Fighting Propaganda with Censorship: A Study of the Ukrainian Ban on Russian Social Media

Yevgeniy Golovchenko (yg@ifs.ku.dk)

Abstract

Many states have become concerned with Russian cyberattacks and online propaganda. The Ukrainian government responded to the information threat in 2017 by blocking access to several Russian websites, including VKontakte, one of the most popular social media websites in Ukraine. By exploiting a natural experiment in Ukraine, I find that the sudden censorship policy reduced activity on VKontakte, despite the fact that a vast majority of the users were legally and technically able to bypass the ban. Users with strong political and social affiliations to Russia were at least as likely to be affected by the ban as those with weak affiliations. I argue that the ease of access to online media — not political attitudes toward the state — was the main mechanism behind the users' response to the ban. These findings suggest that this pragmatic view on the effects of censorship holds, even in the highly politicized military conflict between Russia and Ukraine.

Accepted for publication at *Journal of Politics*.

Keywords: Censorship; propaganda; international conflict; war; social media

How to cite this article: Golovchenko, Yevgeniy. Forthcoming. "Fighting Propaganda with Censorship: A Study of the Ukrainian Ban on Russian Social Media." *Journal of Politics*.

This work is part of the 'Digital Disinformation' and 'DIPLOFACE' research projects, both led by Rebecca Adler-Nissen. The former project (number CF16-0012) is funded by the Carlsberg Foundation, while the latter project (number 680102) is funded by the ERC. The use of data was approved by the University of Copenhagen. This project does not publish the data about VKontakte users due to the websites terms of service. The code used for the analysis and data collection from VKontakte's public API is available here at `https://dataverse.harvard.edu/dataverse/jop`. All supplementary material for this article is available in the online appendix. Replication files are available in the JOP Data Archive on Dataverse (`http://thedata.harvard.edu/dvn/dv/jop`). The empirical analysis has been successfully replicated by the JOP replication analyst.

To what extent does censorship reduce activity on banned media among different societal groups in the context of international conflict? The answer to this question is essential if one is to understand the factors enabling or limiting the capacity of the state to control the flow of online information.

Authoritarian and non-authoritarian states alike use censorship to police cyberspace (Howard et al., 2011; Edwards, 2009). In recent years, an increasing number of European states and tech firms have used it to combat digital disinformation and foreign interference. These information threats are often attributed to the Russian government and portrayed by the authorities as a threat to national security and election integrity (Fiorentino, 2018; BBC, 2017). In this context, Ukraine offers some of the most extreme examples of censorship use as a response to information war and online propaganda from abroad.

In 2017, the Ukrainian government issued an executive order, which forced internet service providers to block access to major Russian websites, including VKontakte, the second most visited social media platform in Ukraine (following YouTube) at the time (Roth, 2017). The Kremlin's control over Russian social media was one of the reasons why the Ukrainian government viewed the VKontakte ban as a national security measure against Russian propaganda and surveillance.

How effective are such interventions? Scholarly literature, ranging from media studies to political science, suggests that censorship may successfully limit the overall access to information (Roberts, 2018; Zhang et al., 2017; Morozov, 2012; Kalathil and Boas, 2010). Censorship can also backfire via the so-called “Streisand effect” by drawing attention to the forbidden information, causing political outrage against the censor and increasing public awareness or interest in the forbidden content (Jansen and Martin, 2015). In addition to this, bans may incentivize users to master circumvention through VPN services, thereby gaining access to even more forbidden platforms (Hobbs and Roberts, 2018). Conversely, the decision not to censor ideologically "undesired" content or the inability to fully implement a ban can in some cases help stabilize authoritarian regimes by providing citizens with highly demanded entertainment and news about societal problems in liberal democracies (Kern and Hainmueller, 2009).

In sum, the existing studies documents diverging instances where censorship backfires in some scenarios and successfully limits information in other cases. Similarly, the research on censorship presents examples where censorship is met by public outrage or discontent (Roberts, 2018; Jansen and Martin, 2015) by some groups and political support by others (Esberg, 2020). Despite of this, the existing literature offers little empirical research regarding the extent to which the same censorship policy may affect opposing societal groups differently; i.e. backfiring in one group while being effective in others. This question is crucial when evaluating the consequences of government censorship. Even if a government succeeds in partially reducing the overall online activity on forbidden media, the ban may backfire if the supporters of the regime are more likely to become less active on the censored platform than the opposition. In other words, the government would risk making the opposition more prevalent on the platform than the supporters of the regime.

The Ukrainian context highlights the importance of this question. If the VKontakte ban reduces social activity among pro-Ukrainian users more than among pro-Russian users, the anti-Kremlin policy would make profiles with close ties to Russia more prevalent in relative terms. From a surveillance point of view, this would give Russian authorities less access to data on pro-Ukrainian users (e.g. activists, Ukrainian volunteers or soldiers). From a propaganda point of view, however, the censorship would backfire by leaving pro-Kremlin propaganda even less contested on VKontakte. Drawing on the Ukraine case, this paper seeks to answer the following question: *What is the effect of the Ukrainian ban on online activity among VKontakte users with close affiliations to Ukraine and Russia?*

I approach the question empirically by using publicly available data from VKontakte and a natural experiment research design to estimate the causal effect of the ban on online activity among different user groups. It is important to distinguish online activity – measured as as number of public posts uploaded by the users on their own respective walls – from mere *access* to the VKontakte profiles, i.e. the ability to log on the platform.

The findings indicate that a vast majority of Ukrainians on VKontakte were able to circumvent censorship by logging back on to the forbidden website – likely through tools like VPN. Nevertheless, the Ukrainian government succeeded in reducing the overall online activity among Ukrainians on the Russian platform. Government attempts at curbing Russian influence have reduced the wall posting activity on VKontakte among users with pro-Russian attitudes at least as much as among pro-Ukrainian users. I find the same pattern when comparing citizens in Ukraine with few social ties to citizens within Russia versus those who are strongly embedded in the Russian social network. The increasing costs of using the banned platform (in terms of time and effort) explain the response to censorship much more than do the *social ties* or *political attitudes* toward the states involved in the armed conflict.

It is important to note that the ban has increased the cost of going online to a relatively small extent when considering the accessibility of free VPN tools in Ukraine and the fact that users are not legally persecuted for circumventing the ban. Even a small increase in the cost of accessing the Russian platform is enough to disrupt online activity among pro-Russian (and pro-Ukrainian users), who would rather shift to cheaper and more accessible alternatives. In other words, the accessibility of the media appears to play a much more important role in the decision to use censored social media than do politics or social ties with citizens in the "hostile" state. The findings are in line with 'accessibility view' view on censorship. The theory emphasizes the costs of accessing forbidden content – and not political attitudes – as the main mechanism behind the effect of censorship on online behavior. This view has been previously used to explain the effects censorship among impatient and relatively apolitical users in China through subtle 'friction' in online access (e.g. by slowing down the connection) (Hobbs and Roberts, 2018; Roberts, 2018). This study suggests that the relatively pragmatic view on censorship holds even in the highly politicized context of the hybrid Russia–Ukraine war (Lanoszka, 2016; Reisinger and Gol'c, 2014).

This article contributes to the literature on censorship by addressing the debates on the driving mechanisms behind large-scale bans (Lorentzen, 2014; Bunn, 2015; Roberts, 2018). Furthermore, this research adds to the burgeoning literature on misinformation (Tucker et al., 2018; Nyhan and Reifler, 2015) and propaganda (Peisakhin and Rozenas, 2018; Slutsky and Gavra, 2017; Stukal et al., 2017) by empirically examining the effects of one of the most radical, large-scale policies to combat online manipulation.

Background

Russian authorities view the internet as a strategically important domain closely tied to national security (Vendil Pallin, 2017). Following a series of legal restrictions in 2014, VKontakte and other Russian websites have been under increasing pressure to hand over private information about its users to FSB, the Russian Federal Security Service. This includes information about administrators behind Euromaidan-related pages (Sanovich, 2017, 12). VKontakte's founder, Pavel Durov, initially attempted to resit the pressure from authorities (Pan, 2017). Eventually, he was pushed from the firm by major shareholders and left Russia (AFP, 2014). Today, VKontakte is predominantly compliant with Russian authorities, who have used the data in multiple criminal cases against individuals, some being prosecuted for anti-government social media posts (Interfaks, 2018; Robinson, 2018).

Ukrainian authorities responded to the Russian government's increasing influence over Russian social media in 2014 by *advising* citizens to delete their accounts on Russian-owned social media websites (Boichak and Jackson, 2019, 13) before proceeding to a censorship policy in 2017. The VKontakte ban is part of a decree that imposed sanctions on 468 organizations and 1,228 individuals, including the Odnoklassniki social media platform and the Yandex search engine. VKontakte, however, was by far the most popular social media platform in Ukraine among all of the sites on the list.

The goal behind the decree is "to protect the national security and territorial integrity of Ukraine" (Dek, 2019). According to Oleksandr Turchynov, the secretary of National Security and Defense Council of Ukraine at the time of the ban, the forbidden websites were used for Russian propaganda, state-driven surveillance and cyberattacks against Ukrainians

(Andrusieczko, 2017). The Ukrainian authorities argue that the ban is an important countermeasure against these information threats following the Russian annexation of Crimea in 2014 and the information war on Ukraine (Kiselyova and Prentice, Kiselyova and Prentice; RFE/RL, 2017). As a result of the annexation, the Ukrainian government could only fully implement the ban in territories under its control and not in Crimea. For the purpose of this article, I refer to the affected region north of the Crimean peninsula as "Mainland".

To this day, Ukraine remains divided by political tension between pro-Ukrainian citizens (consisting of both ethnic Ukrainians and Russians), who support Ukrainian sovereignty, and the pro-Russian minority, who praise and support the Russian Federation or Kremlin-backed separatists in southeast Ukraine (Laruelle, 2014, 2016).

VKontakte is a popular source of entertainment in the post-Soviet space and it has a history of being an important source of pirated music (Popkova, 2019; Kiriya and Sherstoboeva, 2015). Following the Kremlin's turn toward strict Internet regulations, VKontakte has fallen under great pressure and control from Russian authorities (Pan, 2017; AFP, 2014). While most of the content on the platform is unrelated to politics, researchers, journalists, and authorities have argued that VKontakte is also a platform for pro-Russian propaganda (van der Vet, 2019; Dek, 2019; Volchek and Sindelar, 2015), disinformation, Russian surveillance, cyberattacks (Andrusieczko, 2017) and recruitment ground for the separatist movement (DW, 2017). In this sense, the VKontakte users are potentially exposed to pro-Kremlin content, either through their pro-Russian friends and family or their newsfeed.

The Russian surveillance "threat" is relevant for both civilian and military targets. Shklovski and Wulf (2018) find that despite military regulations, Ukrainian soldiers in the war zone use social media (including VKontakte) to search for information and to maintain personal contacts, despite army regulations. This even applies to soldiers who are aware that the enemy may potentially use media surveillance to geolocate Ukrainian positions and to select targets for artillery strikes (Shklovski and Wulf, 2018, 7,10).

The VKontakte ban took place in a time of war but simultaneously also in a relatively democratic context, where citizens still enjoy legal access to a wide range of media outlets and platforms. The ban was not intended to reduce the overall access to social media, but rather to push users away from the Russian VKontakte and toward other media alternatives, such as Facebook, Twitter and Instagram, which are controlled by neither the Russian nor Ukrainian state.

Theory: Mechanisms behind censorship

The literature on censorship describes multiple factors that may cause people to abandon or desist from consuming, sharing, or producing forbidden content as well as instances where censorship "backfires". Overall, this literature can be divided in to at least two strands, each with its own theoretical view on the mechanisms behind censorship. On one hand, censorship may work by increasing the “costs” of accessing forbidden content in terms of time, money, or effort. I refer to this as the “accessibility view” on censorship. On the other hand, censorship can be seen as a signal that calls the citizens to fall in line with government policy, creating fears of reprisal or social exclusion. For the sake of this paper, I will refer to this line of thought as the “political signaling view.” As I will argue below, the two strands do not only explain "successful" censorship, but also unintended outcomes that go against the censorship goals or undermine the censors.

The accessibility view on censorship

The accessibility approach emphasizes the *accessibility* of the forbidden media and the apolitical evaluation of whether accessing content is difficult or not. Instead of banning undesired content entirely, the government may effectively limit user engagement with the content by creating what Roberts (2018) refers to as *friction*: slowing down connections or blocking content in ways that can still be accessed through circumvention tools, such as VPN. By putting up minor barriers to censored content, the government increases the incentives to use more easily available, non-censored alternatives, while simultaneously avoiding the full-scale persecution of the masses and the potential political backlash that may follow (Roberts, 2018; Dickson, 2016).

Numerous studies indicate that citizens in authoritarian states respond to censorship by accessing available content instead (Chen and Yang, 2017; Stockmann, 2013). Interestingly, the literature based on the accessibility view emphasizes that the users' impatience and apolitical use of social media may also backfire – especially if the censored platform cannot be replaced by a similar and freely available alternative to fill in the gap (Hobbs and Roberts, 2018). For instance, Hobbs and Roberts (2018) describe how the abrupt Chinese government ban on the relatively apolitical Instagram increased the online traffic toward prohibited and more political platforms such as Twitter and Facebook. They argue that this pattern is driven by a "gateway effect." The concept refers to a mechanism whereby the motivation to access a newly banned platform opens up for skills (i.e., how to use VPN) that give access to other websites that have long been censored (Hobbs and Roberts, 2018, 623-624).

It is important to stress that this perspective does not explain the response to censorship by referring to the users' political attitudes toward censorship or the government. On the contrary, this strand of literature is in line with the so-called "cute cat theory", which emphasizes that the search for entertainment — not political attitudes — is the main driving force behind the citizens' consumption of online content (Hobbs and Roberts, 2018; Zuckerman, 2014, 624).

There are few studies that test the "accessibility view" on censorship in a democratic context, likely due to the literature's focus on authoritarian or hybrid regimes. The view, however, is highly relevant in democratic or semi-democratic societies. This is the case, precisely because liberal states are ideologically bound to prioritize "softer" tools - such as adjustments to the cost of going online (in terms of time and effort) - over large-scale persecutions and fear of reprisal.

The Political signaling view on censorship

Whereas the "accessibility view" emphasizes the more practical and apolitical aspects of censorship, the "political signaling view" shifts the focus towards the political signal behind censorship as well as the role of political attitudes in the public response towards such policies.

The political-signaling approach sees state-driven censorship and propaganda as a means of demarcating forbidden and undesired conduct. Governments can therefore use censorship as a political signal to socialize citizens into the “desired” behavior – either through implicit political messages, fear, or uncertainty (Stern and Hassid, 2012; Huang, 2015).

For instance, Stern and Hassid (2012) reveal how the Chinese government uses an atmosphere of uncertainty to govern journalists, editors, and lawyers. Not only is the exact line between forbidden and allowed information unclear, it is also constantly changing. This leads to politically engaged individuals creating their own bottom-up explanations for which content is forbidden (and why) and to internalize a practice of self-censorship (Stern and Hassid, 2012, 1240-1241).

This literature strand does not only explain successful censorship, but also instances where such interventions backfire by provoking a political response. The political signals inherent in censorship stand as a possible driving mechanism behind the previously mentioned "Streisand effect". Here, the apparent censorship generates more traffic towards the forbidden content by making people 1) more aware of the forbidden content, 2) more curious and 3) causing public outrage (Jansen and Martin, 2015). While the first two aspects of the Streisand effect are not political per se, the third mechanism generates a backfire effect due to a political response against the censorship policy itself. Using a more theoretical approach, (Shadmehr and Bernhardt, 2015) argue that censorship may cause the population to evaluate the regime more negatively in some instances, because the lack of information may promote a belief that "...there *might* have been bad news that was censored" (Shadmehr and Bernhardt, 2015, 280). In their empirical study, based on news media content data and approval surveys, Gläßel and Paula (2020) find that the German Democratic Republic’s censorship of information about the emigration crisis in 1989 backfired by causing outrage among people who detected state-misinformation through access to Western television.

Hypotheses

In this paper, I use a multi-dimensional approach to analyze user affiliations with Russia and Ukraine. First, I examine the political aspect of the affiliations by distinguishing between users with pro-Russian and pro-Ukrainian attitudes; the former refers to praise and support for Russia, the latter to support of the Ukrainian national state. Second, I capture the social aspect by examining social ties to individuals living in Russia.

My hypotheses and initial theoretical expectations are grounded in a political signaling view for several reasons. Unlike the more covert or subtle 'friction' described in the accessibility literature (using China as a case), the Ukrainian ban against VKontakte was officially announced by the censors and widely discussed by the public. The censorship policy itself was described as a *political* move against Russia and a matter of national security. The government, various journalists and civil society groups framed the use of VKontakte as a political choice, a lack of patriotism or even a part of the Russian information war against Ukraine. Using VKontakte in this heavily politicized context can be interpreted in the Ukrainian public as indifference toward patriotic ideals in times of war, whereas compliance with the ban can be a sign of political loyalty toward Ukraine as a nation-state.

If one sees censorship as a way for the Ukrainian government to signal that using VKontakte is both unpatriotic and undesired by the state, one would expect pro-Ukrainian users to become less active on the Russian platform out of political support for Ukraine and resentment toward the Russian "aggressor state" or due to a fear of being stigmatized as "unpatriotic" by their Ukrainian peers. Following this line of thought, I expect the ban to succeed overall, because the majority of the Ukrainian population are relatively supportive towards Ukraine's sovereignty, while the pro-Russian individuals remain a minority (Arel, 2018). I therefore begin with the following hypothesis:

Hypothesis 1: The censorship reduces online activity among VKontakte users in Mainland.

It is important to note, however, that the accessibility view would predict a similar outcome for different reasons: The ban would reduce the online activity by forcing users to install a

VPN, regardless of their political orientation, and therefore 'slow' down the login procedure or the connection. However, there is an important difference between the two views when it comes to the role of political attitudes. According to the logic in the political signaling view, users with strong Russian affiliations would be less affected by the ban for at least two reasons: Firstly, they may be less affected by the social pressure to reduce their activity on the Russian platform because they are already stigmatized as "pro-Russian" in the context of the armed Russian-Ukrainian conflict. Secondly, the Russian-affiliated users would be more politically outraged by the "anti-Russian" policy and therefore more likely to resist a ban that is heavily embedded in anti-Kremlin sentiment. This theoretical expectation contrast to the more pragmatic accessibility view on censorship, which would predict an equal decline in both groups due to increased costs of going online, regardless of the users' political attitudes towards Russia and Ukraine. Using this view as a point of departure, I formulate the next hypothesis:

Hypothesis 1a: Pro-Ukrainian users are more affected by the censorship than pro-Russian users.

As mentioned earlier, the Ukrainian government sees VKontakte not only as platform for Russian surveillance, but also a pathway of pro-Kremlin propaganda. From a theoretical point of view, VKontakte can be seen as a platform for 'participatory propaganda', where the Ukraine-Russia conflict is socialized by ordinary users (Asmolov, 2019), who actively take part in producing and sharing propaganda with their friends in Ukraine and abroad. The cross-national ties are key in this context. A Ukrainian user with many social ties to individuals in Russia may be heavily embedded in Russian society, either through friends, family, or by having previously lived in Russia. The user likely has high potential exposure (Hjorth and Adler-Nissen, 2019) to pro-Kremlin news, political narratives, and worldviews that are widespread among contacts in Russia (Toal and O'Loughlin, 2017). One would therefore expect these users to be more pro-Russian and more likely to resist the anti-Russian ban, which may essentially jeopardize their online ties to peers in Russia. In line with this, I supplement the hypothesis above with the following:

Hypothesis 1b: Users with fewer social ties to individuals living in Russia are more affected by the censorship than users with many social links to Russia.

According to a report published by NATO StratCom, the overall number of wall posts in Ukraine did fall substantively following the ban (Dek, 2019, 58). It is worth noting that pro-Russian posts increased, which, according to the authors, indicates "users moving to the pro-Russian infosphere" (Ibid.: 51). This alone is in line with my theoretical expectations based on the "political signaling view" above. However, the report does not use an explicit natural-experiment design, nor does it compare the effect of the ban among users with strong and weak affiliations to Russia. The report presents little information on the sampling strategy used to select users or posts, nor does it document the clustering algorithms used to infer the ideology of posts. While the report may offer useful and indicative results, it is difficult to draw causal conclusions on the effect of censorship among different societal groups due to above-mentioned issues. Further systematic inquiry is needed if one is to understand how the effect is mitigated by individual-level factors.

Censorship as an exogenous shock

Before proceeding further, I will first present arguments for why the ban can be seen as an exogenous shock to the online media ecology (i.e., the timing of the ban being largely unrelated to online activity on VKontakte).

The executive order to ban a list of Russian websites, including VKontakte, was signed by Ukrainian President Petro Poroshenko on May 15, 2017 and publicly disclosed the next day, with no prior announcement. Prior to the ban, there had been only few waves of interest in the topic in February, when a Ukrainian public official suggested that the government should ban VKontakte, with no indication of whether this would occur or when (see Figure A1).

In this sense, the sudden ban can be seen as an exogenous shock to the online environment, with the exact timing being largely unanticipated by the public and therefore "near-random" from the perspective of the user. This is important, because knowledge of the ban date would give users the opportunity to prepare: By decreasing their activity in advance (e.g. migrating to other platforms) or increasing activity before the ban in defiance against the policy. This could bias the estimated effect size of the actual ban. Because the exact timing of the actual censorship in 2017 is unexpected by the users, the period immediately before the ban may approximate the counterfactual scenario where the "treatment", in the form of a ban, did not occur. From a research perspective, this enables a natural experiment setting with a relatively clear delimitation between the period before (control) and after the ban (treatment) separated by a narrow window of up to 2 weeks for the policy to be fully enacted.

According to journalists, some users began reporting the effects of the ban as early as the evening of May 17 (UNIAN, 2017), although the ban was not fully implemented among all internet service providers at this stage. Due to technical constraints and privacy concerns, it is not possible for me to pinpoint the exact hour of the ban for each individual user. For pragmatic reasons, I treat 00:01am on May 18 as the beginning of the full-scale implementation. As I will show in the analysis, posting activity fell drastically on this day.

The Ukrainian government cannot implement censorship in all of its regions, because parts of its territory has been annexed by Russia in 2014. I use this spatial variation to distinguish between users who have not been exposed to the Ukrainian ban, as they are located in Russian-occupied Crimea, and those who were likely exposed to the ban, as they are located in "Mainland", i.e. on the non-occupied side of the Crimean border in Kherson Oblast.

I use Crimea for comparison—and not neighboring regions countries like Russia, Belarus or Poland—due to its relative similarity to adjacent Ukrainian regions as well as a clear border separating censored and non-censored territories, unlike the frontline in the war-torn Donbass region (see Appendix G online). It is important to note that the geographical border separating the censored and non-censored users is not random. For this reason, the spatial delimitation does *not* in itself enable a natural experiment research design, unlike the as if-random, temporal “line” separating the censored and non-censored time period. However, the comparison of online activity across geographic space adds both valuable nuance and validation to the comparison of online activity over time.

Data

To compare the change in VKontakte before and after the ban, I collect publicly available data on VKontakte from the company’s own Application Programming Interface (API). All of the user-level data has been collected during August 2018, while the collection of wall posts took place in August-September the same year (see Appendix T online for an overview of the collected data). The data collection consists of six steps.

First, I identify all of the cities in the VKontakte database that are located in administrative regions ("oblast") adjacent to the border separating Crimea and Mainland.

In the second step, I use Google Maps API to automatically locate coordinates for each city. I test the reliability of the automated approach by manually examining the results for each city. Of the 370 cities, only six have been misplaced in the automated process, which I have then corrected manually.

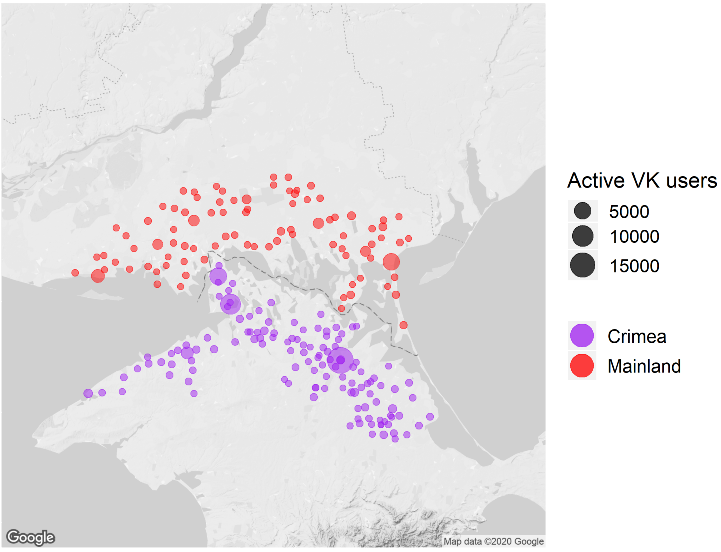
Third, in order to filter out cities located more than 50 km from the border, I use a Geographic Information Systems (GIS) approach to compute the distance between each city and the nearest point at the administrative land border separating Crimea from Mainland. This relatively narrow delimitation is intended to *maximize* the regional similarity between the towns exposed to censorship in the "treatment" region and towns in the "control" region that have not been exposed, while simultaneously keeping the distance bandwidth broad enough to capture a sufficient number of public profiles from different political groups.

In the fourth step, I sample the individuals by using the VKontakte API to return data on up to 1000 users from each city within 50 km of the border. I refer to this initial group of users as the “seed”.

[0] For this purpose, I have used and modified code from the vkR package in R developed by Dmitriy Sorokin: `https://cran.r-project.org/web/packages/vkR/index.html`.

In the fifth step, I develop a *crawler* to initiate a snowball sampling with the seed group as the initial starting point. In this crawl, I first download the meta-data for the friends of the seed group, only keeping those that live in cities within the 50 km distance from the border according to the user's self-registered data. I then download the metadata for the friends' friends, continuing the same procedure six times. Due to the highly interconnected nature of both online and offline networks, this snowballing strategy yields a nearly complete sample of 109,191 users within 50 km of the border (as well as their 3.2 million friends from around the world with self-registered locations). Approximately 65,213 of the users within the 50 km border logged on to the website for the last time at least 30 days prior to the ban or during the period after the ban. Like the rest of the data, the last login dates for the public accounts are collected using VKontakte's API. This information is also a highly visible feature of the public profile and easily accessible by all VKontakte users. In order to ensure that the data is not biased toward users with high posting activity *after* the ban, this sampling step is carried out independently of the users' posting behavior (the data therefore also includes accounts with no public posts). The distribution of these users on both sides of the border between Mainland and Crimea is illustrated in Figure 1.

Figure 1: VKontakte users within 50 km of the Crimean border. All of the users have logged in within 30 days of the ban.



In the final step, I select a smaller sub-sample of 3,553 active users and all 2,127,398 public posts from their respective walls to capture individual-level online activity before and after the ban. I select the individuals using stratified sampling to ensure that I have users in the different political groups from both sides of the border (approximately 590 users per group in in Mainland and 660 in Crimea), as well as randomly selected users from the pool of 65,213 active profiles. I then remove 51 profiles that have posted on average more than three posts daily (26,456 posts in total) in the 90-day period *prior* to the ban (treatment) in order to filter out extreme outlier accounts that are likely automated (i.e. inauthentic bot accounts) (Varol et al., 2017). These profiles account for less than 1.5% of the original sub-sample. The findings in this paper remain robust when also including the hyperactive users. While VKontakte API provides data on an account's last login time, the website does not disclose when the account was created. In order to ensure that the profiles have existed prior to the ban, I further delimit the data to the 3,024 users who have uploaded at least one public post on their own walls prior to the May 18, 2017 implementation date. Descriptive statistics for the final sub-sample used in the analysis are available in the online Appendix A. The findings remain robust also when including users without a single public post prior to the ban (see Appendix Q online). In total, the sub-sample in the analysis section includes 1,067 pro-Russian, 1,112 pro-Ukrainian and 845 users from both sides of the border. While the random users are not necessarily politically neutral, they are likely less pro-Russian or pro-Ukrainian than those who publicly expose their political affiliation. I use the information about the users last login date to examine whether users accessed their VKontakte account after the ban.

The data from VKontakte is used to operationalize user activity, geographic location, social ties to Russia, and political attitudes toward Russia and Ukraine. These variables are described below.

User activity is operationalized as the number of public posts written by the users on their own profile walls. In this sense, *user activity* is different from mere access to an account.

Users may, for instance, circumvent the ban through VPN tools to passively access their accounts without generating any public posts on their walls.

Geographic location is operationalized using the profiles' self-reported information on "Current city" in August 2018.

Social ties to Russia is measured as the proportion of users' friends who have publicly indicated that they live in Russia.

Political attitude toward Russia and Ukraine. I operationalize Pro-Russian users as those who only follow pro-Russian VKontakte community pages, while pro-Ukrainian users are those who only follow pro-Ukrainian communities (see Appendix B-D online for details on sampling and coding of the communities). While the military conflict is often discussed in Russia in terms of an ethnic divide (Shklovski and Wulf, 2018), speaking Russian does note equate with Russian identity, and Russian identity among Ukrainian citizens does not necessarily imply pro-Russian attitudes in the current Russia-Ukraine conflict (Arel, 2018).

Difference in Differences specification

In this section, I will describe the empirical setup for measuring the effect of the ban on online activity. All of the models in this paper are based on individual-level data, where each "user-time" observation represents a number of wall posts uploaded by the individual users on a given day.

I use Difference-in-Differences (DD) approach (Card and Krueger, 1993) to estimate the average effect of the ban on user activity among the individuals in the sub-sample living in Mainland within 50 km of the border.

Minor fluctuations in posting activity are common on both sides of the border, both before and after the ban (see Appendix Figure A2 online). One could reasonably assume that online activity in Mainland would have experienced a small and gradual decline - even if the ban would not have occurred. This means that a simple comparison of posting activity before and after the ban only in areas exposed to censorship (treatment) runs the risk of overestimating the effect size. For these reasons, I introduce users from non-censored Crimea as an additional control.

The advantage of using DD design in this context is that it allows treatment and control groups to be different, since the method is not based on the assumption that treatment (the ban) is randomly assigned. The key argument for using this design is the similarity in time trends between the two groups - that the Crimean trend approximates the counterfactual scenario in Mainland where the ban did not occur.

I specify the DD regression using the following OLS model:

$$y_{it} = \alpha + \beta_1 Ban_{it} + \beta_2 Mainland_{it} + \beta_3 Ban_{it} * Mainland_{it} + \delta X + \varepsilon_{it} \tag{1}$$

In this equation, y_{it} indicates the mean number of posts uploaded by user i during day t, where the earliest implementation date (May 18, 2017) is standardized as 0. Ban_{it} is binary variable indicating whether the posts are uploaded by the user before or after the ban and $Mainland$ reflects whether the user lives in Mainland ($Mainland_{it} = 1$) or Crimea ($Mainland_{it} = 0$). The interaction term, $\beta_3 Ban_{it} * Mainland_{it}$ indicates the main effect of the ban when using Crimea as a control. In the other words, the term reflects the difference between the change in mean number of posts per day in Mainland (treatment group) and Crimea (control group). The main hypothesis predicts that the ban has a negative effect on posting activity ($\beta_3 Ban_{it} * Mainland_{it} < 0$). Additional control variables include total number of VKontakte friends per August 2018 and a dichotomous variable indicating whether the post was written during the weekend as well as the users' self-reported gender.

To test whether the effect is mitigated by political affiliations and social ties to Russia, I create two additional DD models only for pro-Russian and pro-Ukrainian users from Mainland. Here, I replace $Mainland$ with 1) a binary variable indicating wither the user has above median or up to median proportion of VKontakte Friends and 2) political affiliation toward Russia and Ukraine, respectively. If the effect of the ban is mitigated by these factors, the respective interaction terms will be significantly different from zero.

In the final step, I examine the difference between the effect size of the ban among the various groups by introducing Crimea as a control. The literature refers to this approach as the Difference-in-Difference-in-Differences or Triple Differences (DDD) method (Yelowitz, 1995). More specifically, I expand the equation above by introducing a three-way interaction between *Ban*, *Mainland*, and a dichotomous variable, *Pro-Russian*, indicating whether or not the user is pro-Russian:

$$\begin{aligned} y_{it} = \alpha + \beta_1 Ban_{it} + \beta_2 Mainland_{it} + \beta_3 Pro\text{-}Russian_{it} + \beta_4 Ban_{it} * Mainland_{it} \\ + \beta_5 Ban_{it} * Pro\text{-}Russian_{it} + \beta_6 Ban_{it} * Mainland_{it} * Pro\text{-}Russian_{it} + \delta X + \varepsilon_{it} \end{aligned} \tag{2}$$

This triple difference is indicated by the interaction term $Ban_{it} * Mainland_{it} * Pro\text{-}Russian_{it}$, which can be broken down into several parts. The interaction $Ban_{it} * Pro\text{-}Russian_{it}$ shows the difference between the change in mean posting activity among pro-Russian and pro-Ukrainian profiles in Crimea (which is in the reference category). This provides an estimate of how much more pro-Russian users are affected by the ban than the pro-Ukrainian ones. By introducing $Mainland_{it}$ to the interaction term, the estimate shows the difference between the estimate for $Ban_{it} * Pro\text{-}Russian_{it}$ in Mainland (treatment) and Crimea (control). If, for instance, the difference between the pro-Russian and pro-Ukrainian users is identical on *both sides of the border*, the interaction term will be close to zero. I use the same approach by replacing $Pro\text{-}Russian_{it}$ with a binary variable, *Friends in Russia*, which indicates whether the users have above(*Friends in Russia = 1*) or up to median proportion of VKontakte friends in Russia (*Friends in Russia = 0*). Because users in Crimea have on average higher proportion of VKontakte friends in Russia, I compute two separate median values for users in Mainland and Crimea respectively.[1] I use cluster robust standard errors for all of the DD and DDD models in order to take into account the structure of the panel data, where there are multiple "user-time" observations for the same individuals across time.

Last logins

Before proceeding with the analysis of daily posting activity, I will first examine whether users log in to their accounts after the ban or whether they last did so during the ban period.

[1] The findings remain robust also when using one single median value for profiles on both sides of the border.

"Last login" in this case, refers to the last time a non-deleted account has logged in for the last time per August 2018, the data collection period.

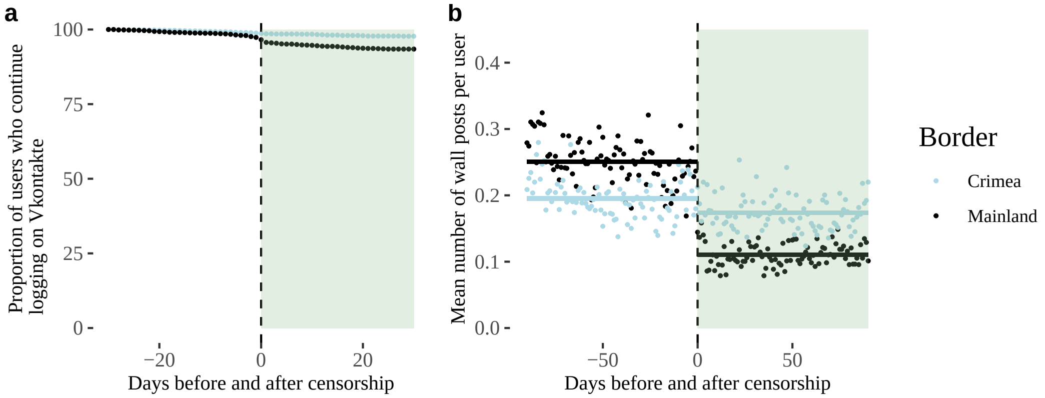
Of the 109,191 accounts from both sides of the border in this study, 65,213 accounts (59.7%) have last logged in during the 30-day period prior to the ban and the data collection date. The remaining 40.3% have logged in for the last time at least 1 month before the ban or earlier.

The findings suggest that the vast majority of the non-deleted accounts continued to log back in *after* the ban. Approximately 94.1% of the 23,506 “active users” in Mainland logged in at least once after the implementation date, and 90.8% continued logging in more than 30 days after the ban. For users in Crimea, the figures are 99.0% and 98.3%, respectively (see Appendix H online for an in-depth comparison between Mainland and Crimea). Of all the 1,433 users in the sub-sample from Mainland, who have posted at least once before the ban, 93.4% continued logging in more than 30 days after the ban, even though the policy remains both legally and technically in place to this day. For pro-Russian, Pro-Ukrainian and random users in Mainland the proportions are 96.4%, 91.8% and 91.4% respectively. The pattern is illustrated in Figure 2a and the online Appendix Figure A3. The difference between the proportions for the pro-Russian and pro-Ukrainian users is statistically significant ($p < 0.05$), however, not substantively large. This indicates that a large percentage of users have found a way of avoiding the censorship, some through freely available VPN services. However, these numbers do not include the change in posting activity among the users who have bypassed censorship. I will therefore turn to an analysis of a smaller sub-sample of users from both sides of the border in order to estimate the impact of the ban on their daily posting activity.

The overall effect of the ban

How did the Mainland posting activity change compared to northern Crimea within just 50 km, where there was no ban? To answer this question, I will use DD to estimate the mean effect of the ban on posting activity in Mainland. In this part of the analysis, I will use a sub-sample of 3,024 profiles from pro-Russian, pro-Ukrainian and random users from both sides of the border. Each of the 547,344 user-day observations reflects the number of posts uploaded by the respective individual on a given day within the bandwidth of 90 days before and after the cutoff date (181 days in total), standardized as 0. The descriptive Figure 2b shows the observed mean number of posts per user for all of the 3,024 users in the sub-sample. The figure illustrates the drastic reduction in the daily posting activity for all Mainland users. In comparison, the ban is accompanied by a relatively small, continuous decline in Crimea.

Figure 2: Last logins and posting activity

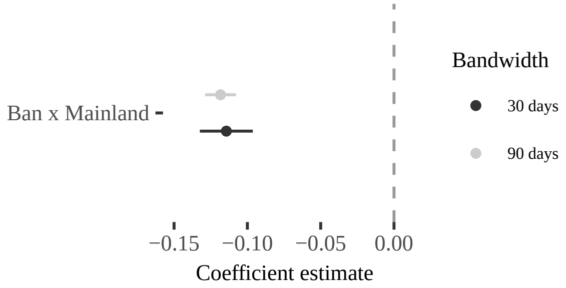


a Proportion of users who have access to their VKontakte profile. The users are considered to have access to VKontakte until their very last login date. The figure is based on the assumption that all of the 3,024 accounts existed throughout the entire period before the censorship. This can be confirmed for at least 98.84% of the accounts (see the description in Figure A3 for details). Note that the decline begins before the ban. This is due to the fact that the figure is based on last logins: i.e. some users login for the last time not knowing that they will be locked out by the sudden ban. **b** Posting activity for all 3,024 accounts. Horizontal lines reflect observed means for the respective periods and groups.

The effect size of the sudden censorship policy in the DD model is reflected by the interaction term *Ban*Mainland* in Figure 3 and online appendix Table I1. The ban reduced the posting activity for the average user by 0.8 posts per week or 0.114 posts per day ($t = -12.4075, p < 0.01$) in Mainland, when using a 30-day bandwidth before and after the ban (61 days in total). In other words, Mainland online activity (treatment group) declined by an additional 0.114 daily mean number of posts for the average user, compared to the change in Crimea (control group).

This effect size is substantively large. The ban decreased the daily mean number of posts per user by 45.42% according to the most conservative estimate (based on 30-day bandwidth). When using the 90-day bandwidth, the decline is equivalent to 47.01% of the pre-ban daily mean number of posts.

Figure 3: Change in mean number of posts per day after the ban (95% confidence intervals)



The figure reflects models 1 and 2 in Table I1, which are based on data from all 3,024 users (including pro-Russian, pro-Ukrainian and Random users)

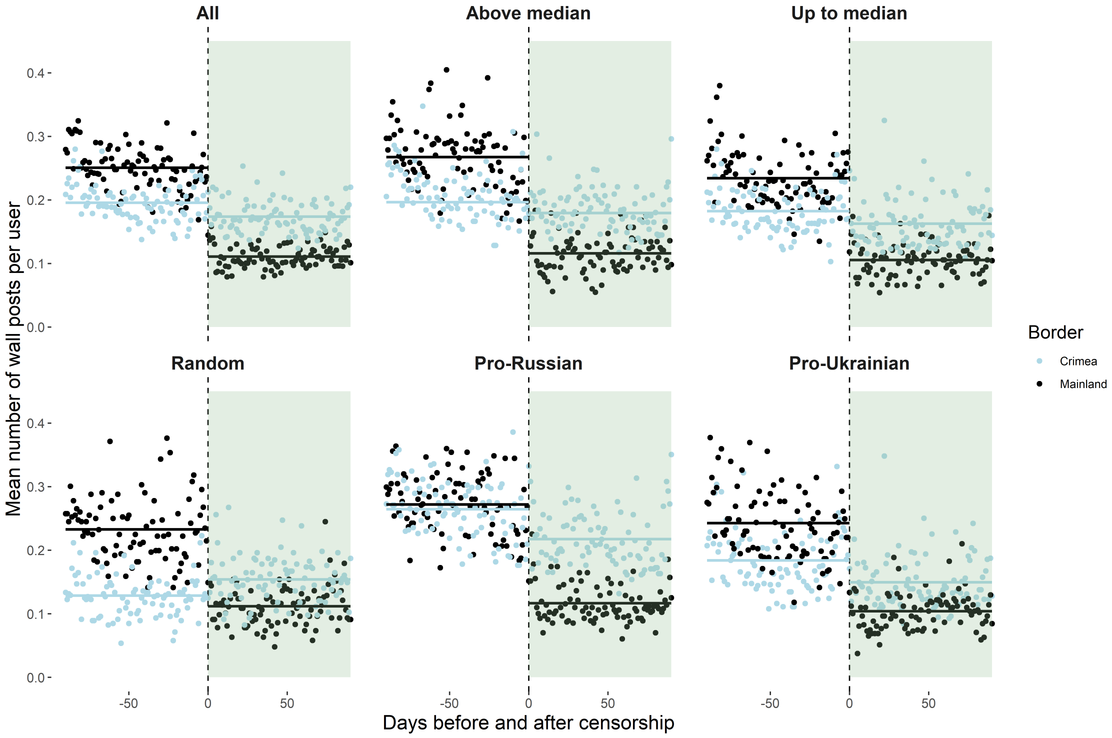
The difference in effect size

The findings above confirm the first overall hypothesis: The ban reduces the online activity on VKontakte in Mainland, also when examining posting activity. This leads to the following question: To what extent is the censorship effect mitigated by different factors?

Figure 4 gives a descriptive overview of the observed mean number of posts per user for each day for the different groups and all of the users combined. The change in posting activity appears strikingly similar when visually comparing user groups with different (political and social) affiliations to Russia and Ukraine. In this section, I will use DD and DDD models to test whether the difference is statistically significant and substantively large when including control variables.

I will now turn to this question by testing Hypothesis 1a, which predicts that the anti-Kremlin ban affects pro-Ukrainian users more than pro-Russian users. This part of the analysis is delimited to pro-Russian and pro-Ukrainian users (without random profiles) in order to enable the comparison between the two groups. The variables of interest for the

Figure 4: Change in mean number of posts per day after the ban (for all users and the split-sample)



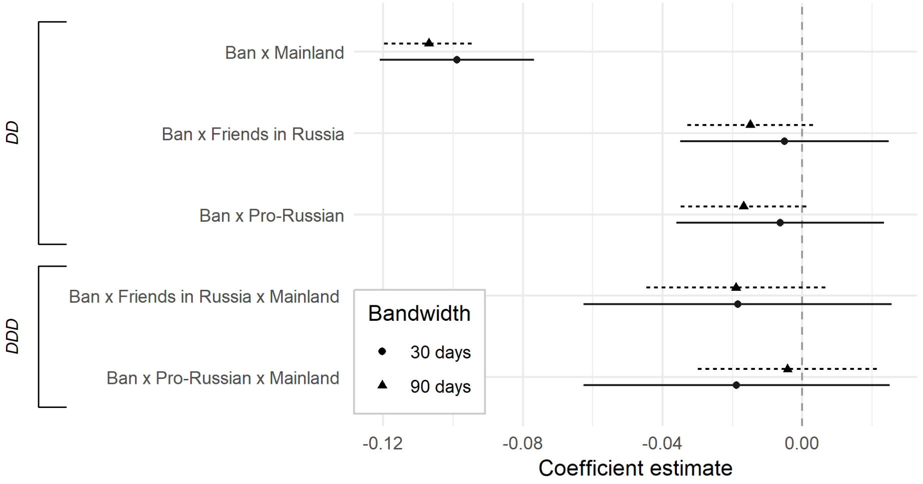
The figure shows posting activity for all users, as well as the split samples: Pro-Russian, Pro-Ukrainian users, as well as those with up to median and above median proportion of friends in Russia. Horizontal lines reflect overall means for the respective periods and user groups.

DD and DDD models are illustrated in Figure 5.

I find no statistically significant difference between the two groups when using DD and a significance level of 5% to examine the data only from Mainland. The difference for the DD models is indicated by the coefficient for the interaction term *Ban*Pro-Russian* in the output in Figure 5 and online appendix Table I2 (Model 3). While the difference is statistically significant at a 10% significance level for the 90-day bandwidth, these estimates indicate that pro-Russian users are *more* affected by the ban than pro-Ukrainian users. Here, the ban reduced the average daily activity among pro-Russian users by 0.017 posts more than their pro-Ukrainian counterparts ($t = -1.8229, p < 0.1$), contrary to the hypothesis.

Similarly, the results show little evidence supporting the hypothesis, when adding Crimea as a control to the DDD model. Here, the difference between pro-Russian and pro-Ukrainian

Figure 5: Change in mean number of posts/day after the ban (95% confidence intervals)



The figure reflects the models in Table I2 and I3. The models are based on data from users with political affiliations (and not random users). The order of the interaction terms is based on the model numbers, so that the first interaction term (top) is based on Model 1 and the fifth interaction term (bottom) is based on Model 5.

users is reflected by the interaction term *Ban*Pro-Russian*Mainland* in Figure 5 and Table I2 (Model 5). According to the most conservative estimates – based on a 90-day bandwidth – the effect size of the ban is 0.004 posts greater among pro-Russian users than pro-Ukrainian ones ($t = -0.3202, p = 0.7488$), contrary to the hypothesis. While this difference appears even greater, when narrowing the bandwidth down to 30 days, it remains statistically insignificant. The data therefore provides little evidence in support of Hypothesis 1a.

I find similar results when testing Hypothesis 1b, which predicts that individuals with fewer social ties to users in Russia are more affected by the ban than those with extensive social ties linking to the country. I examine the effect by comparing users with an above-median percentage of VKontakte friends living in Russian to those with an up-to-median proportion. There is no statistically significant difference in how much the two groups are affected when only using data from Mainland. This is reflected by the *Ban*Friends in Russia* interaction term in Figure 5, when using a bandwidth of 30 and 90 days. The pattern is consistent when introducing Crimea as an additional control in the DDD model, by including the following interaction term: *Ban*Friends in Russia*Mainland.* According to the most conservative DDD estimate, the ban has reduced online activity by 0.0185 posts ($t = -0.8212, p = 0.4115$) more among users with above-median proportion than those with fewer friends in Russia – contrary to the theoretical expectation. However, this difference remains statistically insignificant. These results provide little evidence supporting Hypothesis 1b.

I find similar results when rerunning the models behind Figure 5 with user-level fixed effects (see Figure I1 in the online Appendix) or negative binomial regressions instead of OLS. Furthermore, the findings remain robust when testing for long-term effects by expanding the time period, utilizing Regression Discontinuity in Time models instead of DD, using different independent variables (including strength of ties to Russian users) and reiterating the analysis on users from Kyiv. The results based on DD and DDD remain the same when delimiting the analysis to users who managed to technically circumvent the censorship by logging in at least 30 days after the ban. In other words, the censorship reduced VKontakte activity even among users who were capable of circumventing the ban. I validate the political affiliation variable by examining whether users labeled "Pro-Russian" are more connected to Russia than "Pro-Ukrainian" users through friendship networks, reposts and use of Russian websites, as one would expect. I find that this is the case. I have conducted a placebo test by reiterating the analysis on the same dates one year prior to the ban. By doing so, I show that the results are not driven by reoccurring seasonal events. Appendix J provides an in-depth overview of the robustness and validity tests (available online).

A pragmatic response to censorship

The vast majority of pro-Russian and pro-Ukrainian users continue logging in on the forbidden website despite of the ban. Pro-Russian users are slightly more likely to maintain passive access the platform than their pro-Ukrainian counterparts. However, the difference is not substantively large and they are at least as affected when it comes to online activity. Why do users, who can legally access VKontakte and have the technical know-how to do so, still choose to shift their attention away from the Russian platform?

From the “political signaling” point of view, the decision to either contest or comply with the government ban is driven by political attitudes toward the regime. From this perspective, the ban is effective because it actively signals that using the online services provided by the “aggressor state” is unpatriotic and a potential threat to national security. In other words, the underlying reason for reducing activity on the platform is more political than practical. One would therefore expect pro-Ukrainian individuals and those with fewer social ties to Russia to be more compliant with the anti-Kremlin ban on Russian social media by reducing their online activity. The findings present little evidence for this view; on the contrary, if there is a difference between the two groups, pro-Russian users and those with relatively many social ties to Russia are more likely to comply with the anti-Russian ban by reducing their online activity on VKontakte.

The potential difference between the two groups cannot be explained by fears of legal reprisals or social stigma alone. More than 90% individuals in the sub-sample continued to log in to their public accounts at least one month after the ban without hiding their last log-in date. This means that their friends, strangers, and authorities alike can see that the respective users continued using VKontakte at least one month after the ban. This information is visible even if the users do not write any posts.

Implications

I will argue that the behavioral response to censorship in Ukraine is largely in line with the competing “accessibility view” on censorship, represented by scholars such as Hobbs and Roberts (2018). Seen from this theoretical perspective, the response to censorship is driven by consumer impatience and the increasing costs of going online and to a lesser extent political affiliations with Russia or Ukraine. This view explains why pro-Russian users and those with potential exposure to Russian information networks are no more likely to resist the anti-Russian ban than pro-Ukrainian users or those with few social ties to Russia.

The Ukrainian ban offers a hard case for the practical accessibility view on censorship.

Not only does the government actively politicize Russian media as a means of information warfare against Ukraine, the users are also encouraged to demonstrate loyalty to their country by ceasing to use services from the aggressor state. Yet the findings suggest that the mechanisms described by Hobbs and Roberts (2018) and Roberts (2018) in their study of China offer a powerful explanation even in a heavily politicized context of war, where the forbidden media itself is portrayed as a weapon.

Whereas Hobbs and Roberts (2018) discuss the incentive to acquire the skills to access forbidden websites via VPN, the Ukrainian case shows that the mechanisms behind the accessibility view may hold even when the majority of the politically interested users find a legal way to circumvent censorship free of government persecution. The practically oriented and pragmatic user has an incentive to log on to VKontakte occasionally, either to respond to unanswered messages or to check for new ones. However, free VPN services often add 'friction' by slowing down the connection or providing secure browsing only for a limited traffic volume. For example, at least 3 out of the 5 VPN services that the VKontakte team has publicly recommended for circumventing the Ukrainian ban (Vkontakte, 2017) require individuals to pay a fee for a fast connection and/or unlimited browsing. One of the free options requires individuals to use the Opera browser. This introduces additional obstacles for those who are used to Google Chrome, Firefox, or other widespread browsers. Even in this case, however, some bloggers have complained that their connection began to lag more when using the Opera VPN – right after users from Ukraine began to switch to the service (help-wifi.com, ND).

The cost of logging in increases dramatically, if the user chooses to move away from the free options in order to acquire a faster and more efficient service. A paid VPN service may cost an additional 3 to 7 US dollars per month – a sizable amount in Ukraine. However, VPN clients may potentially complicate the login procedure even if they succeed to provide a fast connection. A pro-Ukrainian resident in Kyiv explained to me in an interview why he no longer uses VKontakte on a daily basis: The VPN on his PC required more clicks and slowed down the connection by a few seconds. Interestingly, the same person could enter VKontakte automatically on his smartphone without extra clicks, likely due to a pre-installed VPN. Despite this, he still decided to stop using VKontakte as his main platform, because he felt it was too troublesome to manage his account only from a smartphone and not being able to do so unhindered through a PC. This evidence is anecdotal and calls for more (qualitative and quantitative) research on the user-experience of banned platforms. Nevertheless, the account emphasizes the potential complexities of managing "forbidden" accounts across multiple devices and operating systems.

Because of the 'friction' created by the censorship policy, the impatient user has an incentive to switch to a cheaper alternative that is unaffected by the ban: Facebook. The American website has many of the same functions as its Russian competitor but can be accessed without VPN. While the data in this paper does not reveal the extent to which the users in the sample have migrated to other social media platforms, aggregated data from online traffic monitors suggests that this may have been the case. According to data from SimilarWeb (an analytics firm), used in NATO StratCom's report, VKontakte's ranking fell from the 1st most visited website in Ukraine on 18th of March 2017 to the 5th most visited site on 14th of August 2018 (Dek, 2019, 42). Facebook, on the other hand, jumped from the 8th to 4th place in the same period. This can be viewed as a great success, seen from the perspective of the Ukrainian government. As mentioned earlier, the purpose behind the ban is not necessarily to reduce all social media behavior as a whole, but to push users towards other media alternatives – away from a social media platform, which provides Russian authorities with surveillance opportunities while showing no commitment to curb pro-Kremlin disinformation campaigns.

Because of these factors, even users with access to the "freedom technology" needed to circumvent online bans may still be highly vulnerable to government censorship. This is important, because it emphasizes that the effects of censorship cannot be evaluated by examining the technical ability of the population to bypass censorship alone. The findings also serve as a reminder to political scientists that practical and relatively pragmatic considerations may play a more important role in terms of how the users respond to online censorship than their political affiliations toward the censoring state or to its foreign “enemy”—even if the opposing political groups are involved in war against each other.

How generalizable are these findings? As mentioned previously, the results are in line with the ’accessibility view’ on information restrictions, which has been used to describe online censorship in China. The findings in this article suggest the pattern can be generalized to a more democratic and politicized context, such as the war in Ukraine. Following the logic of this accessibility view, one could expect a similar pattern in other countries if at least three conditions are met.

Firstly, the ban must be technically implemented in a way that significantly complicates and slows down access to the forbidden source. This is not always the case. For example, the Russian authorities arguably failed to do so in their ban on the Telegram app in 2018 (see Appendix P online). Because Telegram moved its service to Google and Amazon domains (i.e. through "domain fronting"), users could continue accessing the website with relative ease during prolonged periods. The service was officially unblocked in 2020.

Secondly, users must have unhindered access to similar services that can partly replace the banned product. In the case of Ukraine, VKontakte can be replaced with highly similar alternative: Facebook. Conversely, if Russian authorities were to ban Facebook, as they have previously threatened to do (Doffman, 2019), the service could be replaced with VKontakte. In the example above, Telegram was likely difficult to replace in Russia with a more "cooperative" platform, because its core selling point at the time was precisely its unwillingness to share user data with the authorities. If they fail to do so, the impatient users, some of whom are addicted to the product, would have had a much higher incentive to circumvent the ban (Hobbs and Roberts, 2018).

Lastly, I theorize that similar bans may be more effective, if the forbidden platform is already widely known. As pointed by the literature on the Streisand Effect, a ban may backfire by making the general public more curious and aware of the forbidden content. This was less likely during the Ukrainian ban in 2017, precisely because VKontakte, unlike Telegram in Russia during 2018, was already a mundane household name and a part of everyday life for the majority of online users. While these theoretical perspectives may offer a guideline for future enquiries, more cross-national research is needed to empirically examine when and how censorship succeeds in one context and not the other.

Limitations

For ethical reasons, I measures online activity by using only publicly available posts from public accounts, and not private posts or messages.

This paper does not argue that politics do not matter at all in the citizens' response to censorship or that politically engaged users are less affected by the ban than those who are disinterested in politics. Instead, the study is focused on the difference between users with strong and weak (political and social) affiliations with Russia. This question is pivotal to understanding the extent to which the ban has backfired by affecting one affiliation group more than the other.

The findings suggest that users with either pro-Russian or pro-Ukrainian affiliations are on average less affected by the ban than random users from the same region. One possible explanation is that users with political affiliations are on average more active online (see Appendix A online), and therefore more invested in the platform. More research is needed to validate this view. Although Ukraine's measure against the "unpatriotic" Russian platform may have made political users on both sides more prevalent in relative terms, it affected the pro-Russian users at least as much as the pro-Ukrainian ones. It is important to note that the pro-Russian users in the data set publicly follow pro-Russian communities in a time of war and are therefore likely to be relatively extreme in their political attitudes. More research is needed to test whether the findings hold among individuals with only slightly pro-Russian views or individuals who do not have any visible affiliation with Russia in the online realm. Similarly, this research does not exclude that the effects of the ban may vary depending on whether the users access VKontakte through a static connection or a mobile connection on a smartphone.

Furthermore, the study does not shed light on whether the ban has affected the consumption of pro-Kremlin content among Ukrainian users offline or whether it changed the popularity of different political topics (see Dek (2019)). In line with this, the findings do not reveal the extent to which the ban has affected political attitudes among ordinary citizens.

Instead, the results show that the Ukrainian authorities succeeded in achieving an important goal in their anti-propaganda policy: to reduce online activity on Russian social media. From the perspective of the Ukrainian government, this is an important achievement, because it sees Russian social media as a platform for foreign propaganda, surveillance, disinformation, and a national security threat.

The findings in this study are limited to instances where users respond to sudden censorship on social media. Although political and social affiliations have not played a major role in how the users have responded to the ban, it is still possible that these factors influence the consumption of Russian services in the long run, gradually and independent of sudden shocks to the media ecology.

Conclusion

This paper shows that online censorship may work even when users have the technical and legal means to circumvent it. The Ukrainian government succeeded in reducing the overall online activity on VKontakte, which it perceives as a platform for foreign propaganda and a tool for surveillance used against Ukraine in times of war. Even a minor increase in the costs of entering on the banned website is enough to significantly shrink online activity, as long as there are more accessible alternatives.

The data supports the accessibility view on censorship, which argues that the users respond to censorship primarily based on the costs of going online (in terms of time and effort), not their political attitudes toward the regime. The anti-Russian ban, meant as a response to Ukraine's war with Russian separatists, affected pro-Russian users as much as pro-Ukrainian ones in terms of online activity. Similarly, the ban is at least as likely to reduce the activity among users with high levels of potential exposure to pro-Russian content, measured as social ties to Russia, as those with low levels. In other words, the ban did not make VKontakte more pro-Russian in relative terms.

The results are favorable from the perspective of the censor, who wishes to combat foreign propaganda and disinformation by using one of the most drastic countermeasures available. However, this article paints a more concerning picture for those who see censorship as a threat to democracy. This study contributes to the growing evidence suggesting that states can greatly regulate online consumption by setting up minor obstacles between the impatient user and the banned service. The Ukrainian case indicates that this may be true even in a relatively democratic context and among users who already have the technical means to bypass censorship and can do so free of persecution.

Based on these findings, one would expect censorship in other countries to lead to a similar outcome. A sudden, large-scale ban would not make the opposition more predominant on the forbidden platform compared to the pro-regime movement. Similarly, if Russia were to use its newly upgraded censorship infrastructure to ban Facebook to prevent foreign influence, one would expect the ban to be successful from the point of view of the government if Russians were to respond in a similar manner as Ukrainians have. Understanding the mechanisms behind such interventions is important if one is to shed light on how states police cyberspace. This calls for more research on the individual-level factors that may mitigate the effects of censorship, together with cross-national comparative studies of how and when large-scale social media shutdowns are met with political resistance among citizens rather than user compliance.

Acknowledgement

I would like to thank the JOP editors and anonymous reviewers for their detailed comments and invaluable work. I am very grateful to Rebecca Adler-Nissen, David Dreyer Lassen, Sune Lehmann, Margaret E. Roberts, Sarah Oates, Anders Wivel, Mikhail Suslov, Asmus Leth Olsen, Valentina Shapovalova, Frederik Hjorth, Benjamin Carl Krag Egerod, David Szakonyi, Joshua A. Tucker, James Pamment, Alicia Fjällhed, Lene Hansen, Kasper Møller Hansen, Barbara Maria Piotrowska, Arjen van Dalen, Marie Heldgaard Seestedt, Jeppe Vierø, Frederik Kristoffer Kjøller Larsen, Gregory Eady and Simon Polichinel von der Maase for their thorough and helpful feedback. I would also like to thank everyone who discussed this paper during events at APSA, DPSA, SMaPP, SODAS and the PhD seminar at the University of Copenhagen.

References

AFP (2014). Pavel Durov left Russia after being pushed out. *The Economic Times*. Accessed October 26, 2020. shorturl.at/fqF59.

Andrusieczko, P. (2017). Ukraine blocks access to Russian Internet. *Outriders*. Accessed October 26, 2020. https://bit.ly/2QGYW5j.

Arel, D. (2018). How ukraine has become more Ukrainian. *Post-Soviet Affairs 34*(2-3).

Asmolov, G. (2019). The effects of participatory propaganda: From socialization to internalization of conflicts. *Journal of Design and Science*.

BBC (2017). Twitter bans RT and Sputnik ads amid election interference fears. Accessed October 26, 2020. https://bbc.in/33DJ1dc.

Boichak, O. and S. Jackson (2019). From national identity to state legitimacy: Mobilizing digitally networked publics in eastern ukraine. *Media, War & Conflict*.

Bunn, M. (2015). Reimagining repression: New censorship theory and after. *History and Theory 54*(1), 25–44.

Card, D. and A. B. Krueger (1993). Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania. *National Bureau of Economic Research*.

Chen, Y. and D. Y. Yang (2017). 1984 or the Brave New World? Evidence from a Field Experiment on Media Censorship in China.

Dek, Anton, K. K. T. M. (2019). The Effects of Banning the Social Network VK in Ukraine. *NATO StratCom.* Accessed October 26, 2020. https://www.ceeol.com/search/viewpdf?id=791422.

Dickson, B. (2016). *The dictator's dilemma: the Chinese Communist Party's strategy for survival.* Oxford University Press.

Doffman, Z. (2019). Facebook, Instagram And YouTube Will Now Be Banned, Russians Warned. *Forbes.* Accessed October 26, 2020. https://www.forbes.com/sites/zakdoffman/2019/09/13/putin-now-plans-100-facebook-instagram-and-youtube-bans-russians-warned/75f41b7157ff.

DW (2017). Kak na Ukraine sudjat za posti v sotssetjah. *Deutsche Welle.* Accessed October 26, 2020. https://bit.ly/34wo2v0.

Edwards, L. (2009). Pornography, censorship and the internet. *Law and the Internet. Hart Publishing.*

Esberg, J. (2020). Censorship as reward: Evidence from pop culture censorship in chile. *The American political science review 114*(3), 821–836.

Fiorentino, M.-R. (2018). France passes controversial 'fake news' law. *Euronews.* Accessed October 26, 2020. https://bit.ly/3bcPZsm.

Gläßel, C. and K. Paula (2020). Sometimes less is more: Censorship, news falsification, and disapproval in 1989 east germany. *American Journal of Political Science 64*(3), 682–698.

help-wifi.com (N.D.). Pochemu internet s vklychennim vpn "tupit" i "tormozit"? Accessed October 26, 2020. https://help-wifi.com/reshenie-problem-i-oshibok/pochemu-internet-s-vklyuchennym-vpn-tupit-i-tormozit/.

Hjorth, F. and R. Adler-Nissen (2019). Ideological asymmetry in the reach of pro-russian digital disinformation to united states audiences. *Journal of Communication 69*(2).

Hobbs, W. R. and M. E. Roberts (2018). How sudden censorship can increase access to information. *American Political Science Review 112*(3), 621–636.

Howard, P. N., S. D. Agarwal, and M. M. Hussain (2011). When do states disconnect their digital networks? regime responses to the political uses of social media. *The Communication Review 14*(3), 216–232.

Huang, H. (2015). Propaganda as signaling. *Comparative Politics 47*(4), 419–444.

Interfaks (2018). Zhitelya Peterburga sobralis napravit na prinuditelnoye lechenije za anekdot "Vkontakte". *Interfaks*. Accessed October 26, 2020. https://www.interfax.ru/russia/625841.

Jansen, S. C. and B. Martin (2015). The Streisand Effect and Censorship Backfire. *International Journal of Communication 9*.

Kalathil, S. and T. C. Boas (2010). *Open networks, closed regimes: The impact of the Internet on authoritarian rule*. Carnegie Endowment.

Kern, H. L. and J. Hainmueller (2009). Opium for the masses: How foreign media can stabilize authoritarian regimes. *Political Analysis 17*(4), 377–399.

Kiriya, I. and E. Sherstoboeva (2015). Piracy & social change| russian media piracy in the context of censoring practices. *International Journal of Communication 9*, 13.

Kiselyova, M. and A. Prentice. Russian social media site tells Ukrainians how to dodge web block. *Reuters*. Accessed March 11, 2020. https://reut.rs/2UsKW04.

Lanoszka, A. (2016). Russian hybrid warfare and extended deterrence in eastern Europe. *International Affairs 92*(1), 175–195.

Laruelle, M. (2014). Russian Nationalism and Ukraine. *Current History 113*(765), 272–277.

Laruelle, M. (2016). The three colors of Novorossiya, or the Russian nationalist mythmaking of the Ukrainian crisis. *Post-Soviet Affairs 32*(1), 55–74.

Lorentzen, P. (2014). China's strategic censorship. *American Journal of Political Science 58*(2), 402–414.

Morozov, E. (2012). *The net delusion: The dark side of Internet freedom*. PublicAffairs.

Nyhan, B. and J. Reifler (2015). Displacing misinformation about events: An experimental test of causal corrections. *Journal of experimental political science 2*(1), 81–93.

Pan, J. (2017). How market dynamics of domestic and foreign social media firms shape strategies of internet censorship. *Problems of Post-Communism 64*(3-4), 167–188.

Peisakhin, L. and A. Rozenas (2018). Electoral effects of biased media: Russian television in ukraine. *American journal of political science 62*(3), 535–550.

Popkova, A. (2019). From Bootlegging Hollywood to Streaming Battle Rap. *World Entertainment Media: Global, Regional and Local Perspectives*.

Reisinger, H. and A. Gol'c (2014). Hybrid War in Ukraine Russia's Intervention and the Lessons for the NATO. *Osteuropa 64*(9-10).

RFE/RL (2017). Poroshenko Restricts Access To Russian Websites, Social Networks. *RadioFreeEurope/ RadioLiberty*.

Roberts, M. E. (2018). *Censored: Distraction and Diversion Inside Chinas Great Firewall.* Princeton University Press.

Robinson, O. (2018). The memes that might get you jailed in Russia. *BBC*. Accessed October 26, 2020. https://bbc.in/3boWdFp.

Roth, A. (2017). In new sanctions list, Ukraine targets Russian social-media sites. *The Washington Post*. Accessed October 26, 2020. https://wapo.st/2wkQ6mZ.

Sanovich, S. (2017). Computational Propaganda in Russia: The Origins of Digital Misinformation. *Working Paper*.

Shadmehr, M. and D. Bernhardt (2015). State censorship. *American Economic Journal: Microeconomics 7*(2), 280–307.

Shklovski, I. and V. Wulf (2018). The use of private mobile phones at war: Accounts from the donbas conflict. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–13.

Slutsky, P. and D. Gavra (2017). The Phenomenon of Trump's Popularity in Russia: Media Analysis Perspective. *American Behavioral Scientist 61*(3), 334–344.

Stern, R. E. and J. Hassid (2012). Amplifying silence: uncertainty and control parables in contemporary china. *Comparative Political Studies 45*(10), 1230–1254.

Stockmann, D. (2013). *Media commercialization and authoritarian rule in China.* Cambridge University Press.

Stukal, D., S. Sanovich, R. Bonneau, and J. A. Tucker (2017). Detecting Bots on Russian Political Twitter. *Big data 5*(4), 310–324.

Toal, G. and J. O'Loughlin (2017). 'Why Did MH17 Crash?': Blame Attribution, Television News and Public Opinion in Southeastern Ukraine, Crimea and the De Facto States of Abkhazia, South Ossetia and Transnistria. *Geopolitics*, 1–35.

Tucker, J. A., A. Guess, P. Barberá, C. Vaccari, A. Siegel, S. Sanovich, D. Stukal, and B. Nyhan (2018). Social media, political polarization, and political disinformation: A review of the scientific literature.

UNIAN (2017). "kievstar" uzhe blokiruet vkontakte i odnoklassniki (foto). *UNIAN*. Accessed October 26, 2020. https://bit.ly/33GPGDA.

van der Vet, F. (2019). Imprisoned for a 'like'. *Freedom of Expression in Russia's New Mediasphere*.

Varol, O., E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini (2017). Online human-bot interactions: Detection, estimation, and characterization. In *Eleventh international AAAI conference on web and social media.*

Vendil Pallin, C. (2017). Internet control through ownership: the case of russia. *Post-Soviet Affairs 33*(1), 16–33.

Vkontakte (2017). Kak poluchit dostup k vkontakte esli vash provayder ego zablokiroval? *Medium*. Accessed October 26, 2020. shorturl.at/gwNQZ.

Volchek, D. and D. Sindelar (2015). One professional russian troll tells all. *Radio Free Europe/Radio Liberty 25.*

Yelowitz, A. S. (1995). The Medicaid Notch, Labor Supply, and Welfare Participation: Evidence from Eligibility Expansions*. *The Quarterly Journal of Economics 110*(4).

Zhang, A. F., D. Livneh, C. Budak, L. P. Robert Jr, and D. M. Romero (2017). Shocking the crowd: The effect of censorship shocks on chinese wikipedia. *arXiv preprint arXiv:1704.00412*.

Zuckerman, E. (2014). Cute cats to the rescue? *Participatory media and political expression*, 131–154.

BIOGRAPHICAL STATEMENTS

YEVGENIY GOLOVCHENKO is a POSTDOC at the Department of Political Science, University of Copenhagen.