# Social Media and Xenophobia: Theory and Evidence from Russia[*]

**Leonardo Bursztyn**[†] **Georgy Egorov**[‡]
**Ruben Enikolopov**[§] **Maria Petrova**[¶]

January 2024

## Abstract

We study the effect of social media on xenophobic attitudes in Russia. We build a model where social media increases the likelihood of meeting like-minded people locally and in other cities, and where online interactions can be persuasive. We show that social media increases the share of individuals *holding* extreme opinions, but it may also increase the share of people *hiding* these opinions, which calls for proper measurement of attitudes. Empirically, we confirm these predictions by combining data from a survey experiment with data on hate crimes, and exploiting quasi-exogenous variation in city-level social media penetration in Russia. We find that higher city-level social media exposure: i) increased the share of individuals holding xenophobic attitudes (consistent with a persuasion mechanism); ii) reduced people's willingness to openly express xenophobia (consistent with a social image mechanism); iii) led to more hate crimes in cities with higher pre-existing levels of nationalism (consistent with a mechanism of connecting like-minded people locally, thus facilitating the coordination of collective action).

**Keywords:** social media, xenophobia, hate crime, Russia, persuasion, stigma
**JEL Classification: D7, H0, J15**

[*] We are grateful to Garance Genicot, Matthew Gentzkow, Mikhail Golosov, Brian Knight, Andrew Little, Andreas Madestam, Natalya Naumenko, Nancy Qian, Aakaash Rao, Alvaro Sandroni, Daniel Spiro, David Strömberg, Alireza Tahbaz-Salehi, and Ekaterina Zhuravskaya for very helpful discussions, to Danil Fedchenko and Andrei Kim for excellent research assistance, and to numerous seminar participants at AEA, BEROC, BI Norwegian, Chicago Harris, DCPEC Political Economy Seminar, EEA, HSE, Northwestern, Pompeu Fabra, Stockholm, UC-Riverside, Uppsala, NBER Political Economy meeting, and Social Media and Political Participation NYU conference for comments and suggestions. Ruben Enikolopov and Maria Petrova acknowledge financial support from the BBVA Foundation grant.

[†] University of Chicago and NBER, bursztyn@uchicago.edu

[‡] Northwestern University–Kellogg School of Management and NBER, g-egorov@kellogg.northwestern.edu

[§] ICREA-Universitat Pompeu Fabra, Barcelona IPEG, New Economic School, and BSE, ruben.enikolopov@upf.edu

[¶] ICREA-Universitat Pompeu Fabra, Barcelona IPEG, New Economic School, and BSE, maria.petrova@upf.edu

# 1 Introduction

Social media helps people find like-minded individuals: they become part of *communities.*[1] Indeed, already in 2010, Google was indexing 620 million groups on Facebook.[2] In 2019, there were more than 400 million people in groups that they found "meaningful."[3] Individuals find communities based on their interests, no matter how fringe and unusual they are. Social media communities can be a positive force: they might facilitate market transactions and the coordination of leisure activities, for example. However, the power of social media to help find like-minded people can also have negative consequences. As early as 2001, Sunstein (2001) argued that these like-minded discussions online could become a "breeding ground for extremism" (p. 71). Consider individuals holding an extreme opinion, such as hate toward foreigners. For such individuals, it might be particularly difficult to find like-minded people in the real world – indeed, most xenophobes would probably not openly advertise their positions – so social media would be particularly helpful in connecting people sharing such views. Much like social capital (Satyanath et al., 2017), social media communities could therefore have a "dark side."

In this paper, we first build a model in which social media helps individuals find other people with similar opinions locally and in other cities and where online exposure to individuals with a certain opinion can be persuasive. We show that exposure of individuals in a city to social media increases the share of individuals with extreme positions, such as xenophobia. This increased prevalence of extremists can itself increase hate crimes, but this is especially true in cities with a high pre-existing level of xenophobia (because social media's power to connect like-minded people falls on fertile ground) and for crimes with multiple perpetrators (because coordination is particu-

[1] The broad social sciences literature has long suggested that the Internet can make it easier to meet like-minded people (Van Alstyne and Brynjolfsson (2005), Putnam (2000), Sunstein (2001, 2017)), and that social media reinforces this effect (see Barberá (2020) for a recent overview).

[2] See Google Now Indexes 620 Million Facebook Groups.

[3] See Mark Zuckerberg shifted Facebook's focus to groups after the 2016 election, and it's changed how people use the site.

larly important for such crimes, albeit online social groups could provide valuable information or nudging even for crimes committed by single individuals). At the same time, perhaps counterintuitively, the increase in the share of extremists does not necessarily lead to a higher share of people openly agreeing with these extreme opinions; to the contrary, we show that the share of people who *hide* these extreme opinions may *increase*, and it does so for simple functional forms. This is because social media increases the share of extreme views on both sides of the spectrum. Thus, the importance of social image and stigma from expressing fringe opinions is increased. We then apply this framework to examine the effect of social media on xenophobia in Russia and test these theoretical predictions.

The main challenge in identifying a causal effect of social media is that access and consumption of social media are not randomly assigned. We follow the approach from Enikolopov et al. (2020) to overcome this challenge. This approach exploits the history of the main Russian social media platform, *VKontakte* (VK). This online social network, which is analogous to *Facebook* in functionality and design, was the first mover in the Russian market and secured its dominant position with a user share of over 90 percent by 2011. VK was launched in October 2006 by Pavel Durov, who was an undergraduate student at Saint Petersburg State University (SPbSU) at the time. Initially, users could only join the platform by invitation through a student forum of the University, which had also been created by Durov. The vast majority of early users of VK were, therefore, Durov's fellow students of SPbSU. This, in turn, made friends and relatives of these students more likely to open an account early on. Since SPbSU attracted students from across the country, this sped up the propagation of VK in the cities these students had come from. As a result, the idiosyncratic variation in the distribution of the home cities of Durov's classmates had a long-lasting effect on VK penetration. This allows us to use fluctuations in the distribution of SPbSU students across cities as an instrument for the city-level penetration of VK.[4] Our approach enables

[4] To deal with the possibility that cities with a taste for social media were also more likely to send students to SPbSU, we control for the distribution of SPbSU students in cohorts several years older and several years younger than the VK founder.

us to have plausibly exogenous variation at the *city level*. This allows us to consider the effect of social media in finding like-minded individuals both locally and in other cities (individual-level randomization across cities would not allow for that). We thus evaluate the effect of higher VK penetration on both attitudes and hate crimes towards other ethnicities using existing survey data, a newly collected dataset on hate attitudes from a survey experiment we designed and implemented, and data on hate crimes collected between 2007 and 2015 by an independent Russian NGO, *SOVA*.

To test our theoretical predictions, measuring true – as opposed to self-reported – attitudes of individuals toward people of other ethnicities is crucial. In many contexts, self-reported opinions would be a good proxy to truly-held attitudes. In our context, however, the model shows that these do not even need to co-move as a result of the introduction of social media and the ensuing polarization. Thus, to elicit truer attitudes, we designed and conducted an online survey experiment in the summer of 2018, with over 4,000 respondents from 124 cities in Russia.[5] The survey was framed as a study of patterns of usage of social media and the Internet, to which we added our question of interest on ethnic hostility. Given the possibility that a stigma associated with directly reporting xenophobic views in a survey can prevent our respondents from truthfully reporting their attitudes, we use the list experiment technique.[6] Specifically, we were interested in whether respondents agree with a statement (which we borrowed from existing surveys): "*I feel annoyance or dislike toward some ethnicities*."

Using the list experiment, we find a positive effect of social media penetration on elicited ethnic hostility, i.e., the share of respondents that hold xenophobic attitudes. The magnitude of the effect is particularly large in certain subsamples, specifically younger respondents and those with lower levels of education. Numerically, a 10% increase in VK penetration increases the

[5] This survey was pre-registered on the AEA RCT Registry website under entry AEARCTR-0003066.

[6] This is one of the main methods to elicit truthful answers to sensitive survey questions (Blair and Imai, 2012, Glynn, 2013), and it has been shown to perform particularly well in online surveys (Coutts and Jann, 2011). The intuition behind this technique is that the respondents are asked only to indicate the number of statements from a list with which they agree. By adding the statement of interest to a random subgroup of respondents, one can estimate the share of respondents agreeing with this statement without being able to identify who exactly agrees with it. We discuss the procedure in more detail in subsection 3.4.

share of respondents agreeing with the hateful statement in the list experiment by 9.5%, with this magnitude going up to 21.3% for younger respondents and to 17.3% for those with low education (these differences are not statistically significant at conventional levels).

To understand the expression of xenophobia and its associated stigma, we use self-reported hostility. We measure stigma as the difference between the level of elicited ethnic hostility from the list experiment and the share of respondents who answered positively to a direct question about whether they agreed with the statement about xenophobia.[7] This allows us to examine the effect of social media on the expression of ethnic hostility without the cover provided by the list experiment, thus using both the direct question and the list experiment to create a measure of stigma of expressing xenophobic opinions. Consistent with our theoretical model, we find that stigma increased in places with higher social media penetration. We obtain similar results if we use the answers to the same direct question from a much larger, nationally representative face-to-face survey of more than 30,000 respondents conducted in 2011 by one of the biggest Russian survey companies, FOM (*Fond Obschestvennogo Mneniya*, Public Opinion Foundation). Interestingly, in the data from both our survey and this larger survey, we do not see a significant effect of social media penetration on self-reported hostility, consistent with the ambiguous prediction of the theoretical model and further validating the need to use elicitation techniques such as list experiment to understand the impact of social media on truly-held opinions.

Lastly, we use the same instrumental variables approach to show that higher penetration of social media had real-world consequences. Specifically, it led to more ethnic hate crimes, though only in cities with a higher baseline level of nationalist sentiment prior to the introduction of social media. To proxy for baseline local nationalist sentiment, we use the city-level vote share of *Rodina* ("Motherland"), an explicitly nationalist and xenophobic party, in the 2003 parliamentary election, the last one before the creation of VK. Our results suggest that a 10% increase in VK penetration

[7] The direct question was asked to the subjects in the control group (i.e., those randomly assigned to the list not containing the statement about xenophobia) after the list experiment to avoid contamination.

increased hate crimes by 21.7% in cities where Rodina received the most votes but had no effect in cities where Rodina got minimal support. The stronger result on crimes with multiple perpetrators suggests that in addition to polarization, social media may have facilitated the coordination of hate crimes among people willing to commit them.

Our paper contributes to several lines of research. First, our paper builds on the literature on social learning (see Mobius and Rosenblat (2014) and Golub and Sadler (2016) for overviews). The main conceptual difference is that in our model, agents interact to form, in a DeGroot (1974) fashion, their political beliefs rather than learn about a common unknown state of the world.[8] While the social learning literature largely focuses on the convergence of beliefs, one notable exception is Dasaratha et al. (2023), where the state of the world is constantly changing, and beliefs converge to a non-atomistic distribution. In this paper, we make a similar assumption of preference shocks to get a nontrivial steady-state distribution of political preferences in perhaps the simplest possible way.[9]

Second, we contribute to a growing empirical literature on the impact of social media on political attitudes and polarization. Allcott et al. (2020), Mosquera et al. (2020), and Enikolopov et al. (2023b) provide evidence that social media contributes to the increasing polarization of individuals' political opinions.. Gentzkow and Shapiro (2011) finds that online interactions are less segregated than offline interactions with friends, colleagues, family members, or neighbors. In contrast, Halberstam and Knight (2016) shows that the segregation of communications on social media (Twitter) is more pronounced and closer to the segregation in offline interactions. The importance of peer interactions in the formation of political beliefs is also documented in Madestam et al. (2013) and Satyanath et al. (2017).[10]

[8] See also Bisin and Verdier (2000) and Bisin and Verdier (2001) on intergenerational transmission of culture through interaction with parents and other senior members of the society.

[9] In Acemoglu et al. (2013), there is no convergence of opinions on a network due to the presence of "stubborn" agents who do not change their opinions but exercise influence in every period. Unlike shocks, the assumption of stubborn agents would not allow us to study increasing polarization.

[10] Other related papers documenting the effect of outside sources on formation of political opinions include DellaVigna and Kaplan (2007) on the effect of Fox News, Enikolopov et al. (2011) on independent channel NTV in Russia,

Third, we contribute to the literature on social image concerns and social stigma. Earlier work where individuals are judged for their type includes Morris (2001), Bénabou and Tirole (2006), Ali and Bénabou (2020), Ali and Lin (2013), Austen-Smith and Fryer (2005), Bursztyn et al. (2019), Bursztyn et al. (2020). In all these papers, the agent's type is binary (in Bursztyn et al. (2019), it has two binary dimensions). Our paper makes a methodological contribution by introducing a tractable way of modeling social stigma with a continuum of types (which is needed to study political polarization in a meaningful way). The importance of social image concerns has been documented empirically in different contexts; see e.g., DellaVigna et al. (2012) on charitable giving, DellaVigna et al. (2017) on voting decisions, Perez-Truglia and Cruces (2017) on campaign contributions, Bursztyn and Jensen (2015) on schooling choices, Bursztyn et al. (2018) on status goods, and Enikolopov et al. (2018a) on political protests. In Bursztyn et al. (2020), the popularity of a certain opinion (xenophobia) is shown to increase the likelihood that it is expressed, and the people who do so are judged less negatively. In contrast, this paper suggests that the effect of polarization is subtler and that even if radical opinions become more popular, they may, at the same time, become more stigmatized.

Fourth, our work is related to the literature on legacy and social media promoting hate crimes and genocide. In this line of research, most papers document an immediate effect of posts on social media or propaganda on hateful actions. For example, Müller and Schwarz (2021) shows that anti-refugee sentiment on Facebook predicts day-to-day changes in crimes against refugees in Germany. This, however, does not speak to longer-lasting changes in the patterns of hate crime with the arrival of social media and could instead reflect displacements of hate crime towards days with more xenophobic content. Similarly, Müller and Schwarz (2023) find that anti-Muslim hate crimes in the United States have increased in counties with high Twitter user penetration, but only since the start of Donald Trump's presidential campaign; again, these findings are consistent

Ou and Xiong (2021) on propaganda during the Cultural Revolution in China, Cantoni et al. (2017) on the effect of school curriculum, and Adena et al. (2015) on radio in Germany before and after the Nazis' ascent to power.

with Trump's posts nudging users to commit hate crimes rather than fundamentally changing their views of the world. Jiménez Durán et al. (2023) documents that a German law that mandates major social media companies to remove hateful posts lowers the toxicity of social media posts and reduces the number of hate crimes. Other studies that document the effect of political speech or propaganda on violence include Yanagizawa-Drott (2014) on the Rwandan genocide (speeches by key government officials, including Prime Minister Jean Kambanda, were an integral part of RTLM radio propaganda in Rwanda) and DellaVigna et al. (2014) on the war in former Yugoslavia. Unlike our paper, none of these papers document the long-term effects of the proliferation of social media.[11]

Fifth, our paper contributes to a growing literature studying the recent rise of populism and nationalist attitudes. There is evidence that this rise often has deep historical roots; see, e.g., Cantoni et al. (2019b) on factors that contributed to the success of the AfD in Germany and Enke (2020) on communal moral values driving the support for extreme parties. Bursztyn et al. (2020) trace the increase in expression of xenophobia to the 2016 U.S. presidential campaign and election results, highlighting the role of individual politicians and information aggregation in elections. Guriev et al. (2021) demonstrate the role of technology, specifically, that 3G penetration around the globe promoted populist voting and reduced government support. Economic hardship also played a role: Algan et al. (2017) show that the Great Recession triggered a trust crisis and led to higher voting shares of non-mainstream, in particular, populist parties; see also Sartre and Daniele (2022) on a specific case of toxic loans in France. Our paper suggests an important role that social media played in the rise of extreme views, which may explain – along with the Great Recession – the contemporaneous rise of nationalism and extremism on a global scale.

[11] The paper is also related to a more general literature on the political effects of social media. A number of papers provide evidence that social media helps to promote collective action, such as political protests (Acemoglu et al. (2018), Enikolopov et al. (2020), Qin et al. (2021). Social media can also help political mobilization during elections Bond et al. (2012) and reduce corruption by promoting accountability Enikolopov et al. (2018b). It also has an indirect effect by affected reporting strategies of legacy media (Hatte et al., 2020; Sen and Yildirim, 2016)). See Zhuravskaya et al. (2020) for a more detailed overview of this literature.

The rest of this paper proceeds as follows. In Section 2, we present our theory framework and derive predictions. We discuss our identification strategy, data sources, and the survey in Section 3 and present our empirical results in Section 4. In Section 5, we discuss the connection of our framework and our findings to the results in the existing literature. Section 6 concludes.

# 2 Theoretical Framework

We now present a simple model of opinion formation through social interactions and use it to study the effect of penetration of social networks.

## 2.1 Social networks and distribution of preferences

There is a finite number $N$ of cities with a unit continuum of citizens living in each city (the assumption of similar size of cities is adopted for expositional purposes and is not consequential). Time is discrete and infinite and is denoted by $t = 0, 1, 2, \ldots$. Each citizen has a political position over some dimension of interest, such as xenophobia. This political position may be interpreted as ideological or taste-based (e.g., whether the individual likes or hates immigrants) or an opinion about a particular policy (e.g., the number of immigrants to be allowed or the minimal requirements such as education and lack of criminal history that they must satisfy). Importantly, an individual's position can change over time as a result of interactions with other people and due to random shocks. We denote the position of individual $i$ at time $t$ by $x_i^t$. The position at time 0, $x_i^0$, is taken from distribution $F_{n(i)}^0$, where $n(i) \in \{1, \ldots, N\}$ denotes the city where individual $i$ lives; these distribution may be different for different cities, and they are assumed to have finite first (denoted by $\mu_n^0 = \mathbb{E}_{i:n(i)=n} x_i^0$) and second moments, and we assume for simplicity that there is a continuum of individuals at each political position.

The positions $x_i^t$ at time $t \geq 1$ are determined endogenously. Specifically, we assume that in each period $t$, each individual $i$ interacts with other members of the society, and his position evolves as a result of these interactions in the following way. With weight $\omega$, the new position $x_i^t$ incorporates $i$'s prior political position $x_i^{t-1}$; we think of political positions to be relatively stable, and for some comparative statics results, we will assume that $\omega$ is sufficiently close to 1. With complementary weight $1-\omega$, it incorporates the weighted average of prior political positions of other individuals $i$ interacts with, which we denote by $y_i^{t-1}$, with weights described below. Lastly, $i$'s political position is subject to a random shock $\varepsilon_i^t$, which has normal distribution $\mathcal{N}\left(0,\sigma_\varepsilon^2\right)$; these shocks are independent across individuals and time.[12] We thus have the following evolution of opinion of individual $i$:

$$x_i^t = \omega x_i^{t-1} + (1-\omega)\, y_i^{t-1} + \varepsilon_i^t. \tag{1}$$

The people that individual $i$ interacts with are not completely random. Some interactions are offline (in-person), and some are online (using social media); we allow the share of interactions on social media to be city-dependent and denote this share in city $n$ by $\tau_n$. We assume that almost all (i.e., with probability 1) in-person interactions happen within the city where the person lives, and the political positions of these people are drawn randomly from the $F_{n(i)}^{t-1}$; in other words, we effectively assume that the offline social network is uncorrelated with one's own political preferences. In contrast, social media interactions exhibit homophily, a tendency to interact with like-minded people.[13] In fact, it is natural to think of one's contacts in an online social network to include in-person contacts, perhaps friends of those (who are also likely to live in the same city), but also people or even groups of people with the same interests that they could, in principle, find throughout the country. We capture this by assuming person $i$ spends share $h$ of their time on social networks interacting with people with exactly the same political position $x_i^{t-1}$, and these people

[12] The shocks are best thought of as idiosyncratic, but it is easy to amend the model so that these shocks capture the influence of sources that maintain distribution over time. For example, these might come from general human knowledge (say, books that individual $i$ might read in period $t$) or influence by a certain group of individuals (influencers, celebrities, politicians) who have fixed positions that do not evolve over time. Technically, shocks ensure that the positions of individuals do not converge to a point, allowing us to study the distribution of opinions and polarization.

[13] In-person interactions may also show homophily, but it is particularly pronounced in social media. Our results go through as long as social media interactions show stronger homophily, and we assume no homophily in in-person interactions to simplify notation.

are chosen randomly from the entire country.[14] The rest of their online time, share $1-h$, citizen $i$ spends with people drawn randomly from the same city $n(i)$. By the law of large numbers, the weighted political positions of all the people citizen $i$ has interacted with sum up to

$$\begin{aligned} y_i^{t-1} &= \tau_{n(i)}\left(hx_i^{t-1}+(1-h)\,\mu_{n(i)}^{t-1}\right)+\left(1-\tau_{n(i)}\right)\mu_{n(i)}^{t-1} \\ &= \tau_{n(i)}hx_i^{t-1}+\left(1-\tau_{n(i)}h\right)\mu_{n(i)}^{t-1}. \end{aligned} \tag{2}$$

Thus, in the model, higher social media penetration (a higher $\tau_n$) increases the frequency of interaction of those living in city $n$ with like-minded individuals and their weight in the updating process at the expense of smaller exposure to random opinions in the society.

**Lemma 1.** *The distributions of political positions in each city $n \in \{1,\dots,N\}$, $F_n^0, F_n^1, F_n^2, \dots$, converge in distribution to $\mathcal{N}\left(\mu_n, \sigma_n^2\right)$ as $t \to \infty$, where $\mu_n = \mu_n^0 = \int_{-\infty}^{+\infty} x dF_n^0(x)$ is the mean of the initial distribution and $\sigma_n^2$ is given by*

$$\sigma_n^2 = \frac{\sigma_\varepsilon^2}{1-(\omega+(1-\omega)\,\tau_n h)^2}, \tag{3}$$

*which is increasing in $\sigma_\varepsilon^2$, $\omega$, $\tau_n$, and $h$.*

In other words, this model of opinion formation with shocks predicts convergence of the distribution of political positions in every city to a normal one, with the mean given by the mean of the original distribution, whereas the other information about the original distribution is lost over time. The variance of the limit distribution is nontrivial because of persistent shocks to preferences, which prevents full convergence. The more individuals are influenced by people with other opinions, the faster these preference shocks dissipate and the smaller the variance of the limit dis-

[14] For the purposes of modeling opinion formation, it is not important whether the like-minded people an individual meets online are from the same city or not. However, we find it natural to assume that such groups of interests span multiple cities, with a higher representation of individuals from cities where this political opinion is overrepresented. This will be important later when modeling coordination in hate crimes.

tribution. Conversely, interactions with like-minded people (higher $\tau_n h$), as in "echo chambers," slow down the convergence process and result in a limit distribution with a higher variance. Therefore, a higher penetration of social media results in a higher polarization of views in a city while preserving its mean.

## 2.2 Extreme political opinions

We now consider the effect of social network penetration in city $n$, $\tau_n$, on support for extreme opinions. Without loss of generality, we focus on right-wing political positions. Take any cutoff $q$; since in the limit, the distribution of individuals' types is normal and given by Lemma 1, the share of individuals in city $n$ with political opinions $x_i > q$ is given by

$$R_n(q) = \Pr(x_i > q \mid n(i) = n) = \frac{1}{\sqrt{2\pi}\sigma_n} \int_q^{+\infty} \exp\left(-\frac{(x-\mu_n)^2}{2\sigma_n^2}\right) dx.$$

We have the following result.

**Proposition 1.** *Suppose that $q > \mu_n$. Then $R_n(q)$ is higher for higher $\tau_n$. Conversely, if $q < \mu_n$, then $R_n(q)$ lower for higher $\tau_n$.*

This result follows from the single-peakedness of normal distribution. For $q > \mu_n$, the individuals with $x_i > q$ are a minority, and an increase in $\tau_n$ leads to an increase in variance $\sigma_n$, which results in an increase in the popularity of this minority opinion. The effect is the opposite for $q < \mu_n$, where the opinions to the right of $q$ include the median opinion (so effectively a majority); then an increase in social media penetration increases the popularity of the minority (in this case, left-wing opinion) as well. Thus, social media and the resulting polarization enhances support for fringe political opinions.

Naturally, social media can have a stronger effect in the case if people are, in principle, open to other viewpoints. For example, it is straightforward to see from (3) that the effect of $\tau_n$ is stronger if $\omega$ smaller (i.e., $\frac{\partial^2 \sigma_n^2}{\partial \tau_n \partial \omega} < 0$). One can, therefore, expect the effect of $\tau_n$ on $R_n(q)$ is decreasing in $\omega$ as well. While $R_n(q)$ is a non-linear transformation of $q$, we prove the following result.

**Proposition 2.** *For any $q \in (\mu_n, \mu_n + \sigma_n)$ there is $\tilde{\omega} = \tilde{\omega} \in (0,1)$ such that for $\omega \geq \tilde{\omega}$, $\frac{\partial^2 R_n(q)}{\partial \tau_n \partial \omega} < 0$.*

In the statement of this proposition, $q \in (\mu_n, \mu_n + \sigma_n)$ means that individuals with opinions $x_i > q$ are in a minority, but the minority is not extreme, and the cutoff $q$ is within a standard deviation from the median. In the case of a normal distribution, this means that the opinion is held between $F(-1) \approx 16\%$ and $50\%$ of citizens, which is true about the number of people who dislike other nationalities in our setting. For such cases, Proposition 2 states that the positive effect of social media on the number of people supporting a fringe opinion is getting weaker as $\omega$ increases further. This would be the case, for example, for people who are well-read and experienced and for whom their prior opinion carries a lot of weight. In contrast, people with more malleable opinions, say younger or less educated, will experience a stronger effect of social media. We confirm these predictions in our experiment.

## 2.3 Stated support for extreme positions and stigma

To understand the discrepancy between individuals' political positions and their expression, and in our case, the difference between true and reported xenophobia, consider an individual $i$ in city $n$ with position $x_i$ who is asked before an audience (and therefore under social pressure) whether $x_i$ exceeds $q$, where $q$ is some cutoff. Denote the affirmative answer by $d_i = Y$ and the negative answer by $d_i = N$. The individual gets disutility from expressing preferences that are far from his own, or to put it another way, there is a cost of lying. Specifically, if $x_i > q$ and he chooses $d_i = N$, he gets disutility $h(x_i - q)$, where $h(\cdot)$ is an increasing continuous function with $h(0) = 0$; in other words, we assume that egregious lies are more costly than little lies. Similarly, if $x_i < q$ and he chooses $d_i = Y$, he gets disutility $h(q - x_i)$. In both cases, telling the truth does not yield direct utility or disutility.

The individual also cares about social approval. We assume that $i$'s response to the question whether his $x_i$ exceeds $q$ is observed by another random individual in the same city (assuming that it is observed by several or even all individuals leads to a very similar model with similar results). This other individual $j$ will form a posterior belief about the individual $i$'s type. We assume that an individual with political position $x_j$ dislikes individual with position $x_i$ according to a function $g\left(x_j - x_i\right)$; to simplify expressions, we will focus on the case where $g(x) = \gamma x^2$; naturally, we would expect $\gamma$ to be higher for more sensitive topics.[15] Assume that the individual $i$ cares about the observer's expected (dis)approval with intensity $\lambda$. Then, individual $i$ chooses answer $d_i$ to maximize his utility $U_i$ that consists of (negative) direct cost $C_i$ and social cost $S_i$:

$$\begin{aligned} & U_i(d_i) = -C_i(d_i) - S_i(d_i) \\ = \quad & -\mathbf{I}_{\{x_i > q \wedge d_i = N\}} h(x_i - q) - \mathbf{I}_{\{x_i < q \wedge d_i = Y\}} h(q - x_i) \\ & - \int_{-\infty}^{\infty} \mathbb{E}_{-i}\left(\lambda g(x-y) \mid d(x) = d_i\right) \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu_n)^2}{2\sigma_n^2}\right) dy. \end{aligned}$$

The latter term $S_i(d_i)$ captures the expectation of $\lambda g(x_i - y)$ by an observer with position $y$ who knows that individual $i$ chose action $d_i$, and then the expectation is taken over the possible realization of observer's types.

In general, the game admits multiple equilibria because of strategic complementarity of adherence to social norms; however, if individuals are sufficiently averse to lying, the equilibrium is unique. In what follows, we will assume for simplicity that the $h(\cdot)$ is differentiable and such that the equilibrium is unique.[16] Nevertheless, even with multiple equilibria, the comparative statics result would remain true for equilibria with the largest and smallest shares of individuals giving a

[15] Our functional form implies, in particular, that social (dis)approval of individuals with known types is symmetric: individual $i$ likes or dislikes individual $j$ as much as $j$ likes or dislikes $i$. We adopt it for simplicity of exposition; in the real world, it is possible that xenophobic people dislike tolerant ones but not the other way around, or alternatively that tolerant people dislike xenophobes, but xenophobes are indifferent about tolerant people as long as they are not migrants. Genicot (2022) shows that such asymmetry would have important implications for forming social networks.

[16] For equilibrium uniqueness, it is sufficient to require that $h(\cdot)$ is steeper than some linear function for small $x$ and steeper than some quadratic function for large $x$.

particular answer (Milgrom and Shannon, 1994) or for the equilibrium with the highest share of individuals answering truthfully.

**Proposition 3.** *The equilibrium is characterized by a cutoff $z$, such that individuals with $x_i > z$ choose $d_i = Y$ while individuals with $x_i < z$ choose $d_i = N$. Moreover, if $q > \mu_n$, then $z > q$, and if $q < \mu_n$, then $z < q$.*

*Suppose now that $q > \mu_n$. The cutoff $z$ is decreasing in $\mu_n$ and is increasing in $\sigma_n$ and $q$. The equilibrium share of individuals choosing $d_i = Y$ is increasing in $\mu_n$ and decreasing in $q$; the effect of an increase in $\sigma_n$ is ambiguous.*

The first part of Proposition 3 highlights the effect of social stigma: fewer people would admit holding a minority belief than the number of people actually holding it, because some types would cave in to social pressure and misstate their preferences. Since social stigma is the same for any personal belief and the relative direct benefit of answering $Y$ rather than $N$ is increasing in one's type, the equilibrium takes the form of a cutoff. A higher $q$ (a more extreme question) or a lower $\mu_n$ (more tolerant population) makes fewer people willing to agree that their $x_i$ exceeds $q$.

The impact of an increased polarization $\sigma_n$ comes from two effects. On the one hand, it increases the number of people agreeing with a minority opinion. However, a higher polarization effectively increases social image concerns by giving more weight to people with extreme opinions on both sides, and the difference between benefits from adhering to "normal" and "extreme" opinions is therefore also increasing. These opposite effects lead to an ambiguous prediction about the share of people admitting that $x_i > q$. Paradoxically, this means that an increase in support of an extreme opinion driven by growing polarization may lead to a decline in stated support of this extreme opinion. This highlights the importance of differentiating elicited and self-reported opinions for empirical purposes.

Despite this ambiguity, the effect of polarization on social stigma, understood as the difference between the share of people with $x_i > q$ and the share of people admitting it by choosing $d_i = Y$, is positive, at least for some particular cost of lying. The next proposition provides a sufficient condition.

**Proposition 4.** *Suppose that the cost of lying $h(\cdot)$ is linear on $(0, z-q)$. Then, the share of individuals who hold belief $x_i > q$ but do not admit it publicly is increasing in $\sigma_n$.*

The takeaway from this proposition is that social media, by increasing polarization, may have an ambiguous effect on stated support of an extreme position, but the difference between true and stated support is likely to increase. Effectively, social media makes more people willing to hide their extreme opinion, despite the fact that there are more people holding it. Thus, polarization may mean more extreme opinions and a higher stigma of holding these opinions at the same time, and this is exactly what we see in our data.

## 2.4 Political preferences and hate crimes

Consider a simple model of hate crimes. For a person of type $x_i$, committing a crime provides benefit $b = b(x_i)$, where $b(\cdot)$ is a strictly increasing function (naturally, for most people, it would take negative values). The expected cost of committing a crime is due to the probability of getting arrested or hurt, and is denoted $c_1$ if the individual commits a crime alone and $c_2$ if he does so with a partner. It is natural to think that $c_2 < c_1$, as there may be strength in numbers, someone can be a lookout, etc. Not everyone who is willing to commit a crime will necessarily do so every period; we assume that such opportunity arises with probability $\kappa_1$ for crimes with a single perpetrator. For crimes with multiple perpetrators, we assume that every time two individuals willing to commit such a crime meet, offline or online (which happens according to the same process as in the opinion formation model above), they will find an opportunity to commit this crime with probability $\kappa_2$, but only if they live in the same city. Consequently, the amount of single-perpetrator crimes committed in city $n$ is

$$C_1^n = \kappa_1 \int_{b^{-1}(c_1)}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{(x-\mu_n)^2}{2\sigma_n^2}\right) dx,$$

as only individuals with $b(x_i) > c_1$ are willing to commit a crime. The corresponding value for multiple-perpetrator crimes is

$$\begin{aligned} C_2^n &= \kappa_2 \int_{b^{-1}(c_2)}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{(x-\mu_n)^2}{2\sigma_n^2}\right) \times \\ &\left[ (1-\tau_n h) \int_{b^{-1}(c_2)}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{(y-\mu_n)^2}{2\sigma_n^2}\right) dy \right. \\ &\left. + \tau_n h \frac{\frac{1}{\sigma_n}\exp\left(-\frac{(x-\mu_n)^2}{2\sigma_n^2}\right)}{\sum_{k=1}^{N} \frac{1}{\sigma_k}\exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right)} \right] \times dx. \end{aligned}$$

Indeed, for an individual to commit a crime with a partner, it must be either because he meets a random individual (in person or online), and both happen to have positions above $b^{-1}(c_2)$, or because he meets a like-minded person online, and that person is from the same city. We have the following result.

**Proposition 5.** *Suppose that $b^{-1}(\min\{c_1, c_2\}) > \mu_n + 2\sigma_n$ for all n. Then for both types of crimes, the number of that crime in city n is increasing in social media penetration $\tau_n$, in cities with a higher initial level of xenophobia $\mu_n$, and these effects are mutually reinforcing: the effect of social media penetration is stronger if the initial level of xenophobia $\mu_n$ is higher.*

Here, the first condition asserts that both types of crimes are committed by people at least two standard deviations above the median in every city (in the case of a normal distribution, this roughly corresponds to 2% most extreme xenophobes, though this could be relaxed). The proposition states that not only does social media lead to more hate crimes, but that initial xenophobia makes this effect stronger. For single-perpetrator hate crimes, the intuition is that polarization is not likely to have a major effect if there are too few people with radical opinions; in other words, radical ideas proliferate better if there is a more sizeable original seeding. For multiple-perpetrator hate crimes, this intuition applies as well, but there is an additional effect of coordination: social media helps find like-minded people. For a person living in a city with a lot of extremists, more of the people he meets online will be from the same city, and thus, opportunities to commit hate crimes will arise more frequently. In other words, our theory predicts that the positive interaction effect of social media and initial xenophobia is due to the persuasion effect for single-perpetrator crimes and due to both persuasion and coordination effects for crimes with multiple perpetrators.

# 3 Empirical Analysis

The theoretical model above generates several testable predictions. First, we expect the share of people who hold relatively extreme opinions, e.g., xenophobes, to increase following an increase in social media penetration. Second, we expect that the stigma of publicly expressing extreme attitudes, e.g., xenophobic ones, will also go up. Finally, we expect that there is a positive effect of initial xenophobia on hate crimes, which is amplified by social media penetration.

In what follows, we test these predictions using data from Russian cities. Russia is a multiethnic society with more than 180 ethnic groups. Russians make up the largest ethnic group, accounting for over 70% of the population. A sizeable share of the population holds xenophobic views. For example, the openly xenophobic political party *Rodina* got 9.2% of the national vote back in 2003. The share of people openly admitting having xenophobic views was 33% in 2011, according to the nationally representative FOM survey mentioned above.

We designed and implemented a survey experiment to measure both the share of people holding xenophobic attitudes and the stigma of expressing xenophobic attitudes. We elicited the levels of xenophobia in 124 Russian cities using list experiments. We relate this measure of ethnic hostility to social media penetration, using the peculiarities of initial penetration of social media in Russia for identification. We compare the levels of xenophobia elicited from the list experiments to the level of xenophobia that is openly admitted in the survey to measure the stigmatization of xenophobic beliefs and use the same identification approach to relate it to social media penetration.

Finally, we use data on hate crimes to test the predictions of the model about the effect of social media on this form of violence.

## 3.1 Identification strategy

Our empirical strategy for the identification of a causal effect of social media penetration follows the approach in Enikolopov et al. (2020).[17] In particular, we look at the penetration of the most popular social network in Russia, *VKontakte* (VK), which had substantially more users than Facebook throughout the whole period we analyze. For example, in 2011, the midpoint of our hate crime data, VK had 55 million users in Russia, while Facebook had 6 million users. VK was created in the fall of 2006 by Pavel Durov, who was a student at Saint Petersburg State University (SPbSU) at the time. The first users of the network were largely students who studied with Durov at SPbSU. This made their friends and relatives at home more likely to open an account, leading to a faster VK spread in these cities. Network externalities magnified these effects, and, as a result, the distribution of the home cities of Durov's classmates had a long-lasting effect on VK penetration. In particular, the distribution of home cities of the students who studied at SPbSU at the same time as Durov predicts the penetration of VK across cities. This prediction is robust to controlling for the distribution of the home cities of the students who studied at SPbSU several years earlier or later. This effect persists throughout the study period between 2007 and 2016, although the magnitude of the effect decreases over time.

The number of students in Durov's cohort is positively and significantly (at 1% level) related to subsequent VK penetration, while the number of students in older or younger cohorts does not significantly predict VK spread (column (1) of Table 1). We also show that even though VK penetration is correlated with nationalistic party support, future VK penetration does not predict past nationalist party support, either in the reduced form or in the IV specifications in columns (2)-(3) of Table 1.

[17] Using corrected control variables from Enikolopov et al. (2023a).

In the results based on survey data we do not have observations on the dependent variables for all the cities, so we use a two-sample instrumental-variables approach (Angrist and Krueger, 1992, Currie and Yelowitz, 2000,Olivetti and Paserman, 2015), in which we predict social media penetration from the full sample of 625 cities and use the predicted values in the second stage.

The F-statistic in the IV specifications is almost 18, which is higher than traditional Stock and Yogo (2005) thresholds (see the first stage in column (1) of Table 1). However, these thresholds are valid only for homoscedastic errors and cannot be applied to a model with robust or clustered standard errors. If we use a more appropriate methodology developed by Montiel Olea and Pflueger (2013), the effective F-statistic in this specification is 16, which is lower than the threshold of 23 for 10% potential bias and a 5% significance. Since concerns about potential weak instruments cannot be fully dispelled, we follow the recommendation in Andrews et al. (2019) and also report the weak-instrument-robust confidence intervals for each main coefficient of interest. In particular, for the results that use a two-sample instrumental-variables approach, we report weak-instrument-robust confidence sets developed by Choi et al. (2018) for this setting, and for the results that use a standard instrumental-variables approach we report weak-instrument-robust confidence sets developed by Chaudhuri and Zivot (2011) and Andrews (2017), and implemented in Stata by Sun (2018).

## 3.2 City-level data

The data on social media penetration comes from the authors' data collection, similar to the one used in Enikolopov et al. (2020). The sample consists of 625 Russian cities with a population of over 20,000, according to the 2010 Census.[18] To measure social media penetration, we use information on the number of users of the most popular social media service in Russia, VK. In particular, we calculate the number of VK users who report a particular city as their city of residence in 2011, the midpoint for our hate crime data. We summarize the evolution of VK penetration over

[18] The exceptions are Moscow and St. Petersburg, which are excluded from the sample as outliers.

time in Figure B1.

Data on hate crimes comes from the database compiled by the SOVA Center for Information and Analysis.[19] SOVA is a Moscow-based Russian independent nonprofit organization providing information related to hate crimes, which is generally considered to be the most reliable source of information on that issue. The dataset covers incidents of violent hate crime, which include murders, assaults, batteries, and death threats. These data have been collected consistently since 2007, with some incomplete data for 2004-2006. In the analysis, we use data from 2007-2015.

Figure 1 presents information about the number of ethnic hate crimes in the 2007-2015 period across Russian cities in our sample on the map. Table B1 presents more detailed information on the number of victims for each type. Based on the textual description of each incident in the database, we have also manually coded the number of perpetrators for every incident. The average number of recorded hate crimes and hate crime victims has been declining over time (see Figures B2 and B3).

As a measure of nationalist sentiment in a city *before* the creation of the VK social network, we use the vote share of the *Rodina* ("Motherland") party in the parliamentary election of December 2003, the only election this party participated in and the last parliamentary election before the creation of VK. This party ran on an openly nationalist platform. It received 9.2 percent of the vote and got 37 of the 450 seats in the *State Duma*, the lower house of the Russian parliament. We validate that the vote share for the party can serve as a proxy for nationalist sentiment in a city by showing that it is positively and significantly correlated with ethnic hate crime in the subsequent years, as well as with xenophobic attitudes revealed in the pre-existing opinion polls (Table B3).

City-level data on population, age, education, and ethnic composition come from the Russian Censuses of 2002 and 2010. Data on average wages come from the municipal statistics of RosStat, the Russian Statistical Agency. Additional city characteristics, such as latitude, longitude, year of city foundation, and the location of the administrative center, come from the Great Russian

[19] The database can be found at https://www.sova-center.ru/en/database/violence/

Encyclopedia.[20]

## 3.3 Survey data

The data on attitudes towards other ethnicities come from a survey that we conducted in the summer of 2018 in 124 Russian cities from our city-level sample.[21] The survey was administered by a professional marketing firm, *Tiburon Research*, with a representative panel of urban Internet users in Russia. To be able to conduct the list experiment within each city (see the next subsection), we tried to maximize the number of respondents per city, so the survey was not designed to create a representative sample of the cities and was biased towards bigger cities. The resulting median number of respondents per city is 39.[22] The sample consists of 4,447 respondents, of which 2,221 were allocated to the control group and 2,226 to the treatment group in the survey experiment.[23]

We also use data from the MegaFOM opinion poll conducted by one of Russia's leading opinion polling firms, FOM (*Fond Obschestvennogo Mneniya*, Public Opinion Foundation), in February 2011. This is a regionally representative survey of 54,388 respondents in 79 regions of Russia, of which 29,780 respondents come from 519 cities in our city-level sample. In particular, we use information on answers to exactly the same direct question about hostility to different ethnicities

[20] The electronic version of the Encyclopedia can be found at https://bigenc.ru/

[21] The survey questionnaire is available in Appendix C.

[22] On average, the cities in our survey sample were larger than the average city in the hate crime sample. It is possible that for the smaller cities, the effect of social media could be more modest, yet more than half of the Russian population lives in cities with populations above 100,000, and these are also the cities with the highest social media penetration. Thus, we believe our estimates remain relevant from both academic and policy standpoints.

[23] More specifically, we collected the data in two batches, the pilot and the main experiment. As part of the pilot, we surveyed 1,007 individuals from 20 cities. Individuals from this batch were randomized into three groups, with one containing a statement about ethnic minorities as part of the list experiment, another containing a statement about LGBTQ individuals, as well as a control group. As we found no reliable data on hate crimes against LGBTQ individuals, we dropped the second group of 336, leaving us with 671 individuals from the pilot. We surveyed 4,034 individuals from 111 cities as part of the main experiment. In this batch, the cities were randomly chosen by the firm we were working with, and since we had the data on VK penetration for only 105 of these cities, we had to drop 246 observations from six cities. An additional 12 surveys were incomplete, which left us with 3,776 observations from the main part. In most analyses, we pool the two batches together, but our results are robust to only looking at the second batch. The survey was approved by the University of Chicago Institutional Review Board (IRB18-0858) and was pre-registered in the AEA RCT Registry (AEARCTR-0003066).

that we asked in our survey in 2018.

### 3.4 List experiment

To elicit xenophobic attitudes, the survey included a list experiment. This design (also called the "unmatched count" and the "item count technique") was originally formalized by Raghavarao and Federer (1979) and further developed in recent works by Blair and Imai (2012) and Glynn (2013), among others. It is a standard technique for eliciting truthful answers to sensitive survey questions. The list experiment works as follows. First, respondents are randomly assigned to either a control or treatment group. Subjects in both groups are then asked to indicate the number of statements they agree with. In this way, the subjects never reveal their agreement with any particular statement, only the total number of statements (unless the subject agrees with all or none, which is something the experimental design should try to avoid). In the control condition, the list contains a set of statements or positions that are not stigmatized. In the treatment condition, the list includes all the statements from the control list but also adds the statement of interest, which is potentially stigmatized (and in both cases, the positions of statements are randomly rotated). The support for the stigmatized opinion can then be inferred by comparing the average number of statements the subjects agree with in the treatment and control conditions. For recent applications of list experiments in economics, see Cantoni et al. (2019a) and Enikolopov et al. (2020).

In our case, the survey participants were asked the following question: *"Consider, please, whether you agree with the following statements. Without specifying exactly which ones you agree with, indicate just the number of statements that you can agree with."*

The respondents in the control group were given four statements unrelated to the issues of ethnicity.[24] The respondents in the treatment group were given the additional fifth statement: *"I feel annoyance or dislike toward some ethnicities."* Here, we took the exact wording used by

[24] The exact statements were the following: i) Over the week I usually read at least one newspaper or magazine; ii) I want to see Russia as a country with a high standard of living; iii) I know the name of the Chairman of the Constitutional Court of the Russian Federation; iv) Our country has a fairly high level of retirement benefits.

FOM in its survey mentioned above, which has the additional advantage of making our results comparable with the results of the opinion polls by this firm. Respondents in the control group, after answering the question on the number of statements they agreed with (which did not include the statement on ethnicities), were then asked a direct question about annoyance or dislike toward some ethnicities.

# 4 Empirical Results

## 4.1 Elicited hostility

Given the randomization, comparing the mean number of positive answers between treatment and control groups provides a valid estimate of the percentage of respondents who agree with the sensitive statement about having xenophobic attitudes (Imai, 2011). Since our goal is to estimate how answers to this sensitive question are affected by social media penetration, which varies at the city level, we first construct a measure of elicited hostility for each city. In particular, we take the difference between the treatment and control group in each city with respect to the number of statements in the list experiment with which the respondents agree:

$$ElicitedHostility_j \equiv \frac{\sum_{T_{ij}=1} y_{ij}}{\sum_i T_{ij}} - \frac{\sum_{T_{ij}=0} y_{ij}}{\sum_i (1-T_{ij})} \tag{4}$$

where $y_{ij}$ is the number of statements in the list experiment that respondent $i$ in city $j$ agrees with, given the treatment status $T_{ij} \in \{0,1\}$. Next, we estimate the following model:

$$ElicitedHostility_j = \beta_0 + \beta_1 \text{VK}_j + \beta_2 \mathbf{X_j} + \eta_\mathbf{j} \tag{5}$$

where $VK_j$ is social media penetration in city $j$,[25] while $\mathbf{X_j}$ is a vector of control variables that

[25] Following Enikolopov et al. (2020), we measure VK penetration as log(1+#VK users). We use the data on VK penetration in 2011 for several reasons: first, people having a larger VK community in 2011 are likely to be experienced

includes the number of students from the city in the other two five-year student cohorts, those that studied three to seven years earlier than Durov, and those that studied three to seven years later than Durov. It also includes the following socioeconomic controls: the logarithm of the population, the indicator for being a regional or a subregional (*rayon*) administrative center, the average wage in the city, the number of city residents of different five-year age cohorts, the share of the population with higher education in 2010 in each five-year age cohort, the indicator for the presence of a university in the city, ethnic fractionalization, and the logarithm of the number of Odnoklassniki users in 2014.[26]

Proposition 1 predicts that the coefficient $\beta_1$ is positive. The OLS relationship of this equation is likely to be biased, as social media penetration could be correlated with unobserved determinants of ethnic hostility. As discussed above, we instrument social media penetration in city $j$ with the number of students from city $j$ who studied at the SPbSU together with the founder of VK, controlling for older and younger cohorts of SPbSU students from the same city and a number of socioeconomic controls. To cope with a weak instrument problem for the reduced sample of 124 cities, we use two-sample IV in the estimation, predicting VK penetration from the full sample of cities and using the predicted values at the second stage, following Choi et al. (2018). Unfortunately, this method does not allow for weighting of the observations in the second stage, so we cannot take into account the fact that the number of observations in the survey for different cities is different.

In what follows, we also look at the subsamples, paying special attention to the groups more likely to be impressionable, i.e. young respondents (below the median age in the sample, which is 32), and respondents with lower levels of education (below college-level education in our sample). Proposition 2 predicts that the effect of social media on extreme opinions could be stronger for those groups.

social media users by 2018; second, the first stage is stronger for earlier years; third, this is the midyear for our data on hate crimes that we use in the following analysis.

[26] The set of controls is identical to the one used in Enikolopov et al. (2020, 2023a).

Table 2 reports the results of the estimation of equation (5). 95% weak-instrument-robust confidence intervals are reported below the main coefficients. Column (1) implies that, consistent with the theoretical prediction, social media penetration leads to higher levels of ethnic hostility. A 10% increase in social media penetration increases the share of people agreeing with the xenophobic statement by 9.5 percentage points, with the 90% weak-instrument-robust confidence set lying entirely above zero. While we do not detect statistical differences across groups at conventional levels, one can see that the magnitude of this effect is larger for those with lower education (17.3 p.p.) as compared with those with higher education (6.9 p.p.), and for younger (21.3 p.p.) as compared with older (4.3 p.p.).[27] This heterogeneity lends tentative support to the prediction from Proposition 2. Overall, the results in Table 2 suggest that social media indeed increased the share of those holding xenophobic opinions, consistent with the predictions of a social learning model presented above.

## 4.2 Social stigma

The share of respondents who agreed with the xenophobic statement in the list experiment (i.e., the difference between the average number of statements with which respondents in the treatment and control groups agreed) was approximately 38%, while the percentage of respondents who admitted being xenophobic in the direct question was 33%. Thus, not every person agreeing with a xenophobic question in the list experiment was ready to openly admit it in answering a direct question, which is consistent with the existence of social stigma in expressing xenophobic attitudes. To test Proposition 4 of the theoretical model, following our pre-analysis plan, we also report the results of the estimation of the following model:

$$ElicitedHostility_j - SelfReportedHostility_j = \beta_3 + \beta_4 \mathrm{VK}_j + \beta_5 \mathbf{X_j} + \varepsilon_\mathbf{j}. \qquad (6)$$

[27] Note that the number of cities in different columns varies because for some cities we do not happen to have respondents in both treatment groups in all the categories.

where the dependent variable is the difference between elicited hostility and self-reported hostility at the city level (to compute $SelfReportedHostility_j$ at the city level, we average the responses of those from the control group who were asked the direct question about xenophobia later on). Note that Proposition 4 suggests that the coefficient $\beta_4$ should be positive.

The results of this estimation are presented in Panel A of Table 3. The magnitudes indicate that, on average, a 10% increase in social media penetration increases the percentage of people unwilling to admit xenophobia (but actually being xenophobic) by 11.6%. Similar to the results in Table 2, these results are stronger for those who are younger and those with lower levels of education, for whom a 10% increase in VK penetration corresponds to 20.0% and 23.8% increase in stigma, respectively.

The measure of social stigma in Panel A is based on the measure of self-reported xenophobia that comes from our online survey. In Panel B, we estimate the results using the same measure of elicited hostility but a measure of self-reported xenophobia that comes from the face-to-face 2011 FOM survey. The coefficients remain positive and significant for most categories, though the magnitudes are somewhat smaller (e.g., our main average coefficient in column (1) reduces from 1.116 to 0.855). Thus, we conclude that our results do not substantially change depending on the mode of survey.

Overall, the results in Table 3 imply that social media does not decrease the stigma of expression of xenophobic attitudes, but, quite the opposite, it increases this stigma. This confirms our theoretical intuition: increased polarization makes people more xenophobic while simultaneously making its expression costlier in terms of social image because one's actions are judged by people with more extreme opinions. When the increased prevalence of extreme opinions is driven by growing polarization, as is the case with the introduction of social media, it does not necessarily lead to an erosion of the social norms that have historically sanctioned the expression of such opinions, as in Bursztyn et al. (2020).[28]

[28] We also report how VK penetration affects self-reported hate (Table B2). Our theoretical prediction (Proposition

## 4.3 Social media and hate crime

The results so far provide support for our theoretical reasoning. We show that the penetration of social media increases elicited xenophobia. We also find evidence that social media increases the stigma associated with the expression of xenophobia, which implies that the increase in truly held xenophobia does not necessarily lead to a change or erosion of social norms. These findings are consistent with the results of Propositions 1–4 of the model.

In this section, we empirically assess Proposition 5 of the model. Our theoretical reasoning suggests that people who participate in our survey and people who commit hate crimes are different kinds of people; however, we expect that social learning from interactions with like-minded people on social media is similar. In what follows, we aim to test if there is a positive interaction effect for social media penetration and pre-existing nationalism and if this relationship is different for hate crimes with multiple perpetrators. Note that the results in Müller and Schwarz (2023) and Müller and Schwarz (2021) are consistent with our theoretical predictions for the United States and Germany cases, with a caveat. Both of those papers look at the impact of the same national-level content in places with different access to social media, following particular content in a Facebook group in Germany or Trump's xenophobic tweets. In our paper, we look at the long-term impact of localized interactions that our model describes.

Following Proposition 5, our main hypothesis is that social media penetration increases hate crime, and this effect is larger in places with higher levels of pre-existing nationalism. To test if the data supports this claim, we estimate the following model:

$$HateCrime_j = \gamma_0 + \gamma_1 VK_j \times NationalistSupport_j + \gamma_2 \mathbf{X_j} + \varepsilon_\mathbf{j}, \tag{7}$$

3) suggests that the effect of social media penetration is ambiguous here. The results in Table B2 are, indeed, quite different from the results in Table 2: the coefficients are negative, not positive, and are marginally statistically significant in only four out of seven cases (see Panel A of Table B2). These results highlight the importance of using survey experiments or other elicitation methods when working with individual-level data on racism and xenophobia. The results are similar if we use data from the FOM 2011 survey.

where $HateCrime_j$ is a measure of hate crime, which reflects either the total number of victims of hate crimes in city $j$ during the period 2007-2015 or the number of victims of particular types of hate crime (ethnic or non-ethnic crimes, conducted by single or multiple perpetrators). $\text{VK}_j$ is VK penetration in city $i$ in summer 2011. As above, the endogenous variable is instrumented using the number of students from each city who have studied together with the founder of VK, Durov. $NationalistSupport_j$ denotes the votes for the nationalist Rodina party in 2003 and captures the pre-existing level of nationalism (this measure is demeaned to simplify the interpretation of the direct coefficients). $\mathbf{X_j}$ is the same vector of control variables as in the previous specifications. For all specifications, we report weak-instrument-robust confidence sets.

Table 4 reports the results of estimating Equation (7). On average, the effect of social media penetration on hate crime is significantly stronger in cities with higher pre-existing levels of nationalism (column (1)). This is especially true for hate crimes conducted by multiple perpetrators (columns (3) and (6)) and for ethnic hate crimes conducted by a single perpetrator. Numerically, the results imply that the effect of a 10% increase in VK penetration ranges from being close to zero (non-significant with different signs) at the minimum level of nationalist party support to a 21.7% increase in the total number of hate crimes at the maximum level of nationalist support (column (1) of Table 4).[29]

Table 5 reports the results of placebo regressions for hate crime in the period 2004-2006, i.e., *before* the creation of the VK social network. The results indicate no significant positive effect of social media on hate crime even in cities with maximum level of support of the nationalist party, with two out of seven coefficients being negative, rather than positive, and significant at 10% level.

[29] For the sake of completeness, we also estimate the direct effect of social media on hate crime. These results are presented in Table B4 in the Appendix. There is no consistent evidence of a significant effect of VK penetration on hate crime for either ethnic or non-ethnic hate crimes or for crimes conducted by single or multiple perpetrators in the IV specification. At the same time, the confidence intervals do not allow us to rule out large effects. These results imply that social media does not uniformly make its users so hateful that they commit hate crimes; this effect is primarily observed in places with higher levels of pre-existing nationalism. This effect is also consistent with the findings in Müller and Schwarz (2023) that the positive effect of Trump tweets in places with higher Twitter penetration comes from the places with pre-existing nationalistic groups.

These findings are consistent with the premise that social media has a causal effect on hate crime after the creation of VK in the end of 2006, in places with a higher level of nationalistic party support, and there was no significant pre-trend in these places.[30]

A potential concern with this data is that there could be a differential likelihood of recording crimes in cities with different social media penetration in a way that is consistent with our results. For example, hate crimes might be more likely to be covered by legacy media and get recorded in the database if they catch attention on social media. Although we do not have evidence to directly rule out this possibility, we believe that it is highly unlikely that ethnic hate crimes were disproportionately reported in areas with both higher penetration of VK *and* a higher baseline level of nationalist sentiment, *and* especially so for crimes with multiple perpetrators. We do a number of additional tests to ensure that this possibility does not bias our results.

First, we check if the effects that we identify negatively depend on the size of the cities. Arguably, in smaller cities with fewer (if any) traditional news media sources, reporting of hate crimes may be more dependent on whether they were discussed in social media or not. This should make the measurement error in hate crime data larger in smaller cities. However, we find that, on the contrary, if we restrict the sample to cities with populations above certain population thresholds, the magnitudes of the reduced form results only increase (see Table B6).[31] Second, if indeed social media makes hate crimes more visible, this effect is supposed to be growing over time as social media penetration increases (Figure B1). However, the magnitude of the effect of social media, if anything, decreases over time (see Tables B7 and B8). Finally, our survey results on attitude changes are also consistent with social media having an effect beyond the mere reporting of hate

[30] Note that the null results in Table 5 may also be driven by the fact that the data for this time period are incomplete, in contrast to the later years. We also test whether the coefficients are statistically different for the period 2004-2006 and 2007-2015 in a pooled regression (see Table B5). In all specifications except one, either the coefficient for the direct effect of VK penetration or its interaction with the support of the nationalistic party or both are statistically different from each other. Unfortunately, we cannot provide weak-instrument-robust confidence intervals for this specification, as the triple-difference specification turns out to be too demanding and the confidence intervals often become degenerate and consist of a single point.

[31] However, due to the reduction in the sample size the coefficients lose their statistical significance.

crimes.

Overall, the results in Tables 4-5 indicate that social media had a positive effect on hate crime, but only in places where the level of nationalism was already sufficiently high before the creation of social media, which is consistent with the theoretical predictions from Proposition 5. The theory also predicts that the effect for multiple perpetrators should be stronger than the effect for single perpetrators, if the coordination channel is important, and this is exactly what we observe. At the same time, the results are also positive and significant for ethnic hate crimes committed by single perpetrators, which suggests that while social media may have facilitated coordination and thus contributed to hate crime, coordination alone is unlikely to fully explain the impact of social media (though coordination broadly speaking may take other forms, such as providing information on opportunities for committing crime to single perpetrators). Thus, we can conclude that the effect of social media on hate crimes comes from both persuasion and coordination.

# 5 Discussion

In this paper, we focus on the effect of social media on the spread of xenophobic views, the likelihood of their public manifestation, and observable actions that follow from having such views. Our model in Section 2 highlights several effects of social media that are supported by the results of our empirical analysis. The introduction of social media increases the likelihood of interactions with like-minded people and causes polarization, which increases the share of people holding xenophobic views. At the same time, this increase in polarization strengthens the role of social image concerns associated with being open about one's xenophobic views – there is a stronger stigmatization of such views. Thus, we observe the situation in which social media increases the number of people *having xenophobic views* as well as the number of people who are *hiding these views* and not expressing them publicly. These two effects can compensate for each other so that the number of people publicly expressing xenophobic views does not change. In the context of xenophobia, however, actually having such views turns out to be more important than openly admitting them since even not-openly-admitted xenophobia can push people to commit hate crimes either individually or together with like-minded xenophobes. Thus, we observe an increase in hate crimes as a result of higher penetration of social media, but only in places with sufficiently high initial levels of xenophobia.

While simple, we believe that our baseline model can be used to explain a wider range of phenomena and can be applicable as a theoretical basis for the fast-growing empirical studies of social media effects. In what follows, we summarize observational and experimental data from some notable recent works on the impact of social media, and we briefly discuss these papers through the lens of our model.

Allcott et al. (2020) studies the effects of disengagement from Facebook in the weeks prior to the 2018 midterm elections in the U.S. and document, among other things, an increase in offline interactions and a decrease in political polarization among disengaged individuals. These findings mirror our model's predictions for the deactivation rather than the introduction of social media.

Our model highlights that social media interactions increase polarization in the society while preserving the mean opinion, for any initial distribution. This implies, however, that any shocks to the mean opinion on a social platform are likely to have a persistent or even permanent effects. An example of this is documented in Fujiwara et al. (2023), who argue that the early adoption of Twitter by South by Southwest festival participants in 2007 resulted in its relatively liberal content and lowered the Republican voter share in the 2016 Presidential election. This negative effect is particularly remarkable given Donald Trump's extensive use of Twitter in his political campaign.

A recent important study by Levy (2021) documents the experimental effect of exposing individuals on Facebook to opposite political opinions. Our model predicts that this experience would reduce political polarization among affected individuals, and in making them closer to the center, it would improve their perception of people with the opposite political views, consistent with the findings in that paper. At the same time, reduced polarization and social stigma would make the prediction about expressing these views ambiguous. In line with that, even though Levy (2021) finds some evidence for social media decreasing polarization after an increase in exposure to counter-attitudinal content, it does not find an experimental effect on political opinion stated in a survey, which is consistent with our theoretical predictions. In future work, it would be interesting to measure the effect of Facebook on truly held beliefs elicited, for example, using a list experiment.

Moreover, the model may be helpful to understand the effects of Internet penetration more broadly. The polarizing effect of social media, according to our model, stems from social media inhibiting convergence as a result of interactions with people holding different opinions. This means that online experiences where homophily is less pronounced than in the offline world may have the opposite effect of reducing polarization. The influential study by Gentzkow and Shapiro (2011) suggests that consumption of online news is one such example (although social media platforms such as Facebook tend to reintroduce homophily strategically; see Levy (2021)). With such a technology, an increase in Internet usage overall is not necessarily associated with a rise in polarization, consistent, for example, with Asimovic et al. (2021) and Boxell et al. (2017). Guriev et al. (2021) also document a mixed effect of 3G Internet access on polarization: on the one hand, access made the public more precisely informed about corruption (increasing perception of corruption in corrupt countries and decreasing it in countries with low corruption). On the other hand, internet access benefited both left-wing and right-wing populists, consistent with an increase in polarization.

# 6 Conclusion

We study the longer-term, causal effects of exposure to social media on xenophobic attitudes and ethnic hate crimes in Russia, using exogenous variation in the city-level initial penetration of social media. We start with a model where interactions between individuals can be persuasive, social media increases the likelihood of meeting like-minded people, and individuals have social image concerns that affect their willingness to reveal their political preferences. We confirm our theoretical predictions empirically. We show that the introduction of social media increases the share of people holding extreme political opinions while at the same time increasing the number of those who conceal such opinions. However, the growing share of people holding extreme political views is consequential – the introduction of social media leads to an increase in hate crimes, particularly in cities with a higher baseline level of nationalist sentiment as well as for crimes with multiple perpetrators.

Taken together, our findings contribute to a growing body of evidence indicating that social media is a complex phenomenon that has both positive and negative effects on the welfare of people (see also Allcott et al., 2020). These effects need to be taken into account when discussing the policy implications of the recent changes in media technologies, as well as possible government regulation or self-regulation by social media platforms.

Our paper also hints at promising directions for future research. One direction is finding more direct evidence on the effect of social media on polarization, which might help understand whether and when social media may contribute to moderation. More generally, it would be interesting to understand the factors that determine opinion formation, both offline and in online social networks, and the implications for the regulation of social media or their internal policies. Finally, it would be interesting to analyze direct evidence on how social media facilitates coordination in practice by analyzing text content in social media forums and understanding how online discussions lead to offline interactions.

# References

**Acemoglu, Daron, Giacomo Como, Fabio Fagnani, and Asuman Ozdaglar**, "Opinion Fluctuations and Disagreement in Social Networks," *Mathematics of Operations Research*, 2013, *38* (1), 1–27.

_ **, Tarek A. Hassan, and Ahmed Tahoun**, "The Power of the Street: Evidence from Egypt's Arab Spring," *Review of Financial Studies*, 2018, *31* (1), 1–42.

**Adena, Maja, Ruben Enikolopov, Maria Petrova, Veronica Santarosa, and Ekaterina Zhuravskaya**, "Radio and the Rise of the Nazis in Prewar Germany," *Quarterly Journal of Economics*, Nov 2015, *130* (4), 1885–1939.

**Algan, Yann, Sergei Guriev, Elias Papaioannou, and Evgenia Passari**, "The European Trust Crisis and the Rise of Populism," *Brookings Papers on Economic Activity*, 2017, *Fall*, 309–382.

**Ali, S Nageeb and Charles Lin**, "Why people vote: Ethical motives and social incentives," *American Economic Journal: Microeconomics*, 2013, *5* (2), 73–98.

_ **and Roland Bénabou**, "Image versus Information: Changing Societal Norms and Optimal Privacy," *American Economic Journal: Microeconomics*, 2020, *12* (3), 116–64.

**Allcott, Hunt, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow**, "The Welfare Effects of Social Media," *American Economic Review*, March 2020, *110* (3), 629–76.

**Alstyne, Marshall Van and Erik Brynjolfsson**, "Global Village or Cyber-Balkans? Modeling and Measuring the Integration of Electronic Communities," *Management Science*, 2005, *51* (6), 851–868.

**Andrews, Donald W. K.**, "Identification-Robust Subvector Inference," Working Paper 2017.

**Andrews, Isaiah, James Stock, and Liyang Sun**, "Weak Instruments in IV Regression: Theory and Practice," *Annual Review of Economics*, 2019 2019, *11*, 727–753.

**Angrist, Joshua D. and Alan B. Krueger**, "The Effect of Age at School Entry on Educational Attainment: an Application of Instrumental Variables with Moments from Two Samples," *Journal of the American Statistical Association*, 1992, *87* (418), 328–336.

**Asimovic, Nejla, Jonathan Nagler, Richard Bonneau, and Joshua A Tucker**, "Testing the Effects of Facebook Usage in an Ethnically Polarized Setting," *Proceedings of the National*

*Academy of Sciences*, 2021, *118* (25), e2022819118.

**Austen-Smith, David and Roland G Fryer**, "An Economic Analysis of 'Acting White'," *The Quarterly Journal of Economics*, 2005, *120* (2), 551–583.

**Barberá, Pablo**, "Social Media, Echo Chambers, and Political Polarization," in Nathaniel Persily and Joshua A. Tucker, eds., *Social Media and Democracy: The State of the Field, Prospects for Reform*, Cambridge University Press, 2020, pp. 34–55.

**Bénabou, Roland and Jean Tirole**, "Incentives and Prosocial Behavior," *American Economic Review*, 2006, *96* (5), 1652–1678.

**Bisin, Alberto and Thierry Verdier**, ""Beyond the Melting Pot": Cultural Transmission, Marriage, and the Evolution of Ethnic and Religious Traits," *The Quarterly Journal of Economics*, 2000, *115* (3), 955–988.

**_ and _** , "The Economics of Cultural Transmission and the Dynamics of Preferences," *Journal of Economic theory*, 2001, *97* (2), 298–319.

**Blair, Graeme and Kosuke Imai**, "Statistical Analysis of List Experiments," *Political Analysis*, 2012, *20* (1), 47–77.

**Bond, Robert M., Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler**, "A 61-Million-Person Experiment in Social Influence and Political Mobilization," *Nature*, September 2012, *489* (7415), 295–298.

**Boxell, Levi, Matthew Gentzkow, and Jesse M. Shapiro**, "Greater Internet Use is Not Associated with Faster Growth in Political Polarization Among US Demographic Groups," *Proceedings of the National Academy of Sciences*, 2017, *114* (40), 10612–10617.

**Bursztyn, Leonardo and Robert Jensen**, "How Does Peer Pressure Affect Educational Investments?," *The Quarterly Journal of Economics*, 2015, *130* (3), 1329.

**_ , Bruno Ferman, Stefano Fiorin, Martin Kanz, and Gautam Rao**, "Status Goods: Experimental Evidence from Platinum Credit Cards in Indonesia," *Quarterly Journal of Economics*, 2018, *133* (3), 1561–1595.

**_ , Georgy Egorov, and Robert Jensen**, "Cool to Be Smart or Smart to Be Cool? Understanding Peer Pressure in Education," *The Review of Economic Studies*, 2019, *86* (4), 1487–1526.

**_ , _ , and Stefano Fiorin**, "From Extreme to Mainstream: The Erosion of Social Norms," *American Economic Review*, 2020, *11* (110), 3522–3548.

**Cantoni, Davide, David Y. Yang, Noam Yuchtman, and Y. Jane Zhang**, "Protests as Strategic Games: Experimental Evidence from Hong Kong's Antiauthoritarian Movement," *Quarterly Journal of Economics*, May 2019, *134* (2), 1021–1077.

**_ , Felix Hagemeister, and Mark Westcott**, "Persistence and Activation of Right-Wing Political Ideology," *Working paper*, 2019.

**_ , Yuyu Chen, David Y Yang, Noam Yuchtman, and Y Jane Zhang**, "Curriculum and ideology," *Journal of Political Economy*, 2017, *125* (2), 338–392.

**Chaudhuri, Saraswata and Eric Zivot**, "A New Method of Projection-based Inference in GMM with Weakly Identified Nuisance Parameters," *Journal of Econometrics*, 2011, *164* (2), 239 – 251.

**Choi, Jaerim, Jiaying Gu, and Shu Shen**, "Weak-instrument Robust Inference for Two-sample Instrumental Variables Regression," *Journal of Applied Econometrics*, 2018, *33* (1), 109–125.

**Coutts, Elisabeth and Ben Jann**, "Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT)," *Sociological Methods & Research*, 2011, *40* (1), 169–193.

**Currie, Janet and Aaron Yelowitz**, "Are Public Housing Projects Good for Kids?," *Journal of Public Economics*, 2000, *75* (1), 99–124.

**Dasaratha, Krishna, Benjamin Golub, and Nir Hak**, "Learning from Neighbours about a Changing State," *Review of Economic Studies*, 2023, *90* (5), 2326–2369.

**DeGroot, M. H.**, "Reaching a Consensus," *Journal of the American Statistical Association*, 1974, *69*, 118–121.

**DellaVigna, Stefano and Ethan Kaplan**, "The Fox News Effect: Media Bias and Voting," *Quarterly Journal of Economics*, 2007, *122* (3), 1187–1234.

**_ , John A. List, and Ulrike Malmendier**, "Testing for Altruism and Social Pressure in Charitable Giving," *Quarterly Journal of Economics*, 2012, *127* (1), 1.

**_ , _ , _ , and Gautam Rao**, "Voting to Tell Others," *Review of Economic Studies*, 2017, *84* (1), 143.

**_ , Ruben Enikolopov, Vera Mironova, Maria Petrova, and Ekaterina Zhuravskaya**, "Cross-border Media and Nationalism: Evidence from Serbian Radio in Croatia," *American Economic Journal: Applied Economics*, 2014, *6* (3), 103–32.

**Durán, Rafael Jiménez, Karsten Müller, and Carlo Schwarz**, "The Effect of Content Moderation on Online and Offline Hate: Evidence from Germany's NetzDG," *Available at SSRN 4230296*, 2023.

**Enikolopov, Ruben, Alexey Makarin, and Maria Petrova**, "Social Media and Protest Participation: Evidence from Russia," *Econometrica*, 2020, *88*, 1479–1514.

**_ , _ , and _** , "Online Corrigendum to "Social Media and Protest Participation: Evidence From Russia"," *Econometrica*, 2023, *91* (3), 1–24.

**_ , _ , _ , and Leonid Polishchuk**, "Social Image, Networks, and Protest Participation," Technical Report 2018.

**_ , Maria Petrova, , Gianluca Russo, and David Yanagizawa-Drott**, "Have Online Networks Undermined Local Communities? Evidence from Facebook," mimeo 2023.

**_ , _ , and Ekaterina Zhuravskaya**, "Media and Political Persuasion: Evidence from Russia," *American Economic Review*, 2011, *101* (7), 3253–85.

**_ , _ , and Konstantin Sonin**, "Social Media and Corruption," *American Economic Journal: Applied Economics*, 2018, *10* (1), 150–74.

**Enke, Benjamin**, "Moral Values and Voting," *Journal of Political Economy*, 2020, *128* (10), 3679–3729.

**Fujiwara, Thomas, Karsten Müller, and Carlo Schwarz**, "The effect of social media on elections: Evidence from the United States," *Journal of the European Economic Association*, 2023, p. jvad058.

**Genicot, Garance**, "Tolerance and Compromise in Social Networks," *Journal of Political Economy*, 2022, *130* (1), 94–120.

**Gentzkow, Matthew and Jesse M. Shapiro**, "Ideological Segregation Online and Offline," *Quarterly Journal of Economics*, 11 2011, *126* (4), 1799–1839.

**Glynn, Adam N.**, "What Can We Learn with Statistical Truth Serum?: Design and Analysis of the List Experiment," *Public Opinion Quarterly*, 2013, *77* (S1), 159–172.

**Golub, Benjamin and Evan Sadler**, "Learning in Social Networks," in Yann Bramoullé, Andrea Galeotti, and Brian Rogers, eds., *The Oxford Handbook of the Economics of Networks*, Oxford University Press, 2016, pp. 504–542.

**Guriev, Sergei, Nikita Melnikov, and Ekaterina Zhuravskaya**, "3g Internet and Confidence in

Government," *Quarterly Journal of Economics*, 2021, *136* (4), 2533–2613.

**Halberstam, Yosh and Brian Knight**, "Homophily, Group Size, and the Diffusion of Political Information in Social Networks: Evidence from Twitter," *Journal of Public Economics*, 2016, *143*, 73 – 88.

**Hatte, Sophie, Etienne Madinier, and Ekaterina Zhuravskaya**, "Reading Twitter in the Newsroom: How Social Media Affects Traditional-Media Reporting?," *Working paper*, 2020.

**Imai, Kosuke**, "Multivariate Regression Analysis for the Item Count Technique," *Journal of the American Statistical Association*, 2011, *106* (494), 407–416.

**Levy, Ro'ee**, "Social Media, News Consumption, and Polarization: Evidence from a Field Experiment," *American Economic Review*, 2021, *111* (3), 831–70.

**Madestam, Andreas, Daniel Shoag, Stan Veuger, and David Yanagizawa-Drott**, "Do Political Protests Matter? Evidence from the Tea Party Movement," *Quarterly Journal of Economics*, 2013, *128* (4), 1633–1685.

**Mobius, Markus and Tanya Rosenblat**, "Social Learning in Economics," *Annual Review of Economics*, 2014, *6*, 827–847.

**Morris, Stephen**, "Political Correctness," *Journal of Political Economy*, 2001, *109* (2), 231–265.

**Mosquera, Roberto, Mofioluwasademi Odunowo, Trent McNamara, Xiongfei Guo, and Ragan Petrie**, "The Economic Effects of Facebook," *Experimental Economics*, 2020, *23* (2), 575–602.

**Müller, Karsten and Carlo Schwarz**, "Fanning the flames of hate: Social media and hate crime," *Journal of the European Economic Association*, 2021, *19* (4), 2131–2167.

**_ and _** , "From hashtag to hate crime: Twitter and antiminority sentiment," *American Economic Journal: Applied Economics*, 2023, *15* (3), 270–312.

**Olea, Jose Luis Montiel and Carolin Pflueger**, "A Robust Test for Weak Instruments," *Journal of Business & Economic Statistics*, 2013, *31* (3), 358–369.

**Olivetti, Claudia and M Daniele Paserman**, "In the Name of the Son (and the Daughter): Intergenerational Mobility in the United States, 1850-1940," *American Economic Review*, 2015, *105* (8), 2695–2724.

**Ou, Susan and Heyu Xiong**, "Mass Persuasion and the Ideological Origins of the Chinese Cultural Revolution," *Journal of Development Economics*, 2021, *153*, 102732.

**Perez-Truglia, Ricardo and Guillermo Cruces**, "Partisan Interactions: Evidence from a Field Experiment in the United States," *Journal of Political Economy*, 2017, *125* (4), 1208–1243.

**Putnam, Robert D.**, *Bowling Alone: The Collapse and Revival of American Community*, New York: Simon & Schuster, 2000.

**Qin, Bei, David Strömberg, and Yanhui Wu**, "Social media and collective action in China," *Cepr discussion paper no. dp16731*, 2021.

**Raghavarao, Damaraju and Walter T. Federer**, "Block Total Response as an Alternative to the Randomized Response Method in Surveys," *Journal of the Royal Statistical Society: Series B (Methodological)*, 1979, *41* (1), 40–45.

**Sartre, Emilie and Gianmarco Daniele**, "Toxic Loans and the Rise of Populist Candidacies," *Available at SSRN 4261333*, 2022.

**Satyanath, Shanker, Nico Voigtländer, and Hans-Joachim Voth**, "Bowling for Fascism: Social Capital and the Rise of the Nazi Party," *Journal of Political Economy*, 2017, *125* (2), 478–526.

**Sen, Ananya and Pinar Yildirim**, "Clicks Bias in Editorial Decisions: How Does Popularity Shape Online News Coverage?," *Working paper*, 2016.

**Stock, James and Motohiro Yogo**, *Testing for Weak Instruments in Linear IV Regression*, Cambridge: Cambridge University Press, 2005.

**Sun, Liyang**, "Implementing Valid Two-Step Identification-Robust Confidence Sets for Linear Instrumental-Variables Models," *The Stata Journal*, 2018, *18* (4), 803–825.

**Sunstein, Cass R.**, *Republic.Com*, Princeton, NJ, USA: Princeton University Press, 2001.

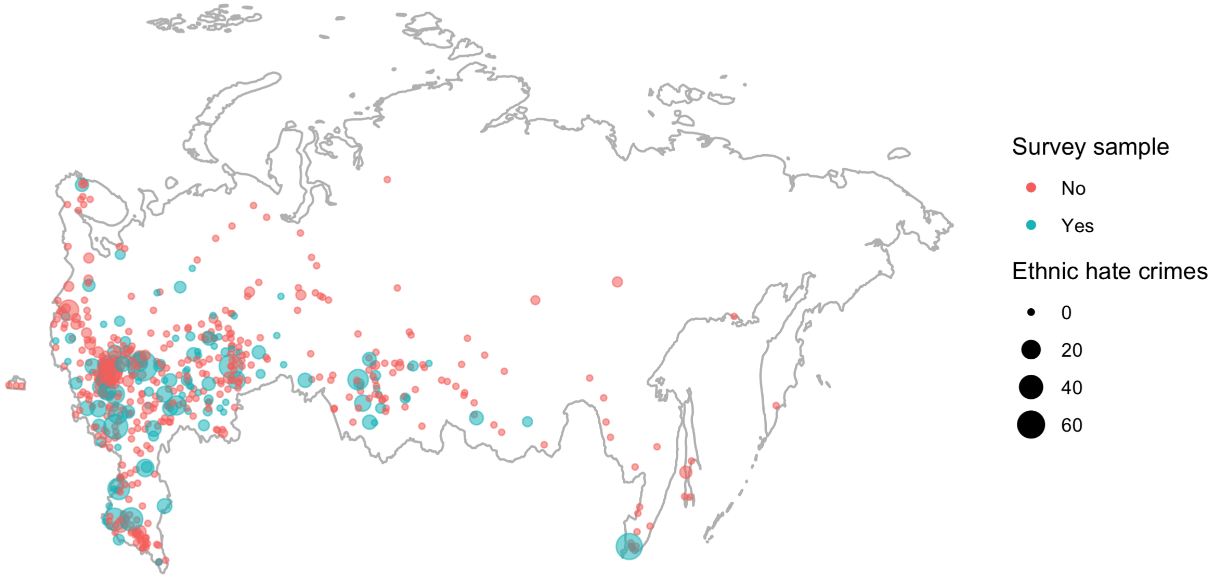
_ , *#Republic: Divided Democracy in the Age of Social Media*, Princeton, NJ, USA: Princeton University Press, 2017.

**Yanagizawa-Drott, David**, "Propaganda and Conflict: Evidence from the Rwandan Genocide," *Quarterly Journal of Economics*, 11 2014, *129* (4), 1947–1994.

**Zhuravskaya, Ekaterina, Maria Petrova, and Ruben Enikolopov**, "Political Effects of the Internet and Social Media," *Annual Review of Economics*, 2020, *12* (1), 415–438.

# Figures and Tables

**Figure 1:** Total Number of Ethnic Hate Crimes by City Across Russia (2007-2015)



*Notes:* The bubble map shows the total number of ethnic hate crimes by city across Russia from 2007 to 2015. Data on ethnic hate crimes comes from the database compiled by the SOVA Center for Information and Analysis.

# Table 1: VK Penetration, SPbSU Student Cohorts, and Nationalistic Party Support

| | Log(Number of VK users, 2011) | Nationalistic party support in 2003 | |
|---|---|---|---|
| | (1) | (2) | (3) |
| Log (SPbSU students), same 5-year cohort as VK founder | 0.144*** | -0.001 | |
| | [0.042] | [0.002] | |
| Log (SPbSU students), one cohort younger than VK founder | -0.046 | -0.001 | -0.001 |
| | [0.031] | [0.001] | [0.001] |
| Log (SPbSU students), one cohort older than VK founder | -0.002 | -0.001 | -0.001 |
| | [0.040] | [0.002] | [0.002] |
| Nationalistic party support in 2003 | 3.951*** | | |
| | [1.192] | | |
| Log(Number of VK users, 2011) | | | -0.010 |
| | | | [0.011] |
| Socioeconomic city-level controls | 625 | 625 | 625 |
| Observations | 0.924 | 0.489 | 0.425 |
| p-value equality all cohorts | 0.003 | 0.951 | |
| p-value equality with young cohort | 0.001 | 0.820 | |
| p-value equality with old cohort | 0.035 | 0.987 | |

*Notes*: Column (1) presents OLS results for the first-stage regression. Column (2) presents OLS results for the reduced form regression. Column (2) presents the results of the IV regression. Robust standard errors in brackets are adjusted by clusters within regions. The unit of observation is a city. The logarithm of any variable is calculated with 1 added inside. Socioeconomic city-level controls include flexible controls for population (5th polynomial), age cohort controls (the number of people aged 25-29, 30-34, 35-39, 40-44, 45-49, 50 and older, in each city according to 2010 Russian Census), education controls (the share of population with higher education overall according to 2002 Russian Census and separately in each of the age cohorts according to 2010 Russian Census, to account for both the levels and the change in education), dummies for regional and county centers, distances to Moscow and St Petersburg, log(average wage), a dummy for the presence of a university, internet penetration in 2011, log(Odnoklassniki users in 2014), and ethnic fractionalization according to 2010 Russian Census. *** p<0.01, ** p<0.05, * p<0.1.

## Table 2: Social Media and Ethnic Hostility, Elicited from List Experiment

| | List Experiment elicited hostility | | | | | | |
|---|---|---|---|---|---|---|---|
| Subsample: | Full Sample | Male | Female | Low Education | High Education | Younger | Older |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Log (Number of VK users, 2011) | 0.950*** | 0.758** | 0.997*** | 1.734*** | 0.693** | 2.130*** | 0.427 |
| Weak Instrument Robust Confidence 95% Sets | ( -.026, 2.037) | (-.585, 2.101) | (-.205, 2.199) | ( .008, 3.699) | (-.354, 1.739) | ( .676, 4.280) | (-.685, 1.539) |
| | [0.277] | [0.343] | [0.307] | [0.501] | [0.267] | [0.548] | [0.284] |
| Socioeconomic city-level controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Younger/Older SPbSU student cohorts | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 124 | 122 | 116 | 124 | 111 | 121 | 116 |

*Notes*: Table presents the results of the two-sample two-stage least squares estimation. The unit of observation is a city. Robust standard errors in square brackets. *** p<0.01, ** p<0.05, * p<0.1. The logarithm of any variable is calculated with 1 added inside. Socioeconomic city-level controls include flexible controls for population (5th polynomial), age cohort controls (the number of people aged 25-29, 30-34, 35-39, 40-44, 45-49, 50 and older, in each city according to 2010 Russian Census), education controls (the share of population with higher education overall according to 2002 Russian Census and separately in each of the age cohorts according to 2010 Russian Census, to account for both the levels and the change in education), dummies for regional and county centers, distances to Moscow and St Petersburg, log(average wage), a dummy for the presence of a university, internet penetration in 2011, log(Odnoklassniki users in 2014), ethnic fractionalization according to 2010 Russian Census, and nationalistic party vote share in 2003 (pre-social media).

Table 3: Social Media and Social Stigma, Elicited from List Experiment

| Panel A. Stigma measured with 2018 survey. | | | | Social Stigma | | | |
|---|---|---|---|---|---|---|---|
| Subsample: | Full Sample | Female | Male | Low Education | High Education | Younger | Older |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Log (Number of VK users, 2011) | 1.116*** | 0.935** | 1.149*** | 2.002*** | 0.739** | 2.379*** | 0.356 |
| Weak Instrument Robust Confidence 95% Sets | ( .077, 2.355) | (-.510, 2.381) | (-.059, 2.463) | ( .280, 4.160) | (-.424, 1.902) | ( .759, 4.772) | (-.745, 1.456) |
| | [0.316] | [0.369] | [0.335] | [0.550] | [0.297] | [0.611] | [0.281] |
| Socioeconomic city-level controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Younger/Older SPbSU student cohorts | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 124 | 122 | 116 | 124 | 111 | 121 | 116 |
| Panel B. Stigma measured with 2011 FOM survey. | | | | Social Stigma | | | |
| Subsample: | Full Sample | Female | Male | Low Education | High Education | Younger | Older |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Log (Number of VK users, 2011) | 0.855*** | 0.419 | 0.965*** | 1.147*** | 0.796*** | 2.025*** | 0.362 |
| Weak Instrument Robust Confidence 95% Sets | (-.184, 1.894) | ( -.820, 1.657 | (-.210, 2.140) | (-.387, 2.680) | (-.323, 1.914) | ( .434, 4.124) | (-.640, 1.364) |
| | [0.265] | [0.316] | [0.300] | [0.391] | [0.285] | [0.536] | [0.256] |
| Socioeconomic city-level controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Younger/Older SPbSU student cohorts | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 118 | 116 | 111 | 118 | 105 | 115 | 110 |

*Notes*: Table presents the results of the two-sample two-stage least squares estimation. The unit of observation is a city. Robust standard errors in square brackets. *** p<0.01, ** p<0.05, * p<0.1. The logarithm of any variable is calculated with 1 added inside. Socioeconomic city-level controls include flexible controls for population (5th polynomial), age cohort controls (the number of people aged 25-29, 30-34, 35-39, 40-44, 45-49, 50 and older, in each city according to 2010 Russian Census), education controls (the share of population with higher education overall according to 2002 Russian Census and separately in each of the age cohorts according to 2010 Russian Census, to account for both the levels and the change in education), dummies for regional and county centers, distances to Moscow and St Petersburg, log(average wage), dummy for the presence of a university, internet penetration in 2011, log(Odnoklassniki users in 2014), ethnic fractionalization according to 2010 Russian Census, and nationalistic party vote share in 2003 (pre-social media).

Table 4: Social Media, Hate Crime, and Pre-Existing Nationalism. Period: 2007-2015.

| | Log (# of hate crimes) | | | Log (# of ethnic hate crime) | | | Log (# of non-ethnic hate crime) | | |
|---|---|---|---|---|---|---|---|---|---|
| | *total* | *single perpetrator* | *multiple perpetrators* | *total* | *single perpetrator* | *multiple perpetrators* | *total* | *single perpetrator* | *multiple perpetrators* |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Log (number of VK users), 2011 x Nationalist Party Support in 2003 | 10.034** | 5.334 | 9.324** | 8.889** | 4.575* | 8.066** | 6.356 | 1.373 | 5.313 |
| Weak Instrument Robust Confidence 95% Sets | ( 2.955, 20.653) | ( 2.391, 17.106) | ( 2.709, 19.246) | ( 2.524, 18.438) | ( .141, 13.443) | ( 1.520, 17.885) | ( 2.852, 20.371) | (-2.346, 6.952) | ( 2.565, 16.304) |
| | [4.334] | [3.604] | [4.050] | [3.897] | [2.715] | [4.008] | [4.291] | [2.277] | [3.365] |
| Log (number of VK users), 2011 | 0.052 | 0.195 | 0.004 | -0.012 | 0.250 | -0.149 | 0.289 | 0.039 | 0.363 |
| Weak Instrument Robust Confidence 95% Sets | (-.430, .774) | (-.209, 1.004) | (-.720, .729) | (-.455, .653) | (-.098, .947) | (-.859, .324) | ( -.232, 1.331) | ( -.250, .471) | ( -.078, 1.243) |
| | [0.295] | [0.247] | [0.296] | [0.271] | [0.213] | [0.290] | [0.319] | [0.176] | [0.269] |
| Nationalist Party Support in 2003 | 4.578** | 1.685 | 4.322* | 4.297** | 0.942 | 4.418** | 1.475 | 0.516 | 0.676 |
| | [2.259] | [1.227] | [2.218] | [1.968] | [0.878] | [2.124] | [1.531] | [0.949] | [1.214] |
| Log (SPbSU students), one cohort younger than VK founder x Nationalist Party Support in 2003 | -7.107 | -4.764 | -6.613 | -6.289 | -4.191 | -4.634 | -6.164 | -1.608 | -5.903 |
| | [5.568] | [4.102] | [5.128] | [4.835] | [3.221] | [4.864] | [5.233] | [2.572] | [4.462] |
| Log (SPbSU students), one cohort older than VK founder x Nationalist Party Support in 2003 | -2.178 | 0.203 | -3.370 | -1.625 | 0.216 | -2.887 | -1.709 | 0.428 | -1.790 |
| | [2.823] | [2.282] | [2.630] | [2.358] | [1.628] | [2.282] | [2.444] | [1.408] | [1.895] |
| Socioeconomic city-level controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 625 | 625 | 625 | 625 | 625 | 625 | 625 | 625 | 625 |
| Kleibergen-Paap F-statistics | 6.449 | 6.449 | 6.449 | 6.449 | 6.449 | 6.449 | 6.449 | 6.449 | 6.449 |
| Full Effect at min level of Nationalist Party | -0.429 | -0.060 | -0.443 | -0.438* | 0.031 | -0.536** | -0.016 | -0.027 | 0.108 |
| Weak Instrument Robust Confidence 95% Sets | (-1.160, .058) | (-.576, .455) | ( -1.165, .039) | (-1.035,-.039) | (-.243, .441) | (-1.194,-.098) | (-.501, .711) | ( -.468, .414) | (-.298, .718) |
| Full Effect at max level of Nationalist Party | 2.168** | 1.320 | 1.971** | 1.863* | 1.215* | 1.552 | 1.629 | 0.328 | 1.483* |
| Weak Instrument Robust Confidence 95% Sets | ( .452, 4.742) | ( -.181, 4.323) | ( .349, 4.402) | ( .259, 4.269) | ( .032, 3.580) | (-.093, 4.018) | ( .747, 5.158) | (-.579, 1.689) | ( .782, 4.289) |

*Notes*: Table presents the results of the two-sample two-stage least squares estimation. The unit of observation is a city. Robust standard errors in square brackets. *** p<0.01, ** p<0.05, * p<0.1. The logarithm of any variable is calculated with 1 added inside. Socioeconomic city-level controls include flexible controls for population (5th polynomial), age cohort controls (the number of people aged 25-29, 30-34, 35-39, 40-44, 45-49, 50 and older, in each city according to 2010 Russian Census), education controls (the share of population with higher education overall according to 2002 Russian Census and separately in each of the age cohorts according to 2010 Russian Census, to account for both the levels and the change in education), dummies for regional and county centers, distances to Moscow and St Petersburg, log (average wage), a dummy for the presence of a university, internet penetration in 2011, log(Odnoklassniki users in 2014), and ethnic fractionalization according to 2010 Russian Census.

Table 5: Social Media, Hate Crime, and Pre-Existing Nationalism. Placebo Estimates. Period: 2004-2006

| | Log (# of hate crimes) | | | Log (# of ethnic hate crime) | | | Log (# of non-ethnic hate crime) |
|---|---|---|---|---|---|---|---|
| | *total* | *single perpetrator* | *multiple perpetrators* | *total* | *single perpetrator* | *multiple perpetrators* | *total* |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Log (number of VK users), 2011 | | | | | | | |
| x Nationalist Party Support in 2003 | -2.903* | -1.119 | -1.784 | -2.292 | -1.119 | -1.173 | -0.611* |
| Weak Instrument Robust Confidence 95% Sets | (-7.218,-.027) | (-5.056,-.332) | (-5.515, 1.947) | (-6.478, .499) | (-5.056,-.332) | -4.770, 2.424) | (-1.501, .279) |
| | [1.761] | [0.964] | [1.523] | [1.709] | [0.964] | [1.468] | [0.363] |
| Log (number of VK users), 2011 | -0.092 | -0.062 | -0.030 | -0.010 | -0.062 | 0.052 | -0.083** |
| Weak Instrument Robust Confidence 95% Sets | (-.423, .239) | (-.291, .052) | (-.235, .277) | (-.310, .291) | (-.291, .052) | (-.131, .328) | (-.151, .054) |
| | [0.135] | [0.070] | [0.126] | [0.123] | [0.070] | [0.112] | [0.042] |
| Nationalist Party Support in 2003 | -0.838 | -0.253 | -0.586 | -0.935 | -0.253 | -0.683 | 0.097 |
| | [0.697] | [0.222] | [0.596] | [0.676] | [0.222] | [0.593] | [0.134] |
| Socioeconomic city-level controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| their interaction with Nationalistic Party Support, 2003 | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 625 | 625 | 625 | 625 | 625 | 625 | 625 |
| Kleibergen-Paap F-statistics | 6.449 | 6.449 | 6.449 | 6.449 | 6.449 | 6.449 | 6.449 |
| Full Effect at minimal level of Nationalist Party | 0.047 | -0.008 | 0.055 | 0.100 | -0.008 | 0.109 | -0.053 |
| p-value for the effect at minimum | (-.157, .352) | ( -.163, .084) | (-.128, .330) | (-.087, .381) | ( -.163, .084) | (-.066, .370) | (-.141, .005) |
| Full Effect at maximum of Nationalist Party Support | -0.705 | -0.298 | -0.407 | -0.493 | -0.298 | -0.195 | -0.212** |
| p-value for the effect at maximum | (-1.799, .025) | (-1.168,-.081) | (-1.052, .562) | (-1.541, .205) | (-1.168,-.080) | (-.797, .708) | (-.383, .045) |

*Notes*: Table presents the results of the two-sample two-stage least squares estimation. The unit of observation is a city. Robust standard errors in square brackets. *** p<0.01, ** p<0.05, * p<0.1. The logarithm of any variable is calculated with 1 added inside. Socioeconomic city-level controls include flexible controls for population (5th polynomial), age cohort controls (the number of people aged 25-29, 30-34, 35-39, 40-44, 45-49, 50 and older, in each city according to 2010 Russian Census), education controls (the share of population with higher education overall according to 2002 Russian Census and separately in each of the age cohorts according to 2010 Russian Census, to account for both the levels and the change in education), dummies for regional and county centers, distances to Moscow and St Petersburg, log(average wage), dummy for the presence of a university, internet penetration in 2011, log(Odnoklassniki users in 2014), and ethnic fractionalization according to 2010 Russian Census.

# Supplementary Appendix

# (Not For Publication)

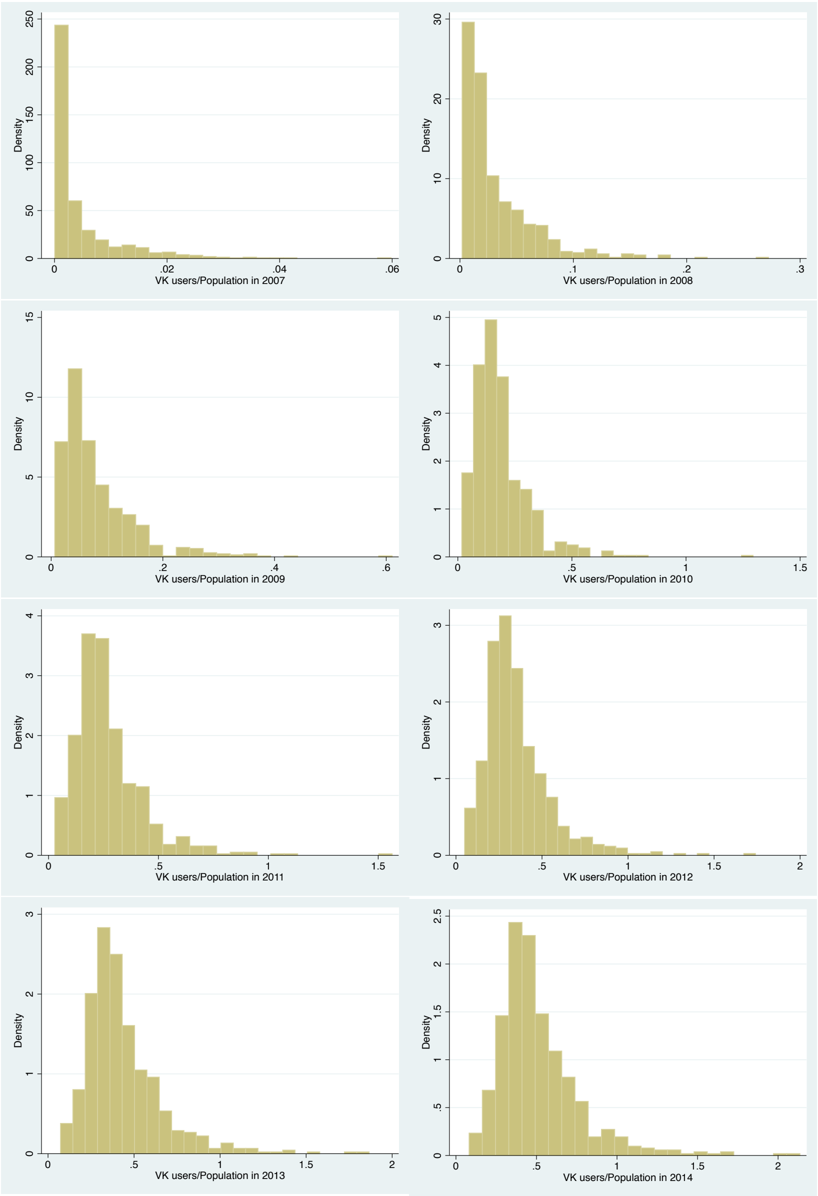
## A Self-Reported Hostility and Social Media

In this section, we present the results of the estimation of equation (5) with reported hostility on the left-hand side. Proposition 3 of our theoretical model predicts that the effect of social media on self-reported hostility is ambiguous. Consistent with this prediction, we find that the effect of social media on reported hostility (Table B2, Panel A) is negative, not positive, and is statistically significant in four out of seven columns. These findings go in stark contrast with Table 2 and highlight that it is important to use various elicitation techniques to study the impact of social media on potentially sensitive opinions.

We also replicate the findings of Table B2 using the data from the 2011 FOM survey, which is a regionally representative survey of 54,388 respondents in 79 regions of Russia, of which 29,780 respondents come from 519 cities in our hate crime sample (Panel B of Table B2). We find that the coefficients remain negative but are only statistically significant for one out of seven categories. These findings are, again, consistent with the prediction from Proposition 3 of our theoretical model.

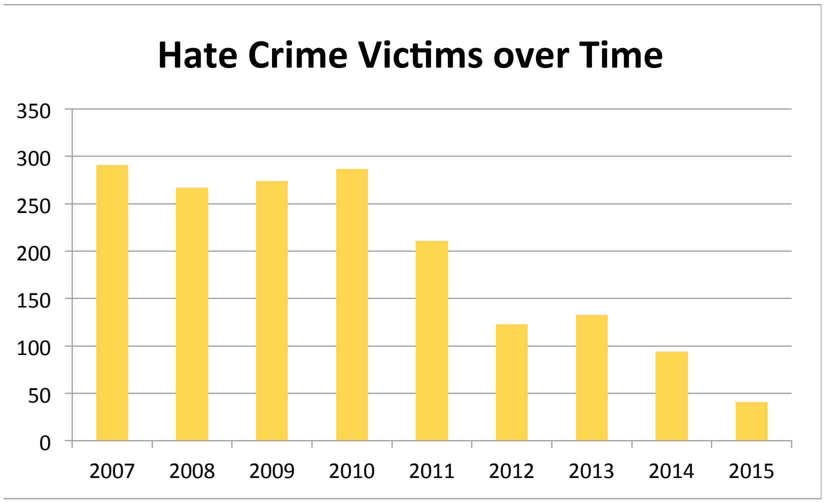
# B Appendix Figures and Tables

**Figure B1:** VK Penetration over Time for 2007-2014



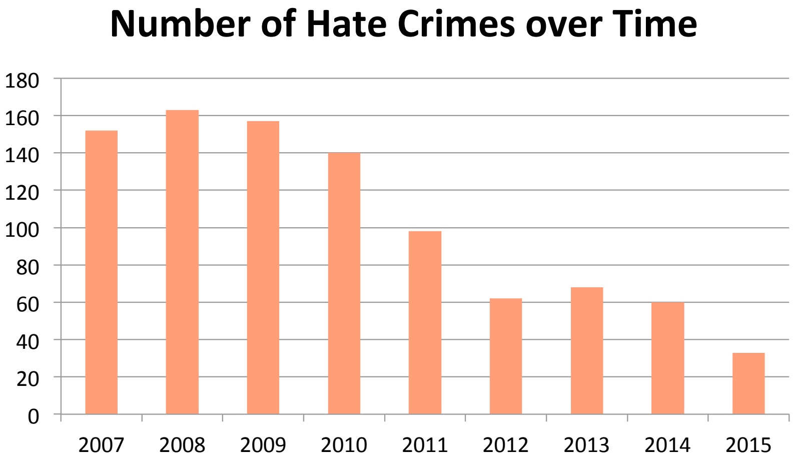
*Notes*: Histograms of VK user population from 2007 to 2014. Data on VK penetration comes from Enikolopov et al. (2020)

**Figure B2:** Hate Crime Victims Over Time



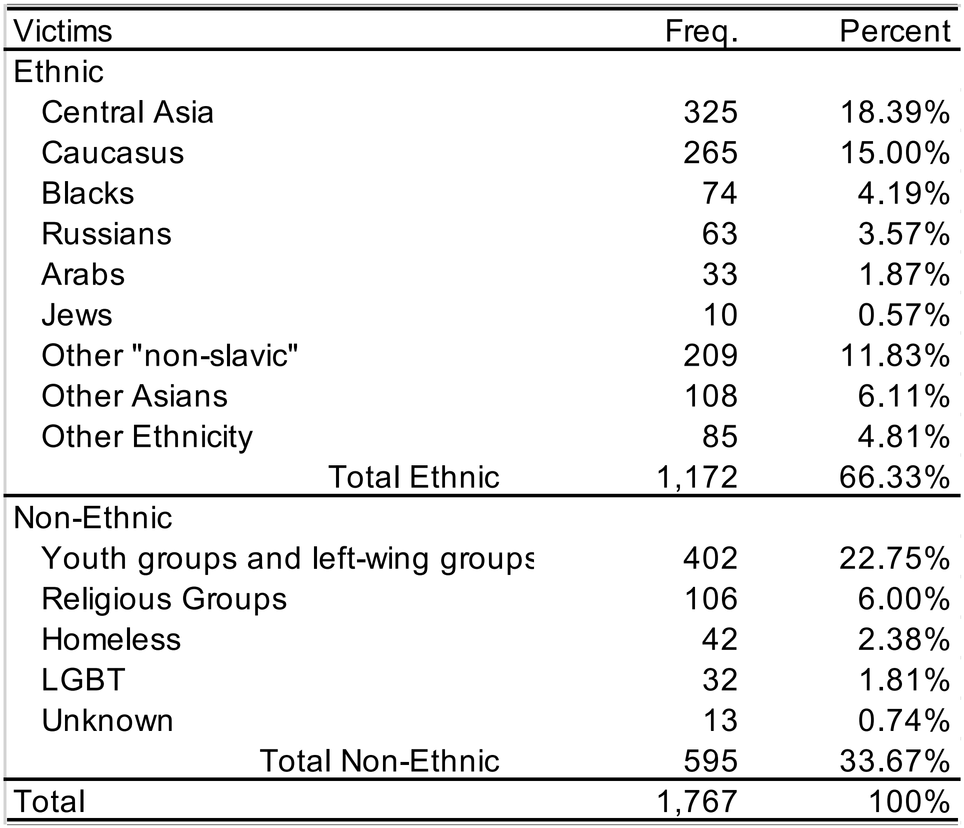
*Notes*: Number of recorded hate crimes from 2007 to 2015. Data on hate crimes comes from the database compiled by the SOVA Center for Information and Analysis.

**Figure B3:** Number of Hate Crimes Over Time



*Notes*: Number of recorded hate crime victims from 2007 to 2015. Data on hate crimes comes from the database compiled by the SOVA Center for Information and Analysis.

Table B1: Number of Victims by Type



| Victims | Freq. | Percent |
|---|---|---|
| Ethnic | | |
| Central Asia | 325 | 18.39% |
| Caucasus | 265 | 15.00% |
| Blacks | 74 | 4.19% |
| Russians | 63 | 3.57% |
| Arabs | 33 | 1.87% |
| Jews | 10 | 0.57% |
| Other "non-slavic" | 209 | 11.83% |
| Other Asians | 108 | 6.11% |
| Other Ethnicity | 85 | 4.81% |
| Total Ethnic | 1,172 | 66.33% |
| Non-Ethnic | | |
| Youth groups and left-wing groups | 402 | 22.75% |
| Religious Groups | 106 | 6.00% |
| Homeless | 42 | 2.38% |
| LGBT | 32 | 1.81% |
| Unknown | 13 | 0.74% |
| Total Non-Ethnic | 595 | 33.67% |
| Total | 1,767 | 100% |

*Notes*: Number of hate crime victims by ethnic and non-ethnic characteristics. Data on hate crimes comes from the database compiled by the SOVA Center for Information and Analysis, 2004-2015.

## Table B2: Social Media and Self-Reported Hostility to Other Ethnicities.

| Subsample: | All | Male | Female | Low Education | High Education | Younger | Older |
|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Panel A: Measured with 2018 survey | | | | | | | |
| Log (Number of VK users, 2011) | -0.153* | -0.142 | -0.197* | -0.256** | -0.069 | -0.217** | -0.060 |
| Weak Instrument Robust Confidence 95% Sets | (-.460, .154) | (-.542, .258) | (-.625, .232) | (-.671, .159) | (-.465, .326) | (-.623, .189) | (-.431, .311) |
| | [0.078] | [0.102] | [0.109] | [0.106] | [0.101] | [0.103] | [0.095] |
| Nationalistic Party Support, 2003 | 0.564 | 2.519*** | -0.675 | 1.082* | -0.032 | 0.272 | -0.449 |
| | [0.498] | [0.675] | [0.654] | [0.601] | [0.681] | [0.696] | [0.585] |
| Socioeconomic city-level controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Younger/Older SPbSU student cohorts | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 125 | 124 | 119 | 125 | 116 | 123 | 121 |
| Panel B: Measured with 2011 FOM survey | | | | | | | |
| Log (Number of VK users, 2011) | 0.010 | -0.032 | 0.045 | 0.055 | -0.318*** | -0.005 | 0.054 |
| Weak Instrument Robust Confidence 95% Sets | (-.127, .148) | (-.232, .167) | (-.121, .211) | (-.112, .222) | (-.750, .113) | ( -.222, .212) | (-.101, .209) |
| | [0.035] | [0.051] | [0.042] | [0.043] | [0.110] | [0.055] | [0.040] |
| Nationalistic Party Support, 2003 | 0.668** | -0.075 | 1.235*** | 0.643** | 1.967*** | 0.175 | 0.544* |
| | [0.263] | [0.391] | [0.307] | [0.309] | [0.618] | [0.366] | [0.285] |
| Socioeconomic city-level controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Younger/Older SPbSU student cohorts | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 118 | 118 | 118 | 118 | 118 | 118 | 118 |

*Notes*: Table presents the results of the two-sample two-stage least squares estimation. The unit of observation is a city. Robust standard errors in square brackets. *** p<0.01, ** p<0.05, * p<0.1. The logarithm of any variable is calculated with 1 added inside. Socioeconomic city-level controls include flexible controls for population (5th polynomial), age cohort controls (the number of people aged 25-29, 30-34, 35-39, 40-44, 45-49, 50 and older, in each city according to 2010 Russian Census), education controls (the share of population with higher education overall according to 2002 Russian Census and separately in each of the age cohorts according to 2010 Russian Census, to account for both the levels and the change in education), dummies for regional and county centers, distances to Moscow and St Petersburg, log(average wage), dummy for the presence of a university, internet penetration in 2011, log(Odnoklassniki users in 2014), ethnic fractionalization according to 2010 Russian Census, and nationalistic party vote share in 2003 (pre-social media).

## Table B3: Nationalistic Party Support and Measures of Xenophobia

| | Log (# of hate crimes) | Log (# of ethnic hate crimes) | Log (# of non-ethninc hate crimes) | Self-reported hostility to other ethnicities |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Nationalist Party Support in 2003 | 1.520** | 1.834*** | -0.044 | 0.496** |
| | [0.669] | [0.654] | [0.346] | [0.225] |
| Population and voting controls | Yes | Yes | Yes | Yes |
| Observations | 625 | 625 | 625 | 27,696 |
| R-squared | 0.493 | 0.439 | 0.382 | 0.012 |

*Notes*: Unit of observation is a city in columns (1)-(3) and an individual respondent in column (4). Robust standard errors are clustered by region in brackets in columns (1)-(3) and at the city level in column (4). Self-reported hostility from FOM survey. Population controls include flexible controls for population (5th polynomial). Electoral controls include votes for United Russia, KPRF, LDPR, Yabloko, SPS, and votes against all. *** p<0.01, ** p<0.05, * p<0.1.

# Table B4: Social Media and Hate Crime. Specification without Interactions. Period: 2007-2015

| | Log (# of hate crimes) | | | Log (# of ethnic hate crime) | | | Log (# of non-ethnic hate crime) | | |
|---|---|---|---|---|---|---|---|---|---|
| | *total* | *single perpetrator* | *multiple perpetrators* | *total* | *single perpetrator* | *multiple perpetrators* | *total* | *single perpetrator* | *multiple perpetrators* |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Log (number of VK users), 2011 | -0.106 | 0.107 | -0.148 | -0.150 | 0.173 | -0.270 | 0.170 | 0.014 | 0.254 |
| Weak Instrument Robust Confidence 95% Sets | (-.776, .564) | (-.263, .633) | (-.837, .540) | (-.728, .377) | (-.153, .672) | (-1.010, .306) | (-.355, .971) | (-.327, .419) | (-.146, .983) |
| | [0.294] | [0.197] | [0.302] | [0.254] | [0.174] | [0.277] | [0.279] | [0.164] | [0.238] |
| Nationalist Party Support in 2003 | 2.199 | 0.579 | 1.785 | 2.264* | -0.015 | 2.416* | -0.397 | 0.257 | -1.105 |
| | [1.636] | [0.964] | [1.550] | [1.372] | [0.695] | [1.465] | [1.317] | [0.682] | [1.157] |
| Log (SPbSU students, one cohort younger) | -0.091* | -0.050* | -0.069 | -0.130*** | -0.035 | -0.113** | 0.037 | -0.025 | 0.057* |
| | [0.049] | [0.027] | [0.049] | [0.044] | [0.024] | [0.047] | [0.036] | [0.023] | [0.032] |
| Log (SPbSU students, one cohort older) | 0.041 | 0.045* | 0.025 | 0.032 | 0.013 | 0.033 | 0.022 | 0.035* | -0.003 |
| | [0.044] | [0.024] | [0.043] | [0.040] | [0.022] | [0.039] | [0.033] | [0.020] | [0.030] |
| Socioeconomic city-level controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 625 | 625 | 625 | 625 | 625 | 625 | 625 | 625 | 625 |
| Kleibergen-Paap F-statistics | 13.586 | 13.586 | 13.586 | 13.586 | 13.586 | 13.586 | 13.586 | 13.586 | 13.586 |
| Effective F-statistics (Montiel Olea and Pflueger 2013) | 15.904 | 15.904 | 15.904 | 15.904 | 15.904 | 15.904 | 15.904 | 15.904 | 15.904 |
| Montiel Olea-Pflueger threshold for 10% worst case bia | 23.109 | 23.109 | 23.109 | 23.109 | 23.109 | 23.109 | 23.109 | 23.109 | 23.109 |
| Montiel Olea-Pflueger threshold for 20% worst case bia | 15.062 | 15.062 | 15.062 | 15.062 | 15.062 | 15.062 | 15.062 | 15.062 | 15.062 |

*Notes*: Unit of observation is a city. Robust standard errors clustered by region in brackets. *** p<0.01, ** p<0.05, * p<0.1. The logarithm of any variable is calculated with 1 added inside. Socioeconomic city-level controls include logarithm of the population according to 2010 Russian Census, age cohort controls (the number of people aged 25-29, 30-34, 35-39, 40-44, 45-49, 50 and older, in each city according to 2010 Russian Census), the share of the population with higher education in each of the age cohorts according to 2010 Russian Census, a dummy for the regional center, log(average wage in 2011), a dummy for the existence of a university in a city, log(Odnoklassniki users in 2014), and ethnic fractionalization according to 2010 Russian Census.

# Table B5: Social Media and Hate Crime. Panel Specification. Period: 2007-2015

| | Log (# of victims of hate crime) | | | Log (# of victims of ethnic hate crime) | | | Log (# of victims of non-ethnic hate crime) | | |
|---|---|---|---|---|---|---|---|---|---|
| | *total* | *single perpetrator* | *multiple perpetrators* | *total* | *single perpetrator* | *multiple perpetrators* | *total* | *single perpetrator* | *multiple perpetrators* |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Log (number of VK users), 2011 x Nationalist Party Support in 2003 x Dummy for the 2004-2006 period | -2.296 | 0.201 | -2.917 | -2.928 | 0.170 | -3.618* | -0.028 | 0.249 | 0.126 |
| | [2.223] | [0.827] | [2.243] | [2.077] | [0.744] | [2.087] | [1.283] | [0.597] | [1.293] |
| Log (number of VK users), 2011 x Nationalist Party Support in 2003 x Dummy for the 2007-2015 period | 7.098** | 3.218** | 6.197* | 7.332** | 2.618** | 7.152** | 3.339 | 0.772 | 2.064 |
| | [3.342] | [1.613] | [3.416] | [2.894] | [1.290] | [2.872] | [2.616] | [1.331] | [2.673] |
| Log (number of VK users), 2011 x Dummy for the 2004-2006 period | -0.314* | -0.046 | -0.306* | -0.215 | 0.035 | -0.259 | -0.135 | -0.072 | -0.049 |
| | [0.180] | [0.081] | [0.182] | [0.154] | [0.074] | [0.162] | [0.147] | [0.064] | [0.141] |
| Log (number of VK users), 2011 x Dummy for the 2007-2015 period | 0.186 | 0.218 | 0.083 | 0.141 | 0.271* | -0.032 | 0.265 | 0.020 | 0.351 |
| | [0.267] | [0.168] | [0.264] | [0.238] | [0.141] | [0.242] | [0.257] | [0.126] | [0.239] |
| Socioeconomic city-level controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Cohorts of SPbSU students, older and younger, their interaction with Nationalistic Party Support, 2003, intereacted with period dummies | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 1,250 | 1,250 | 1,250 | 1,250 | 1,250 | 1,250 | 1,250 | 1,250 | 1,250 |
| p-value for the equality of interaction effects of Log (number of VK users) and Nationalist Party Support for two periods | 0.053 | 0.153 | 0.068 | 0.017 | 0.148 | 0.012 | 0.335 | 0.755 | 0.586 |
| p-value for the equality of direct effects of Log (number of VK users) for two periods | 0.006 | 0.013 | 0.033 | 0.034 | 0.009 | 0.185 | 0.011 | 0.221 | 0.010 |

*Notes*: Unit of observation is a city. Robust standard errors clustered by region in brackets. *** p<0.01, ** p<0.05, * p<0.1. The logarithm of any variable is calculated with 1 added inside. Socioeconomic city-level controls include logarithm of the population according to 2010 Russian Census, age cohort controls (the number of people aged 25-29, 30-34, 35-39, 40-44, 45-49, 50 and older, in each city according to 2010 Russian Census), the share of the population with higher education in each of the age cohorts according to 2010 Russian Census, a dummy for the regional center, log(average wage in 2011), a dummy for the existence of a university in a city, log(Odnoklassniki users in 2014), and ethnic fractionalization according to 2010 Russian Census.

# Table B6: Social Media and Hate Crimes. Reduced form Estimates. Specification with Interactions for Different Thresholds of the City Size

| | Log (# of hate crimes) | | | Log (# of hate crimes) | | | Log (# of hate crimes) | | |
|---|---|---|---|---|---|---|---|---|---|
| | *total* | *single perpetrator* | *multiple perpetrators* | *total* | *single perpetrator* | *multiple perpetrators* | *total* | *single perpetrator* | *multiple perpetrators* |
| **Cities above 50K** | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Log (SPbSU students, Durov's cohort) x Nationalist Party Support in 2003 | 6.887* | 4.566* | 5.940 | 5.940* | 3.662* | 5.417 | 4.764 | 0.789 | 3.801 |
| | [3.937] | [2.656] | [3.969] | [3.564] | [1.896] | [3.656] | [3.409] | [2.253] | [2.504] |
| Log (SPbSU students, Durov's cohort) | -0.044 | -0.017 | -0.026 | -0.047 | 0.003 | -0.053 | 0.019 | -0.006 | 0.049 |
| | [0.066] | [0.039] | [0.070] | [0.052] | [0.031] | [0.056] | [0.060] | [0.035] | [0.051] |
| Nationalist Party Support in 2003 | 0.298 | 0.665 | -0.283 | 0.416 | 0.546 | -0.089 | -0.491 | -0.017 | -0.495 |
| | [1.201] | [0.682] | [1.108] | [1.283] | [0.582] | [1.126] | [0.861] | [0.467] | [0.690] |
| Socioeconomic city-level controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 323 | 323 | 323 | 323 | 323 | 323 | 323 | 323 | 323 |
| **Cities above 75K** | | | | | | | | | |
| Log (SPbSU students, Durov's cohort) x Nationalist Party Support in 2003 | 7.289 | 6.099 | 6.454 | 7.012 | 5.021* | 6.639 | 5.871 | 0.777 | 4.637 |
| | [5.288] | [3.879] | [5.244] | [4.892] | [2.628] | [4.859] | [3.952] | [3.306] | [2.827] |
| Log (SPbSU students, Durov's cohort) | -0.043 | -0.006 | -0.027 | -0.040 | 0.021 | -0.064 | 0.022 | -0.003 | 0.062 |
| | [0.088] | [0.060] | [0.089] | [0.073] | [0.046] | [0.077] | [0.072] | [0.049] | [0.063] |
| Nationalist Party Support in 2003 | -1.546 | 2.791 | -3.030 | -2.397 | 2.113 | -4.128 | 0.242 | -0.019 | 0.341 |
| | [4.910] | [4.064] | [4.705] | [5.447] | [3.417] | [4.848] | [3.359] | [2.575] | [2.899] |
| Socioeconomic city-level controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 200 |
| **Cities above 100K** | | | | | | | | | |
| Log (SPbSU students, Durov's cohort) x Nationalist Party Support in 2003 | 7.289 | 6.933 | 6.032 | 6.381 | 7.001* | 5.288 | 7.184 | -0.823 | 7.059 |
| | [6.054] | [4.750] | [6.453] | [6.031] | [3.524] | [5.950] | [4.866] | [3.984] | [4.268] |
| Log (SPbSU students, Durov's cohort) | 0.043 | 0.007 | 0.072 | 0.033 | 0.013 | 0.020 | 0.075 | 0.034 | 0.105 |
| | [0.119] | [0.082] | [0.125] | [0.103] | [0.071] | [0.110] | [0.104] | [0.063] | [0.099] |
| Nationalist Party Support in 2003 | -0.372 | 4.394 | -2.148 | 0.232 | 5.303 | -3.361 | -1.096 | -1.993 | 0.205 |
| | [6.107] | [5.225] | [5.814] | [6.613] | [4.539] | [6.125] | [4.547] | [3.379] | [4.291] |
| Socioeconomic city-level controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 158 | 158 | 158 | 158 | 158 | 158 | 158 | 158 | 158 |

*Notes*: Table presents the results of the two-sample two-stage least squares estimation. The unit of observation is a city. Robust standard errors in square brackets. *** p<0.01, ** p<0.05, * p<0.1. The logarithm of any variable is calculated with 1 added inside. Socioeconomic city-level controls include flexible controls for population (5th polynomial), age cohort controls (the number of people aged 25-29, 30-34, 35-39, 40-44, 45-49, 50 and older, in each city according to 2010 Russian Census), education controls (the share of population with higher education overall according to 2002 Russian Census and separately in each of the age cohorts according to 2010 Russian Census, to account for both the levels and the change in education), dummies for regional and county centers, distances to Moscow and St Petersburg, log(average wage), a dummy for the presence of a university, internet penetration in 2011, log(Odnoklassniki users in 2014), and ethnic fractionalization according to 2010 Russian Census.

# Table B7: Social Media and Hate Crimes. Specification with Interactions. Specification By Period

| | Log (# of hate crimes) | | | Log (# of hate crimes) | | | Log (# of hate crimes) | | |
|---|---|---|---|---|---|---|---|---|---|
| | *total* | *single perpetrator* | *multiple perpetrators* | *total* | *single perpetrator* | *multiple perpetrators* | *total* | *single perpetrator* | *multiple perpetrators* |
| **2007-2009** | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Log (number of VK users), 2011 | | | | | | | | | |
| x Nationalist Party Support in 2003 | 8.267** | 2.075 | 8.272** | 8.189** | 2.282 | 7.743** | 4.033*** | 0.130 | 3.836** |
| Weak Instrument Robust Confidence 95% Sets | ( 1.921, 20.959) | (-.621, 7.467) | ( 1.801, 21.214) | ( 1.936, 20.694) | (-.295, 7.436) | 1.46398, 20.3017) | ( 1.76112, 13.1213 | (-1.154, 2.055) | ( 1.615, 12.721) |
| | [3.886] | [1.651] | [3.962] | [3.828] | [1.578] | [3.845] | [2.782] | [0.786] | [2.720] |
| Log (number of VK users), 2011 | 0.136 | 0.127 | 0.180 | -0.012 | 0.044 | 0.022 | 0.445** | 0.104* | 0.400** |
| Weak Instrument Robust Confidence 95% Sets | (-.377, .906) | (-.135, .521) | (-.352, .978) | (-.514, .741) | (-.349, .437) | (-.484, .780) | ( .052, 1.231) | (-.003, .317) | ( .019, 1.162) |
| | [0.314] | [0.161] | [0.326] | [0.307] | [0.160] | [0.309] | [0.241] | [0.065] | [0.233] |
| Nationalist Party Support in 2003 | 1.291 | 0.094 | 1.247 | 2.101 | 0.472 | 1.901 | -0.562 | -0.354 | -0.385 |
| | [1.449] | [0.594] | [1.466] | [1.514] | [0.625] | [1.423] | [1.130] | [0.311] | [1.107] |
| Full Effect at minimal level of Nationalist Party Support | -0.260 | 0.028 | -0.216 | -0.404* | -0.066 | -0.350 | 0.252 | 0.098* | 0.216 |
| p-value for the effect at minimum | (-.806, .286) | (-.188, .350) | (-.604, .366) | (-.960,-.034) | (-.384, .147) | (-.906, .207) | (-.098, .952) | (-.010, .259) | (-.122, .893) |
| Full Effect at maximum of  Nationalist Party Support | 1.880** | 0.565 | 1.924** | 1.715** | 0.525 | 1.654* | 1.296*** | 0.131 | 1.209*** |
| p-value for the effect at maximum | ( .136, 4.494) | (-.187, 2.067) | ( .150, 5.474) | ( .015, 4.265) | (-.206, 1.622) | (-.056, 4.220) | ( .700, 3.680) | (-.186, .607) | ( .627, 3.539) |
| **2010-2012** | | | | | | | | | |
| Log (number of VK users), 2011 | | | | | | | | | |
| x Nationalist Party Support in 2003 | 7.091** | 2.476 | 6.180** | 6.251** | 3.238* | 3.776 | 1.491 | -0.741 | 2.284 |
| Weak Instrument Robust Confidence 95% Sets | ( .528, 16.936) | (-1.338, 10.104) | ( .573, 14.590) | ( .102, 15.476) | ( -.247, 10.209) | (-1.769, 12.093) | (-2.010, 8.493) | (-2.718, 2.225) | (-.610, 8.071) |
| | [4.018] | [2.335] | [3.433] | [3.765] | [2.134] | [3.395] | [2.143] | [1.211] | [1.772] |
| Log (number of VK users), 2011 | 0.092 | 0.122 | 0.050 | 0.151 | 0.179 | -0.007 | -0.073 | -0.022 | -0.022 |
| Weak Instrument Robust Confidence 95% Sets | (-.309, .693) | (-.169, .702) | (-.509, .610) | (-.248, .750) | (-.088, .713) | (-.575, .372) | (-.380, .387) | (-.182, .218) | (-.286, .373) |
| | [0.245] | [0.178] | [0.228] | [0.244] | [0.164] | [0.232] | [0.188] | [0.098] | [0.161] |
| Nationalist Party Support in 2003 | 4.088** | 1.358* | 3.108** | 3.052** | 1.208* | 2.077 | 1.585* | 0.041 | 1.456** |
| | [1.698] | [0.733] | [1.578] | [1.515] | [0.664] | [1.379] | [0.821] | [0.425] | [0.709] |
| Full Effect at minimal level of Nationalist Party Support | -0.248 | 0.003 | -0.246 | -0.149 | 0.024 | -0.188 | -0.145 | 0.013 | -0.132 |
| p-value for the effect at minimum | (-.842, .148) | (-.227, .463) | (-.826, .141) | (-.701, .219) | (-.194, .459) | (-.678, .139) | (-.402, .242) | (-.122, .216) | (-.372, .229) |
| Full Effect at maximum of  Nationalist Party Support | 1.587* | 0.644 | 1.353** | 1.469* | 0.862* | 0.790 | 0.241 | -0.178 | 0.459 |
| p-value for the effect at maximum | (-.003, 3.972) | (-.364, 2.660) | ( .008, 3.371) | (-.057, 3.758) | (-.055, 2.695) | (-.619, 2.198) | (-.700, 2.124) | (-.700, .603) | (-.301, 1.980) |
| **2013-2015** | | | | | | | | | |
| Log (number of VK users), 2011 | | | | | | | | | |
| x Nationalist Party Support in 2003 | 1.540 | 2.021 | -0.124 | 0.460 | 0.570 | 0.332 | 1.525 | 1.645 | -0.145 |
| Weak Instrument Robust Confidence 95% Sets | (-5.141, 5.994) | (-.833, 6.302) | (-9.074, 4.352) | (-7.894, 4.637) | (-.957, 2.860) | (-8.408, 4.703) | (-1.488, 6.043) | (-.890, 5.447) | (-1.766, 2.287) |
| | [2.727] | [1.747] | [2.740] | [2.558] | [0.935] | [2.676] | [1.844] | [1.552] | [0.993] |
| Log (number of VK users), 2011 | -0.203 | 0.077 | -0.292 | -0.167 | 0.112** | -0.251 | -0.022 | -0.013 | -0.045 |
| Weak Instrument Robust Confidence 95% Sets | (-.884, .138) | (-.127, .384) | (-.966, .045) | (-.798, .149) | ( .004, .275) | (-.900, .073) | (-.369, .325) | (-.300, .274) | ( -.267, .103) |
| | [0.208] | [0.125] | [0.206] | [0.193] | [0.066] | [0.198] | [0.142] | [0.117] | [0.091] |
| Nationalist Party Support in 2003 | 1.641* | 0.167 | 1.589* | 1.116 | -0.485 | 1.627* | 0.594 | 0.617 | 0.105 |
| | [0.993] | [0.759] | [0.885] | [0.845] | [0.465] | [0.864] | [0.837] | [0.811] | [0.392] |
| Full Effect at minimal level of Nationalist Party Support | -0.277 | -0.020 | -0.286 | -0.189 | 0.085 | -0.267 | -0.095 | -0.092 | -0.038 |
| p-value for the effect at minimum | (-.735, .029) | (-.371, .215) | ( -.813,-.023) | (-.564, .062) | (-.029, .256) | (-.735,-.033) | (-.489, .167) | (-.451, .148) | (-.228, .089) |
| Full Effect at maximum of  Nationalist Party Support | 0.122 | 0.503 | -0.318 | -0.070 | 0.233 | -0.181 | 0.299 | 0.334 | -0.075 |
| p-value for the effect at maximum | (-1.587, 1.261) | (-.1517, 1.486) | (-2.686, .865) | (-2.274, 1.033) | ( -.133, .781) | (-2.521, .989) | (-.394, 1.340) | (-.210, 1.149) | (-.731, .581) |

*Notes*: Table presents the results of the two-sample two-stage least squares estimation. The unit of observation is a city. Robust standard errors in square brackets. *** p<0.01, ** p<0.05, * p<0.1. The logarithm of any variable is calculated with 1 added inside. Socioeconomic city-level controls include flexible controls for population (5th polynomial), age cohort controls (the number of people aged 25-29, 30-34, 35-39, 40-44, 45-49, 50 and older, in each city according to 2010 Russian Census), education controls (the share of population with higher education overall according to 2002 Russian Census and separately in each of the age cohorts according to 2010 Russian Census, to account for both the levels and the change in education), dummies for regional and county centers, distances to Moscow and St Petersburg, log(average wage), a dummy for the presence of a university, internet penetration in 2011, log(Odnoklassniki users in 2014), and ethnic fractionalization according to 2010 Russian Census.

# Table B8: Social Media and Hate Crimes. Reduced form Estimates. Specification with Interactions. Specification By Period

| | Log (# of hate crimes) | | | Log (# of hate crimes) | | | Log (# of hate crimes) | | |
|---|---|---|---|---|---|---|---|---|---|
| | *total* | *single perpetrator* | *multiple perpetrators* | *total* | *single perpetrator* | *multiple perpetrators* | *total* | *single perpetrator* | *multiple perpetrators* |
| **2007-2009** | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Log (SPbSU students, Durov's cohort) x Nationalist Party Support in 2003 | 5.138** | 1.255 | 5.125** | 5.144** | 1.416 | 4.852** | 2.365* | 0.042 | 2.258* |
| | [2.127] | [0.935] | [2.152] | [2.083] | [0.900] | [2.100] | [1.343] | [0.515] | [1.315] |
| Log (SPbSU students, Durov's cohort) | -0.006 | 0.012 | 0.000 | -0.026 | -0.001 | -0.020 | 0.051* | 0.014* | 0.045* |
| | [0.038] | [0.020] | [0.041] | [0.038] | [0.020] | [0.040] | [0.026] | [0.008] | [0.026] |
| Nationalist Party Support in 2003 | -1.460* | -0.231 | -1.334* | -1.199 | -0.263 | -1.092 | -0.416 | 0.003 | -0.338 |
| | [0.850] | [0.240] | [0.785] | [0.785] | [0.225] | [0.701] | [0.331] | [0.121] | [0.315] |
| **2010-2012** | | | | | | | | | |
| Log (SPbSU students, Durov's cohort) x Nationalist Party Support in 2003 | 4.417* | 1.508 | 3.860* | 3.867* | 1.965* | 2.373 | 0.963 | -0.457 | 1.442 |
| | [2.238] | [1.287] | [2.007] | [2.214] | [1.118] | [2.098] | [1.329] | [0.792] | [1.042] |
| Log (SPbSU students, Durov's cohort) | -0.008 | 0.010 | -0.012 | 0.002 | 0.016 | -0.012 | -0.015 | -0.001 | -0.010 |
| | [0.034] | [0.021] | [0.033] | [0.032] | [0.019] | [0.029] | [0.026] | [0.013] | [0.023] |
| Nationalist Party Support in 2003 | 1.630 | 0.852* | 0.849 | 1.159 | 0.624* | 0.549 | 0.704 | 0.248 | 0.460 |
| | [1.021] | [0.464] | [0.914] | [0.811] | [0.322] | [0.762] | [0.525] | [0.219] | [0.416] |
| **2013-2015** | | | | | | | | | |
| Log (SPbSU students, Durov's cohort) x Nationalist Party Support in 2003 | 1.043 | 1.239 | 0.032 | 0.351 | 0.316 | 0.303 | 0.965 | 1.038 | -0.074 |
| | [1.733] | [1.025] | [1.616] | [1.619] | [0.543] | [1.634] | [1.107] | [0.914] | [0.630] |
| Log (SPbSU students, Durov's cohort) | -0.033 | 0.005 | -0.041* | -0.025 | 0.014 | -0.036 | -0.008 | -0.007 | -0.006 |
| | [0.027] | [0.019] | [0.023] | [0.024] | [0.009] | [0.023] | [0.021] | [0.018] | [0.012] |
| Nationalist Party Support in 2003 | 0.232 | -0.332 | 0.490 | 0.278 | -0.270 | 0.509 | -0.099 | -0.088 | -0.013 |
| | [0.452] | [0.222] | [0.511] | [0.477] | [0.167] | [0.520] | [0.193] | [0.132] | [0.129] |
| Socioeconomic city-level controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 625 | 625 | 625 | 625 | 625 | 625 | 625 | 625 | 625 |

*Notes*: Table presents the results of the two-sample two-stage least squares estimation. The unit of observation is a city. Robust standard errors in square brackets. *** p<0.01, ** p<0.05, * p<0.1. The logarithm of any variable is calculated with 1 added inside. Socioeconomic city-level controls include flexible controls for population (5th polynomial), age cohort controls (the number of people aged 25-29, 30-34, 35-39, 40-44, 45-49, 50 and older, in each city according to 2010 Russian Census), education controls (the share of population with higher education overall according to 2002 Russian Census and separately in each of the age cohorts according to 2010 Russian Census, to account for both the levels and the change in education), dummies for regional and county centers, distances to Moscow and St Petersburg, log(average wage), a dummy for the presence of a university, internet penetration in 2011, log(Odnoklassniki users in 2014), and ethnic fractionalization according to 2010 Russian Census.

# C Translated Survey Script

## Questionnaire

**Q0. Which city have you been living in for the last 6 months?**
**List of cities**

**Q1. How often do you use social networks?**
**One answer**

| | |
|---|---|
| **1** | Not at all **[skip to question 3]** |
| **2** | Once a month or less |
| **3** | Once a week |
| **4** | Every day or almost every day |
| **5** | Several times a day |
| **6** | I'm using social networks nonstop |

**Q2. Which of the following social networks do you use?**
**Several answers possible + rotation**

| | |
|---|---|
| **1** | VKontakte |
| **2** | Facebook |
| **3** | Odnoklassniki.ru |
| **4** | LiveJournal |
| **5** | Twitter |
| **98** | Other **(please specify)** |

**Q3. Which websites do you visit most often?**
**One answer**

| | |
|---|---|
| **1** | News and analytics websites |
| **2** | Social networks |
| **3** | Games and entertainment websites |
| **4** | Online stores |
| **5** | Search engines |
| **98** | Other |
| **99** | Unsure |

**Q4. On social networks, do you use your real name or an alias?**
**One answer**

| | |
|---|---|
| 1 | Real name |
| 2 | An alias, for privacy concerns |
| 3 | An alias, but for a reason other than privacy concerns |

**Q5. How many friends/followers do you have in social networks?**
**One answer**

| | |
|---|---|
| 1 | Less than 10 |
| 2 | 10-100 |
| 3 | 100-250 |
| 4 | 250-500 |
| 5 | 500-1000 |
| 6 | More than 1000 |

**Q6. Do you agree with the statement “I get a lot of important news from social networks”?**
**One answer**

| | |
|---|---|
| 1 | Agree |
| 2 | Somewhat agree |
| 3 | Somewhat disagree |
| 4 | Disagree |

**Q7. Do you agree with the statement “Social networks help me find people with similar interests”?**
**One answer**

| | |
|---|---|
| 1 | Agree |
| 2 | Somewhat agree |
| 3 | Somewhat disagree |
| 4 | Disagree |

**Q8. Do you agree with the statement “In social networks, people are more sincere than in real life”?**
**One answer**

| | |
|---|---|
| 1 | Agree |
| 2 | Somewhat agree |
| 3 | Somewhat disagree |
| 4 | Disagree |

**Q9. To what extent do you trust information in social networks?**
**One answer**

| | |
|---|---|
| 1 | Completely trust **[skip to question 10]** |
| 2 | Somewhat trust **[skip to question 10]** |
| 3 | Somewhat distrust **[skip to question 11]** |
| 4 | Completely distrust **[skip to question 11]** |

**Q10. Why do you trust information in social networks?**
**Several answers possible**

| | |
|---|---|
| 1 | People are more sincere in social networks than in real life |
| 2 | In social networks one can find a variety of opinions |
| 3 | Certain information is only available in social networks |
| 98 | Other reason **(please specify)** |

**Q11. Why do you distrust information in social networks?**
**Several answers possible**

| | |
|---|---|
| 1 | Many users deliberately spread incorrect information |
| 2 | Many users unwittingly spread incorrect information |
| 3 | Many users play the fool and write rubbish |
| 98 | Other reason **(please specify)** |

**Q12. In social networks, how often do you encounter:**
**[scale: A. Very often, B Often, C Occasionally, D Rarely, E Never]**
**Rotation of statements, one answer**

| | |
|---|---|
| 1 | Personal insults |
| 2 | Obviously incorrect information |
| 3 | Extremist statements |
| 4 | Propaganda of violence |
| 5 | Religious propaganda |
| 6 | Pornography |

**Q13. Which modern technology do you use to organize gatherings with friends or acquaintances?**
**Several answers possible + rotation**

| | |
|---|---|
| 1 | Yes, video calls (e.g., Skype) |
| 2 | Yes, messengers embedded in social networks (VKontakte, Facebook, etc) |
| 3 | Yes, standalone messengers (WhatsApp, Telegram, ICQ, etc) |
| 4 | Yes, blogs or public posts in social networks |
| 5 | Yes, SMS (short text messages sent over the phone) |
| 6 | Yes, phone calls |

THERE ARE TWO RANDOMIZED CELLS.

CELL 1 [QUESTION Q14_1]

Q14_1. **Please think, which of the following statements you agree with. Without telling which particular statements you agree with, please specify the number of statements you agree with.**

THE ANSWER IS A NUMBER BETWEEN 0 AND 5, ROTATION

| | |
|---|---|
| 1 | Each week I usually read at least one newspaper or magazine |
| 2 | I want Russia to be a country with high living standard |
| 3 | I know the name of the Chairman of the Constitutional Court of the Russian Federation |
| 4 | I feel annoyance or dislike toward some ethnicities |
| 5 | Retirement benefits in our country are sufficiently high |

CELL 2 [QUESTIONS Q14_2, 15, 16, 17]

Q14_2. **Please think, which of the following statements you agree with. Without telling which particular statements you agree with, please specify the number of statements you agree with.**

THE ANSWER IS A NUMBER BETWEEN 0 AND 4, ROTATION

| | |
|---|---|
| 1 | Each week I usually read at least one newspaper or magazine |
| 2 | I want Russia to be a country with high living standard |
| 3 | I know the name of the Chairman of the Constitutional Court of the Russian Federation |
| 4 | Retirement benefits in our country are sufficiently high |

Q15. **Do you feel annoyance or dislike toward some ethnicities?**
One answer

| | |
|---|---|
| 1 | Yes |
| 2 | No |

Q16. **In your opinion, which percentage of the survey participants from your city answered "Yes" to the previous question? If your answer is the most accurate, you will get an additional 100 rubles.**
Enter a number with a percentage sign – restrict from 0 to 100

Q17. **How certain are you in your answer to the previous question?**
SLIDER FROM 0 (COMPLETELY UNCERTAIN) TO 10 (COMPLETELY SURE)

**QUESTIONS ON GENDER AND AGE ARE ASKED ON THE TECHNICAL PAGE "CIRCLE", SURVEY RESTRICTED TO PEOPLE 18 – 55 YEARS OF AGE**

**S3. Please specify your education.**
**One answer**

| | |
|---|---|
| 1 | Incomplete secondary |
| 2 | Secondary |
| 3 | Vocational |
| 4 | Incomplete higher |
| 5 | Higher |
| 6 | Doctorate |
| 99 | Not sure |

**S4. Please specify your occupation (your position).**
**One answer**

| | |
|---|---|
| 1 | Director, deputy director |
| 2 | Division head (of a branch, shift, department) |
| 3 | Specialist with a higher education (medical doctor, teacher, sales manager, engineer, etc) |
| 4 | Mid-level employee (secretary, salesperson, security, driver, etc) |
| 5 | Creative work (photographer, artist, actor, etc) |
| 6 | Small business (owner of a business or individual entrepreneur) |
| 7 | Technical or service personnel |
| 8 | Worker |
| 9 | Military |
| 10 | Student |
| 98 | Other **(please specify)** |

**S5. How would you describe your family's current financial well-being?**
**One answer**

| | |
|---|---|
| 1 | Not enough money even for food |
| 2 | Enough money for food, but purchasing clothes is problematic |
| 3 | Enough money for food and clothes, but purchasing a TV, a fridge or a washer would be difficult |
| 4 | Enough money for major appliances, but we would not be able to buy a new car |
| 5 | Enough money for everything except expensive purchases like a country house or an apartment |
| 6 | No material difficulties. Can afford to buy a country house or an apartment if necessary |

# D Theory Proofs

**Proof of Lemma 1.** First of all, plugging (2) into (1) we get

$$x_i^t = \left(\omega + (1-\omega)\,\tau_{n(i)}h\right) x_i^{t-1} + (1-\omega)\left(1-\tau_{n(i)}h\right)\mu_{n(i)}^{t-1} + \varepsilon_i^t. \tag{8}$$

Taking the expectation of both sides and using that $\mathbb{E}x_i^t = \mu_{n(i)}^t$, we get

$$\begin{aligned} \mu_{n(i)}^t &= \mathbb{E}x_i^t = \left(\omega + (1-\omega)\,\tau_{n(i)}h\right)\mathbb{E}x_i^{t-1} + (1-\omega)\left(1-\tau_{n(i)}h\right)\mu_{n(i)}^{t-1} \\ &= \mathbb{E}x_i^{t-1} = \mu_{n(i)}^{t-1}, \end{aligned}$$

and therefore $\mathbb{E}x_i^t = \mathbb{E}x_i^0 = \mu_{n(i)}^0 = \mu_{n(i)}$ for each $t$.

We can now iteratively plug in $x_i^{t-1}, x_i^{t-2}, \ldots$ into (8) to get

$$\begin{aligned} x_i^t = & \left(\omega + (1-\omega)\,\tau_{n(i)}h\right)^t x_i^0 \\ & + (1-\omega)\left(1-\tau_{n(i)}h\right)\textstyle\sum_{k=1}^t \left(\omega + (1-\omega)\,\tau_{n(i)}h\right)^{k-1}\mu_{n(i)} \\ & + \textstyle\sum_{k=1}^t \left(\omega + (1-\omega)\,\tau_{n(i)}h\right)^{k-1}\varepsilon_i^{t-k+1}. \end{aligned} \tag{9}$$

Since $\left(\omega + (1-\omega)\,\tau_{n(i)}h\right) \in (0,1)$, the first term converges to 0 in probability as $t \to \infty$. The second term equals

$$(1-\omega)\left(1-\tau_{n(i)}h\right)\frac{1-\left(\omega + (1-\omega)\,\tau_{n(i)}h\right)^t}{1-\left(\omega + (1-\omega)\,\tau_{n(i)}h\right)}\mu_{n(i)} = \mu_{n(i)} - \left(\omega + (1-\omega)\,\tau_{n(i)}h\right)^t\mu_{n(i)},$$

which converges to $\mu_{n(i)}$ in probability. Now the last term in (9) is a sum of $t$ independent normal variables, and thus the sum is also normal. Its mean is zero, and its variance equals

$$\textstyle\sum_{k=1}^t \left(\left(\omega + (1-\omega)\,\tau_{n(i)}h\right)^{k-1}\right)^2\sigma_\varepsilon^2 = \dfrac{1-\left(\omega + (1-\omega)\,\tau_{n(i)}h\right)^{2t}}{1-\left(\omega + (1-\omega)\,\tau_{n(i)}h\right)^2}\sigma_\varepsilon^2.$$

This latter term converges to $\sigma^2$ defined by (3) as $t \to \infty$, which implies that the sum converges to $\mathcal{N}\left(0,\sigma^2\right)$ in distribution. Since the last term in (9) converges to $\mathcal{N}\left(0,\sigma^2\right)$ in distribution, and the sum of the first two converges to a constant $\mu_{n(i)}$ in probability, we have that $x_i^t$ converges to $\mathcal{N}\left(\mu,\sigma^2\right)$ in distribution as $t \to \infty$. The comparative statics results are straightforward, which completes the proof. ■

**Proof of Proposition 1.** We have:

$$\begin{aligned} R_n(q) &= \frac{1}{\sqrt{2\pi}\sigma_n}\int_q^{+\infty}\exp\left(-\frac{(x-\mu_n)^2}{2\sigma_n^2}\right)dx = \left[\frac{x-\mu_n}{\sigma_n}=y\right] \\ &= \frac{1}{\sqrt{2\pi}}\int_{\frac{q-\mu_n}{\sigma_n}}^{+\infty}\exp\left(-\frac{y^2}{2}\right)dy. \end{aligned}$$

Now,

$$\frac{\partial R_n(q)}{\partial\sigma_n} = \frac{1}{\sqrt{2\pi}}\frac{q-\mu_n}{\sigma_n^2}\exp\left(-\frac{(q-\mu_n)^2}{2\sigma_n^2}\right),$$

which is positive if $q > \mu$ and negative otherwise. Since by Lemma 1 $\sigma_n$ is increasing in $\tau_n$, the result follows immediately. ■

**Proof of Proposition 2.** Using the formula in the proof of Proposition 1 and substituting $\sigma_n$ from (3), we have

$$\begin{aligned} R_n(q) &= \frac{1}{\sqrt{2\pi}}\int_{\frac{q-\mu_n}{\sigma_n}}^{+\infty}\exp\left(-\frac{y^2}{2}\right)dy \\ &= \frac{1}{\sqrt{2\pi}}\int_{\frac{q-\mu_n}{\sigma_\varepsilon}\sqrt{1-(\omega+(1-\omega)\tau_n h)^2}}^{+\infty}\exp\left(-\frac{y^2}{2}\right)dy. \end{aligned}$$

Differentiating with respect to $\tau_n$, we have

$$\begin{aligned} \frac{\partial R_n(q)}{\partial\tau_n} &= \frac{1}{\sqrt{2\pi}}\frac{q-\mu_n}{\sigma_\varepsilon}\frac{h(1-\omega)(\omega+(1-\omega)\tau_n h)}{\sqrt{1-(\omega+(1-\omega)\tau_n h)^2}} \\ &\quad \times\exp\left(-\frac{1-(\omega+(1-\omega)\tau_n h)^2}{2}\left(\frac{q-\mu_n}{\sigma_\varepsilon}\right)^2\right). \end{aligned}$$

Differentiating again with respect to $\omega$ and simplifying, we have

$$\frac{\partial^2 R_n(q)}{\partial \tau_n \partial \omega} = K \times \frac{h(1-\omega)(1-\tau_n h)}{\left(1-(\omega+(1-\omega)\tau_n h)^2\right)^{\frac{3}{2}}} \times \exp\left(-\frac{1-(\omega+(1-\omega)\tau_n h)^2}{2}\left(\frac{q-\mu_n}{\sigma_\varepsilon}\right)^2\right),$$

where

$$\begin{aligned} K &= -1+3(1-\omega)(1-\tau_n h)-(1-\omega)^2(1-\tau_n h)^2 \\ &\quad + \left(\frac{q-\mu_n}{\sigma_\varepsilon}\right)^2 \left(1-(\omega+(1-\omega)\tau_n h)^2\right)(\omega+h\tau_n-\omega h\tau_n)^2 \\ &= -1+3(1-\omega)(1-\tau_n h)-(1-\omega)^2(1-\tau_n h)^2 + \left(\frac{q-\mu_n}{\sigma_n}\right)^2 (\omega+h\tau_n-\omega h\tau_n)^2. \end{aligned}$$

Notice that $\frac{\partial^2 R_n(q)}{\partial \tau_n \partial \omega}$ has the same sign as $K$. If $\omega = 1$, $K = \left(\frac{q-\mu_n}{\sigma_n}\right)^2 - 1$, which is negative if $q \in (\mu_n, \mu_n+\sigma_n)$. By continuity, for any such $q$ there is $\tilde{\omega} < 1$ such that for all $\omega \in (\tilde{\omega}, 1)$, $K < 0$, and in this case $\frac{\partial^2 R_n(q)}{\partial \tau_n \partial \omega} < 0$. This completes the proof. ■

**Lemma 2.** *Let $F(\cdot)$ and $f(\cdot)$ be the c.d.f. and p.d.f. of the standard normal distribution. Then:*

*(i) $\phi(x) = \frac{xf(x)}{F(x)(1-F(x))}$ is increasing in x;*

*(ii) for $x > 0$, $\frac{\phi(x)}{x^2}$ is decreasing in x;*

*(iii) for $x > 0$, $\frac{d\phi(x)}{dx} < \frac{2\phi(x)}{x}$;*

*(iv) for $x > 0$, $\frac{d\phi(x)}{dx} < 2\sqrt{\phi(x)+1}$;*

*(v) for $0 < x < \min\{1, y\}$,*

$$\frac{\frac{\phi(y)}{x} - \phi'(y)}{\frac{\phi(y)}{y-x} - \phi'(y)} + \exp\left(\frac{(y-x)(y+x)}{2}\right) - 1 > 0,$$

*provided that the denominator is positive;*

*(vi) for $0 < x < y$,*

$$\frac{-\phi'(y)}{\frac{\phi(y)}{y-x} - \phi'(y)} + \exp\left(\frac{(y-x)(y+x)}{2}\right) - 1 < 0,$$

*provided that the denominator is positive.*

**Proof of Lemma 2.** The proofs of these mathematical statements are quite long and cumbersome and are available upon request. The reader can easily verify these claims graphically. ■

**Proof of Proposition 3.** First of all, define

$$H(y) = \begin{cases} h(y) & \text{if } y \geq 0; \\ -h(-y) & \text{if } y < 0; \end{cases}$$

then $H(y)$ is a strictly increasing odd function. It is easy to see that the difference in direct costs of an individual with position $x_i$ to give answer $Y$ as compared to $N$ to the question whether $x_i$ exceeds $q$ equals $B(x_i) = C_i(Y) - C_i(N) = H(q - x_i)$. Indeed, if $x_i < q$ then saying $N$ is costless whereas the cost of saying $Y$ is $h(q - x_i)$; if $x_i > q$ then saying $Y$ is costless while the cost of saying $N$ is $h(x_i - q) = -H(q - x_i)$, so the difference is $\tilde{h}(q - x_i)$ in this case as well.

Let us show that if in an equilibrium individual $i$ with type $x_i$ weakly prefers $d_i = Y$, then any individual $k$ with type $x_k > x_i$ strictly prefers $d_i = Y$. This follows immediately from that the social cost of the individuals does not depend on their type, and the differences in the direct costs equal $H(q - x_i)$ and $H(q - x_k)$, respectively. Since $H(\cdot)$ is strictly increasing, the difference for agent $k$ is smaller, so the decision $d_i = Y$ involves less cost and the resut follows. This implies, in particular, that every equilibrium must take the form of a cutoff $z$, with individuals with type $x_i > z$ choosing $d_i = Y$ in equilibrium, wherease those with type $x_i < z$ choosing $d_i = N$.

Let us now take a closer look at the social costs $S_i(N)$ and $S_i(Y)$ given the cutoff $z$. We have

$$\begin{aligned} S_i(N) &= \int_{-\infty}^{\infty} \mathbb{E}_{-i}(\lambda g(x-y) \mid x < z) \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) dy \\ &= \int_{-\infty}^{\infty} \frac{\int_{-\infty}^{z} \lambda \gamma (x-y)^2 \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx}{F\left(\frac{z-\mu}{\sigma}\right)} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) dy \\ &= \frac{1}{F\left(\frac{z-\mu}{\sigma}\right)} \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) K(x)\, dx, \end{aligned}$$

where the term

$$K(x) = \lambda\gamma \int_{-\infty}^{\infty} (x-y)^2 \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) dy$$

captures the social cost an individual whose type is known to be $x$ from interacting with a random individual $y$. Our assumption that $g(\cdot)$ is quadratic allows us to compute this integral explicitly:

$$\begin{aligned} K(x) &= \lambda\gamma \int_{-\infty}^{\infty} \left((x-\mu)^2 + (y-\mu)^2 - 2(x-\mu)(y-\mu)\right) \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) dy \\ &= \lambda\gamma\left((x-\mu)^2 + \sigma^2\right). \end{aligned}$$

Thus,

$$\begin{aligned} S_i(N) &= \frac{\lambda\gamma}{F\left(\frac{z-\mu}{\sigma}\right)} \int_{-\infty}^{z} \frac{(x-\mu)^2+\sigma^2}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \\ &= \frac{\lambda\gamma}{F\left(\frac{z-\mu}{\sigma}\right)} \left(\int_{-\infty}^{z} \frac{(x-\mu)^2}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx + \sigma^2 F\left(\frac{z-\mu}{\sigma}\right)\right) \\ &= \lambda\gamma\sigma^2 + \frac{\lambda\gamma\sigma^2}{F\left(\frac{z-\mu}{\sigma}\right)} \int_{-\infty}^{\frac{z-\mu}{\sigma}} \frac{t^2}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt \\ &= \lambda\gamma\sigma^2 \left(1 + \frac{F\left(\frac{z-\mu}{\sigma}\right) - \frac{z-\mu}{\sigma} f\left(\frac{z-\mu}{\sigma}\right)}{F\left(\frac{z-\mu}{\sigma}\right)}\right) = \lambda\gamma\sigma^2 \left(2 - \frac{\frac{z-\mu}{\sigma} f\left(\frac{z-\mu}{\sigma}\right)}{F\left(\frac{z-\mu}{\sigma}\right)}\right), \end{aligned}$$

where we used the fact that $\frac{d}{dx}(F(x) - xf(x)) = x^2 f(x)$. We can similarly find

$$S_i(Y) = \gamma\sigma^2 \left(2 + \frac{\frac{z-\mu}{\sigma} f\left(\frac{z-\mu}{\sigma}\right)}{1 - F\left(\frac{z-\mu}{\sigma}\right)}\right),$$

and therefore

$$S(x_i) = S_i(Y) - S_i(N) = \lambda\gamma\sigma^2 \frac{\frac{z-\mu}{\sigma} f\left(\frac{z-\mu}{\sigma}\right)}{F\left(\frac{z-\mu}{\sigma}\right)\left(1 - F\left(\frac{z-\mu}{\sigma}\right)\right)} = \lambda\gamma\sigma^2 \phi\left(\frac{z-\mu}{\sigma}\right).$$

In equilibrium, the individual with type $z$ is indifferent between choosing $Y$ and $N$. For this type,

$$\begin{aligned} U(x_i) &= U_i(Y) - U_i(N) = -(C_i(Y) - C_i(N)) - (S_i(Y) - S_i(N)) \\ &= -H(q-z) - \lambda\gamma\sigma^2\phi\left(\frac{z-\mu}{\sigma}\right) \\ &= H(z-q) - \lambda\gamma\sigma^2\phi\left(\frac{z-\mu}{\sigma}\right). \end{aligned}$$

It is straightforward to check, using the results of Lemma 2 about function $\phi(\cdot)$, that for the examples in footnote 16 this function $U_i(Y) - U_i(N)$ is monotonically increasing in $z$ from $-\infty$ to $\infty$ and therefore has a unique root. It is straightforward to check that if $q = \mu$, then this root is $z = q$. Now, since $U_i(Y) - U_i(N)$ is decreasing in $q$, it must be that for $q > \mu$ we have $z > q$ and for $q < \mu$ we have $z < q$.

In what follows, assume that $q > \mu$. Since $\phi(\cdot)$ is an increasing functions, $U_i(Y) - U_i(N)$ is increasing in $\mu$, and as noted above it is decreasing in $q$. Furthermore, the latter term may be rewritten as $\lambda\gamma(z-\mu)^2 \frac{1}{y^2}\phi(y)$, and by property (ii) of Lemma 2, this term is decreasing in $y$ and therefore increasing in $\sigma$, which implies that $U_i(Y) - U_i(N)$ is decreasing in $\sigma$. Consequently, the equilibrium cutoff $z$ is increasing in $q$ and $\sigma$ and decreasing in $\mu$.

These results imply the following about the equilibrium share of types above $z$, which is equal to $\rho = 1 - F\left(\frac{z-\mu}{\sigma}\right)$. If $q$ increases, then $\rho$ decreases, because $z$ is increasing in $q$. Similarly, if $\mu$ increases, then $\rho$ increases. The comparative statics with respect to $\sigma$ is ambiguous, because $\frac{z-\mu}{\sigma}$ may increase or decrease (since $z$ is increasing in $\sigma$), and one can easily construct examples with with positive and negative effects. ■

**Proof of Proposition 4.** Let $\chi = 1 - F\left(\frac{q-\mu}{\sigma}\right)$ and, as in the previous proof, let $\rho = 1 - F\left(\frac{z-\mu}{\sigma}\right)$; let us prove that $\chi - \rho$ is unambiguously increasing in $\sigma$. Indeed, we have $\chi - \rho =$

$F\left(\frac{z-\mu}{\sigma}\right)-F\left(\frac{q-\mu}{\sigma}\right)$. Then:

$$\begin{aligned}
\frac{d}{d\sigma}(\chi-\rho) &= f\left(\frac{z-\mu}{\sigma}\right)\frac{d}{d\sigma}\left(\frac{z-\mu}{\sigma}\right)-f\left(\frac{q-\mu}{\sigma}\right)\frac{d}{d\sigma}\left(\frac{q-\mu}{\sigma}\right)\\
&= f\left(\frac{z-\mu}{\sigma}\right)\frac{d}{d\sigma}\left(\left(\frac{z-\mu}{\sigma}\right)-\left(\frac{q-\mu}{\sigma}\right)\right)+\left(f\left(\frac{z-\mu}{\sigma}\right)-f\left(\frac{q-\mu}{\sigma}\right)\right)\frac{d}{d\sigma}\left(\frac{q-\mu}{\sigma}\right)\\
&= f\left(\frac{z-\mu}{\sigma}\right)\frac{d}{d\sigma}\left(\frac{z-q}{\sigma}\right)+\frac{q-\mu}{\sigma^2}\left(f\left(\frac{q-\mu}{\sigma}\right)-f\left(\frac{z-\mu}{\sigma}\right)\right)\\
&= f\left(\frac{z-\mu}{\sigma}\right)\left(\frac{d}{d\sigma}\left(\frac{z-q}{\sigma}\right)+\frac{q-\mu}{\sigma^2}\left(\exp\left(\frac{(z-q)(z+q-2\mu)}{2\sigma^2}\right)-1\right)\right).
\end{aligned}$$

To compute the first term, we use that $z$ is defined by $H(z-q)-\lambda\gamma\sigma^2\phi\left(\frac{z-\mu}{\sigma}\right)=0$, and then by the implicit function theorem we get:

$$\begin{aligned}
\frac{d}{d\sigma}\left(\frac{z-q}{\sigma}\right) &= \frac{1}{\sigma}\frac{dz}{d\sigma}-\frac{z-q}{\sigma^2}\\
&= \frac{1}{\sigma}\frac{2\lambda\gamma\sigma\phi\left(\frac{z-\mu}{\sigma}\right)-\lambda\gamma(z-\mu)\phi'\left(\frac{z-\mu}{\sigma}\right)}{H'(z-q)-\lambda\gamma\sigma\phi'\left(\frac{z-\mu}{\sigma}\right)}-\frac{z-q}{\sigma^2}\\
&= \frac{1}{\sigma^2}\frac{2\lambda\gamma\sigma^2\phi\left(\frac{z-\mu}{\sigma}\right)-\lambda\gamma\sigma(z-\mu)\phi'\left(\frac{z-\mu}{\sigma}\right)-(z-q)H'(z-q)+(z-q)\lambda\gamma\sigma\phi'\left(\frac{z-\mu}{\sigma}\right)}{H'(z-q)-\lambda\gamma\sigma\phi'\left(\frac{z-\mu}{\sigma}\right)}\\
&= \frac{1}{\sigma^2}\frac{2\lambda\gamma\sigma^2\phi\left(\frac{z-\mu}{\sigma}\right)-\lambda\gamma\sigma(q-\mu)\phi'\left(\frac{z-\mu}{\sigma}\right)-(z-q)H'(z-q)}{H'(z-q)-\lambda\gamma\sigma\phi'\left(\frac{z-\mu}{\sigma}\right)}\\
&= \frac{1}{\sigma^2}\frac{2\lambda\gamma\sigma^2\phi\left(\frac{z-\mu}{\sigma}\right)-\lambda\gamma\sigma(q-\mu)\phi'\left(\frac{z-\mu}{\sigma}\right)-H(z-q)}{\frac{1}{z-q}H(z-q)-\lambda\gamma\sigma\phi'\left(\frac{z-\mu}{\sigma}\right)}\\
&= \frac{1}{\sigma^2}\frac{\lambda\gamma\sigma^2\phi\left(\frac{z-\mu}{\sigma}\right)-\lambda\gamma\sigma(q-\mu)\phi'\left(\frac{z-\mu}{\sigma}\right)}{\frac{1}{z-q}\lambda\gamma\sigma^2\phi\left(\frac{z-\mu}{\sigma}\right)-\lambda\gamma\sigma\phi'\left(\frac{z-\mu}{\sigma}\right)}\\
&= \frac{1}{\sigma}\frac{\phi\left(\frac{z-\mu}{\sigma}\right)-\frac{q-\mu}{\sigma}\phi'\left(\frac{z-\mu}{\sigma}\right)}{\frac{\sigma}{z-q}\phi\left(\frac{z-\mu}{\sigma}\right)-\phi'\left(\frac{z-\mu}{\sigma}\right)},
\end{aligned}$$

where we used $(z-q)H'(z-q)=H(z-q)$ given the linearity assumption. The problem then reduces to proving that

$$\frac{\phi\left(\frac{z-\mu}{\sigma}\right)-\frac{q-\mu}{\sigma}\phi'\left(\frac{z-\mu}{\sigma}\right)}{\frac{\sigma}{z-q}\phi\left(\frac{z-\mu}{\sigma}\right)-\phi'\left(\frac{z-\mu}{\sigma}\right)}+\frac{q-\mu}{\sigma}\left(\exp\left(\frac{(z-q)\,(z+q-2\mu)}{2\sigma^2}\right)-1\right)>0.$$

Replacing $\frac{q-\mu}{\sigma}=x$ and $\frac{z-\mu}{\sigma}=y$, this is equivalent to

$$\frac{\frac{\phi(y)}{x}-\phi'(y)}{\frac{\phi(y)}{y-x}-\phi'(y)}+\exp\left(\frac{(y-x)\,(y+x)}{2}\right)-1>0,$$

which holds by Lemma 2 part (v). ■

**Proof of Proposition 5.** Denote $s=b^{-1}(c)$; then the measure of single-perpetrator crimes committed in a given period in city $n$ is

$$\begin{aligned} C_1^n &= \kappa_1\int_s^\infty \frac{1}{\sqrt{2\pi}\sigma_n}\exp\left(-\frac{(x-\mu_n)^2}{2\sigma_n^2}\right)dx \\ &= \kappa_1\int_{\frac{s-\mu_n}{\sigma_n}}^\infty \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right)dy. \end{aligned}$$

We will prove the result for $C_1^n$; the proof for multiple-perpetrator crimes $C_2^n$ is analogous.

We notice that $\tau_n$ affects $C_1^n$ through $\sigma_n$ only. Differentiating with respect to $\tau_n$, we have

$$\frac{\partial C_1}{\partial\tau_n}=\kappa_1\frac{1}{\sqrt{2\pi}}\frac{s-\mu_n}{\sigma_n^2}\exp\left(-\frac{(s-\mu_n)^2}{2\sigma_n^2}\right)\frac{\partial\sigma_n}{\partial\tau_n},$$

which is positive whenever $s>\mu_n$. We also have

$$\frac{\partial C_1}{\partial\mu_n}=\kappa_1\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(s-\mu_n)^2}{2\sigma_n^2}\right),$$

which is also positive. Lastly,

$$\frac{\partial^2 C_1}{\partial\mu_n\partial\tau_n}=\kappa_1\frac{1}{\sqrt{2\pi}}\frac{(s-\mu_n)^2-\sigma_n^2}{\sigma_n^4}\exp\left(-\frac{(s-\mu_n)^2}{2\sigma_n^2}\right)\frac{\partial\sigma_n}{\partial\tau_n},$$

which is positive, provided that $s > \mu_n + \sigma_n$. This completes the proof. ■