezuela and Russia demonstrate that a regime can handle threats of mass protest nonviolently, even in the presence of independent media outlets and uncensored internet services (Munger et al., 2015).

In this paper, I argue that some media freedom can actually benefit a dictator. I identify a previously unstudied effect of the exposure to free media on dissidents' decisions to take to the street. While not all autocratic regimes allow at least some degree of media freedom, I suggest that the ones that do, can exploit it in order to suppress political protest. Specifically, I argue that autocrats may be able to blunt the momentum of waves of anti-government protests by staging large pro-government rallies. At first glance, this assertion appears puzzling because dissidents are typically cognizant that the dictator can pay or intimidate citizens to participate in such pro-regime rallies. However, if mobilizing citizens to march in support of the dictator is sufficiently costly, extremely unpopular dictators will find it more cost-effective to spend their resources in other ways. Observing a pro-government rally, citizens will rationally infer that the dictator may be not as unpopular as they had thought and so may revise their estimates of the dictator's popularity upward. However, for citizens who do not observe the pro-government rally directly to make such inferences, they must believe the reports concerning it that they receive. Thus, I argue that an autocrat benefits from partially free media that can report observable events truthfully, but cannot conduct independent journalistic investigations.

To test this hypothesis, I employ covariate-balance propensity score techniques (Imai and Ratkovic, 2014) and a difference-in-differences design to compare cities of Russia that received broadcasts of the independent radio station *Echo of Moscow* with those that did not during the 2011-2012 protests in Russia. Studying the relations between actions of media, of dissidents, and of government is challenging because these relations are highly endogenous. Incumbents can affect the level of media freedom and so can complicate the coordination of dissidents, while the latter can protest against censorship and demand media independence.

Nonetheless, unique features of the political and media landscape in Russia make it possible to attempt identification of a causal relationship between media freedom and protest. First, while, media freedom is restricted in Russia, *Echo of Moscow* was allowed to broadcast quite freely for several possible reasons. Because the radio station was owned by the state company *Gazprom-Media*, businesses were not afraid to use it for advertising despite its critical coverage of the authorities. Thus the radio station was commercially successful and received many regional franchise requests. The board of directors set an informal threshold for the minimum regional radio audience size necessary for the company to accept a franchise request, and this size determined the revenues earned by the station from commercials. Thus, in contrast to a typical independent political outlet, the local availability of *Echo of Moscow* was subject to socio-economic and not political determinants. I confirm this empirically by showing that the only statistically significant predictors of the radio station's presence in a region are socio-economic and geographic. This result partially justifies my assumption that exposure to this radio station was as good as as-if random, conditional on the propensity of *Echo of Moscow* to enter local markets.

Second, the Russian government drastically changed its tactics toward the opposition in the middle of this wave of protests, whose scale and intensity grew rapidly during the last month of 2011. Nonetheless, before the early 2012 appointment of hard-liner Vyacheslav Volodin to the position of vice-head of the Kremlin's administration, the authorities preferred not to focus on the unrest and to instead treat it as a minor event. This soon changed drastically. The intensity of protest activity declined as the Presidential elections of March 2012 approached, the number of protesters decreased significantly. As survey data show, only determined radicals continued to take to the streets to protest (Smyth et al., 2015). The Kremlin's new cardinal Volodin switched the government's mild tactics to more aggressive ones. On the day of planned nationwide anti-government demonstrations (February 4th), a massive pro-government rally was organized on *Poklonnaya* Hill in Moscow (Smyth, Sobolev and Soboleva, 2013*a*). As a major source of information for protesters, *Echo of Moscow*, emphasized that this rally was much larger than the anti-government demonstrations occurring simultaneously. Subsequently, rumors circulated that many of the tens of thousands of citizens rallying for the government were actually employees of state-owned enterprises and organizations who had been pressured to participate rather than sincere supporters of President Vladimir Putin. Being a radio station, *Echo of Moscow* was not a producer of investigative journalism and so establishing the approximate number of genuine Putin supporters participating in the rally from its reports was difficult.

In this paper, I test whether credible reports on the relative sizes of pro-government and anti-government demonstrations in the absence of detailed journalist investigations produced by independent media appeared to discourage dissidents from taking to the streets. To do so, I compare the number of protests and the protest turnout in those regional capitols of Russia exposed to *Echo of Moscow* broadcasting with those that were not but which nonetheless satisfied or were close to satisfying the requirements needed for acceptance of a franchise request. Overall, most of the regional capitols experienced reductions in the number of protests and in protest. However, whereas those regional capitols with no exposure to *Echo of Moscow* saw mean number of protests declined from 2.9 to 1.7 and mean protest turnout decrease from .73 to .54 participants per thousand citizens, those with exposure experienced a 3.9 to 1.6 decline in mean number of protests and a a .88 to .42 participants per thousand citizens decline in mean protest turnout. When adjusted for the propensity scores, results suggest that in cities with no exposure to *Echo of Moscow* number of protests and protest turnout decreased on average by 1.5 protests and by .21 participants, respectively. In capitols exposed to *Echo of Moscow* number of protests and protest turnout decreased on average 3 protests and by .57 participants per thousand citizens, respectively.

Overall, the results suggest that when reporting on a government that seeks to create an image of invincibility (Magaloni and Wallace, 2008), independent media outlets can unintentionally strengthen the dictator's position. Such media outlets can effectively play a "bad joke" on the opposition, because they can discourage moderates from participating. Some scholars of Russian politics suggest that among the major reasons for the defeat of the resistance campaign was the fact that after the Presidential elections (and especially after the start of *the Bolotnaya Square* case)[1] moderates left the protest movement (Volkov, 2012). Efforts to create an image of invincibility can be less effective in the absence of credible media outlets. This may explain why Muamar Gadaffi's regime in Libya and the soviet government in the late years of the USSR were unable to awe opposition activists by means of large-scale pro-regime rallies. In both cases, most activists did not take the reports on such rallies broadcast by propaganda sources very seriously (Morris, 2014).

This paper contributes to several literatures. First, it speaks to the literature on the political mobilization (Miner, 2012; Adena et al., 2015; Peisakhin and

[1] A criminal case by the Russian Investigative Committee on the counts of alleged massive riot and alleged violence against police during the March of the Millions on May 6, 2012.

Rozenas, 2014; Yanagizawa-Drott, 2014) and persuasion (Enikolopov, Petrova and Zhuravskaya, 2011; Gehlbach and Sonin, 2014) effects of media. The studies closest to my research are Yanagizawa-Drott (2014) and Peisakhin and Rozenas (2014). The former investigates the effect of state radio propaganda on casualties of the genocide in Rwanda in 1994. The latter finds that the availability of Russian analog television signals raised electoral support for pro-Russian parties and candidates in the 2014 presidential and parliamentary elections in the Ukraine. In contrast to studies that largely focused on the effects of biased news from state-controlled media, I show that sometimes credible reports sent by independent media outlets can be an even more efficient instrument to discourage opposition than can state propaganda.

Second, the paper speaks to the literature on the role that free media plays in autocracies. Studies focus primarily on the ways that autocrats can use free media to increase their regime's performance. They do it generally via gathering information on low-level officials (King, Pan and Roberts, 2013; Egorov and Sonin, 2014) or by producing transparent information on the state of affairs and thus reducing the risk for capital investment (Hollyer, Rosendorff and Vreeland, 2014*a*,*b*). These studies assume a trade-off between the benefits of the free information flow and increased risks of social unrest. I find that this relationship is not always zero-sum. Under certain conditions, increased media freedom can be associated with a lower risk of mass protest.

Third, this study is also related to the literature on the evolution of strategies of authoritarian survival (Magaloni and Wallace, 2008; Munger et al., 2015). Recent studies of Guriev and Treisman (2015*a*) and Gunitsky (2015) find autocratic regimes of the 21st century to be less violent than their predecessors. In the new century, electoral falsifications, bribing and censoring the private press or corrupting online bloggers are cheaper and more efficient means of bolstering the regime's legitimacy than classic repressions. My results are in line with this account. In fact, Russian authorities were able to reduce the number of people taking to the street without any significant cases of violence. Journalists, public commentators, and even leaders of the opposition emphasized the exceptional politeness of policemen.[2]

Finally, the contribution of this paper is limited in scope. First, it does not offer a general theory of collective action but only studies a role of partial media freedom in political survival of autocrats. Inevitably, it ignores the "free-rider" problem and focuses only on the problem of coordination – not because the former is less important, but because the role of media is more pronounced in solving the latter. Empirically, I compare the success of protests in the regional capitols exposed and those not exposed to independent media broadcasting. Second, it does not argue that the Russian government strategically used *Echo of Moscow* to forestall anti-regime collective actions in 2011-2012. Instead, it suggests that, under certain conditions, exposure to credible reports of independent media can discourage potential protesters from taking to the streets.

The remainder of the paper is organized as follows. Section 2 contains clarifications and a numerical example of the theory. Section 3 provides background information. Section 4 describes the data, hypotheses and the identification strategy. Section 5 presents the empirical results and addresses potential concerns and factors that could bias the results. Section 6 concludes.

2.2 Theory and numerical example

In this section, I develop a toy example of how independent and state-controlled media affect dissidents' beliefs on incumbent's popularity using a Bayesian approach. I limit my to a case where the incumbent is able to organize a large-scale pro-government rally. If the rally is not a big one, dissidents cannot infer whether anyone else participated except for pro-regime stalwarts, and thus it makes no

[2] The only significant exception was the *Bolotnaya* Square demonstration (May 6th, 2012). However, this protest happened after the time period that is in focus of my study.

sense for him to spend his resources on it. I consider a non-polarized society, i.e., the majority of citizens are neither radical dissident, nor pro-regime stalwarts but are somewhere in between. Thus, if a large-scale rally takes place, dissidents learn that moderates also took to the street. The question that remains is whether moderates are true-supporters of the regime or are ,in fact, bribed or coerced to participate.

In the every day life of autocratic countries, the extent of media freedom depends highly on the incumbent's decisions. In this section however, I simplify the case by assuming the exogenous nature of media freedom. This assumption is consistent with my identification strategy that suggests conditional independence of the exposure to independent media.Moreover, the results hold if the incumbent is allowed to suppress media strategically, in case the suppression is costly. I address these issues in more detail in the Conclusion.

Setup. Consider a game with three players: *Dissident* (strategic), *Incumbent* (non-strategic), and *Moderate* (non-strategic). *Incumbent* organizes a pro-regime rally. *Moderate* can either show up at the rally or not show up, $a = \{S, \neg S\}$. If *Moderate* supports *Incumbent*, she always shows up at the rally. If she does not support *Incumbent*, the latter may propose a bribe sufficient in size to persuade her to show up despite her lack of enthusiasm for *Incumbent.* She accepts the bribe with probability $P(bribe) < 1$. *Dissident* does not observe whether or not *Moderate* supports the incumbent, but he has a prior probability estimate of this, $P_{prior}(support) < 1$.

Since *Dissident* knows that *Moderate* will participate in the rally if either (a) she supports the *Incumbent* or (b) she does not support the incumbent but has accepted a bribe to participate, *Dissident* also has a prior estimate of the probability that *Moderate* will participate:

$$P_{prior}(S) = P_{prior}(support) + [1 - P_{prior}(support)] \times P(bribe).$$

The protest succeeds with a probability of one minus the probability that

Moderate supports the *Incumbent*, $1 - P(support)$. The expected utility of *Dissident* from the protest is:

$$U_d = [1 - P_{prior}(support)] \times A - P_{prior}(support) \times C,$$

where A is a victory prize, and C are the costs of failure (e.g., retribution).

Suppose now that *Dissident* can directly observe if *Moderate* showed up at the pro-government rally. *Dissident* follows Bayes rule in updating his beliefs on *Incumbent's* popularity, $P(support)$:

$$\begin{aligned}
P(support|S) &= \frac{P(S|support) \times P_{prior}(support)}{P_{prior}(S)} = \\
&= \frac{P(S|support) \times P_{prior}(support)}{P(S|support) \times P_{prior}(support) + P(bribe) \times [1 - P_{prior}(support)]} \\
&= \frac{P_{prior}(support)}{P_{prior}(support) + P(bribe) \times [1 - P_{prior}(support)]}.
\end{aligned}$$

Given that, by assumption $P(bribe < 1)$ and $P_{prior}(support) < 1$,

$$\frac{P_{prior}(support)}{P_{prior}(support) + P(bribe) \times [1 - P_{prior}(support)]} > P_{prior}(support),$$

i.e., if *Dissident* observes that *Moderate* showed up at the rally, he updates his estimates of *Moderate's* support for *Incumbent* upwards. Even though there is a positive probability that moderate has been bribed to attend the rally, *Dissident* still increases his estimate of the regime's popularity after observing that *Moderate* took to the street.

Now suppose that *Dissident* does no observe the rally directly but instead receives a signal from the media. I assume that the media may either be biased — in which case it always reports that *Moderate* rallied ($P_{biased}(signal = S) = 1$), whether she did or not — or unbiased — in which case it reports the truth with probability c. Type of media is common knowledge. One can think that c measurers the media's credibility. Clearly, if the media is biased, *Dissident* will pay no attention to these reports:

$$P(S|signal = S) = \frac{P_{biased}(signal = S|S) \times p_{prior}(S)}{P_{biased}(signal = S)} = \frac{1 \times p_{prior}(S)}{1} = p_{prior}(S).$$

However, if the media is unbiased, she will update her beliefs as follows:

$$\begin{aligned} P(S|signal = S) &= \frac{P_{biased}(signal = S|S) \times p_{prior}(S)}{p(signal = S)} \\ &= \frac{c \times p_{prior}(S)}{c \times p_{prior}(S) + (1-c)(1-p_{prior}(S))}. \end{aligned}$$

This equation allows conditions to be identified when $P(S|signal = S) > p_{prior}(S)$:

$$\begin{aligned} \frac{c \times p_{prior}(S)}{c \times p_{prior}(S) + (1-c)(1-p_{prior}(S))} &> p_{prior}(S) \\ c \times p_{prior}(S) &> p_{prior}(S) \times [2c \times p_{prior}(S) + 1 - c - p_{prior}(S)] \\ c &> 2c \times p_{prior}(S) + 1 - c - p_{prior}(S) \\ 2c[1 - p_{prior}(S)] &> 1 - p_{prior}(S) \\ c &> \frac{1}{2}. \end{aligned}$$

This result shows that a media report will be more likely to lead to an increase in the belief that *Moderate* actually took to the street if the credibility of the media is relatively high. Given the signal from the media, *Dissident* calculates the posterior probability that *Moderate* supports the incumbent by adjusting for media freedom:

$$\begin{aligned} P(support|signal = S) &= P(support|S) \times P(S|signal = S) \\ &= \frac{P_{prior}(support)}{P_{prior}(support) + P(bribe) \times [1 - P_{prior}(support)]} \\ &\times \frac{c \times p_{prior}(S)}{c \times p_{prior}(S) + (1-c)[1 - p_{prior}(S)]}. \end{aligned}$$

In the next paragraph I show how $P(support|signal = S)$ relates to $P_{prior}(support)$, $P(bribe)$, and c.

Comparative Statics. Consider how the extent of media freedom affects *Dissident's* posterior beliefs on *Incumbent's* popularity with respect to his prior beliefs $P_{prior}(support)$, the probability that *Moderate* accepts bribe $P(bribe)$, and the credibility of the independent media outlet c. I set other parameters to particular values for the sake of simplicity.

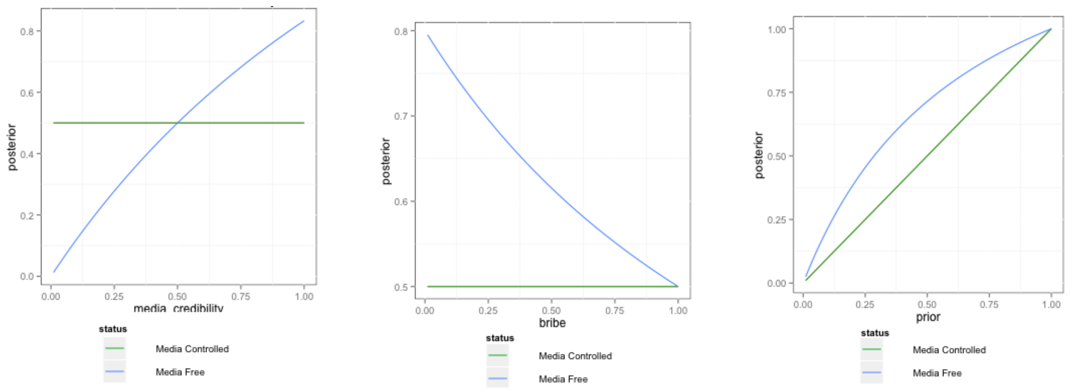
Figure 1a shows the comparative statics for $P(support|signal = S)$ with respect to the credibility of the independent media outlet given $P_{prior}(support) = 1/2$ and $P(bribe) = 1/5$. Two main sources can explain the credibility of independent media; (lack of) professionalism, and dependency from opposition leaders. First, nobody believes even independent reports if journalists are known to be corrupt or unprofessional. Second, dissidents do not trust reports if the media outlet plays up revolutionary leaders who are ready to strike at any cost. The figure depicts that the posterior support is higher under media freedom only if the level of the media's credibility is greater then 1/2.

This result can also be derived in a general case:

$$\begin{array}{rcl} P(support|signal = S) & > & P_{prior}(support) \Leftrightarrow \\ \dfrac{P_{prior}(support)}{p_{prior}(S)} \times \dfrac{c \times p_{prior}(S)}{c \times p_{prior}(S) + (1-c)[1-p_{prior}(S)]} & > & P_{prior}(support) \Leftrightarrow \\ 1 - P_{prior}(S) & > & \dfrac{1-c}{c}[1-P_{prior}(S)] \Leftrightarrow \\ c & > & \dfrac{1}{2}. \end{array}$$

As long as the media is credible (*c ¿ 1/2*), *Dissident* raises his estimate of support for the *Incumbent* when he receives a report that *Moderate* rallied for the *Incumbent* from the unbiased media.

Figure 1b shows comparative statistics for $P(support|signal = S)$ with respect to $P(bribe)$ given $P_{prior}(support) = 1/2$ and c=1/8. Basically, $P(bribe)$



(a) Media credibility

(b) Size of the bribe

(c) Prior beliefs

Figure 2.1: Comparative statics

reveals the size of *Incumbent's* budget. If *Incumbent* has enough resources, he can propose a huge bribe, ensuring that *Moderate* always accepts it. Thus, as *Moderate* always shows up (either due to her support or due to the bribe), *Dissident* learns nothing about Incumbent's popularity even if media is independent.

Figure 1c shows the comparative statics for $P(support|signal = S)$ with respect to $P_{prior}(support)$ given $P(bribe) = 1/5$ and $c = 1/8$. The figure outlines several important results. First, except for the extreme cases when $P_{prior}(support)$ equals to 1 or 0, the independent media generates a higher value of posterior belief in *Incumbent's* popularity than the biased media. Second, the size of the effect of the independent media ($P_{posterior}$ - P_{prior}) peaks when the *Dissident*'s uncertainty on *Incumbent's* popularity is the highest, i.e., $P_{prior}(support) = 1/2$. Finally, in a particular number of cases, the independent media can crucially change *Dissident's* behavior as it transforms his prior beliefs that *Incumbent* is unpopular ($P_{prior}(support) < 1/2$) to the opposite posterior beliefs ($P(support|signal = S)$), while state-controlled media does not change $P_{prior}(support)$

Numerical example. Following is a numerical example that illustrates this main result.

Let $P_{prior}(support) = 1/3$, $P(bribe) = 1/5$ and $c = 1/8$, then

$$P(support|S) = \frac{1 \times \frac{1}{3}}{1 \times \frac{1}{3} + \frac{1}{5} \times \frac{2}{3}} = \frac{1}{3} / \frac{7}{15} = \frac{1}{3} \times \frac{15}{7} = 5/7.$$

If media is controlled by the state, then $P(S|signal) = \frac{1 \times 7/15}{1} = \frac{7}{15}$. Thus, *Dissident* does not change his beliefs:

$$\begin{aligned} P(support|signal = S) &= P(support|S) \times P(S|signal = S) \\ &= \frac{5}{7} \times \frac{7}{15} = \frac{1}{3} = P_{prior}(support). \end{aligned}$$

If media is independent, it reports on the rally with probability

$$\begin{aligned} P(S|signal = S) &= \frac{.8 \times 7/15}{.8 \times 7/15 + .2 \times 8/15} \\ &= \frac{7}{9}. \end{aligned}$$

Dissident changes his beliefs:

$$\begin{aligned} P(support|signal = S) &= P(support|S) \times P(S|signal = S) \\ &= \frac{5}{7} \times \frac{7}{9} = \frac{5}{9}. \end{aligned}$$

Thus, after observing a report from state-controlled media *Dissident* believes that *Incumbent* is popular with probability $1/3 < 1/2$, but with probability $5/9 > 1/2$ if media is independent. This example shows that, under certain conditions, independent media may indeed crucially change *Dissident's* beliefs about *Incumbent's* popularity.

2.3 Brief history of media in Post-Soviet Russia

Vladimir Putin's crusade against Russian media. According to the Freedom House Foundation, the media have not been free in Russia since at least

2003. Except for a limited number of newspapers and magazines, all significant media outlets in Russia were either directly controlled by the state or were owned by oligarchs from Vladimir Putin's inner circle (Gehlbach, 2010*b*). One of Putin's main concerns upon becoming president was gaining control over the major media outlets, particularly TV channels. During Yeltsin's rule, major media empires were under the control of a few oligarchs who actively used them for lobbying their own business and political interests.

Within a short period of time, Putin was able to seize the commanding heights of the media industry (Gehlbach, 2010*b*). His most significant action was the attack on the *ORT* TV station of Boris Berizovsky and the *Media-Most* corporation of Vladimir Gusinsky. After the selective application of tax and criminal law to the company, the invasion of its premises by tax police, the direct pressure of the Ministry of Press, Radio, and Television, and also boardroom intrigue, *Media-Most* collapsed. The leading source of non-state broadcasting, and the only privately-owned TV station with a national reach, became the property of the government-controlled energy company *Gazprom* (Becker, 2004). The new owner completely changed the staff and editorial policy of the channel to become more supportive toward the government. Similar things happened to Boris Berezovsky's *ORT*, and both oligarchs were eventually forced into exile.

Exception to the rule: *Echo of Moscow*. As *Echo of Moscow* was a part of *Media-Most*, the radio service also became the property of *Gazprom*. However, the editorial policy of the station and the team of journalists within the company did not change. Being the oldest post-soviet media outlet, and known worldwide as one of Russia's last bastions of free media, the radio service was allowed to continue broadcasting to audiences across Russia for the following possible reasons.

First, as Editor-in-Chief Alexei Venediktov has pointed out, *Echo of Moscow* serves as a useful tool to refute Western criticism of Russia's lack of freedom of speech, as the Kremlin points to *Echo of Moscow* whenever countries in the West criticize press freedom in Russia .A second reason for the Kremlin's tolerance of Echo of Moscow was that it acts as a safety valve for discontented groups. Even though the station is held in high regard by the country's intelligentsia, it has little influence over the voting masses. Two facts supports this claim: first, according to *TNS Gallup*, the outlet's audience is extremely loyal. For more than half of its listeners, *Echo of Moscow* is the only - or at least the major - radio station. In addition, the radio station itself might not be a factor in affecting political preferences; instead, its audience already consists of those who hold negative attitudes toward the government. Alternatively, the Kremlin's tolerance could also be explained by the relatively small size of the radio station's audience compared to those of TV channels.

Given the existing results in the political economy literature, independent media outlets may help to fill the informational vacuum generated by incentives of subordinates to not report bad news to an autocrat (Wintrobe, 1998). For instance, several journalists have mentioned that the highest-level politicians in Russia are among *Echo of Moscow*'s regular listeners (Barabanov, 2009).

Finally, the informal relations between the station's Editor-in-Chief, Alexey Venediktov, and Vladimir Putin may be the basis of the perception that *Echo of Moscow* is untouchable. In a series of interviews, Venediktov has mentioned that, in the past, he had engaged in hours of informal talks with Vladimir Putin. In addition, Putin's press secretary, Dmitry Peskov often provides the station with exclusive commentaries.

Although the exact reasons for *Echo of Moscow*'s survival are unknown, givent that it is an anti-government radio station, there is a key difference between this radio station and other independent media outlets that specialize in political news. While most of the other outlets suffer from a lack of profits and must depend on wealthy donors (e.g., *Novaya Gazeta* newspaper and *The New Times* magazine depend on Alexander Lebedev and Irena Lisnevskaya respectively), *Echo of Moscow* is an exceptionally profitable company and has paid dividends to its shareholders every year since 1998. In contrast, *Finam.FM* radio station, founded in 2008, rapidly acquired a sizable audience in Moscow, but then, in 2013, the authorities exerted pressure on the station's owners to discontinue three of its programs. Eventually, due to these pressures and insufficient revenues, the owners ceased broadcasting altogether and sold the outlet.

Because businesses were not afraid to use *Echo of Moscow* for advertising and because the station was commercially successful, it received franchise requests from most of the other regions in Russia. Consequently, the station's board of directors then set the informal minimum entrance requirements for the acceptance of regional franchise requests.[3] Thus, in contrast the a typical independent political outlet, the regional presence of *Echo of Moscow* was subject to economic, rather than political determinants. By the beginning of the most recent wave of post-electoral protests, 42 Russian cities were exposed to *Echo of Moscow*'s broadcasting. This exposure was most likely random among the cities that met *Echo of Moscow*'s entrance requirements for franchises at least somewhat closely.

The typical contract between *Echo of Moscow* and regional broadcasters states that the latter can use daytime hours for ads, announcements and local programs, but that evening and morning air time belongs to the Moscow office. This fact crucial for this study, as most reports on the *Poklonnaya Hill* rally were delivered during the evening broadcast of 4th February 2012. In fact, according to records of that evening's broadcast, the radio station reported on the size of both the pro-government and the anti-government demonstrations during each news release from 6 pm to 10 pm. The numbers of participants included in these reports were provided by the demonstrations' organizers, the police, radio station journalists, and independent experts. All of these reports suggested that more participants took to the street in order to support Vladimir Putin than to

[3] According to the author's interviews with the journalists and managers of *Echo of Moscow*

demonstrate against him. While the true scale of the two collective actions is now unknown, the listeners could be assumed to trust these reports, as journalists of *Echo of Moscow* were critical of the government and, thus, unlikely to be incentivized to report the attendance as favorable toward Putin.

Importantly, that according to the survey of the protesters against the government, non-state controlled radio stations – and *Echo of Moscow* in particular – were a main source of political information for them.

Report, but do not investigate: free media in an unfree environment. An important feature of Russia's media environment is the absence of the means required to conduct investigative journalism. Even though *Echo of Moscow* and other independent media are allowed to criticize the government and report on anything that might interest their audiences, journalists lack the opportunities, rights, and legal protections needed to undertake effective investigations.

David Remnick's prominent article on *Echo of Moscow* underlines that although the station is able to broadcast opinions critical of the government, it falls short in conducting thorough investigations. As an example, he cites an interview with one of Russia's most famous journalists and commentators, Yulia Latynina, who admits that investigative work is nearly impossible in Russia:

> *The basic problem is that you cannot really expect, in a regime like that of Marcos or Duvalier, to get solid information into your hands about bank accounts, ... Everyone looks the other way. This is not a dictatorship — no one should exaggerate and compare it to the Soviet Union — but in an authoritarian regime you can't conduct an effective investigation the way you can in a democratic regime.* (Remnick, 2005, 15)

This opinion meshes closely with Gehlbach's (2010*a*) account, suggesting that media freedom in Russia is at the intermediate level. The latter is essential for the empirical strategy of this study, as the theory suggests that, without full freedom of the press, citizens could not infer whether moderates genuinely supported the incumbent or were bribed.

Media and Russian protests 2011-2012. The meetings in protest of the falsification of the parliamentary and presidential elections of 2011-12 were the largest in Russia since the collapse of the Soviet Union in 1991. The largest of these protests took place in the months following the parliamentary elections. The aftermath of the 2011 elections was in sharp contrast to that of previous elections. Although most observers viewed the levels of fraud in the 2011-2012 and the 2007-2008 elections as roughly equivalent, the latter resulted in no mass protests. Even experienced leaders of the opposition expected the ruling party, *United Russia*, to receive the majority of votes and so foresaw no social unrest, especially on such a large scale. However, *United Russia*'s unexpectedly low official results for (49% of the vote) constituted the shock that triggered the mass protests (Hale, 2011).

The scale of the protests had been increasing since the parliamentary elections, with at least five thousand Muscovites taking to the streets in the early evening of December 5, 2011, to voice their dissatisfaction with the results of the parliamentary elections. In the following two months, Putin's Russia experienced the unexpected rise of the opposition movement. Six days after the *Chistiproudny Boulevard* meeting, at least sixty thousand protesters rallied in *Bolotnaya Square*. Two weeks later (on December 24,) this number had increased to around one hundred twenty thousand. These protests occurred not only the capitol cities of Moscow and Saint-Petersburg, but also in most of the regions. Figure 2.2 depicts the number of anti-government demonstrations from December 2011 until May 2012.

More than 500 mass protests took place in almost all Russia's regions.[4] With

[4] Based on the author's calculations

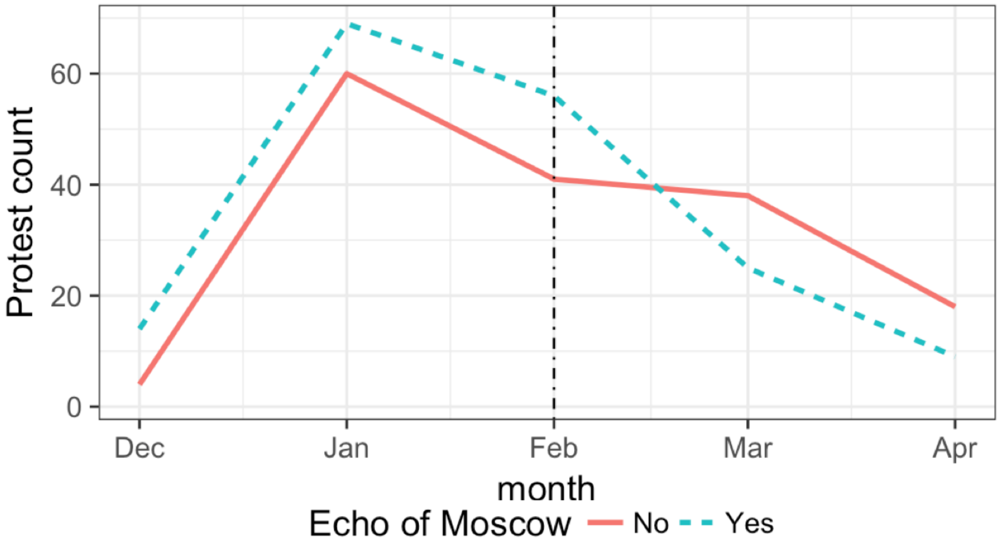


Figure 2.2: Protests in Russia's cities

the day of the presidential election approaching, however, the number of people taking to the streets in protest began to decrease, and the number of anti-government protesters had been declining steadily ever since the parliamentary elections.

Figure 2.2 also shows the patterns of protest intensity in cities that were exposed and were not exposed to *Echo of Moscow* broadcasts. As can be seen, before the major pro-government rally on February 4, these patterns were essentially the same, although the baseline amount of protests was higher for the first group (i.e., those exposed to *Echo*'s broadcasts). However, following the rally, the patterns changed. While the number of protests in the cities that were exposed to the station's broadcasting fell dramatically, this number remained almost constant in the other cities. Then, following the announcement of Vladimir Putin's victory in Presidential elections in March, 2012, protests faded away everywhere.

2.4 Hypotheses, data and empirical strategy

2.4.1 Hypotheses

The theory developed in this paper is consistent with a belief that, in general, the presence of independent media increases the ability of the opposition to mobilize dissidents. However, it also adds a significant nuance: if the incumbent is able to organize a large-scale support rally, then independent media reports can reduce both the probability of following protests and the size of the protest turnout.

Thus, I test two main hypotheses:

1. After a pro-regime rally takes place, the size of the the protest turnout declines to a greater extent in cities exposed to independent media reports about the size of the pro-regime rally than in cities not exposed to such independent media reports*;*

2. After a pro-regime rally takes place, the number of anti-government protests declines to a greater extent in cities exposed to independent media reports about the size of the pro-regime rally than in cities not exposed to such independent media reports.

2.4.2 Data

Outcomes of interest: Protest count and protest size. While the 2011-2012 protests in Russia were a reaction to a single event (the falsification of the parliamentary election results), their scale and frequency varied greatly across time and city. I measure this variation using a protest-event dataset based on the reports from the *NaMarsh.ru* website, which aggregates information drawn from various sources: a network of regional correspondents, the printed press, and online newsreels. Despite the fact that the website is maintained by opposition groups and thus is potentially biased in its reporting of protest events, scholars of Russia's politics believe that the reports it contains accurately capture temporal and spatial protest trends and that it corresponds with national and regional public opinion polls gauging support for and activism in protests (Lankina and Voznaya, 2015). I validate the *NaMarsh.ru* data with reports from the archive of the Russian Institute for Collective Action (ICA). This NGO publishes regular updates about individual opposition protest events across Russia, mostly those involving social claims or those linked to independent trade unions, anti-globalist movements, and other non-mainstream left-wing groups. Although the total number of reports by ICA is smaller than that of NaMarsh.ru, these data are widely used in studies of Russian politics (Clement, 2008; Robertson, 2010; Teague, 2011; Robertson, 2013).

While the complete dataset records more than 7400 opposition events across Russia from 2007 to 2017, I only use data concerning protests that took place within forty days of the February 4, 2012, pro-government rally on *Poklonnaya Hill* in Moscow; there were 251 and 145 of these before and after the date, respectively. Because of the paucity of Russian city-level socio-economic data in Russia, I consider only *regional capitols* that experienced protests in the forty-day time span. That almost eighty percent of *Echo of Moscow*'s branches were located in capitols of regions at that time partly justifies that choice. I measure a change in protest turnout for each city as the difference between the largest opposition protest before and the largest opposition protest after the pro-regime rally on *Poklonnaya Hill*, weighted by the city's population . The forty-day time span was chosen so that demonstrations which took place on the major days of the protest (December $24th$ and March $5th$) appear in the sample. To obtain the second outcome of interest, I calculate the difference in the number of protests in a city before and after the rally on *Poklonnaya Hill* within forty-day time span. Aside from the small amount of data employed directly in testing the hypotheses, I used additional data points to check the assumption of parallel trends

Explanatory variable: Exposure to Echo of Moscow reports. I use a

	Echo of Moscow		Total
	-	+	
No Protest	6	0	6
Protest	43	33	76
Total	49	33	82

Table 2.1: Exposure of regional capitols to *Echo of Moscow* in Russia

binary variable to indicate the exposure of the city to *Echo of Moscow* reports on the day of a major pro-government rally, information I collected from *WebArchive.org*, which contains a copy of the radio station website *Echo.Msk.ru* for the 2011-2012 protest period. At that time, *Echo of Moscow* broadcasted in 42 cities, nine of which (Kinesma, Obninsk, Pereslavl'-Zalesskiy, Rybinsk, Severodvinsk, Tol'jatti, Vyborg, Zelenzogorsk, Zeleznogorsk-Ilimskiy) were not regional capitols and so are not being considered in my study. Because the franchise request approach does not apply to the city of Moscow, I excluded it from my analysis as well.

Table 2.1 shows that 76 regional capitols experienced protests in the forty-day time span, and 33 of them were exposed to *Echo of Moscow.* Anti-government demonstrations occurred in 43 out of 49 of the regional capitols lacking *Echo of Moscow* broadcasting.

2.4.3 Empirical strategy

How to test the validity of rules in accepting franchise requests. As the first stage of my analysis, I check whether the exposure to *Echo of Moscow* was most likely as good as random among the cities that met *Echo of Moscow*'s entrance requirements for franchises -- that is, if the actual exposure of Russian cities to the radio station broadcasting is consistent with the rule described by its management in accepting franchise requests. To account for as many potential confounders as possible I employ LASSO selection approach, collecting a broad set of covariates to model the probability of a city's being exposed to *Echo of Moscow.* Most of the covariates come from the Russian State Agency of Statistics (*GosKomStat*). To identify the economic factors determining the station's local presence, I use local GDP per capita, private capital flows, the unemployment rate, average wage, economic inequality, the size of labor force, proportion of educated people in the labor force, and the number of automobiles per capita. I also use available geographic variables that could also have contributed to the cost of entry, including the distance of the city from Moscow and the mean January and July temperatures (as of 2010).

In addition, I include a set of socio-demographic indicators to account for the size of the potential radio audience and its consumption behavior: local population size, the share of adult Internet and personal laptop users, the share and density of fixed and mobile phones coverage.

One of the challenges of my approach is that local exposure to the *Echo of Moscow* radio station could also have been subject to political determinants. In such case, the causal effect cannot be estimated if both exposure to the station's broadcasts and the scale of protests were functions of local political regimes. To mitigate this problem, I obtained from the Central Electoral Committee of Russia the official electoral scores of the ruling *United Russia* party in parliamentary elections (2003, 2007, 2011) and the vote shares of Vladimir Putin and Dmitry Medvedev in the presidential rallies of 2004 and 2008, respectively. In addition, I also employed estimates of electoral fraud in the 2011 parliamentary elections presented from Kobak, Shpilkin and Pshenichnikov (2012).

Identification. The identification strategy of this study is based on the assumption that exposure to *Echo of Moscow*'s broadcasts was most likely as good as if random among the cities that met *Echo of Moscow*'s entrance requirements for franchises at least somewhat closely. As the actual indicators that were used by radio station management to accept franchise requests were, and still are, to estimate the propensity of a city to meet *Echo of Moscow*'s entrance requirements for franchises, I use city-level predictors of local exposure to *Echo of Moscow.* If the quasi-randomness assumption is valid, then conditional on a city's propensity to be exposed to *Echo of Moscow*, the causal effect τ can be recovered with a difference-in-differences estimator:

$$\hat{\tau}|propensity = \{E[Y(1)|D = 1] - E[Y(0)|D = 1]\} - \{[E[Y(1)|D = 0] - E[Y(0)|D = 0]\},$$

where:

$E[Y(1)|D = 1]$ is the expected size of the protest in a city with *Echo of Moscow* after the pro-regime rally, $E[Y(0)|D = 1]$ is the expected size of the protest a city without *Echo of Moscow* after the rally, $E[Y(1)|D = 0]$ is the expected size of the protest in a city with *Echo of Moscow* before the rally, $E[Y(0)|D = 0]$ is the expected size of the protest in a city without *Echo of Moscow* before the rally.

The same estimator can be used to recover causal effect of media reports on the protest turnout.

2.5 Results

2.5.1 Does a city hear *Echo of Moscow*?

In this section, I identify factors that determined the local availability of *Echo of Moscow*, i.e., if actual exposure of regional capitols to the stations was consistent with the rule followed by its management in accepting franchise requests. As the potential confounders are abundant and the number of regional capitols is relatively small, I employ a LASSO approach (Tibshirani, 1996) to select the predictors of the station's urban presence. LASSO regression minimizes the sum of squared errors with a bound on the sum of the coefficients' values. Because the results of LASSO-modeling depend on initial values, I bootstrap the LASSO estimators of the regression parameters with one thousand bootstraps and re-

	Exposure to *Echo of Moscow*
	Frequency of selection
Population of Region Capitol 2009	963
Regional Capitol Distance from Moscow	980
Temperature in January 2009	961
Observations	76

Table 2.2: Results of Bootstrap LASSO

tained those variables that were selected in at least 95% of the estimated models (Chatterjee and Lahiri, 2011).

Table 2.2 display selected predictors resulting from the estimated LASSO models. It shows the selection frequencies of each significant predictor, including regional capitol's population, the mean temperature in January, and the distance of the capitol to Moscow. A capitol city's population could be regarded as a valid proxy for the size of the *Echo of Moscow*'s audience. Moreover, a regional capitol's distance from Moscow also seems to be a reasonable predictor, as it can be partly associated with the cost of organizing joint broadcasting of the regional company and the central office of *Echo of Moscow.* Finally, the importance of the temperature in January may be related to the non-linearity in statistical relations between the outcome of interest and a regional capitol's distance from Moscow.

The important result of for this analysis is that political variables, such as the results of the presidential (2004, 2008) and the parliamentary (2003, 2007, 2011) elections or the levels of electoral fraud do not predict local availability of *Echo of Moscow.* This result addresses the concern that the local availability of *Echo of Moscow* follows political reasons.

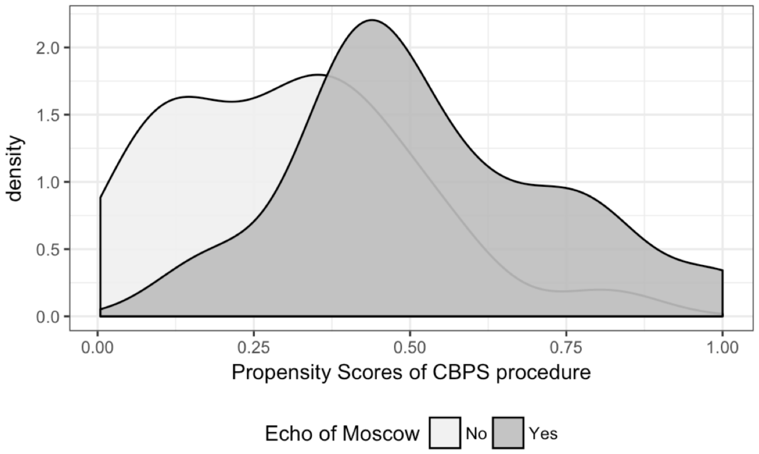


Figure 2.3: Covariate-balance propensity scores for city's exposure to *Echo of Moscow*

2.5.2 Effect of *Echo of Moscow* reports on protest activity

In the unadjusted sample, regional capitols with no exposure to *Echo of Moscow* experienced a decline in mean protest turnout and mean number of protests from .73 to .54 participants per thousand citizens and from 2.9 to 1.7 protests, respectively. At the same time, in capitols exposed to *Echo of Moscow,* mean protest turnout and mean number of protests dropped from .88 to .42 participants per thousand citizens and from 3.9 to 1.6 protests, respectively.

I employ three variables whose selection I described in the previous section to conduct covariate-balance propensity score algorithm (Imai and Ratkovic, 2014). Figure 2.3 shows the distribution of propensity scores for both groups of cities.

Next, I regress two outcomes of interest on exposure to *Echo of Moscow* with and without inverse propensity score weights. Table 2.3 reports the main results of this study. In the unadjusted sample, only the effect on the change in the number of protests appears to be significant (Columns 2 and 4). In the models that use inverse propensity score weighting *Echo of Moscow* is shown to have a large and significant effects on both measures of protest activity (Columns 1 and 3). These models suggest that, in cities with no exposure to *Echo of Moscow,*

	Change in Protest Turnout^{+}		Change in Protest Count	
	IPW^{++}	Unadjusted	IPW	Unadjusted
Echo of Moscow	-0.36**	-0.23	-1.51***	-1.31**
	(0.15)	(0.19)	(0.52)	(0.58)
Constant	-0.21	-0.20	-1.56**	-1.814***
	(0.20)	(0.17)	(0.74)	(0.54)
Observations	76	76	76	76
Log Likelihood	−100.5	−93.8	-164.8	-152.8
Akaike IC	211.1	197.7	339.6	315.7

Note: Standard errors in parentheses,*p<0.1; **p<0.05; ***p<0.01,
$^{+}$Protesters per 1000 citizens,$^{++}$Inverse Propensity Weighting.

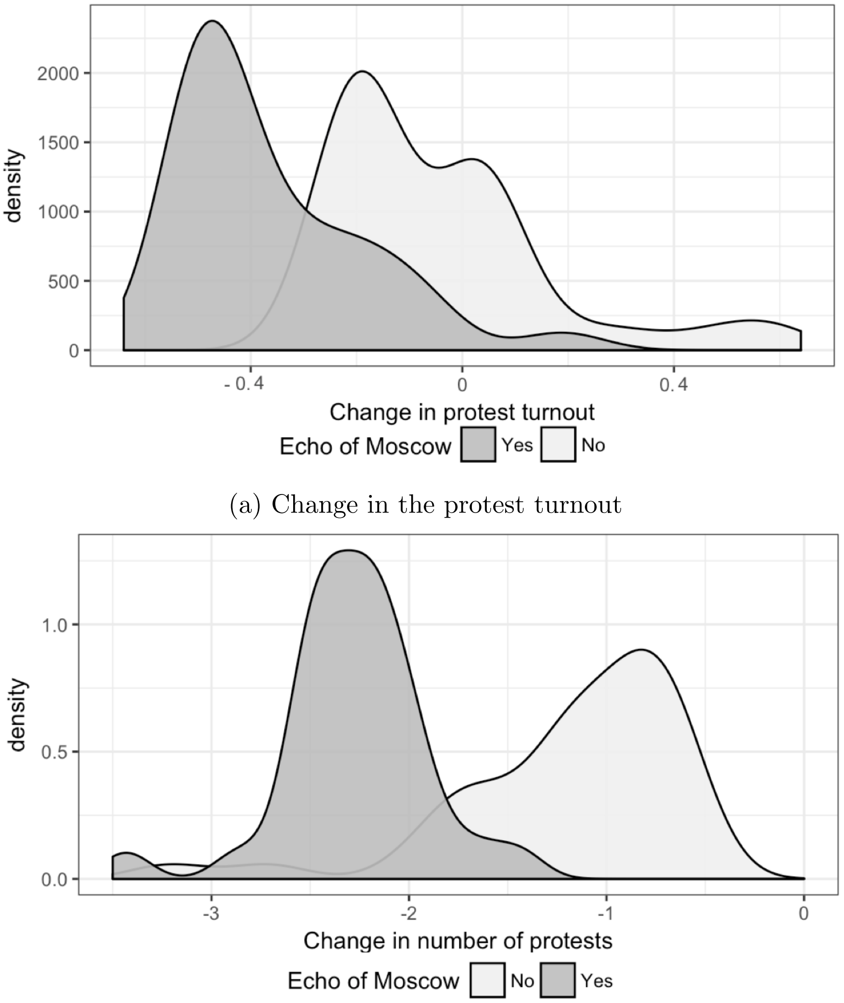
Table 2.3: Effect of local availability of *Echo of Moscow* on change *in protest activity*

protest turnout and number of protests decreased on average by .21 participants per thousand citizens and by 1.5 protests, respectively. In capitols exposed to *Echo of Moscow,* turnout and number of protest decreased, on average, by .57 (-.36 + -.21) participants per thousand citizens and by 3 (-1.5 + -1.5) protests, respectively.

Figure 2.4 illustrates the patterns of change for both number of protests and protest turnout in adjusted sample.

2.5.3 Threats to validity

Censored or missing data. The size of the effect can be influenced by the fact that some regional authorities stopped issuing permits for demonstrations in the post-treatment period. This could explain both missing protest events and cases of extreme changes in the size of protest. In these cases, protesters either didn't take to the street at all, or constituted a small proportion of radicals.



(a) Change in the protest turnout

(b) Change in the number of protests

Figure 2.4: Protest change

The extreme example is Yaroslavl Oblast, where the size of the protest dropped from 1500 to 50 citizens. I account for this possibility by investigating cases of a suspiciously drastic change in the attendance rate. In 8 out of 11 cases the permission to hold a meeting was granted by the authorities, but in 3 of them it was common knowledge that officials tried to prevent the meeting by choosing an inadequate location or date for it. Results presented in section 5.2 do not change after omission of these 3 cases.

Regression to the mean. The fact that before the pro-government rally on February 4th the protest dynamics in both cities with and without exposure to *Echo of Moscow* followed parallel trends (see fig. 2.2) suggests that regression to the mean did not take place and did not account for the results of analysis.

2.6 Conclusion

This paper suggests that exposure to independent (partially-free) media can have a demobilizing effect on dissidents if combined with the aggressive tactics of building an image of overwhelming support for the autocrat. It is commonly believed that the existence of independent media increases the ability of opposition to mobilize. But if the incumbent has enough financial resources and organizational capacities to launch large-scale rallies of supporters, then independent media reports can reduce the number of mobilized protesters.

Data from recent protests in Russia suggests that in capitol cities exposed to *Echo of Moscow* radio station, the number of protests and protest turnout decreased more than in the rest of the capitol cities following the major pro-government rally in Moscow. The results of this study suggest that in cities with no exposure to *Echo of Moscow* protest turnout and number of protest on average decreased by .21 participant per thousand citizens and by 1.5 protests, respectively. In capitols exposed to *Echo of Moscow* turnout and number of protest on average decreased by .57 participant per thousand citizens and by 3 protests, respectively.

Note, that the results do not show that incumbents strategically allow independent media to exist. They only show that under certain conditions, greater media freedom can enable autocrats to forestall anti-regime collective action more effectively. At the same time, it is likely that the actual level of media freedom is likely to be endogenous to the incumbent's perception of the risk of being overthrown (VonDoepp and Young, 2012; Guriev and Treisman, 2015*a*) and to the strength of his regime (Geddes and Zaller, 1989; Stein, 2012), and, thus, to the incumbent's popularity.

CHAPTER 3

How to catch a troll: detection of paid political commentators on social media

3.1 Introduction

In the first two decades of the 21st century, online technologies have become a significant part of a nation's political life. Both in democracies and autocracies, the Internet has been actively used to mobilize public support. This should not be a surprise. By the end of the century's second decade, people around the globe mostly consume news and political information via the world wide web (Ahmed and Cho, 2019; Santana and Dozier, 2019; Wang and Loo, 2019). Though online technologies do not only improve and ease information-sharing and social mobilization. Politicians actively use these new channels of communication at their benefit.

For instance, the international community suspected some authoritarian governments of employing online commentators (trolls) against their domestic and foreign competitors. Politicians can use trolls in many ways: from distracting users' from public policy failures to threatening political opposition or media (Roberts, 2018; Zannettou et al., 2019). Companies, democratic governments, and academics have been working hard trying to detect paid political commentators on the Internet. Why is this task important?

Social media platforms try to prevent consumer churn by banning accounts that produce malicious or misleading information (Broniatowski et al., 2018).

Democratic governments try to protect national elections from foreign influences (Faris, 2019). Academics work on this task for several reasons. They include: studying paid political behaviors (their means, scope, goals, and targets) and testing hypotheses about their potential influence. In general, our capacity to recognize paid political agents is essential for understanding how mechanisms of social coordination and information control work in the 21st century.

To assess the scope of trolls' activity and potential impact on users and voters, scholars of politics and policy experts need a toolkit to recognize them. Most of the studies in this field focus on potentially suspicious accounts of social media. In general, researchers look for commonalities among paid online commentators. In contrast, this is the first study to my knowledge that explicitly compares the behavioral patterns of trolls with regular users on social media.

I focus on the case of Vladimir Putin's regime in Russia and on the behavior of several hundred Internet trolls who published pro-government posts and comments on social media platforms during 2014-15. The data include leaked documents from Russia's so-called *Troll Factory* and interviews with journalists and experts. Most importantly, following the publication of these leaked documents, I collected a complete body of the text of almost 357,604 trolls' posts on social media platform *LiveJournal* (*LJ*). I complement this collection with a random sample of four thousand *LJ* users who published 464,678 posts during the same time.

I approach the task of catching paid online commentators in three steps.

First, I consider situations where a regular user meets an account on social media and suspects if it belongs to a troll. Building on the previous studies, I assume that regular users only pay attention to the textual content of suspicious accounts. I approximate how a regular user recognizes troll accounts by estimating the topical profiles of both users and trolls. I find it unlikely for a regular user to identify paid political commentators successfully.

Second, I consider this task from the social science perspective. A researcher has more opportunities to find distinctions between trolls and regular users. I ask if her ability to observe macro patterns in behaviors of two groups helps to identify essential differences between them. I conclude that the social scientist should be able to detect a significant portion of such differences. At the same time, the statistical models that only use such predictors have a low predictive capacity.

Finally, I provide a reverse-engineering perspective on this task. I train a set of classification models to distinguish between random users and trolls. I then use a group of metrics to select the model with the best performance. In the end, I determine the most important predictors of the troll accounts by calculating their variance importance factor.

The paper attempts to make several contributions.

First, it shows that, in contrast to conventional wisdom, paid online commentators do not always spend most of their working hours by writing posts about politics. In the case of Vladimir Putin's Russia, most of the posts published by trolls accounts did not have any political content.

Second, the results suggest that trolls mimic regular users. They calibrate their behavior by adjusting the topical profiles of their accounts in the direction of profiles of regular users of social media. I suggest that trolls try to present themselves as regular users to build a reputation on social media. First, they need such a reputation to conceal their actual status and to avoid digital platform aggregators from banning their accounts. Second, trolls use this reputation to engage in online discussions actively. Trolls must stay undetected to influence such conversations.

Third, the paper contributes to the debate on malicious misinformation and regulation of digital platforms. Companies, governments, and academics develop sophisticated methods to detect state-sponsored political commentators on the

Internet. Most of these methods are based on a combination of arbitrarily chosen criteria, often including the country of origin of the account's email address or phone number, usage of specific characters (e.g., Cyrillic alphabets), and specific keywords in the message. I show that such methods may be unable to identify a significant proportion of paid political commentators. They are aware of the risks and try hard to hide their troll identity. At the same time, the results suggest that leaked data on troll accounts can be successfully used for training classification models.

Finally, online behavior is the focus of this paper, but it talks to a much broader conversation on how autocrats manage political dissent. In response to increased transparency of the digital era, authoritarian leaders tend to use tools of information control that are hard to observe. While using trolls can be less efficient than censorship or propaganda, it also has the smallest risks to be detected and to backfire.

The remainder is organized as follows. The next section reviews existing approaches in detection of paid online commentators. Section 3 describes data-collection and data-processing steps in the study. Section 4 introduces the results. Section 5 concludes.

3.2 Identification of paid online agents: a review of existing studies

While paid online agents became famous, mostly due to political scandals, the problem of malicious misinformation was raised by the operators of digital online platforms decades ago. Commercial companies such as *Amazon*, *eBay,* and *Yelp* have been facing a problem of fraudulent reviews from the beginning of their foundation (Luca and Zervas, 2016). A typical internet troll is a social media user who deliberately tries to offend others by posting specific comments, photos, videos, or the other forms of online content. In contrast, review trolls are much harder to identify as they try to mimic the behavior of real users. Governments recruit paid trolls for a wide range of political purposes. They can follow different patterns of online behavior from sending inflammatory messages to masking their troll identity. Modern studies of online misinformation distinguish supervised and unsupervised methods of identification of paid online agents.

Unsupervised methods of malicious misinformation detection suggest that a researcher should pre-specify critical characteristics of such behavior and make a guess whether the user's behavior expresses these characteristics. For example, during the 2016 U.S. presidential campaign in the U.S., *Twitter* used the following criteria to determine whether *Trolls Factory* in Saint-Petersburg operated a suspicious account. First: whether the account was registered in Russia. Second: whether the user created his account with a Russian phone carrier or a Russian email address. Third: whether the user's display name contains Cyrillic characters. Forth: whether the user frequently tweets in Russian. Fifth: and whether the user has logged in from any Russian IP address. The effectiveness of criteria-based approaches was questioned in several contexts (Nikiporets-Takigawa, 2013). Indeed, the absence of the ground truth does not allow one to estimate the performance of such classification algorithms. For example, in 2017, *Twitter* appeared in the center of the scandal; the majority of the social media users in Bulgaria were identified as trolls and were banned from the platform for a short time (Persily, 2017).

Supervised methods of malicious misinformation detection require researchers to have information on the true identity of a social media account. The latter is possible in cases where paid agents do not mask their identity. For example, Munger et al. (2015) find that, in Venezuela, public officials actively tweeted non-political messages to shift the public agenda and to reduce the share of dissidents tweeting about the impending protest events of 2014. Multiple intelligence leaks have extended the applications of machine learning for paid agents detection. Keller et al. (2017) report that during the South Korean 2012 presidential race, the National Intelligence Service actively used more than one hundred accounts on *Twitter* to wage a campaign in favor of the eventual winner. Moreover, they identify three different groups of accounts that targeted specific social media audiences. Miller (2017) investigates how regional administrations monitor "public opinion emergencies" and use paid trolls to alter the public perception of the authorities. King, Pan and Roberts (2017) study pro-government commentators in the Chinese blogosphere and find that those paid bloggers spent time celebrating different aspects of Chinese social life while not necessarily engaging in a political debate. As one can see the studies described above look at the behavioral patterns specific to paid online accounts. At the same time, they do not ask another important question: what is a systematic difference between troll and non-troll accounts?

Sanovich, Stukal and Tucker (2017) use a combination of supervised and unsupervised methods to build their classification model of "political bots" (automated scripts) on *Twitter* in Russia. At the first step, authors sampled *Twitter* users and manually labeled them according to their classification scheme.Next stage, they trained their prediction model and applied it to a test sample of social media users on *Twitter.* The main finding of the paper suggests that paid political bots generate the majority of the political content on *Twitter* in Russia. Meanwhile, this result should be taken cautiously as the validity of "selected features of political bots" cannot be tested.

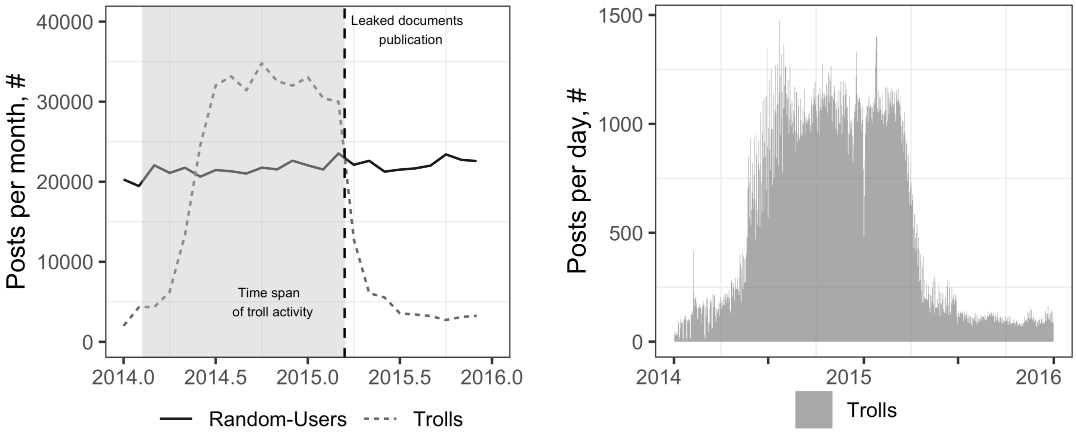
This paper develops a set of classification models to detect paid troll accounts. It attempts to improve the current understanding of the behavior of paid online agents. To do it, I first use documents leaked by investigative reporters of independent outlet *Novaya Gazeta* from the so-called *Troll Factory* in Saint Petersburg. In March 2015, *Novaya Gazeta* published a list of account names. Owners of these accounts had been tasked with actively leaving comments on the blog platform *LiveJournal.* Follow-up investigations showed the existence of a vast industry of paid commentators in Russia. These trolls might not only be engaged in fighting political opposition in Russia but also may be operating against other countries. Among the most famous ones was a promotion of fake news about a severe explosion at a processing plant in Louisiana. I use the list of account names from leaked documents as a "ground truth." Second, instead of asking about the homogeneity of the behavioral patterns within the group of trolls, I explore the difference between random accounts and troll accounts. This approach assumes the absence of the troll accounts in random accounts. In case this assumption is violated, the predictive accuracy of the resulting models should be treated as a conservative one.

3.3 Data

3.3.1 Data collection

Following the publication of the list of paid commentators on *LJ*, I collected all the posts published by their accounts (357,604 posts in total). Most of these posts were dated back to 2014 and 2015. It is noteworthy that trolls registered publicly-opened accounts so that their posts would be available for news aggregators and regular social media users. I complemented this collection with a random sample of two thousand *LJ* users who published 464,678 posts during the same time period.

The temporal perspective on the data sheds light on several important details (see Figures 3.1a and 3.1b). Most owners registered their accounts at the end of 2013. Their activity level remained low at the beginning of 2014, but, in March, it suddenly intensified to ten times higher than previously (going from 2,900 posts in January to 33,000 posts in March). It then remained stable for a year. Owners of accounts disappeared the day after the newspaper published leaked documents. At the same time, around 3 percent of troll accounts continued publishing posts until 2016. I explain this behavior as follows. The digital platform allows users to



(a) Months

(b) Days

Figure 3.1: Activity of social media accounts

schedule the publication dates of their posts. Employees of the *Troll Factory* were required to write a certain number of posts and comments daily, and some were apparently making composing posts in advance. These patterns sharply contrast with the behavior of regular users. The latter stayed stable throughout this time. Moreover, their accounts were not affected by the release of the troll list. These details suggest that those accounts leaked and published by *Novaya Gazeta* represented a group of users that was very distinct from the overall population of users on the digital platform.

Figure3.1b also makes apparent another essential detail. The intensity of published posts by troll accounts remained stable until the release of the journalist investigation. However, one can see several spikes, i.e., days when paid commentators published significantly more posts than average. A closer look at the content of these posts reveals that they coincide with aggravations of the political conflict between Ukraine and Russia. Apart from this conflict, 2014 and 2015 represent an economic stagnation period marked by declining oil prices, rising food and consumer goods prices, and intensive government propaganda. That is what a regular social media user could expect to read about in the content generated by troll accounts. I restricted the analysis to the period when troll

Indicator	Trolls	Regular Users
Posts, # (total)	357,604	464,678
Posts, # (average)	164	99
Posts per day, # (average)	3.5	1.1
Comments to posts, # (total)	9,110,665	8,588,032
Comments to posts, # (average)	26	19
Troll comments to posts, # (total)	197,670	50,406
Troll comments to posts, # (average)	3.4	0.12

Table 3.1: Descriptive statistics

accounts were active, i.e., from 2014 to the beginning of 2015 (see the shaded area on Figure 3.1).

Table 3.1displays additional indicators that can be used to distinguish troll accounts from regular user accounts. As can be seen, the average troll account published 64% more posts than the average regular user (164 and 99 posts, respectively). Moreover, on an average day, trolls published almost four times more posts than other social media users (3.5 and 1.1, respectively). This difference is to be expected. Indeed, regular users published posts on their own, whereas paid commentators were required to write a fixed number of posts during their workday.

It is noteworthy that, on average, troll posts attracted more comments than those published by regular users (26 and 19, respectively).[1] Although the overall number of comments is higher for troll posts, a significant number of them either belonged to other trolls or to the author herself. Indeed, the average troll post attracted more than three the number of comments from other trolls. Why would trolls comment on each other's posts? There are two complementary explanations. First, by commenting on each other, trolls increase the relative popularity

[1] I exclude the author's comments from the calculation

of posts they publish, thus improving the chances that news and social media aggregators identify the topics of these posts and assign them a relatively higher rank in the daily news ratings. The latter could have two effects, thus possibly increasing the chances that social media users would read troll promoted story. It also might force out other topics (including politically sensitive ones) from the top of the ratings, decreasing the chance of the further diffusion and discussion of undesirable news. Second, trolls might comment on each other's posts in order to build a reputation in the eyes of regular users. Indeed, the fact that a published post attracts many comments sends a signal that its author is a credible social media user.

Finally, on average, trolls left 0.12 comments per post published by a regular user. Thus if trolls left just one comment per post, they would target one post in ten. Meanwhile, the trolls rarely left a sole comment on the targeted post; such posts attracted 3.8 troll comments per post. To summarize, 3 percent of randomly sampled posts had comments left by trolls.

The next section describes how the available information was processed for further analysis.

3.3.2 Data processing

Although this paper studies specific patterns of user behavior, most of the data were collected at post level. Below I discuss data processing details. Each post has the following features: text of the post, date and time of posting, author's name, and the text of comments to the post and the account names of the users who left them. For further analysis, I processed the data at post level and aggregated quantities of interest at the account-level. For example, using the features above, I calculated the following quantities of interest: average length of a post, average time of posting, average number of posts, topics of the posts, etc.

While a researcher can efficiently operate with these quantities, it is unreasonable to expect that an ordinary user of social media would be able to spend a significant amount of time and effort to collect all available data on a suspicious account. What is reasonable to expect is that the overall content, i.e., the typical topics of a user's posts and their topical composition, could appear suspicious. In other words, a user should be able to ask the following question: "What does the owner of a suspicious account write about?"

In this study, I use topic modeling to approximate individual perception of the content of the posts published by suspicious accounts. The standard topic model algorithm describes each text as a mixture of topics. Each topic represents a linear combination of words frequently used together across all the documents of the corpus (Blei, Ng and Jordan, 2003; Roberts et al., 2014).

Dimensionality reduction. Topic models show robust and stable results on the corpora of texts which use relatively similar vocabulary. The accuracy and robustness of the topic modeling algorithms decreases with the rise in unique words used in the corpus of a document. This is an example of the "dimensionality curse" (Asuncion et al., 2009). To mitigate this problem, scholars conduct several preprocessing steps involving the raw collection of words in a corpus. These steps usually include tokenization, lemmatization, stemming, and elimination of stop-words. Tokenization includes splitting the text into words, lowercasing, and removing punctuation. Lemmatization suggests that words in the third person are changed to the first person, and verbs in the past and future tenses are changed to present tense. Stemming is a reduction of words to their root form. Researchers conduct lemmatization and stemming by using dictionaries. While these two steps are considered crucial, they often fail to succeed when dealing with corpora generated by Internet users (Liu et al., 2016) because of two problems. First, individuals make typos, and there are no editors to proofread the texts they post. Second, a social media user often intentionally changes the form of the word to stress her attitudes toward the discussed subject. For example, skeptics of the Russian government often write Vladimir Putin's last name as "Puten." Existing

Operation	Standard algorithm	Alternative algorithm
	Unique words in the corpus	
Raw count	2,862,893	
Tokenization	1,802,186	
Lemmatization	953,254	
Stemming	774,201	
Morphological prediction		178,414
Stop-words elimination	772,631	176,844

Table 3.2: Dimensionality reduction of the corpus of posts

Note: the list of stop-words is provided in additional material.

dictionaries do not account for intentional and unintentional typos and so often fail to solve the "dimensionality curse" problem effectively.

I employ an alternative strategy when dealing with the dimensionality of a corpus of social media posts. After conducting tokenization, I input the words into the morphological analyzer *Mystem* developed by Yandex, a Russian analog of Google. This analyzer is trained to guess the original form of the word using typos made by Yandex users in their search queries.

Table 3.2 compares the performance of the standard approach with an approach developed in this paper. While the standard procedure reduced the unique words in the corpus of posts by a factor of 4, the alternative algorithm reduced the initial count by a factor of 16. The resulting corpus of words produced by the latter appeared to be more than 4.5 times smaller than the one provided by the former (772,631 and 176,844, respectively). By construction, the improvement in solving the dimensionality problem increased the stability and robustness of the topic modeling algorithm that I used in the next step of data processing.

The optimal number of topics. Topic modeling assumes a fixed amount of topics pre-specified by the researcher. However, the actual number of such topics

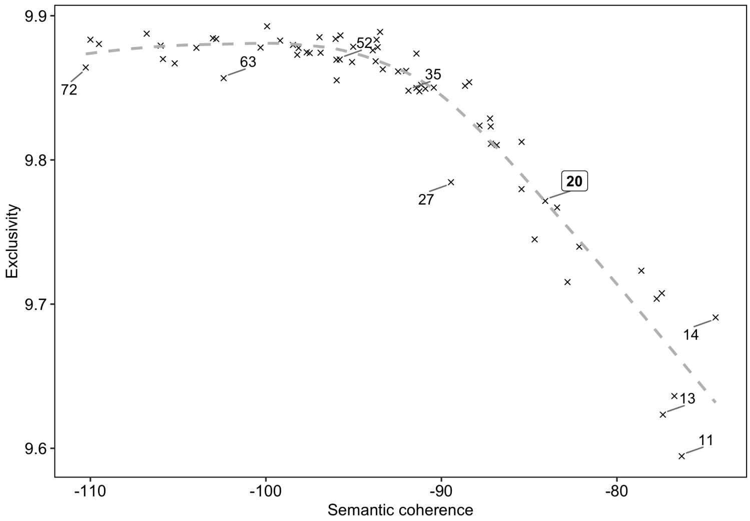


Figure 3.2: Models with different number of topics

is always unknown. There is also no best answer to the number of topics that are appropriate for a given collection of texts. Several studies do report sets of metrics that can aid a researcher in selecting this number Wallach et al. (2009); Mimno et al. (2011); Taddy (2012); Roberts et al. (2014). These include exclusivity, semantic coherence, residuals, and held-out likelihood. Here I employed the first two metrics. Bischof and Airoldi (2012)define exclusivity as the relative rate at which the most frequently used words on a given topic are used in other topics. The aggregate exclusivity index is high when most of the identified topics represent combinations of words significantly distinct from each other. Complete exclusivity is achieved if each topic is represented by a unique combination of words.

I used exclusivity and semantic coherence scores, as recent studies have found that they are correlated with the semantic quality of a topic as judged by human annotators. Thus, finding the topics using these metrics should deliver the best possible approximation of human judgment without individual evaluations or external reference corpora Mimno et al. (2011). Semantic coherence is maximized when the most probable words in a given topic frequently co-occur in the same documents. At the same time, extremely high semantic coherence reflects that ubiquitous words dominate estimated topics and have low exclusivity. Thus, the choice between exclusivity and semantic coherence represents a trade-off.

I solved this trade-off by finding a model where a marginal change in exclusivity equaled a marginal change in semantic coherence. First, I estimated a set of models while varying the number of topics from 5 to 75. Next, I calculated the metrics of interest for each of the models. Finally, I plotted the resulting curve. Figure 3.2 shows that models with 19 to 20 topics cameclose to satisfying the optimization criteria, while the model with 20 topics was the closest one. I employed this model in further. analysis

Topics description. Topic models do not supply the user with substantive names of the topics. Instead, the researcher is supposed to label a topic using the most critical (or frequently used) words that represent it. For example, one of the estimated topics is represented by the words *artist, art, exhibition, create, drawing, draw, portrait, painting, picasso, work.* I label this topic "Fine arts." Table 3.3 shows three of the most frequently used words in each topic along with their labels. One can see that six of the topics are politically sensitive. These include the "National economy," "Prices," "Conflict with Ukraine," "International affairs," "War," and "Rule of Law." I classified these words into two groups: "National economy," and "Prices" comprised the first group, which I called "Economic issues," and the other four topics comprise the second one, which I entitled "Political issues." Further, I described these topics in detail (A complete description of all topics is provided in additional material).

Recall that the collected data consist of posts from 2014 and early 2015. In Russia, this was a period of political conflict, with Ukraine related to the Russian annexation of Crimea and the military conflict in the eastern regions of Ukraine. The conflict provoked western countries (including the U.S.) to impose economic sanctions on Russia. The latter coincided with ongoing economic stagnation, declining oil prices, rising food, and consumer goods prices. In line with expectations, the "Economic issues" group of topics reflects these socio-economic

Topic	Most frequent words	Topic	Most frequent words
National economy	economy, Putin, sanction	Health	doctor, medicine, help
Prices	price, salary, cost	Family	child, man, woman
Conflict with Ukraine	Ukraine, Crimea, war	Trip	trip, flight, camera
International affairs	Europe, America, refugee	Fine arts	artist, paint, exhibit
War	war, military, army	Beauty	skin, makeup, dress
Rule of law	law, convicted, citizen	Fiction	book, read, author
History	Russia, emperor, tsar	Mode of life	sleep, smoke, home
Cooking	oil, add, taste	Blogging	post, blogger, repost
Movies	film, role, play	Hello world	good, morning, friend
Tour	city, tour, museum	Travel	road, sea, place

Table 3.3: Identified topics in the social media posts

Note: see the full description of topics in the Additional material

processes. The "National economy" topic was characterized by such words as *economy, billion, oil, sanction, investment, state.* Posts where this topic prevails mostly discuss the poor performance of the economy, its dependence on the export of natural resources, and the negative consequences of economic sanctions. They also often mentioned Russian authorities (*president*, *Putin*, *bureaucracy, central bank*) and state monopolies (such as *Gazprom*). While the "National economy" topic often related problems with the national economy to government policies, posts with the prevailing topic "Prices" mostly consisted of personal complaints about harsh economic conditions of life in Russia. In general, this topic mostly excluded the discussion of macroeconomic indicators and public policies. Instead, it is represented by such words as *euro, money, ruble, price, buy, work, salary, cost, cash.* The authors who write about this topic rarely mention the national government or specific politicians in relation to personal problems. At the same time, they often express their concerns about the uncertainty of their economic well-being in the future and discuss ways to preserve their financial savings.

"Political issues" represented a more diverse group of topics than "Economic issues." One of them ("Rule of law") focused on domestic politics, whereas three others focused on international agenda. Two topics ("War" and "Conflict with Ukraine") refer to military issues, while two others ("International affairs" and "Rule of law") discuss humanitarian and civic problems. "Conflict with Ukraine" accurately caught the essential issues in Russia-Ukraine relations between 2014 and 2015. Apart from references to *Ukraine*, *Russia*, *conflict*, and, *war,* the keyword combinations of the topic almost entirely related to epicenters of the conflict in the East of Ukraine (*Donbass, Donetsk, Debaltseve*) and the name of major Ukrainian politicians (*Poroshenko, Yanukovych, Saakashvili, Akhmetov*). The topic "War" referred to a general discussion of military strategy and military conflicts of the past. Its word representations included *army, military, soldier, officer, general, lieutenant, prisoner, tank.* It is noteworthy that social media users who engaged with this topic rarely discussed specific historical episodes or related their speculations to contemporary conflicts. A significant portion of these discussions compared differences in strategy and military tactics of the Soviet Union and Germany during WWII. Posts with this topic often used such words as *Soviet, German, USSR.* Topic "Rule of law" focused on the domestic political agenda and mostly consisted of stories about non-governmental organizations, human rights activism, and criminal convictions in the regions and cities of Russia. The authors of this topic often linked existing problems of the civil society with local politics, including elections. The following words characterized this topic: *law, citizen, organization, region, federation, criminal, trial, conviction, elections, deputy, chairman, police.* Topic "International relations" caught the actual episodes of politics in Europe and the Middle East. It usually referred to regions of an ongoing armed conflict, while avoiding discussion of its military aspect. Instead, posts on this topic discussed problems of refugees and the influx of migration from the Middle East to European countries. The most frequently used words on this topic included *refugee, migrant, resident, Arabic,*

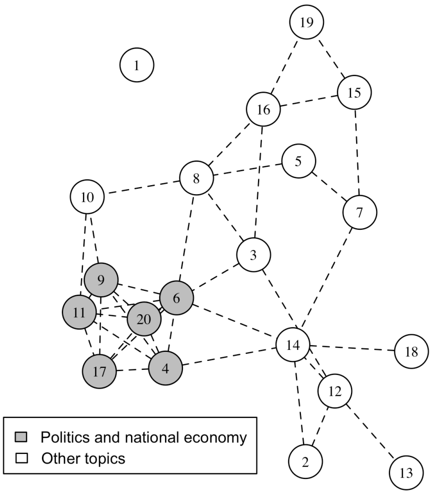


Figure 3.3: Graphical display of topic correlations

Note: 4 - Prices, 6 - War, 9 - National economy, 11 - Rule of law, 17 - Conflict with Ukraine, 20 - International Affairs.

Muslim, Europe, America, Turkey, Syrian. In general, the identified set of political topics displayed a significant overlap with the actual agenda of media in Russia (Yablokov, 2015; Field et al., 2018).

Both groups of topics described above should be of particular interest to the researcher. Indeed, previous studies have found paid online commentators to engage with this kind of informational content (Gunitsky, 2015; Stukal et al., 2017; Tucker et al., 2018). Interestingly, the majority of the identified topics had no relation with political or economic issues (14 out of 20). Instead, most of them referred to personal interests, hobbies, and lifestyles (their detailed description is provided in the additional material). One explanation suggests that troll accounts try to promote specific political topics in favor of the government while a typical social media user is primarily concerned with individual problems or entertainment content. Indeed, identified topics include traveling, movies, fine arts, cooking, etc. I check this idea in Section 3.4.

The credibility of the estimation. How credible are the estimated topics? Several pieces of indirect evidence justify whether it is reasonable to trust the results of the topic modeling. First, identified topics are consistent with common beliefs about actual events and socio-economic processes in Russia (Yablokov, 2015; Field et al., 2018). Second, apart from this consistency, metrics of exclusivity and semantic coherence provide additional evidence. For example, while both "National economy" and "Prices" are explicitly related to "Economic issues," these two topics are comprised of entirely different combinations of words and thus are highly exclusive from one another. Take another example. Do the topics "War" and "Conflict with Ukraine" display considerable overlap? From the descriptions provided above, one can see that these two topics describe distinct events and thus display very different words. Topic "War" has no reference to any ongoing military conflict and primarily describes details of military tactics of nations during WWII. In contrast, though sharing few words with "War," topic "Conflict with Ukraine" almost entirely consists of words that describe geographic and political contexts of Russia-Ukraine relations.

One could also check whether different topics co-exist in the same publications. Figure 3.3 represents the topical correlations in the corpus of social media posts. In this context, topics are correlated if they frequently appear in the same posts. While topics about politics and economy in Russia are highly exclusive and semantically coherent, it is interesting to note that these topics correlate with each other (see the full mapping from topics to their indices in additional material). In other words, if a post covers the negative consequence of sanctions for the Russian economy, there is a relatively small chance that another politically sensitive topic such as human rights violations would be covered as well. At the same time, social media users in the sample did not mix political and non-political topics. This fact represents additional indirect evidence of the credibility of the results of the data processing I conducted.

The next section describes the results of this study.

3.4 Results

3.4.1 Internet-user perspective: Can a social media consumer identify a paid troll by looking at his account?

The existing literature on media consumption helps to formulate expectations about the behavior of Internet users. Typically, media consumers are overwhelmed with available information and so usually have very minimal attention to focus on particular pieces of information (Conover and Feldman, 1984; Popkin, 1994; Hamilton, 2004). In her decision to put some effort in collecting specific information, a consumer relies on the expected costs and benefits the data can provide (Lupia et al., 1998). Even if consumers realize the benefits of collecting accurate information, they usually have a hard time evaluating collected information and making inferences from it. Zaller (1992, p.18) summarizes: media consumers are, for the most part, rationally ignorant and spend little time investing in information. Indeed, imagine that a social media user would like to investigate a suspicious account and determine if it belongs to a troll. Even if her analysis provides a correct conclusion, it is unlikely that this knowledge provides her with any non-negligible benefit.

In line with the studies on media consumption, Roberts (2018) finds social media users to have a highly elastic demand for collecting online information. In the case of political news, she finds that small increases in its cost sharply decrease the probability that the user actually consumes it. For example, the Chinese government reduced the national traffic on Google by 80% just by throttling access to the company's services. In other words, users stopped using Google's services because their first attempt to access Google's website failed one time out of four.

Following the results of the previous studies, I suggest several expectations about the behavior of social media users.

First, users do not collect all available data on a suspicious account. Most digital platforms allow collecting many details of a specific account (text of posts, their date, time, length, any comments to them, and so on). I assume that a user of social media does not put the effort into collecting these details. Still, instead, she pays attention to the most easily observable parts of the account (i.e., texts of its posts) and uses them to form a general impression of their author. Moreover, this user could be expected to move to the next post once she has identified the primary topic of the current post. In other words, by reviewing the suspicious account, she should be able to describe what topics its owner usually covers in his posts. She also should be able to establish if specific topics dominate most of these posts.

Second, in contrast to researchers, the regular user does not conduct statistical tests. Indeed, while she can notice if a topic dominates the content of the account's posts, it is highly unlikely that she can compare topical profiles of other users in twenty-dimensional space or conduct tests of their distributional equivalence.

Finally, I expect the regular user to know the actual average topical composition in the population of non-troll accounts. Specifically, she would be expected to see the set of topics and their respective shares in the posts of an average non-troll account. This final expectation is not genuinely realistic. It suggests that a hypothetical social media user has more information and is less myopic than one can imagine given the existing evidence across the globe (Pfeffer, Zorbach and Carley, 2014; Roberts, 2018). It also suggests the presented results to be conservative. Thus if this study finds users to be incapable of identifying trolls given the provided expectations, one should be even more confident regarding the inability of regular users to determine if a suspicious account belongs to a troll in actual cases.

Distribution of topics. I start my discussion of the analysis by comparing the topical profiles of an average troll account with the account of an average *LJ*

	Political issues	Economic issues	Other topics	Total
Trolls	38	7	55	100
Random users	36	5	59	100

Table 3.4: Groups of topics, %

user. Table 3.4 reports the relative shares of each group of topics. First, one can see that the aggregated topical composition of averaged accounts is very similar. It is surprising given the common belief that paid online commentators produce content that is very distinct from that of regular users (e.g., posts referencing specific topics or persona). Second, regular users publish a lot of content related to politics. More than one-third of published information discussed the topics of the "Political issuers" group. The latter is significantly more than one could expect, given the existing studies of Internet behavior and media consumption. Third, troll accounts do not only publish posts on politics or the economy. Indeed, posts that cover political and economic issues constitute a minor part of the content generated by troll accounts. Why would that be the case? A potential explanation suggests that while troll and non-troll profiles can look similar at the aggregate level, they produce distinct content within each of the three groups of topics. I investigate this possibility below.

Figure 3.4 depicts the distribution of the dominant topics of posts for the topic group and the type of social media account. Three out of four topics in the "Political issues" group ("War," "Rule of law," and "International politics") have almost identical coverage between troll accounts and those of regular users (see Figure 3.4a). It is noteworthy that neither ordinary users nor trolls spend their time posting on the issues of international relations. The share of such posts in the total amount of publications is around 1 percent in both groups. The last topic of this group ("Conflict with Ukraine") does not fit the pattern of the rest. Both groups publish a significant amount of posts that cover the topic. At the

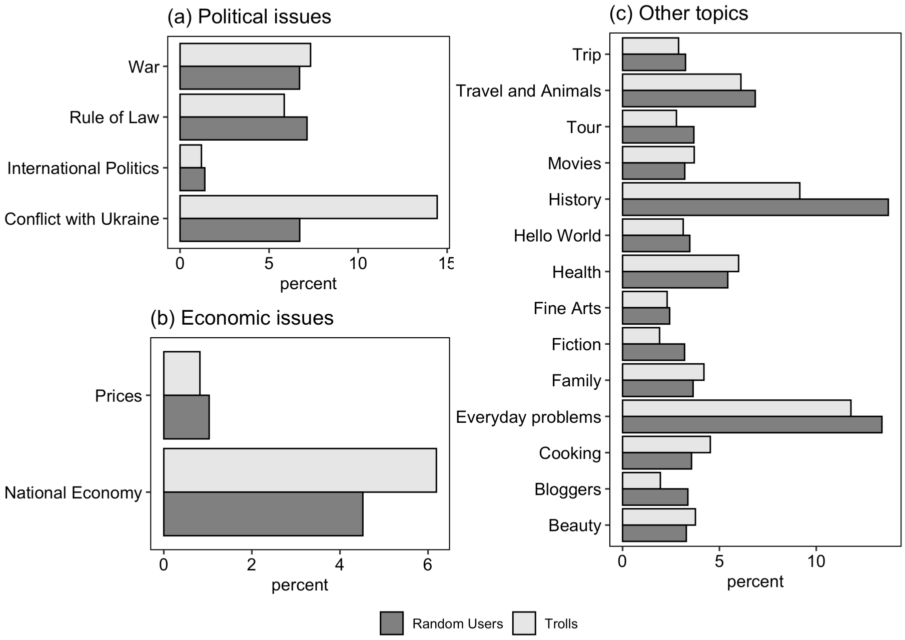


Figure 3.4: Distribution of dominant topics of posts, %

same time, the relative share of posts on this topic for troll accounts is twice the size of this share for non-trolls. By this point, that is the only evidence that fits expectations formed by the previous studies. Figure 3.4b describes the topical distribution within the "Economic issues" group. While both types of accounts equally cover the topic "Prices," trolls appear to publish more posts about the national economy than regular users (6% and 4.4%, respectively). Finally, Figure 3.4c shows the percentages for "Other topics." It is extremely noteworthy how close trolls are to regular users in covering these topics. The topic "History" constitutes the only noticeable difference. This topic could be referred to as both groups ("Other topics" and "Political issues") although I classified it as in the former. Indeed, authors of these posts rarely link the described events to the current political context of Russia or promoted specific ideological messages. It is also interesting that regular users published more posts on this topic than trolls (14% and 8%, respectively).

To summarize, while there is a difference in the proportion of content that covers the topic "Conflict with Ukraine," the overall topical composition of profiles is indeed very similar. Given the expectations described at the beginning of this paragraph, I conclude that a regular user would have a hard time trying to determine if a suspicious account does, indeed, belong to a paid online commentator.

One could also argue that, although trolls and regular users published posts on political or economic issues at the same rate, they might cover these topics from different ideological positions. Thus, a regular user could check if a suspicious account belonged to a paid commentator by looking at the specific words that this commentator used. For example, posts of regular users could be more critical toward the Russian government and could contain such words as *annexation, aggression, cost.* At the same time, authors who supported the national government could describe the case as the failure of the federal government of Ukraine or as a conspiracy against Russia organized by the West (Kuzio, 2018; Borenstein, 2019). Such posts would frequently use the following words: *USA, NATO, plot, Georgia, Saakashvili.* One would then expect the topic model algorithm to identify two different topics that reference the same events. This appears to not to be true, however. Indeed, even models with more than 20 topics did not identify such cases.

Non-topical features. Other pieces of information on specific posts are available to the regular user, and these were listed previously. These include the date, time, and length of a post, comments made by other users, etc. In contrast to the overall content of a specific blog, these are much harder to collect and analyze. It is also important to note that, even if a regular user would be provided with such data, there is no explicit expectation that it could help her to distinguish trolls from users. Should a troll publish a lot of posts per day? Should such posts be particularly lengthy? Does it matter if others comment on such posts? It is unclear how such data could help a typical user to classify accounts without knowing the actual account types. At the same time, a researcher who knows the types of some social media accounts can make use of such data, and I describe this further in the next two sections.

3.4.2 Social science perspective: Global patterns of troll behavior

A researcher has more opportunities to study distinctions between trolls and non-trolls than does a regular user of social media. These include his ability to calculate sophisticated metrics and to conduct statistical tests. The most crucial opportunity comes from the information on the actual type of accounts in the dataset. The latter, of course, depends on the assumption that no paid online commentators appeared in the group of randomly sampled users. Indeed, imagine that most of the accounts on *LJ,* in fact, belong to trolls. In this case, randomly sampled accounts would also belong to trolls. There would then be no difference in the behavioral patterns of the two groups, and no results of that study would be considered credible.

For further analysis, I assumed that the actual share of trolls in the total population of *LJ* accounts was negligible. The latter is partially corroborated by the fact that most investigations reported the total number of troll accounts on a specific platform (e.g., *Facebook*, *Twitter*, *VK.com*, or *LJ*) as limited to several hundred (Broniatowski et al., 2018). At the same time, *LJ* has around 40 million registered users, with 50 percent of its traffic generated by Russian users. To conduct the comparison, I also assumed that an account that was randomly drawn from the population of all Cyrillic *LJ* accounts did not belong to a troll. Though unlikely, it is still possible that trolls constitute not a negligible but rather a minor fraction of accounts on that digital platform. If that were the case, then the resulting comparison should return conservative results about the existing differences between the two groups of users in the sample.

Concerning the content of the posts published on the Internet, the researcher can operate over more detailed data than can regular users. First, he can calculate a frequency with which each user used every word. Second, using the

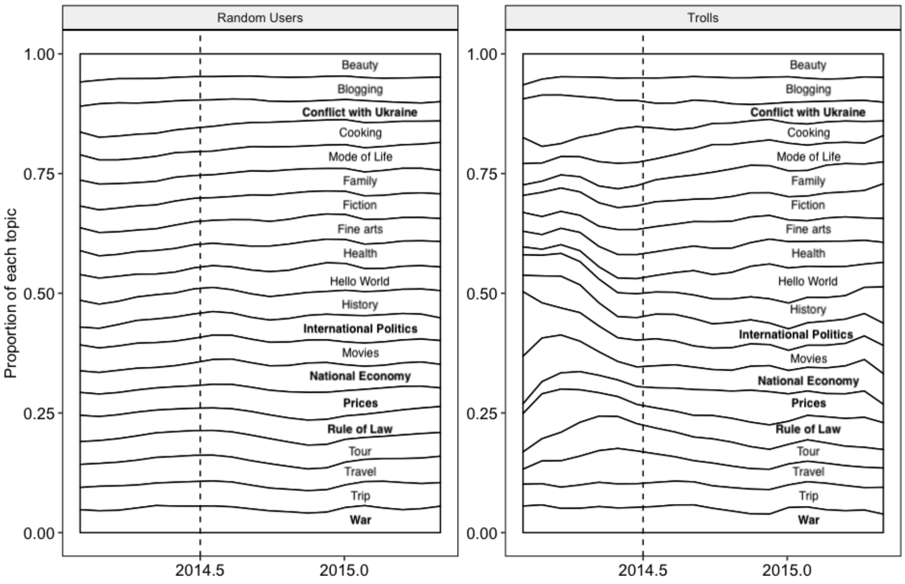


Figure 3.5: Topic prevalence over time

automated methods of text analysis, not only can he determine the dominant topic but he can also estimate a "complete" topical composition of each post. For example, a post could cover three topics ("International politics," "National economy," and "Rule of law") in the following respective proportion: 60 percent, 30 percent, and 10 percent. In plain language, this post would mostly describe international political agenda, probably in the context of economic sanctions imposed on Russia by western countries, with one or two references to imperfect domestic law enforcement practices. Third, considering the overall amount of collected data, one could also introduce a temporal dimension in the analysis to trace the dynamics of the topical composition of posts published by social media accounts.

Figure 3.5 represents how the topical composition of the posts changed over time. It is easy to see the differences between the two groups. The topical composition remained mostly stable over 18 months. Most of the topics received almost the same coverage in the first and in the last months observed (January 2014 and June 2015, respectively). In contrast, troll accounts were formidably vulnerable over time. As is easy to see during the first six months after their registration, a few topics dominated troll accounts, including “Conflict with Ukraine,” “International politics,” “National economy,” “Prices,” “Rule of law,” and some non-political topics, although, with time, this disproportional coverage almost disappeared. After the first six months in the field, topical representation of paid posts became almost stable and more equal. The latter can also be observed within some topical groups. For instance, the “Economic issues” group was virtually entirely presented by the “National economy” topic up until the middle of 2014. Trolls did not touch the problem of high prices and the harsh economic conditions of Russians. However, closer to 2015, the relative proportions of these topics in the troll-account topical profile approached those in the regular-user topical profile. The latter equally covered both topics most of the time.

What might explain the trends discovered above? One explanation suggests that trolls had been trying to mimic regular users. Indeed, the topical profiles of the two groups had been steadily approaching one another over time. This increased similarity was due to the changes in the group of troll accounts, and not vice versa. Given the observed trends in Figure 3.5, it is reasonable to suggest that trolls were actively calibrating their online behavior and completed this task within six months. Trolls started their online activities by posting a lot of information regarding political and economic issues, but they had been adjusting their behavior by introducing more and more posts on non-political topics. Thus, their topical profiles became very much like those of regular over time.

Interestingly, trolls were engaging not only with political content at the beginning of 2014. In the first two months, they devoted a significant part of their publications to “Movies.” A potential explanation suggests that, during a calibration period, employers required them to produce posts on a particular non-political topic as a training task. The latter could be used to assess the overall quality of the new employees or to train them.

In general, statistical analysis found the suggestion of trolls mimicking regular

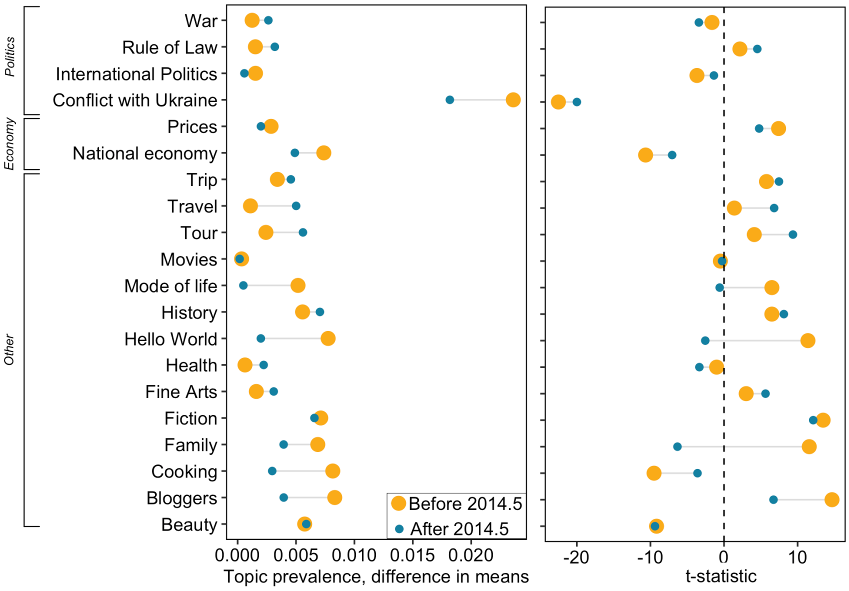


Figure 3.6: Topic prevalence over time

Note: Topics are: Beauty, Bloggers, Conflict with Ukraine, Cooking, Everyday problems, Family, Fiction, Fine Arts, Health, Hello World, History, International Politics, Movies, National Economy, Prices, Rule of Law, Tour, Travel and Animals, Trip, War users to be consistent with the data. To avoid potential imbalances, I subset the data at the topic-post level. For each topic, I sampled ten thousand topic estimates from both troll and non-troll accounts separately for two periods (before and after mid-2014, respectively). I conducted difference-in-means comparison for each of the topic.

Figure 3.6 represents the results of multiple comparisons of topical profiles. The left part of the figure depicts the difference in mean topical prevalence between paid commentators and regular users concerning each topic. Large dots represent this difference before mid-2014, and the small dots afterward. With time, most of the imbalanced topics significantly reduced the mean difference. For instance, "Conflict with Ukraine" is the topic with the most substantial difference in coverage between the two groups. With time, trolls adjusted the coverage of this topic by reducing it by a factor of 2.5 (see Figure 3.5), although this imbalance remained statistically significant even for later periods. The right part of Figure 3.6 shows the change in the statistical significance of the differences in the mean prevalence of topics. With time, most of the t-statistics had been tightening around zero. In other words, the calculated differences between the two groups of social media accounts became less significant. Indeed, the absolute value of averaged t-statistics reduced from 2.8 to 2.1.

While it is reasonable to suggest that trolls were trying to mimic regular users, some of the topical and non-topical features (see Table 3.1) in their behavior remained statistically distinct from the reference group. Are these features useful for the classification of a user's type? I continued my analysis by running a battery of logistic regressions to predict if the account belonged to a troll.

I started by aggregating available features at the account level. I then conducted regression analysis on three different samples: observations before mid-2015, observations after that date, and the pooled sample of all observations. As I analyzed at the user level, the number of observations in the pooled sample remained almost the same as those in the separate samples. I did not include the

	Dependent variable:		
	Account Status = 'Troll'		
	Before	After	Pooled
	2014.5	2014.5	sample
National economy	2.23	0.42	1.87
	(1.39)	(0.65)	(.95)
Prices	-10.61***	-5.24***	-7.79***
	(2.71)	(2.31)	(1.73)
Rule of law	-2.51**	-4.86***	-3.81***
	(1.27)	(1.19)	(0.87)
Conflict with Ukraine	2.91***	1.98*	2.46***
	(0.86)	(1.01)	(0.64)
International politics	1.42	0.78	1.09
	(2.21)	(1.91)	(1.43)
War	-3.45***	-1.22	-0.97
	(1.29)	(1.15)	(0.85)
Non-topical covariates	+	+	+
Observations	2,348	2,561	2,615
Sensitivity	0.82	0.80	0.79
Specificity	0.35	0.32	0.33
Balanced accuracy	0.59	0.56	0.57

Table 3.5: Predicting account status with logistic models

"Other topics" group in the model to reduce the multicollinearity problem.

Table 3.5 reports the results of these regression analyses. The significance of some topical predictors is lower for the second sample (after 2014.5) compared to the first one (before 2014.5). These include: "National Economy," "Conflict with Ukraine," and "War." The table also reports the quality of the in-sample predictive capacity of the models. Two details are worthy of note. First, the balanced accuracy of predictions from the second sample is the lowest among all three (0.56 against 0.57 and 0.59). The latter represents another result consistent with the idea that trolls adjust their behavior and mimic regular users. Second, the overall predictive capacity of the model remains low. For example, the predictive capability of the model from the third column ("Pooled sample") is 0.57. In other words, its predictions are better than the toss of a coin by slightly more than 10 percent.

Why do the estimated models have such weak predictive capacity? Several factors play a role here. First, while most of the topical covariates were identified as highly significant, the actual differences in the topical profiles of the two groups remained tiny. The functional form assumed in the logistic regression placed substantial constraints on the model's ability to produce accurate predictions. For instance, consider an hour when a post was published. Imagine that trolls were active only two times a day, from 8 am to 10 am and from 10 pm to 8 pm. At the same time, regular users published their posts independently of hours. The curve of the *logit*-model would not fit the two spikes in the activity of paid commentators and thus would not be able to distinguish them from social media users. The enormous number of available features that could potentially be decisive in identifying account type comprised another constraint. Third, the models above do not consider potential interaction between features, although such interactions could be plausible in the classification task of this study. The next sections introduce possible solutions to these issues.

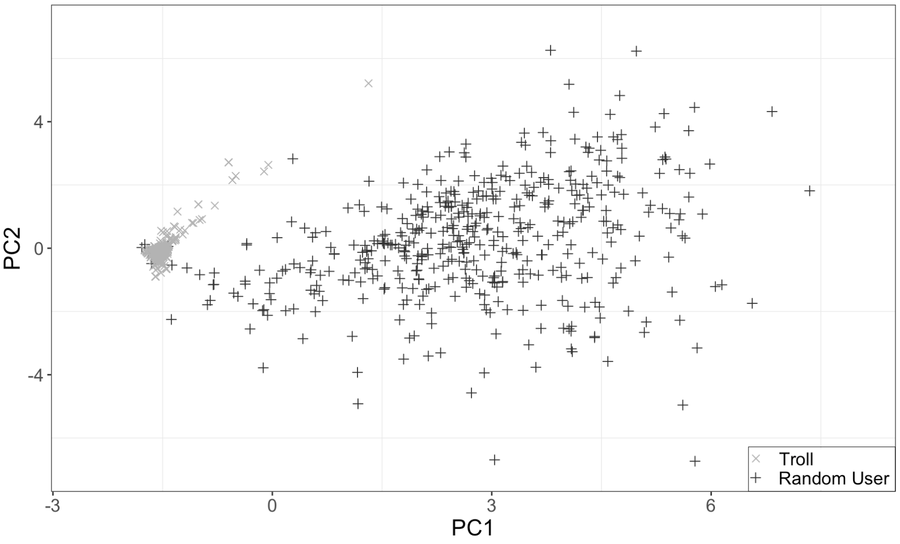


Figure 3.7: Trolls VS regular social media users: principal component analysis

3.4.3 Reverse-engineering perspective: Machine-learning methods of classification of social media accounts

In this section, I apply a machine learning approach to the problem of identifying troll accounts. Machine-learning methods aim to maximize the quality of prediction. On the backside, these methods produce results that can be hard to interpret concerning predictors. However, some of these can report the overall weight of a specific predictor by calculating its variance importance factor.

I conducted this comparison in three steps. First, I extracted features from the posts. Second, based on the latter, I trained a set of classification models to distinguish between the randomly sampled *LJ* accounts and accounts belonging to the leaked list of trolls on a sub-sample of data (i.e., a training set). I applied the trained model to another sub-set of data that was not used for the training (i.e., a test set). Finally, I compared the performance of the trained model across a set of metrics.

I extracted features from the posts described above as follows. First, I took the estimates of topical profiles from section 3.3.2. Secondly, 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 a *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 topic probabilities, mean values of TF-IDF components, and 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 account-level features.

I used machine-learning methods 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 3.7 depicts the results of performing a principal components analysis on the space of the first two principal components. As can easily be seen, most trolls are located far away from random users. The troll group has a much smaller variance 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 using several machine learning methods, including 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 do this, I randomly split the data into training and test sets. The training dataset was used to perform a *grid search* (Hsu et al., 2003) over the hyper-parameter space of each model with five-fold cross-validation. Finally, I applied trained models to the test data to evaluate

Model	Recall	Precision	F1-score	Accuracy	ROC
Logistic Regression	0.924	0.746	0.825	0.870	0.884
SVM (l)	0.793	0.646	0.712	0.787	0.789
SVM (g)	0.750	0.873	0.807	0.881	0.848
Naive Bayes *(g)*	0.924	0.545	0.685	0.718	0.770
Naive Bayes *(m)*	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 3.6: Performance of classification models

each model's performance.

Table 3.6 displays the statistics measuring the performance of the various classification models. The overall improvement in the quality of prediction is considerable – half of the trained models correctly classified 8 out of 10 posts. The models showed a balanced accuracy of 90 percent or higher. Finally, with a 96 percent precision and a 92 percent accuracy, *the random forest* appears to be the most efficient classification model.[2]

Apart from the accuracy of prediction, the *random forest* model identified the most critical features distinguishing trolls from random users. Concerning word usage, trolls more frequently used such terms as *"USA,"* "*America,*" and "*Obama,*" whereas random users were more likely to use the words "*Sanctions,*" "*Crimea,*" and "*Putin.*" Moreover, trolls and non-trolls differed significantly in the timing of their posts. Random users rarely published posts between 2 am and 12 pm. To conclude, modern statistical methods that use combinations of textual and timing features can correctly distinguish paid online commentators from regular social media users with very high accuracy.

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

3.5 Conclusion

In the last decade, authoritarian governments have been suspected of employing online commentators. They could be used to perform multiple tasks, ranging from distracting media users' attention away from news about public policy failures to threatening opposition activists and independent journalists.

To assess the scope of their activity and potential impact on users and voters, scholars of politics and policy experts need a toolkit to identify these trolls. Most of the studies in this field focus on potentially suspicious accounts of social media. To my knowledge, this is the first study explicitly comparing the behavioral patterns of paid online commentators with regular users on social media.

It attempts to make several contributions to the literature.

First, in contrast to conventional wisdom, paid online commentators do not always spend most of their working days writing posts about politics. In the case of Vladimir Putin's Russia, most of the posts published by troll accounts were completely apolitical.

Second, the results suggest that trolls mimic regular users. They do not only post a lot of politically irrelevant information, but they also calibrate their behavior by adjusting the topical profiles of their accounts in the direction of profiles of regular users of social media.

Third, while Internet trolls are good at hiding their troll identities from other users, modern statistical tools are able to identify them with a high degree of accuracy. While trolls try to mask themselves as regular users, some of their behavioral patterns differ sharply from those of ordinary-citizen users. The methods I employed allowed me to distinguish trolls from ordinary users with 96 percent precision. Thus, although companies, national governments, and academics can develop sophisticated techniques to detect state-sponsored political commentators on the Internet, most are still employing a combination of arbitrarily chosen criteria. My study suggests that models that employ "ground truth" data without pre-specified metrics can deliver predictions with a high degree of accuracy.

Why do trolls mimic regular users? Why do they spend a significant amount of time publishing posts on politically irrelevant topics? One possibility suggests that trolls invest in building reputations among regular users. Such reputations can help them hide their troll identities so as to accomplish several purposes. First, trolls need to conceal their identities to avoid Internet aggregators from blocking the accounts. This approach by paid online commentators suggests that they maintain their reputations in order to be able promote pro-government messages from time to time. Second, this reputation can be required if trolls use their accounts to engage in online discussions. To influence the content of such conversations, owners of these accounts must remain undetected. I study this possibility in the next part of my dissertation project.

CHAPTER 4

How pro-government "trolls" influence online conversations in Russia

4.1 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 (2017) 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. To date, researchers have focused on developing tools to identify paid online actors, their target groups, and the scale of their Internet presence. The research 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”.

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, 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.

The major goal of the paper is 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, but only by the political discussions that primarily referenced Putin'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 engagement in social media. Existing studies of the political role of social media have tended to primarily focus on 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, 2017; Keller et al., 2017; Miller, 2017; Sanovich, Stukal and Tucker, 2017) have established that 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.

4.2 Political astroturfing as a tool of information control

4.2.1 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

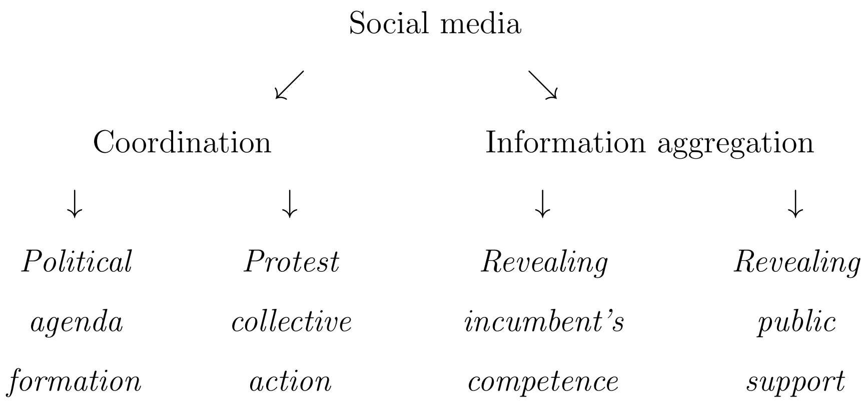

Social media
Coordination
Information aggregation
Political agenda formation
Protest collective action
Revealing incumbent's competence
Revealing public support

Figure 4.1: Political effects of social media

social media in political discussion has become 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, Figure 4.1). 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.

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 4.2). The goal of censorship

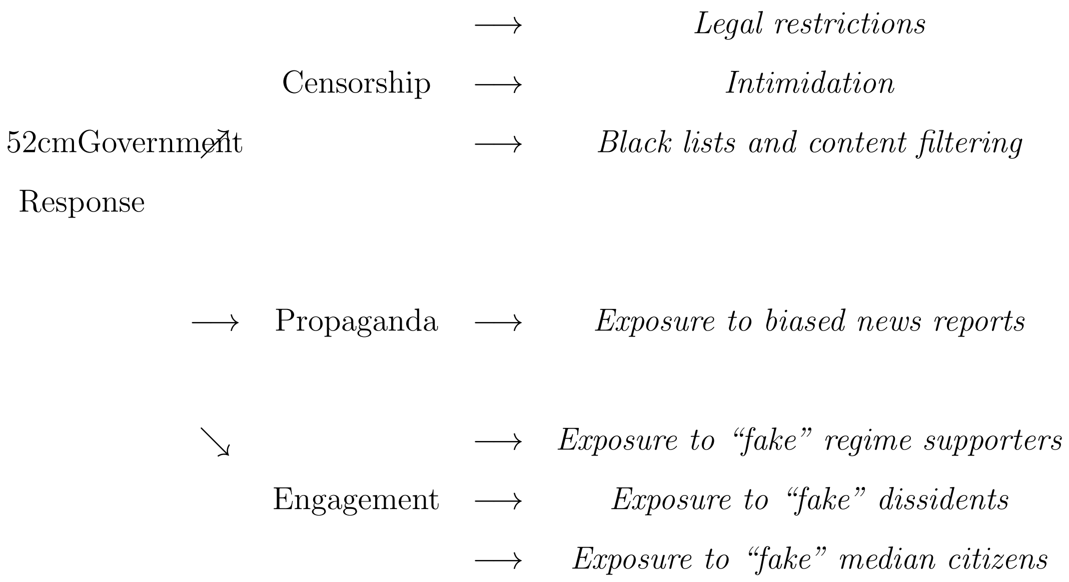

52cmGovernment
Response
Censorship
Legal restrictions
Intimidation
Black lists and content filtering
Propaganda
Exposure to biased news reports
Engagement
Exposure to "fake" regime supporters
Exposure to "fake" dissidents
Exposure to "fake" median citizens

Figure 4.2: Government responses to political threats of social media

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. 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 mean 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.

4.2.2 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 (2017) 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 (2017), Sanovich, Stukal and Tucker (2017) to develop a classification of trolls' goals, targets, and tactics.

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.

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.

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. Former 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 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 U.S. ... ?" (for example, "What about U.S. interference in the domestic affairs of other countries?"). The topic then naturally shifts away from discussion of the Russian government. Some examples of *diversion* and *promotion* include the following. Diversion happens if a conversation about corruption in the Russian government shifts toward a discussion of the IQ-levels of the participants in the current talk. Promotion happens if a conversation about corruption in the Russian government shifts toward a discussion about corruption in the opposition. Another example of promotion suggests that s conversation about Russia's alleged support for the insurgencies in Eastern Ukraine shifts toward a discussion of the legitimacy of the U.S. 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 one designates 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 shift the conversation toward discussion of positive aspects of the Crimea annexation.

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

4.3.1 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 Putin's 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, 2016). 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 in Russia. This, in turn, has caused Russian politicians and especially those in 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 (Sundstrom and Henry, 2016). In addition, state attempts to control the media are also viewed as preventing citizens from converting private grievances into public ones (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 percent 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*b*) pointed 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, Litvinenko and Bodrunova (2013) showed 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. Koltsova and Shcherbak (2015) 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.

4.3.2 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. Soldatov and Borogan (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*. 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 famous ones was promotion of fake news about a serious explosion at a processing plant in Louisiana.

The fact that paid commentators appear on *LJ* 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, *LJ* 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.

4.4 Data

4.4.1 Data collection

Following the publication of the list of 700 paid commentators on *LJ*, 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 name for any social media, including *LJ*. Its search range was limited to the last thousand 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. 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.

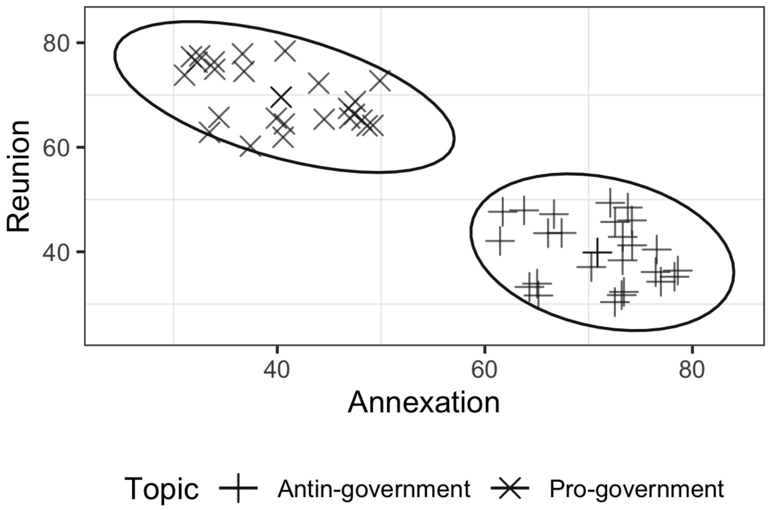


Figure 4.3: A hypothetical example of LDA classification

4.4.2 Post classification and processing of conversations

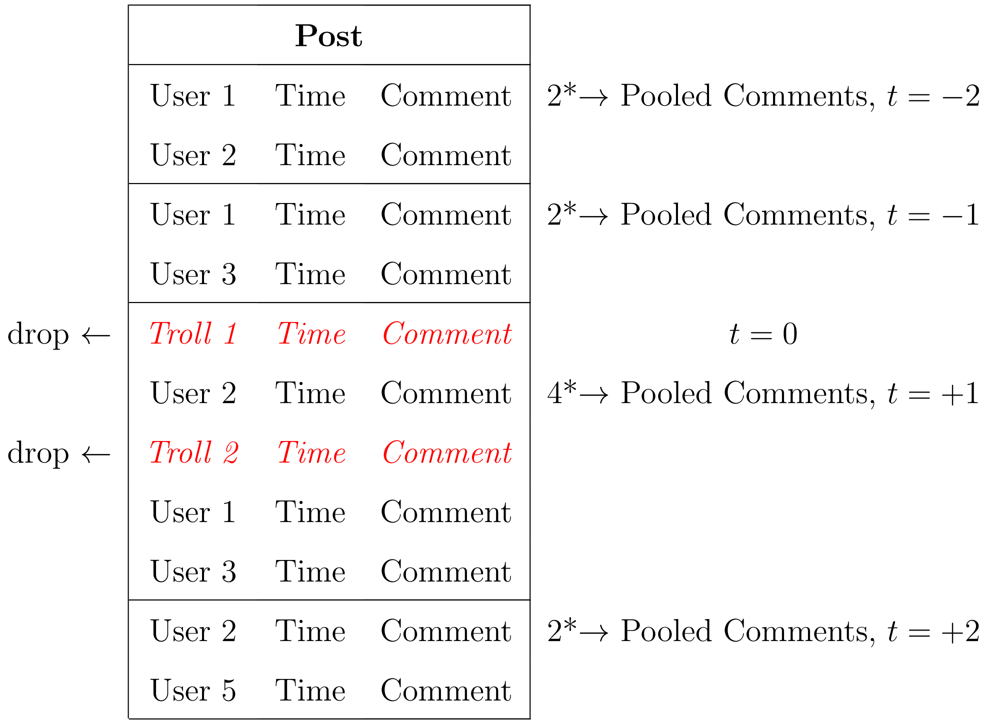
Automated data classification with Latent Dirichlet Allocation. Several parts of this study rely on automatic text classification using latent Dirichlet allocation (LDA), a generative statistical model that allows sets of texts to be described by their propensity to clusters (topics) (Blei, Ng and Jordan, 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 4.3 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 represent the propensities of a given conversation to each of the two topics.

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 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 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. Thus, 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, twenty



	Post			
	User 1	Time	Comment	2*→ Pooled Comments, $t = -2$
	User 2	Time	Comment	
	User 1	Time	Comment	2*→ Pooled Comments, $t = -1$
	User 3	Time	Comment	
drop ←	*Troll 1*	*Time*	*Comment*	$t = 0$
	User 2	Time	Comment	4*→ Pooled Comments, $t = +1$
drop ←	*Troll 2*	*Time*	*Comment*	
	User 1	Time	Comment	
	User 3	Time	Comment	
	User 2	Time	Comment	2*→ Pooled Comments, $t = +2$
	User 5	Time	Comment	

Figure 4.4: Processing online conversations: an example

slices (ten before and ten after the troll intervention) were generated. Figure 4.4 provides an example of the implementation of this algorithm with ordered 30-minute slices of conversations as units in this analysis.

4.4.3 Measurement

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.[1]

[1] 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.

Recall the updated example from the previous section. 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 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.

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 4.1). 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 65 to 85 percent within the topic mixture.

Two dependent variables are used in my analysis: propensity of a slice of a conversation to cover the anti-government topic and propensity of a slice of a conversation to cover a pro-government topics.

4.5 Research design and identification strategy

The focus of this research is 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 con-

	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 4.1: Anti-government and pro-government topics in online conversations.

versations, 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. 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 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

The first troll's comment

↓

$$Comment_1 \qquad Comment_2 \qquad Comment_3 \qquad Comment_4 \qquad Comment_5$$

$$\overbrace{acontrafactual} \qquad\qquad \overbrace{atreatedbinofcomments^{\circ}}$$

Figure 4.5: A plausible contrafactual for "treated" slices under the narrow time frame assumption

time frame, the set of comments appearing before the troll intervention constituted a contrafactual (see Figure 4.5).

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. 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.

4.6 Results

Figures 4.6 and 4.7 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

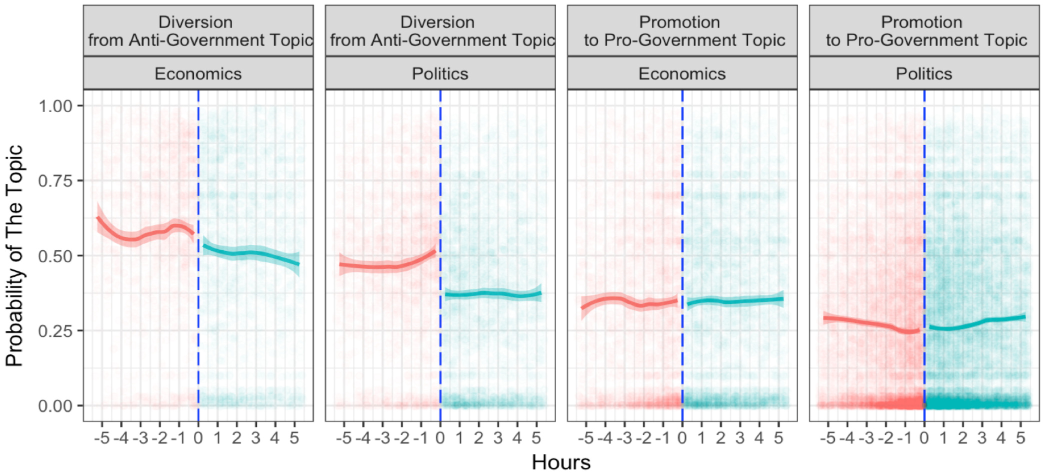
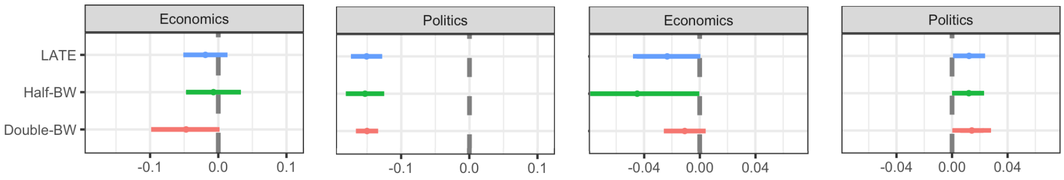


Figure 4.6: Troll interventions in online conversations



(a) Diversion from economic discussions

(b) Diversion from political discussions

(c) Promotion in economic discussions

(d) Promotion in political discussions

Figure 4.7: Effects of trolls' interventions on online conversations

statistically significant but negligible. Troll interventions increase the propensity of a conversation to cover a pro-government topic by about one percent. Trolls were successful in diverting discussions from purely political topics but had no effect on discussions on the national economy. Discussions on poor economic growth, unemployment, or price inflation seemed not to have been responsive to troll interventions.

4.7 Robustness and threats to validity

4.7.1 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

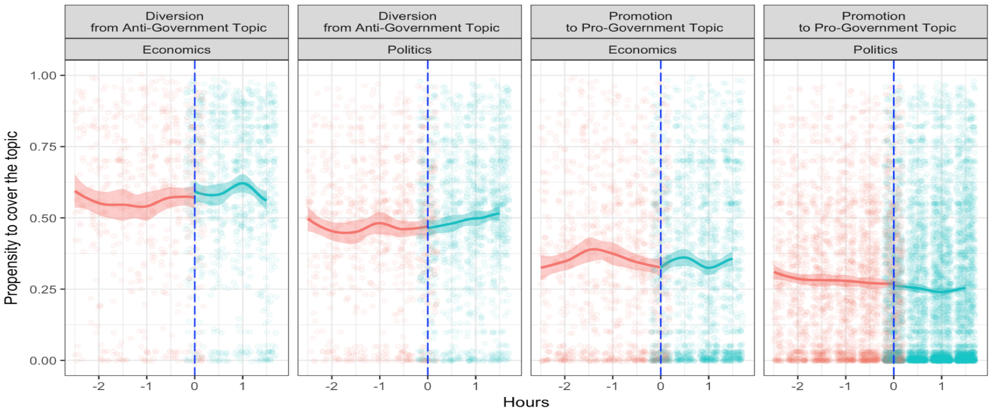


Figure 4.8: A random user's intervention effect on the evolution of conversations targeted by trolls

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 .

4.7.2 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.

Evidence from journalist investigations. Media investigations suggest that the published list was exhaustive. For example, Lyudmila Savchuk, a former

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 4.2: Potential working schedule of troll shifts

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. The leaked list of trolls was divided into four shifts named for the shift's supervisor. 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 4.2 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 list published approximately the same number of comments during each day and night of the week.

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.

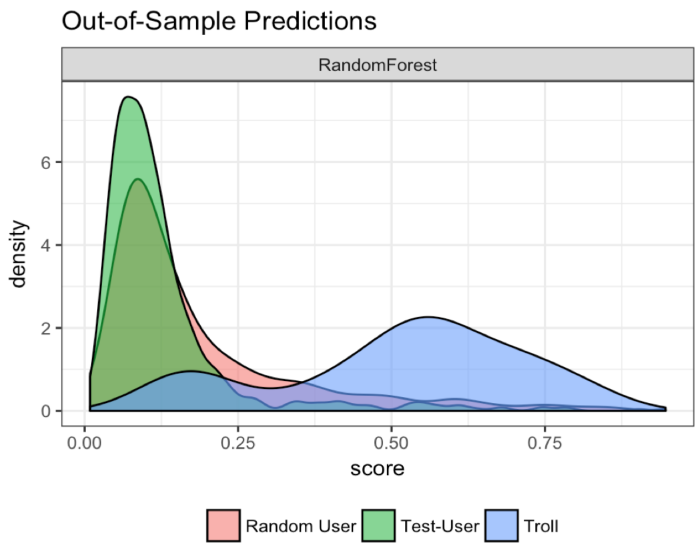


Figure 4.9: Propensity of trolls, random users, and participants of targeted conversations to be a troll

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. To do it, I applied the classification model trained in the part 2 of my dissertation project to a random sample of 650 participants of targeted discussions. First, I collected collected their posts. I then 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 4.9 displays the distribution of the propensities of trolls, random users, and participants in targeted conservations.

Figure 4.9 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 percent. This evidence lends credibility to the hypothesis that the list of troll accounts published by *Novaya Gazeta* was, in fact, exhaustive.

4.8 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 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.

CHAPTER 5

Conclusion

In this dissertation, I attempt to advance the study on a repertoire of tools authoritarian leaders employ to maintain political control. I focus on on non-democratic government hiring of agents to impersonate ordinary citizens and engage online and offline with members of the political opposition. My dissertation contributes to the scholarly discussion in several ways.

First, it speaks to the literature on the political mobilization and persuasion. In contrast to studies that largely focused on the effects of biased news from state-controlled media (Enikolopov, Petrova and Zhuravskaya, 2011; Miner, 2012; Adena et al., 2015; Peisakhin and Rozenas, 2014; Yanagizawa-Drott, 2014), I show that sometimes credible reports sent by independent media outlets can be an even more efficient instrument to discourage opposition than can state propaganda.

Second, it adds to the literature on the problem of information control. As previous studies have shown, paid commentators primarily target regular users and create informational noise so as to complicate these users' access to news that is potentially dangerous to the regime (Munger et al., 2015; King, Pan and Roberts, 2017; Keller et al., 2017; Miller, 2017; Roberts, 2018). In contrast, I find that paid Internet trolls can also be used to target political activists employing very different tactics, masking their troll identity, and infiltrating conversations on social media with messages that induce ideological divergence.

Third, my work makes a methodological contribution by developing a framework for analyzing paid political engagement in social media. Specifically, it proposes a method for estimating the causal effect of troll interventions on politically charged online discussions. In contrast to 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 choose to target a discussion after observing the direction of its movement. Moreover, the framework I propose can be generalized to study the behavior of paid online agents in other contexts.

Finally, my results contribute to the debate on malicious misinformation and regulation of digital platforms. Companies, national governments, and academics develop sophisticated methods to detect state-sponsored political commentators on the Internet, but most of these methods are based on a combination of arbitrarily chosen criteria, often including the country of origin of the account's email address or phone number, usage of specific characters (e.g., cyrillic alphabets), and specific keywords in the message. My study shows that such methods may be unable to identify a significant proportion of paid political commentators. These commentators are apparently aware of the risks and try hard to hide their troll identity. At the same time, my research demonstrates that analysis of leaked data like that I obtained can successfully identify behavioral patterns that effectively distinguish paid commentators from regular users of digital platforms.

CHAPTER 6

Additional material

6.1 Echo of Moscow: List of locations

1. Abakan, Hakasiya
2. Barnaul, Altay kray
3. Blagoveshhensk, Amur oblast
4. Volgograd, Volgograd oblast
5. Vologda, Vologda oblast
6. Vyborg, Leningrad oblast
7. Ekaterinburg, Sverdlovsk oblast
8. Irkutsk, Irkutsk oblast
9. Kazan, Tatarstan
10. Kamensk-Uralskiy, Sverdlovsk oblast
11. Kineshma, Ivanovo oblast
12. Kirov, Kirov oblast
13. Lipeck, Lipeck oblast
14. Mahachkala, Dagestan
15. Moscow

16. Nizhnevartovsk, Tyumen oblast

17. Obninsk, Kaluga oblast

18. Orenburg, Orenburg oblast

19. Omsk, Omsk oblast

20. Perm, Perm kray

21. Penza, Penza oblast

22. Pereslavl-Zalesskiy, Yaroslavl oblast

23. Rybinsk, Yaroslavl oblast

24. Rostov-na-Donu, Rostov oblast

25. Saratov, Saratov oblast

26. Samara, Samara oblast

27. Severodvinsk, Arhangelskoblast

28. Saint-Petersburg

29. Surgut, Khanty-Mansi

30. Tambov, Tambov oblast

31. Tolyatti, Samara oblast

32. Tomsk, Tomsk oblast

33. Tula, Tula oblast

34. Tyumen, Tyumen oblast

35. Ulan-Udye, Respublika Buryatiya

36. Ufa, Respublika Bashkortostan

37. Chelyabinsk, Chelyabinskoblast

38. Yaroslavl, Yaroslavl oblast

39. Zheleznogorsk-Ilimskiy, Irkutskaya oblast

40. Zelenogorsk, Krasnoyarskiy kray

6.2 Paid commentators and regular users on social media

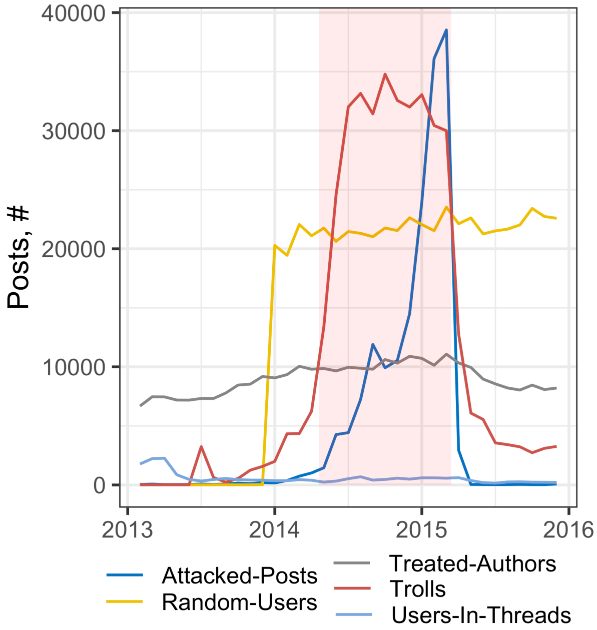


Figure 6.1: Activity of social media accounts: all groups

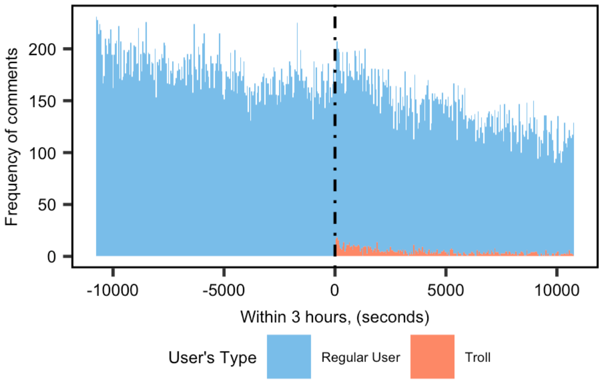


Figure 6.2: Online conversations: frequency of comments

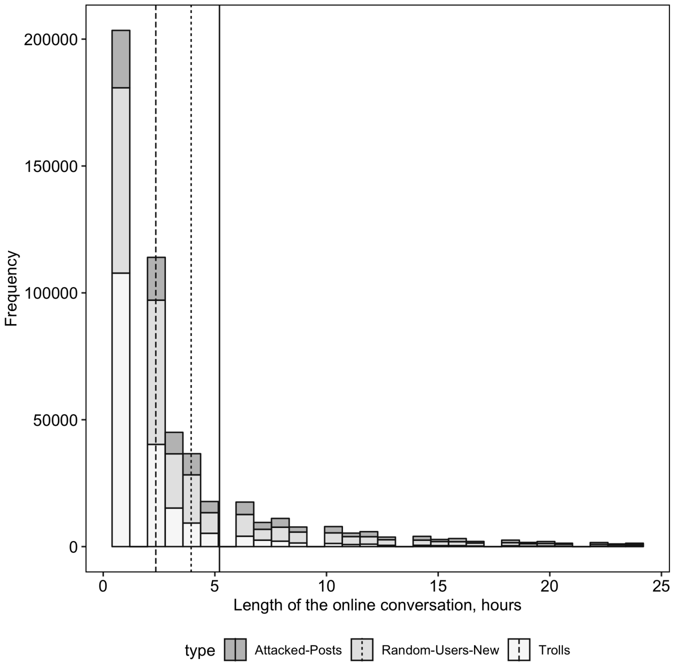


Figure 6.3: Length of online conversations

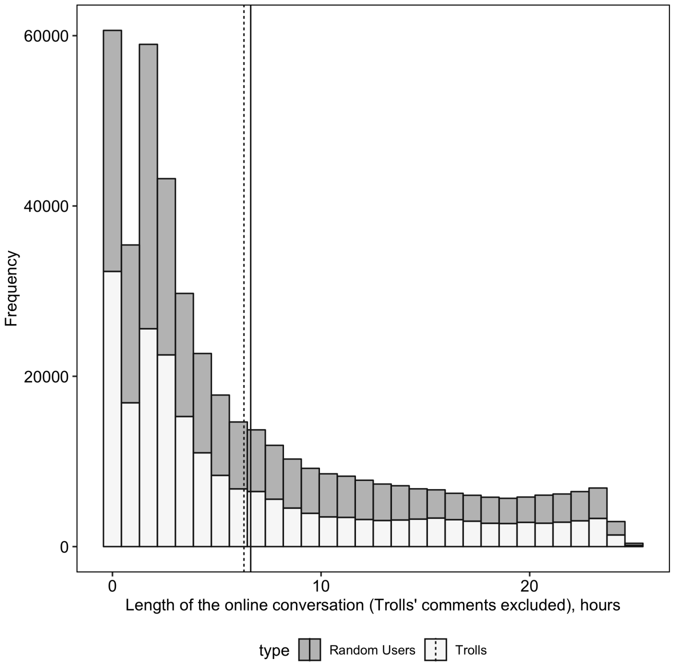


Figure 6.4: Length of online conversations: trolls excluded

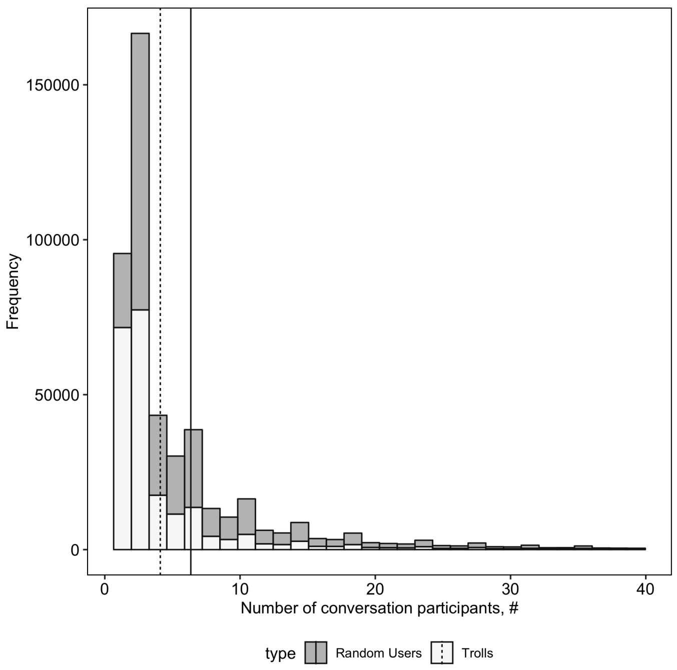


Figure 6.5: Participants of online conversations

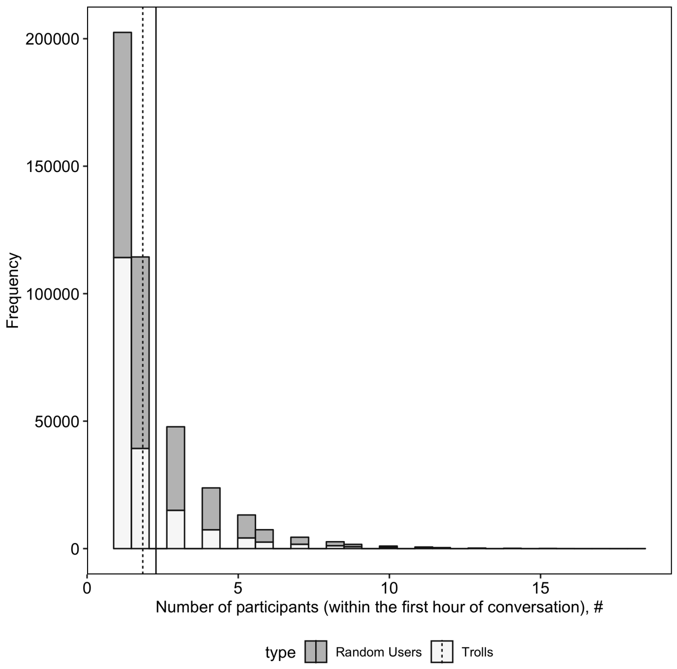


Figure 6.6: Participants of online conversations: first hour

6.3 Topic models: Scores

Topic 1: *Cooking*

Highest Prob: oil, add, taste, water, recipe, dough, sugar

FREX: creamy, oven, garlic, dish, sour cream, fry, salad

Lift: creamy, agar, eggplant, knead, breast, oven, custard

Score: creamy, whisk, fry, recipe, oven, dough, butter

Topic 2: *Movies*

Highest Prob: film, game, role, first, movie, music, play

FREX: film, series, cinema, actor, director, music, Olympic

Lift: acting, jennifer, clip, nagiev, producer, single, avva

Score: film, Ava, director, actor, TV series, Olympics, actress

Topic 3: *Tour*

Highest Prob: city, building, house, place, museum, part, build

FREX: building, tour, architect, architecture, build, museum, station

Lift: depot, dace, cathedral, restore, diesel locomotive, stage, church

Score: building, museum, monument, diesel locomotive, cathedral, park, temple

Topic 4: *History*

Highest Prob: russian, people, russia, great, saint, church, land

FREX: Orthodox, Emperor, Christ, Tsar, Patriarch, Prince, Russia

Lift: heretic, constantinople, liturgy, reverend, ragnar, alexis, metropolitan

Score: church, emperor, orthodox, patriarch, saint, prince, russian

Topic 5: *Health*

Highest Prob: human, most, scientist, doctor, brain, help, organism

FREX: organism, drug, substance, energy, medicine, scientist, doctor

Lift: evolutionary, metabolism, digestion, drug, calcium, mole, iron

Score: drug, organism, doctor, scientist, substance, vitamin, patient

Topic 6: *War*

Highest Prob: war, military, army, soviet, ussr, german, soldier

FREX: general, German, soldier, officer, tank, camp, army

Lift: division, prisoner of war, dachau, militia, rohlin, prisoner, lieutenant

Score: war, army, military, general, soldier, army, german

Topic 7: *Family*

Highest Prob: child, man, woman, man, life, understand, live

FREX: man, woman, child, sex, parent, feeling, school

Lift: introvert, anal, reason, sex, teacher, marriage, self-esteem

Score: woman, man, child, sex, parent, girl, person

Topic 8: *Trip*

Highest Prob: airplane, machine, speed, car, flight, first, camera

FREX: speed, flight, space, flight, plane, smartphone, device

Lift: take-off, gradient, sensor, light, snow blower, chassis, diaphragm

Score: airplane, airport, asis, clickable, rocket, galaxy, boeing

Topic 9: *National economy*

Highest Prob: Russia, country, Putin, Russian, president, state, politics

FREX: economy, economic, Obama, oil, sanction, billion, Putin

Lift: investment, trillion, central bank, trump card, Gazprom, Eurasian, nomenclature

Score: Putin, economy, Russia, president, Obama, economic, billion

Topic 10: *Prices*

Highest Prob: money, ruble, price, buy, work, company, receive

FREX: ruble, goods, amount, value, salary, money, euro

Lift: cost, Russian fund, cash, gratuity, Krasnodar, ruble, salary

Score: ruble, money, goods, cost, price, euro, shop

Topic 11: *Rule of law*

Highest Prob: law, region, organization, russia, employee, citizen

FREX: bulk, deputy, employee, law, chairman, organization, criminal

Lift: marijuana, convicted, rector, sex-enlightenment, pre-trial detention center, search, extremism

Score: bulk, deputy, state, federation, police, chairman, elections

Topic 12: *Fine arts*

Highest Prob: artist, ring, art, work, painting, exhibition, create

FREX: artist, drawing, draw, portrait, exhibition, ring, dragon

Lift: tonio, consuelo, picasso, exupery, camilla, artist, judith

Score: artist, tonio, exhibition, ring, drawing, consuelo, picasso

Topic 13: *Beauty*

Highest Prob: skin, color, hair, girl, clothes, face, dress

FREX: leather, hair, dress, linen, cosmetics, skirt, mask

Lift: blouse, makeup, scrub, firmness, deodorant, palette, perfume

Score: Skin, Cream, Hair, Moisturize, Gel, Dress, Fragrance

Topic 14: *Fiction*

Highest Prob: book, history, read, author, language, word, write

FREX: book, poet, writer, novel, last name, andrey, verse

Lift: Koba, Tartary, Yesenin, literary, smooth, King, absinthe

Score: book, writer, verse, novel, work, tartary, library

Topic 15: *Mode of life*

Highest Prob: know, home, understand, hand, go, sit, think

FREX: sleep, call, home, sit, man, bed, mom

Lift: smoke, nod, rout, daddy, fuck, alarm clock, granny

Score: mom, sleep, apartment, bed, phone, morning, girlfriend

Topic 16: *Travel*

Highest Prob: water, road, sea, place, cat, dog, tree

FREX: cat, lake, dog, beach, sea, wind, shore

Lift: photo story, camping, chan, cat, umbrella, tundra, sandy

Score: cat, lake, dog, beach, shore, tree, river

Topic 17: *Conflict with Ukraine*

Highest Prob: Ukraine, Ukrainian, Russia, power, Crimea, war, Kiev

FREX: Donbass, Maidan, Donetsk, Poroshenko, Ukrainian, Yanukovych, Ukraine

Lift: Akhmetov, Debaltseve, Donetsk, Maidan, Obs, Rada, Saakashvili

Score: Ukraine, Ukrainian, Donbass, Poroshenko, Maidan, Yanukovych, Donetsk

Topic 18: *Blogging*

Highest Prob: write, photograph, friend, post, photo, make, blog

FREX: blog, comment, post, blogger, internet, post, link

Lift: ban, browser, repost, reposter, repost, top-end, friend

Score: blog, photo, blogger, photo, post, repost, magazine

Topic 19: *Hello world*

Highest Prob: new, good, friend, holiday, morning, love, good

FREX: holiday, day off, mood, gift, congratulations, joy, kind

Lift: mimosa, panin, albina, New Year, holiday, New Year, congratulate

Score: holiday, day off, morning, gift, congratulate, mood, calendar

Topic 20: *International affairs*

Highest Prob: country, europe, turkey, resident, live, refugee, america

FREX: Turkey, Israel, Refugee, Migrant, Turkish, Arabic, Muslim

Lift: Griboedov, Jihad, Syrian, migrant, Jordan, Khach, Israel

Score: Refugee, Turkey, Syria, Israel, Migrant, Muslim, Griboedov

6.4 Topic models: List of stop words

abo
about
above
absent
act
active
activity
adaptat
adje
affairs
after
after all
again
against
ale
all
all i
alo
alone
already
always
am
an
an object
and
another

besides
between
blow
both
boules
bouv
bud
bula
bulo
busy
but
buti
buva
by
by me
by whom
call
can
cannot
cat
century
childish
children
city
civil
click

directed
do
do not be
does
doesn't
dogo
doing
don't
down
duh
dumb
during
age
dyakoy
each
eat
eating
educationally
eg
eight
eighteen
eighteenth
eighth
eleven
eleventh
especially

g
game
gave
get out
give up
good
grave
great
grinding
groups
guilty
had
hadn't
hail
has
hasn't
have
haven't
having
he
he speaks
he'd
he'll
he's
health
hello

if a
im
important
in
in and
in general
including
inf
information
informational
infusion
interaction
interesting
into
is
is necessary
is possible
is related
isn't
it
it seems
it was
it's
it's better
its
itself

any
are
are eating
aren’t
around
as
as if
at
at once
at the bottom
ate
axis
b
back
bagato
bases
be
because
been
before
beginning
being
below

clod
close
cofe
com
competition
complex
concepts
continuously
could
couldn’t
cream
cultural
curry
d_isno
dali
dan
dava
day
de
den
deprived
did
didn’t

even
every
everyone
everything
everywhere
exactly
eye
far away
features
few
finally
finding
first
five
for
for nothing
form
formed
friend
from
from everywhere
from here
further

her
here
here’s
hers
herself
highly
him
himself
his
history
house
how
how’s
however
html
http
i
i’d
i’ll
i’m
i’ve
ice
if

what for
what’s
when
where
whether
which
while
who
who
who’s
whom
why
why’s
will
will be
will take
win
with
without
women
won
won’t
zvidusil

mainly
many
material
may
mayge
me
mean
men
mercilessly
meter
methodically
mi
mig
millions
milyoniv
min
mine
minutes
moghi
mogti
more
more beautiful
need to
needed
neither
neridko
never

nikudi
nine
nineteen
nineteenth
ninth
nizh
no
no way
nor
not
not allowed
not possible
nothing
nowhere
numerous
nx
ny
n
ny
occupation
odnak
of
of all
one thing
only
or
organ
osta

ours
ourselves
out
over
own
p'eteen
p'eteenths
past
people
per
pestilence
petits
places
plan
please
populated
power
preparation
pretty
promoted
provided
public
qualities
shouldn't
skill
skin
so
societies

sometimes
soundly
speak up
speaking
special
state
still
stink
such
suddenly
support
taken
ten
tenth
than
that
that's
the
the basics
the beginning
the class
the eighth
the important
they're
they've
this
those
through

to all
to be able
to finish
to me
to mean
to my
to navigate
to one
to our
to which
to whom
together
too
total
total
town
trained
tue
turned
twelfth
twelve
twentieth
twenty
was
wasn't
waters
we
well

more important
moscow
most
mothers
moved
must
mustn’t
my
myself
my
mzh
name
name is
nasa
nationally
navshcho
naybilsh
near
necessary
new
nibi
nikoli

of course
of cultures
of names
of our
of persons
of the month
of the twelve
of the world
of the year
of times
off
offense
often
on
on the neck
on the way
once
one
one day
other
ought
our

quantities
question
ready
really
recently
responsible
s
said
same
scientific
second
sent
several
shan’t
she
she’d
she’ll
she’s
should
some
someday
zvidsi

the inhabitant
the internet
the organization
the region
the time
the world
their
theirs
them
themselves
then
there
there is
there’s
these
they
they are
they’d
they’ll
throughout
time
to

two
under
united
until
up
uphill
uplands
upstairs
us
usual
usually
v50
very
via
vid
view
vsm
vdsotkv
wait
were
weren’t
what

REFERENCES

Adena, Maja, Ruben Enikolopov, Maria Petrova, Veronica Santarosa, and Ekaterina Zhuravskaya. 2015. ”Radio and the rise of the Nazis in prewar Germany.” The Quarterly Journal of Economics 130(4):1885–1939.

Ahmed, Saifuddin, and Jaeho Cho. 2019. ”The Internet and political (in) equality in the Arab world: A multi-country study of the relationship between Internet news use, press freedom,and protest participation.” New Media Society 21(5):1065–1084.

Becker, Jonathan. 2004. ”Lessons from Russia a neo-authoritarian media system.” European Journal of Communication 19(2):139–163.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. ”Latent Dirichlet Allocation.” Journal of Machine Learning research 3:993–1022.

Bond, Robert M, Christopher J Fariss, Jason J Jones, Adam D I 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.

Borenstein, Eliot. 2019. Plots against Russia: Conspiracy, and fantasy after socialism. Cornell University Press.

Broniatowski, David A, Amelia M Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C Quinn, and Mark Dredze. 2018. ”Weaponized health communication: Twitter bots, and Russian trolls amplify the vaccine debate.” American Journal of Public Health 108(10):1378–1384.

Chatterjee, Arindam, and Soumendra Nath Lahiri. 2011. ”Bootstrapping lasso estimators.” Journal of the American Statistical Association 106(494):608–625.

Clement, Karine. 2008. ”New social movements in Russia: a challenge to the dominant model of power relationships?” Journal of Communist Studies, and Transition Politics 24(1):68–89.

Conover, Pamela Johnston, and Stanley Feldman. 1984. ”How people organize the political world: A schematic model.” American Journal of Political Science. 28(1):95-126

Edmond, Chris. 2013. ”Information manipulation, coordination, and regime change.” Review of Economic Studies 80(4):1422–1458.

Egorov, Georgy, and Konstantin Sonin. 2014. Incumbency advantage in nondemocracies. Technical report National Bureau of Economic Research.

Enikolopov, Ruben, Alexey Makarin, and Maria Petrova. 2015. ”Social Media and protest participation: Evidence from Russia.” Technical report National Bureau of Economic Research.

Enikolopov, Ruben, Maria Petrova, and Ekaterina Zhuravskaya. 2011. ”Mediaå and political persuasion: Evidence from Russia.” The American Economic Review 101(7):3253–3285.

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.” Berkman Center Research Publication.

Faris, Robert. 2019. "Cyberwar: how Russian hackers, and trolls helped elect a president: What we don't, can't, and do know." Perspectives on Politics 17(3):884–886.

Field, Anjalie, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. "Framing, and agenda-setting in russian news: a computational analysis of intricate political strategies." arXiv preprint arXiv:1808.09386 .

Geddes, Barbara, and John Zaller. 1989. "Sources of popular support for authoritarian regimes." American Journal of Political Science 33(2):319–347.

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

Gehlbach, Scott. 2010. "Reflections on Putin and the media." Post-Soviet Affairs 26(1):77–87.

Gehlbach, Scott, and Konstantin Sonin. 2014. "Government control of the media." Journal of Public Economics 118:163–171.

Greene, Samuel A. 2014. Moscow in movement: Power and opposition in Putin's Russia. Stanford University Press.

Gunitsky, Seva. 2015. "Corrupting the cyber-commons: Social media as a tool of autocratic stability." Perspectives on Politics 13(01):42–54.

Guriev, Sergei, and Daniel Treisman. 2015. "How modern dictators survive: An informational theory of the new authoritarianism." Technical report National

Bureau of Economic Research.

Hale, Henry E. 2011. "The Putin machine sputters: First impressions of the 2011 Duma election campaign." Russian Analytical Digest (106):2–8.

Hamilton, James. 2004. All the news that's fit to sell: How the market transforms information into news. Princeton University Press.

Hassanpour, Navid. 2014. "Media disruption, and revolutionary unrest: Evidence from Mubarak's quasi-experiment." Political Communication 31:1–24.

Hollyer, James, Peter Rosendorff, and James Raymond Vreeland. "Transparency, protest, and autocratic instability." American Political Science Review 109.4 (2015a): 764-784.

Hollyer, James, Peter Rosendorff, and James Raymond Vreeland. "Why do autocrats disclose? Economic rransparency and inter-elite politics in the shadow of mass unrest." Journal of Conflict Resolution, 63(6), 1488-1516.

Imai, Kosuke, and Marc Ratkovic. 2014. "Covariate balancing propensity score." Journal of the Royal Statistical Society 76(1): 243-263

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.

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(2):326–343.

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

King, Gary, Jennifer Pan, and Margaret E Roberts. 2017. ”How the Chinese government fabricates social media posts for strategic distraction, not engaged argument.” American Political Science Review 111(3):484–501.

Kobak, Dmitry, Sergey Shpilkin, and Maxim Pshenichnikov. 2012. ”Statistical anomalies in 2011-2012 Russian elections revealed by 2D correlation analysis.”. Technical report.

Koltsova, Olessia, and Andrey Shcherbak. 2015. ”LiveJournal Libra!: The political blogosphere, and voting preferences in Russia in 2011–2012.” New Media Society 17(10):1715–1732.

Kuzio, Taras. 2018. ”Russia–Ukraine crisis: The blame game, geopolitics, and national identity.” Europe-Asia Studies 70(3):462–473.

Lankina, Tomila, and Alisa Voznaya. 2015. ”New data on protest trends in Russia's regions.” Europe-Asia Studies 67(2):327–342.

Litvinenko, Anna, and Svetlana Bodrunova. 2013. New media, and the political protest: The formation of a public counter-sphere in Russia of 2008-12. Taylor Francis.

Liu, Huan, Fred Morstatter, Jiliang Tang, and Reza Zafarani. 2016. ”The good, the bad, and the ugly: uncovering novel research opportunities in social media mining.” International Journal of Data Science, and Analytics 1(3-4):137–143.

Lohmann, Susanne. 1994. ”The dynamics of informational cascades.” World Politics 47(1):42–101.

Lorentzen, Peter. 2013. ”Regularizing rioting: Permitting public protest in an authoritarian regime.” Quarterly Journal of Political Science 8(2):127–158.

Luca, Michael, and Georgios Zervas. 2016. ”Fake it till you make it: Reputation, competition, and Yelp review fraud.” Management Science 62(12):3412–3427.

Lupia, Arthur, Mathew McCubbins, Lupia Arthur, and Others. 1998. The democratic dilemma: Can citizens learn what they need to know? Cambridge University Press.

Marshall, Monty, and Keith Jaggers. 2016. ”Polity IV project: Political regime characteristics, and transitions, 1800-2015.”.

McFaul, Michael, and Elina Treyger. 2004. Between Dictatorship, and Democracy: Russian Post-Communist Political Reform. The Brookings Institution Press.

Mickiewicz, Ellen. 2008. Television, power, and the public in Russia. Cambridge University Press.

Miner, Luke. 2012. The unintended consequences of Internet diffusion: Evidence from Malaysia. Technical report. New Economic School.

Morris, Laura. 2014. ”Contextualizing the power of social media: Technology, communication, and the Libya crisis.” First Monday 19(12).

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.”.

Nikiporets-Takigawa, Galina. 2013. ”Tweeting the Russian protests. Digital icons.” Studies in Russian, Eurasian, and Central European New Media 9:1–25.

Oates, Sarah. 2006. Television, democracy, and elections in Russia. Routledge.

Peisakhin, Leonid, and Arturas Rozenas. 2014. ”Electoral mobilization with biased media: The influence of Russian television in Ukraine.”.

Persily, Nathaniel. 2017. ”The 2016 US election: Can democracy survive the Internet?” Journal of Democracy 28(2):63–76.

Pfeffer, Jurgen, Thomas Zorbach, and Kathleen Carley. 2014. ”Understanding online firestorms: Negative word-of-mouth dynamics in social media networks.” Journal of Marketing Communications 20(1-2):117–128.

Popkin, Samuel. 1994. The reasoning voter: Communication, and persuasion in presidential campaigns. University of Chicago Press.

Remnick, David. 2005. ”The translation wars.” The New Yorker 7:14–17.

Roberts, Margaret. 2018. Censored: distraction, and diversion inside China’s Great Firewall. Princeton University Press.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. 2014. ”Structural topic models for open-ended survey responses.” American Journal of Political Science 58(4):1064–1082.

Robertson, Graeme. 2013. ”Protesting putinism: The election protests of 2011-2012 in broader perspective.” Problems of Post-Communism 60(2):11–23.

Robertson, Graeme. 2010. The politics of protest in hybrid regimes: managing dissent in post-communist Russia. Cambridge University Press.

Sanovich, S., Stukal, D. and Tucker, J.A., 2018. “Turning the virtual tables: Government strategies for addressing online opposition with an application to Russia.” Comparative Politics 50(3): 435-482.

Santana, Arthur D, and David M Dozier. 2019. ”Mobile devices offer little in-depth news: Sensational, breaking, and entertainment news dominate mobile news sites.” Journalism Practice 4:1–22.

Smyth, Regina, Anton Sobolev, and Irina Soboleva. 2013. ”A well-organized play.” Problems of Post-Communism 60(2):24–39.

Smyth, Regina, Irina Soboleva, Luke Shimek, and Anton Sobolev. 2015. Defining common ground: The language of mobilization in Russian protests. In Systemic,and non-systemic opposition in the russian federation. Ashgate. pp. 51–76.

Soldatov, Andrei, and Irina Borogan. 2017. The Red Web: The Kremlin’s wars on the Internet. Perseus Books.

Stein, Arthur. 2012. "The unraveling of support for authoritarianism: The dynamic relationship of media, elites, and public opinion in Brazil, 1972-82." The International Journal of Press and Politics 18(1):85–107.

Stukal, Denis, Sergey Sanovich, Richard Bonneau, and Joshua A Tucker. 2017. "Detecting bots on Russian political Twitter." Big Data 5(4):310–324.

Sundstrom, Lisa McIntosh, and Laura A Henry. 2016. Russian Civil Society. Me Sharpe.

Taddy, Matt. 2012. "On estimation, and selection for topic models". Artificial Intelligence and Statistics 1:1184–1193.

Teague, Elizabeth. 2011. "How did the Russian population respond to the global financial crisis?" Journal of Communist Studies and Transition Politics 27(3-4):420–433.

Tibshirani, Robert. 1996. "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society 58(1):267– 288.

Treisman, Daniel. 2018. The new autocracy: Information, politics, and policy in Putin's Russia. Brookings Institution Press.

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.

Volkov, Denis. 2012. "The protesters and the public." Journal of Democracy

23(3):55–62.

VonDoepp, Peter, and Daniel J Young. 2012. ”Assaults on the fourth estate: Explaining media harassment in Africa.” The Journal of Politics 75(01):36– 51.

Wang, Bo, and Becky Loo. 2019. ”The hierarchy of cities in Internet news media, and Internet search: Some insights from China.” Cities 84:121–133.

Wintrobe, Ronald. 1998. The political economy of dictatorship. Cambridge university press.

Yablokov, Ilya. 2015. ”Conspiracy theories as a Russian public diplomacy tool: The case of Russia Today (RT).” Politics 35(3-4):301–315.

Yanagizawa-Drott, David. 2014. ”Propaganda and conflict: Theory and evidence from the Rwandan genocide.” The Quarterly Journal of Economics 129(4):1947–1994.

Zaller, John. 1992. The nature, and origins of mass opinion. Cambridge university press.