IFN Working Paper No. 1350, 2020

Misogynistic and Xenophobic Hate Language Online: A Matter of Anonymity

Emma von Essen and Joakim Jansson

Research Institute of Industrial Economics
P.O. Box 55665
SE-102 15 Stockholm, Sweden
info@ifn.se
www.ifn.se

Misogynistic and xenophobic hate language online: a matter of anonymity*

Emma von Essen†, Joakim Jansson§¶

Abstract

In this paper, we quantify hateful content in online civic discussions of politics and estimate the causal link between hateful content and writer anonymity. To measure hate, we first develop a supervised machine-learning model that predicts hate against foreign residents and hate against women on a dominant Swedish Internet discussion forum. We find that an exogenous decrease in writer anonymity leads to less hate against foreign residents but an increase in hate against women. We conjecture that the mechanisms behind the changes comprise a combination of users decreasing the amount of their hateful writing and a substitution of hate against foreign residents for hate against women. The discussion of the results highlights the role of social repercussions in discouraging antisocial and criminal activities.

Keywords: online hate, anonymity, discussion forum, machine learning, big data
JEL: C55, D00, D80, D90

*We want to thank our research assistant for excellent work. Furthermore, we thank Lina Eklund, Björn Tyrefors Hinnerich, Jonas Vlachos, David Strömberg, Lena Hensvik, Matthew Gentzkow, David Yang, Dan-Olof Rooth, Björn Öckert, Abdulaziz Abrar Rashid, seminar participants at the Department of Economics at Stockholm University, the Research Institute of Industrial Economics, the Swedish Institute for Social Research, Stockholm University, the Dep. of Political Science, Uppsala University, participants at AASLE 2019 and the Association of Internet Researchers 2018 for fruitful comments. Financial support from the Swedish Research Council is also gratefully acknowledged.

†Swedish Institute for Social Research, Stockholm University, Stockholm, Sweden
‡Department of Economics and Business, Aarhus University, Denmark
§Dep. of Economics and Statistics, Linnaeus University, Sweden
¶Research Institute for Industrial Economics, Stockholm, Sweden.

I. Introduction

As online public discussions have become a larger part of our political lives, hate, harassment, and threats have become a growing democratic concern (Cheng et al. 2017). A vast majority of people in the U.S. report that they believe online harassment is a problem, and four in ten state that they have experienced harassment (Duggan 2014). In extreme cases, individuals have posted information with strong misogynistic and xenophobic hate language before performing acts of violence, inspiring others to write more hateful comments. Two well-known examples are Anders Behring Breivik in Norway 2011 and the recent Christchurch mosque shootings in New Zealand 2019. The problem is not these extreme cases but the massive amount of hateful language that can crowd out information (Glaeser 2005). Hate online may thus distort an individual's political and economic decisions, such as whether to vote or participate in political party work. Much online hate is written by individuals hiding behind anonymous usernames, i.e., showing no identifying information. The anonymity of the writer implies that the individual cannot be held accountable for what he or she writes—this is an issue when dealing with hate online. However, anonymity can enable freedom of speech by protecting writers from social and governmental repercussions (Froomkin 2017). Understanding how anonymity affects hate online is thus vital to finding efficient policies for online discussions, e.g., of the regulation and allocation of resources in a democratic society. This paper investigates how anonymity affects hate against females and foreign residents in online discussions of political topics, exploring the effect on the general share of hate faced by readers as well as the effect on those who create hateful content.

To this end, we combine a machine-learning prediction model of hate with a standard difference-in-difference (DD) strategy (Mullainathan and Spiess 2017). To predict hate, we scraped text from a large Swedish anonymous discussion forum called Flashback, similar to the U.S.-based Reddit. Flashback is one of the most visited discussion sites in Sweden, with more than one million registered accounts. Anonymity is a requirement of the forum, and their motto is "True freedom of expression". The discussions at Flashback are arranged in subforums that range from politics, sexual preferences and drug abuse to electronics and family relationships. Flashback categorizes each subforum into discussion threads, and within each thread, a member can contribute by adding a post. The threads comprise posts (or entries) written by users.[1] Our text comes from three subforums containing the largest political discussions: domestic politics, immigra-

[1] As we only observe a user when that user writes something, we have an unbalanced panel.

tion and feminism.[2] These subforums contain controversial topics more often associated with hateful speech than those of other subforums (Cheng et al. 2017).

Posts from a random subset of the threads were manually classified on the basis of whether each post contained hate and of the group or individual towards whom the hate was directed. With the classified posts, we developed a machine-learning model, a logistic lasso, to predict hateful content, particularly misogyny and xenophobia, between 2012 and 2016. The predicted values were then used as the outcomes of DD models. In September 2014, the identities of most of the accounts registered before March 2007 were unexpectedly obtained by journalists, and the journalists publicly exposed the legal identities of a handful of Flashback users along with their hateful writings. Users registered before March 2007 become the treated group, since after September 2014, they ran a risk of having their writings and identities exposed, while users registered after March 2007 become the control group. The event and the news of the exposed users were discussed in traditional media as well as on the forum. Even if the actual risk of being exposed was low, the costs were high. The costs were manifested by the exposed individual haters losing their jobs, friends and family. A decrease in anonymity thus brings the threat of facing social repercussions.

Our empirical estimates show that decreased anonymity—an increased risk of being exposed—leads to a lower share of hate in general. The models of hate towards specific groups show that the share of xenophobic content decreases, while the share of misogyny increases. The decrease in the share of xenophobic hate stems from individuals who had a high share of xenophobic language before the event, who both substituted xenophobic hate for misogynistic hate and decreased their general activity (writing) on the forum.

A growing amount of literature in economics focuses on the consequences for economic decisions of the spread of political information in social and traditional media (Engelberg and Gao 2011, Acemoglu, Hassan, and Tahoun 2017, Qin, Strömberg, and Wu 2017). For example, spreading propaganda through social and other media can trigger acts of violence (Bhuller et al. 2013, Yanagizawa-Drott 2014, Adena et al. 2015, Chan, Ghose, and Seamans 2016, Bursztyn et al. 2019), but it can also increase coordination when mobilizing political protest (Enikolopov, Makarin, and Petrova 2019, Zhuravskaya, Petrova, and Enikolopov 2019) and curbing corruption (Enikolopov, Petrova, and Sonin

[2]Examples of threads in the domestic politics forum include "Keywords for a political alliance", in which participants discuss how all Swedish politicians have similar attitudes towards gender quotas and immigration. In another thread entitled "The number of sick days has dropped more than 50%", discussants focus on paid sick leave and people being forced to work while ill. In the Feminism subforum, the threads include discussions of how feminists affect Swedish politics, for example, "Do feminists want gender equality?".

2018). A contemporaneous debate focuses on how false information on social media affects political decisions. Allcott and Gentzkow (2017) find that false news stories in the U.S. 2016 presidential election favored the Republican presidential candidate, possibly leading to political decisions being made based on false information. False or obscuring information or hostile comments online often bear no author name.

Psychological research suggests that anonymity lowers the perception of the possible social repercussions of breaking a norm or committing a crime (Postmes, Spears, and Lea 1998, Suler 2004). On a platform discussing job opportunities under anonymous usernames, Wu (2018) finds gender biases in the evaluation of individuals' professional careers. The results suggest that social media is not a separate universe—it is interlinked with offline social structures, such as sexism and racism. Anonymity online correlates positively with cyberbullying, cyberhate and aggressive speech; e.g., Suler (2004), Moore et al. (2012), and Van Royen et al. (2017). A study in system sciences, Cho, Kim, and Acquisti (2012), finds a lower number of swear words and slanderous comments when anonymous discussion forums in Korea forced users to use real names. Our study investigates anonymity through a different mechanism than a law. We explore decreased anonymity in terms of the increased risk of being publicly exposed and socially punished.

The economic literature on anonymity and information is scarce and diverse; theoretical and empirical research suggests that less privacy can lead to benefits or losses for welfare depending on the context (Acquisti, Taylor, and Wagman 2016).[3] Hansen, McMahon, and Prat (2018) use a machine-learning approach to show that transparency affects the expression of opinions in policy deliberations at the U.S. Federal Reserve through both increased discipline in discussions and higher conformity in the discussions. The net effect suggests that increased transparency creates a more informative monetary policy debate. A recent working paper indicates that in regard to transparency and expressing political opinions, the wedge between expressing an opinion in public vs in private (anonymous) depends on the social norms related to that specific opinion. If xenophobic hate is socially accepted, there will be little difference in xenophobic expressions between public and private discussions (Bursztyn, Egorov, and Fiorin 2017).

Early theoretical papers on career concerns claimed that greater transparency and more information about the agent improved accountability and were never detrimental to the principal (Holmström 1979). Later research, however, showed that revealing more information about the agent can be harmful to the principal (Holmström 1999,

[3]Transparency can include elements of both anonymity and privacy. Privacy implies concealing or revealing an individual's personal information and actions, such as his or her medical records. In contrast, anonymity implies concealing the identity of the individual, such as the names on the medical records.

Prat 2005). Ali and Bénabou (2016) explicitly models transparency in a principal-agent model of public-good provision. Transparency affects the aggregate provision of public good through agents' concern for their reputations, i.e., social images. If xenophobic and misogynistic hate online is seen as a public bad (i.e., pollution), the model predicts a decrease in hateful content given a negative shock to transparency, since individuals care about their reputations and fear social consequences.[4]

The rest of the paper is structured as follows: Section II describes some background and the data collection process, while Section V concerns the empirical strategy. Section III describes the data and the predictions obtained from the machine learning model, Section VI discusses our main findings, and Section VII describes potential mechanisms. In the final section, we conclude the paper.

II. Background

Flashback, the discussion forum we study, is today one of the most visited Internet pages in Sweden and is well known among users of other similar international forums and message boards, such as Reddit.[5] According to a recent survey by Davidsson, Palm, and Melin Mandre (2018), 33% of the Swedish population state that they use Flashback, and the share is larger among men than women (40% vs 26%). Flashback had more than one million registered accounts in 2018. According to Alexa, the average Flashback user spends approximately seven minutes per visit and, on average, goes to seven pages per visit.

The posts in a discussion are visually displayed in chronological order by time of entry. A post is displayed together with the time of entry, username, number of current posts by that user, user registration date and sometimes a self-selected picture. When clicking on a thread, the first twelve posts are automatically shown on the screen. The user can click once to go to the end of the discussion thread or follow the discussion by reading all the posts.

A post can automatically include quotations from other posts in the same thread. Discussions take place within a thread and never across threads. All posts are saved and publicly visible. A user can never delete a posted message, even by deleting his or her account. Each subforum has users with moderator status supervising the discussions using internal rules (netiquette). If a user breaks any of the rules, a moderator can give

[4]Section B in the Appendix provides a brief outline of the model as a theoretical framework using the context of our paper.

[5]Alexa ranks it as the 23rd most visited in Sweden, and 5214 in the world; https://www.alexa.com/siteinfo/flashback.org, access 2017-01-01.

a warning and temporarily or permanently exclude the user from the forum. Moderators can also lock discussion threads for further posts. However, the general rule at Flashback is that anything should be allowed to be expressed. Almost all warnings, temporary or permanent bans and locked threads are made due to what is labeled "off-topic" content—when a discussion deviates from the original topic. It is not primarily controversial topics or users that are subject to moderation but rather when someone disturbs the general discussion by writing something irrelevant. It is uncommon for a user to either receive a warning or be subject to exclusion due to, for instance, outright racism. A simple search at the forum reveals this quite clearly. In total, there have been 36 instances in which users have been banned for writing hateful posts against minorities, while the number of those banned for advertisement is 623. The number of users banned for misusing multiple accounts is 421, and for infringing on copyright, the number is 280. Furthermore, there are no indications of moderators changing their behavior during the period we studied. Flashback did not, for example, change their internal rules for the moderators.

There is no alternative to Flashback with a similarly large variety of topics across and within subforums in Swedish. If users wish to leave Flashback and still discuss similar topics online, they have to migrate to more topic-specific message boards. Flashback is also (in)famous for its focus on the anonymity of the end-users as a way of promoting freedom of speech. Flashback has been associated with publishing all sorts of comments, including hate speech, and has become an important arena for testing opinions and collecting information. According to Swedish law, hate speech is prohibited. The term is defined as publicly making statements that threaten or express disrespect for a group regarding its race, skin color, ethnic origin, faith or sexual orientation. However, it is not forbidden by law to write in a hateful way based on a person's gender. During the 1990s, several members of extreme right-wing movements were convicted of racial hate speech (Lööw and Nilsson 2001).[6]

II.A. Event: Treatment and Control

On September 10, 2014, Swedish and international media unexpectedly revealed that at least one-third[7] of the accounts registered at Flashback before March 2007, was obtained

[6] In 1998, the law increased the responsibility of the publishers of online message boards to remove hate speech content. However, the publisher of Flashback is registered in the U.S. and is not affected by that part of the law.

[7] According to the journalists, they had information on all the accounts registered before March 2007. However, the owner of Flashback claimed that the journalists only had one-third of those registered before March 2007.

by a group of journalists called Researchgruppen. To open an account at Flashback, a user needs an email address and a username. The information that the group of journalists had was a list of email addresses that Flashback users had used to open accounts. To find the identities of the individuals, the email addresses from Flashback were matched with email addresses used at other internet sites, for example, CDON, from which other personal data, such as their Swedish social security number, were added. The group of journalists publicly exposed, in traditional Swedish media, the identity of four individuals and their hateful writings on Flashback. Two politicians who were exposed lost their current jobs. The same group of journalists had previously publicly exposed users who posted xenophobic content on other Swedish discussion sites.

The event received a great deal of publicity in national and international traditional media as well as social media.[8] A file containing the email addresses that Researchgruppen had access to was posted publicly online. The owner of Flashback could see that the email addresses were sorted based on the usernames, implying that Researchgruppen had access to this information as well. The owner of Flashback then declared in a thread that those who had registered before March 2007 were the ones at risk. In the same thread, several users expressed their disappointment and outrage that this had happened and said that they might leave Flashback.

The event is shown in Figure Ia, which displays data from Google Trends for the weekly relative search frequencies of the terms Flashback and Researchgruppen between 2013 and 2017. There are two visible spikes in the graph: the first corresponds to the week during which the initial media revelation occurred, and the second spike corresponds to the second week of February 2015, when the specific identities were exposed. Panel Ib displays a similar graph using the monthly number of news articles in Sweden containing the words Flashback and Researchgruppen. Again, the same two spikes are clearly visible, in September 2014 and February 2015. In this paper, we consider the first spike as the start of the treatment period. From September 2014 onwards, the treatment group (users registered before 2007) ran an increased risk of having their identities and the content they had written on Flashback publicly exposed. Users registered after March 2007 did not run the same risk. The differences between the control and treatment groups in terms of how writers reacted to the event will be interpreted as the impact of the change in anonymity. If users in the control group also believed that their risk of being exposed increased after the event and changed the way they wrote on Flashback, then there are spill-over effects. This, however, only reduces the

[8] Chen, Adrian, 2014, The Troll Hunters, MIT Technological Review, https://www.technologyreview.com/s/533426/the-troll-hunters./

effect that we wish to capture. If the event triggered migration to other forums from the treatment group, then we have selection out of the treatment group. We are able to capture whether a user stopped writing at Flashback after the event but not whether the user chose to write on another forum. First, there are no equivalent alternative forums in Swedish that offer users anonymity to the same extent as Flashback. Second, previous research suggests that migration across subforums does not cause changes in hateful content (Chandrasekharan et al. 2017).

III. Data

The data used in the study come from text-based messages (called posts) written on the discussion forum Flashback. Using a custom-built script in Python, we scraped all posts in three forums—feminism, domestic politics and immigration—from the time each respective forum started until January 2017. In the first step, a research assistant (hence forth RA) manually classified posts from a random subset of the threads. We then used the manually coded data to find machine learning models that predicted hateful posts. Text discussions are produced on a massive scale every day online, and today, many platforms employ methods that automatically detect hate speech and offensive language (Davidson et al. 2017). Most automatic hate speech detectors are applied to English-language text, however. Using a machine learning approach, we derived three simple prediction models for automatically detecting hateful content in Swedish: general hate, xenophobic hate and misogynistic hate. In the second step, we employed these models on the full data set, giving us a universe of posts across continuous time until the end of 2016. In terms of individual users, we thus had an unbalanced panel, as we did not observe every user at every minute. This procedure vastly increased our sample size, allowing us to carry out a better analysis of our research question.[9]

III.A. Prediction Models Using the Classified Data

In the first step, we randomly selected 100 threads in each forum, and then an RA classified the first twelve and last five posts in each thread. The randomization was implemented at the thread level because we wanted to classify whether the initial hateful content was followed by more or less hateful posts and whether a debate occurred criticizing previous posts. The RA received instructions from us with definitions of the

[9] In the online appendix Section D.C., we produce similar results to our main estimates using only a subset of the data.

main classifications of content types—hateful content, threatening content, and aggressive content—and the group towards whom hateful content was directed—females and feminists, foreign residents, and others. The final random subset contained 4040 classified posts divided equally across the three forums. Through a process called stemming, some of the ending characters were removed, and we also deleted all stop words and numbers. The online appendix describes the classification of hateful content and the subset data in more detail.

In line with methodological practice (James et al. 2013), we randomly split the classified data set into a training set of 2812 posts (observations)—approximately 70 percent—and a test set of 1206 posts. To this training data we then applied a logistic lasso[10], which is a machine learning algorithm equivalent to the standard log-likelihood function for logistic regression with an added penalty term. In essence, it chooses the words that are the best predictors of hate by balancing the bias-variance trade-off. For further details on the prediction process, the logistic lasso and the weighting scheme, please see the online appendix Section D.A.. For further details on the evaluation of the quality of the model predictions, for which we used the test set, please see Section D.B. in the online appendix.

As an example, Table I displays all words and their associated coefficients from the logistic lasso prediction model of hate directed at anyone. The first word is *arab*, and this is the same in Swedish as in English. The second word, *blatt*, comes from the Swedish racial slur word blatte, which is a derogatory word for someone with a dark skin tone. *Dumm* most likely comes from different versions of dumb in Swedish. *Hor* probably comes from hora, which translates to whore, *lill* comes from lilla or lille, which are typically used to belittle someone. *Miljon* means a million and might refer to the cost of a political process, such as immigration, or to the Million Programme, a Swedish public housing project from the '60s and '70s. The word *muslimsk* means Muslim, *parasit* means parasite, *patetisk* translates to pathetic and *rån* means robbery. Lastly, *what* most likely refers to the English word. Overall, the words selected by the logistic lasso seem to conform with words connected to groups that are often targeted by cyberhate and offensive language offline: women and foreign residents (Citron 2014). The coefficients are all positive, which may indicate that the most common mode of discussion is without any hateful content. The levels of the coefficients are not particularly useful to discuss, since the logistic lasso produces biased estimates. We identified two additional models: one predicting hate against foreign residents and one predicting hate against women.

[10]We also ran a support vector machine model on the coded data. The lasso made better predictions, with fewer incorrect and more correct classifications; see the online appendix.

There is some overlap between the words chosen by the logistic lasso for the respective models of hate against anyone, hate against foreign residents and misogyny. The words arab, muslimsk and blatt are also found in models predicting hateful content against foreign residents, whereas hor is the only overlapping word for the model predicting misogyny. For the full set of words selected by the algorithm for hate against foreign residents and misogyny, see Tables A.6 and A.7 in the Appendix. [11]

We used single words as the primary features for classification, which can lead to misclassification since words can have different meanings in different contexts. To include some degree of context, we tried using bigrams, or word pairs, occurring in a sequence. However, N-grams typically have issues related to the distance between relevant words (Chen et al. 2012). Thus, we used pairs of words, rather than single words, weighted by their term frequncy–inverse document frequency as inputs for the logistic lasso. However, this did not improve the classifier's prediction performance, and we thus used the single-word approach.

When evaluating the performance of the predictions of the logistic lasso, we focus on maximizing the sum of the true positive rate (sensitivity) and the true negative rate (specificity).[12] Intuitively, this is a trade-off between type I and II errors, where we strive to minimize the sum of the two. The attenuation bias of our estimates of the effect of anonymity on hate decreases as the sum of type I and II errors decreases, implying that we are in effect minimizing the attenuation bias of our treatment effect. Comparing the result of our prediction model with the actual manual coding results in the test set of the data, we obtain estimates of the degrees of type I and II errors. More specifically, equation 1 shows the relationship between the estimated treatment effect ($\bar{\beta}$) and the true treatment effect without the attenuation bias ($\tilde{\beta}$):

$$\tilde{\beta} = \frac{\bar{\beta}}{(1 - p_{01} - p_{10})}, \tag{1}$$

where p_{01} is the false positive rate and p_{10} is the false negative rate. We use these estimates of the error rates to take the attenuation bias into account in section IV.B. and in online appendix. We obtain remarkably similar results between the random subset of the data and the full sample with our basic DD specifications.

Hate against anyone has a true positive rate of approximately 0.214, while the true

[11] We could not find machine learning models with reasonable precisions for aggression and threat due to a lack of observations.

[12] Since all our outcomes are heavily skewed towards zero, focusing on maximizing accuracy will not yield fruitful predictions, as the best accuracy will typically be attained by predicting all posts to be non-hateful.

negative rate is 0.957, implying that our prediction still makes less than 5 percent of type I errors for this model. For hate against women and feminists, the algorithm allows for a higher false positive rate, thus giving us a true negative rate of 0.855 and a true positive rate of 0.644. Finally, hate against foreign residents displays a true positive rate of 0.465 and a true negative rate of 0.902 (the derivations and full set of results can be found in the online appendix). The model using the classified data seems to predict well; hate against anyone is our noisiest measure, and misogyny is the most precise. In sum, this suggests that it is possible to find a prediction model for hateful messages in Swedish and that the precision of the prediction improves for hate towards a particular group rather than general hate.

III.B. The Prediction-Based Full Data Set

In the second step, we use the three prediction models to detect hateful posts in the full data set; i.e., the data include non-classified posts. Here, we restrict ourselves to the period from January 1, 2012, until December 31, 2016 (when the data from all threads end). Starting in 2012 balances the time before the event, two years and eight months, and after the event, two years and four months. Since the control group decreases as we move further back in time, as it comprises all users registered after March 2007, we do not use any data before January 1, 2012. The data comprise an unbalanced panel of posts, and the coefficient of the share of hate reflects the probability that a post is hateful conditional on the fact that the post is written. We cannot create a balanced panel, i.e., create an observation for each user in our data set at every point in time, since this would require us to collapse the data to a specific time level, and we observe posts being made by users continuously. Additionally, many users of Flashback only read the content and never contribute to the written discussion, and the share of hateful posts reflects the hate they face when reading the discussions online. In other words, our data also represent the information users consume through Flashback.

Table II shows the summary statistics for the full data using the three prediction models. In the full data, there are 1,984,224 posts written by 48,672 users spread out among 29,425 threads. Eight percent are predicted to be hateful against anyone, 14 percent are predicted to be misogynistic, and 16 percent of all posts are predicted to have hateful content against foreign residents. In the manually classified data (shown in Table D.12 in the online appendix), hate against foreign residents and misogyny are, by construction, parts of the share of general hate. This is not true for the full data since we use separate logistic lasso models for each type of hate. The lower share of hate against

anyone thus reflects our relatively poor prediction model for this particular outcome.

IV. Descriptive Results

The prediction models highlight that when entering Flashback, in the subforum where political topics are discussed, 8-16 % of the posts contain hate. Sixteen out of 100 posts contained hateful content against foreign residents. An average user looks at approximately 7 pages per day according to Alexa, and each page displays 12 posts, implying that the typical user sees 84 posts per day and approximately 13.5 contain hate against foreign residents.

IV.A. Who Are the Producers of Hate?

To understand the data, we first want to explore who produces hateful content. In this section, we focus on hate against anyone in order to capture the distribution of all types of hate. We first look at the distribution of the number of posts across percentiles of users over the relevant period 2012-2016. The distribution of activity—see Figure IIa—suggests that we have three types of users: i) users who write only one post during the period (approximately 25 %), ii) users who write between 4 and 688 posts (approximately 74 %), and iii) users who write many posts (approximately 0.1 %). The number of hateful posts per user is also skewed. Figure IIb shows that most users write zero hateful posts, the top 5 percent write approximately 10 hateful posts and the top 0.1 percent (approximately 400 users) write approximately 271 hateful posts in the relevant period. In Figure IIc, we see a similar distribution of the share of hateful posts—again, the bottom 50th percentile have a share of hateful entries that is zero, while for the 75th percentile through the 99th percentile, a fraction of 0.08 to 0.5 of their posts have hateful content. The very top includes users who only write hateful posts, but these are users producing very few posts in total.

A question in the literature is whether users who produce hate also are the most frequent writers or whether anyone can write hateful content. Cheng et al. (2017) argue that anyone can become a hater in an environment with offensive language, suggesting that there is little to no relationship between these properties. Our data confirm a weak relationship between the number of entries and the share of hateful entries a user writes. Table A.4 in the Appendix displays the raw correlations, and Figure A.6 in the Appendix shows a binned scatter plot, where we see that haters are found across the full distribution of frequencies of writing.

Producing hate does not seem to depend on being more or less active as a user. In this respect, all users produce hate. The argument that anyone can become a hater relies on the idea that hate begets hate—a hateful post from one individual triggers hate from the next (Cheng et al. 2017). Our data suggest a similar pattern. If the first entry in a thread contains hate, then the probability that the following posts are hateful increases by 3 percentage points (shown in column 8 in Table A.4 in the appendix). To further explore this idea, we use the data from the RA; we asked the RA to indicate which post responded to which. Writing a hateful post leads to a 20 percent increase in getting a hateful reply. This pattern holds true for all combinations of hate except for hate against foreign residents and misogyny. Writing a misogynistic post brings a 6.6 percentage point lower probability that a response is hateful against foreign residents, while there is no relationship found between an initial hateful post against foreign residents and a reply that is misogynistic. See the figures in Table A.3 in the appendix.

Even if hate begets hate, it can stem from an animus against one particular group or against all groups. We thus look at the overlap between the models. In general, we find that the prediction model for hate in general has more overlap with hate against foreign residents. Approximately one out of four posts that contain xenophobia is also classified as hate against anyone, while the number is one in twenty for misogyny. A discussion post can contain hateful content in general or hate directed towards foreign residents, females or both. Hate against foreign residents and misogyny seem to mainly exist in separate posts, as only 1 percent of the posts are classified as both. The individuals producing hateful content could, however, still be the same. Looking at the individual users' shares of hate, we find a weak negative relationship between general hate and misogyny, as we would expect, and a large positive relationship between the individual share of hate and hate against foreign residents. Thus, it seems that different users write xenophobic hate and misogynistic hate. Additionally, xenophobic hate and misogynistic hate do not overlap; if anything, there is a negative correlation between them. Table A.4 in the appendix shows the full set of correlations. Splitting the outcomes by subforum reveals that misogyny and xenophobia primarily exist on different subforums. Figure IIIa shows that most predicted hate against anyone seems to occur in the immigration forum, where every tenth post contains hateful content, and the least predicted hate seems to occur in the feminist and domestic policy forums. Figures IIIc and IIIb show that misogyny is most prominent in the feminist forum, and hate against foreign residents is most prominent in the immigration forum.

Hate seems to be produced by all types of users, both more and less active users. We do not find evidence of a set of users that browse subforums and produce hate against

all types of groups. Hate seems to trigger more hate in the heat of the discussion, and misogyny and xenophobia seem to be produced by different haters present in different forums.

IV.B. Are Anonymity and Hate Related?

Both hateful content and non-hateful content are produced under the assumption that the user is anonymous at Flashback. However, as we discussed briefly in the introduction and in more detail in section II, the event decreased the sense of being anonymous for the treatment group, i.e. the users registered before march 2007. Table III displays the summary statistics in the pre- and post-event periods for both the treatment and the control group. There are fewer users in the period after the event compared to before, and there are fewer threads and posts. The percent changes in the number of users and number of entries before and after the reform are larger in the treatment group than in the control group. The simple differences in the share of hate between the two groups before and after give us $(0.07 - 0.09) - (0.08 - 0.08) = -0.02$ for hate against anyone, $(0.12 - 0.11) - (0.13 - 0.15) = 0.03$ for misogyny and $(0.15 - 0.18) - (0.16 - 0.17) = -0.02$ for hate against foreign residents. This suggests that hate against foreign residents and hate against anyone decreased, while misogyny increased.

Adjusting for the false classifications, using equation 1 and estimates from the online appendix, we obtain an estimated effect of $-0.02/(1 - 0.786 - 0.043) \approx -0.12$ for hate against anyone, which is the same number we obtained from the manually classified RA data. Hate against foreign residents provides an estimate of $-0.02/(1 - 0.535 - 0.098) \approx -0.06$, which is also similar to the manual classification estimate of -0.06. Finally, for hate against women and feminists, our adjusted estimate becomes 0.03/(1-0.356-0.145)=0.06, which has the opposite sign and is much larger than the estimate of -0.02 from the manually classified data.

V. Empirical Strategy: The Effect of Anonymity on Hate

To investigate whether a decrease in anonymity affects hate in discussions online, we use a DD strategy with the predictions from the three machine learning models as outcomes. Users registered before March 2007 are thus treated, and those registered after form the control group. September 10, 2014, is the date that begins the post-period in our DD

setting. The empirical strategy is formally summarized in equation 2.

$$Y_{ptg} = \alpha + \beta Treated_g * Post_t + \theta Post_t + \gamma Treated_g + \varepsilon_{ptg} \tag{2}$$

Y_{ptg} is the outcome variable, which is a dummy for whether a post p contains hateful content, hateful content against foreign residents or misogyny at time t and whether the writer belongs to group $g \in \{treated, control\}$.[13] $Treated_g$ is a dummy variable taking the value 1 if the post was written by an individual belonging to the treated group of early registered users, $Post_t$ is a dummy taking the value 1 if the post was written after the event and ε_{ptg} is the error term. β thus measures the treatment effect of the change in anonymity, i.e., the increased probability of having one's identity and hateful writings exposed publicly.

The underlying identifying assumption is thus that the treated and control groups would have had similar trends in the absence of treatment. In our setting, however, both assignment to treatment and the pre- and post-periods are functions of time. In addition to the standard event graphs (Angrist and Pischke 2008), we also conduct several robustness checks regarding the registration date, such as dropping all users who registered after the event and controlling for a linear trend in the user start date. We also make sure to not use data from too far back in time in order to separate the treatment and control assignment from the assignment to the pre- and post-periods; see the discussion in Section III.B.. Our identification does not give us the general equilibrium effect since we only estimate the treatment effect for the treated part of the population. However, we are interested in how individuals respond to decreased anonymity and not only in how the entire forum is affected.

As usual, the estimation of standard errors is potentially problematic for us, as we have a DD setting with only one treatment group and one control group (Bertrand, Duflo, and Mullainathan 2004, Donald and Lang 2007, Conley and Taber 2011). Since our treatment only changes once at the control-treatment group level, we estimate the standard errors using a two-step approach according to Pettersson-Lidbom and Thoursie (2013). We first aggregate the data at the group level by week using equation 3.

$$\bar{Y}_{tg} = \alpha + \beta Treated_g * Post_t + \theta Post_t + \gamma Treated_g + \varepsilon_{tg}, \tag{3}$$

where $\bar{Y}_{tg} = \sum_{p=1}^{N} Y_{ptg}/N_g$. We note that we can rewrite equation 3 as the difference

[13] We use continuous time, down to the minute.

between the two groups, $g = control, treated$:

$$\bar{Y}_{t1} - \bar{Y}_{t0} = \Delta Y_t = \pi + \beta Post_t + \Delta\varepsilon_t. \tag{4}$$

Estimating equation 4 with the Newey-West estimator gives us standard errors adjusted for both correlations within the treatment and control groups and serial correlations, thus giving us the correct standard errors. In practice, we use the Newey-West estimator with 4 lags, which is equivalent to a month of autocorrelation.[14]

This approach does however not lend itself very well to performing additional robustness checks with the inclusion of control variables. In our baseline specification, we thus cluster the standard errors at the user level, which tends to, if anything, overestimate the magnitude of the standard errors. [15]

VI. Difference-in-Difference Results

Figure IV shows the differences in the shares of hateful posts between the treatment and the control group, along with 95 percent confidence intervals, per quarter before and after the event (Angrist and Pischke 2008). Panel IVa provides the estimates for hate against anyone, panel IVb the share of hate against foreign residents and panel IVc the share of misogyny. None of the graphs appear to display any real trend over the entire pretreatment period in terms of the coefficient estimates, which implies that our identification strategy seems credible.

Focusing on panel IVa, hate against anyone, it is difficult to observe any clear causal effect. In contrast, panel IVb—hate against foreign residents—shows a clear pattern. The pre-period estimates are close to zero or slightly positive, with no trend over the entire period, while in the post-period, we see a sharp drop to a negative coefficient in the first post-reveal quarter, and the estimates then remain on this level for the rest of our data with the exception of the second quarter after the initial reveal.[16] In line

[14]We have experimented with more lags for the Newey-West estimator without seeing any major changes in the magnitudes of the standard errors.

[15]We randomly draw threads and not posts for the logistic lasso prediction model and would thus like to cluster the standard errors at the thread level to take this into account. However, in our main specifications, we utilize the predicted values from the full data, thus making the need for clustering at the thread level less obvious. We do, however, in Table A.2, provide the results of the baseline DD model, where we cluster the standard errors at both the individual and thread levels, with no discernible impact on the standard errors.

[16]One possible explanation for the deviation is the attack on the French publication Charlie Hebdo, which took place during the beginning of the second quarter. Note that the majority of exposed users had not yet been publicly exposed when this quarter started, and thus some of the more hateful users might have taken the risk of returning to write hatefully regarding this extraordinary event.

with our descriptive statistics in Section III.B., we see the reverse effect of anonymity on misogyny—a positive effect in the post-period, although the pattern is slightly noisy. Although different prediction models are used, hate against anyone is largely a combination of hate against foreign residents and misogyny, and since the effect goes in opposite directions for these two underlying measures, it is not surprising that we do not see a clear effect for hate against anyone.

In Tables IV–VI, we present the DD estimations using equation 2. Column 1 of each table depicts the basic model with standard errors clustered at the user level. Table IV shows the effect of the shock: the decrease in hateful posts against anyone is 1.5 percentage points lower in the treatment group than the control group, with a baseline level of 8 percent. Column 1 in Table V indicates that the effect on hate against foreign residents is significant at 2.9 percentage points, while the estimate in Table VI shows that the threat of exposure increases the amount of misogyny by 3.2 percentage points. In the control group, there was no significant change in the amount of hate against anyone, nor was there a difference in hate between the treatment and control groups in the pretreatment period, as shown by the coefficients in row two and three. However, for hate against foreign residents, we see a minor significant change in the post-period for the control group. One interpretation for this is that it is due to spill-over effects, i.e., that the users in the control group are more cautious in their writing due to the event. The coefficient is much smaller than our estimated treatment effect, and if anything, it would bias our estimate downwards. For misogyny, we see a significant difference between the treated and the control group in the pre-period as well as a decrease for the control group in the share of misogyny in the post-period. The difference between the groups in the pre-period could make our estimates functional form dependent. However, since we are using a binary outcome variable, we only have one functional form that we can choose, and thus this should be no major concern. Our estimate of the change in misogyny is however in part driven by a decrease in the control group. Our interpretation of this is that the DD design filters out any general trend in hateful discussions, in particular misogyny in this case.

For the average reader of Flashback, the average share of hate against foreign residents implies that 16 out of 100 posts contain hateful content—the effect of the event implies that individuals in the treatment group decreased their hate against foreign residents to approximately 13 out of 100 posts. The average daily user looks at approximately 7 pages according to Alexa. Each page displays 12 posts, implying that the typical daily user sees 84 posts per day. Before the event, they saw approximately 13.5 posts containing hate against foreign residents each day. The treatment effect implies

that after the event, the average user saw approximately 11 entries with hate against foreign residents (we only observe the treatment effect on the treated group and not the general equilibrium effect).

The next seven columns in each of the three tables show the results of various robustness checks. Column 2 collapses the data by week and uses the Newey-West estimator with 4 lags to adjust the standard error more appropriately—see Section V. The standard error in column 2 decreases to approximately half the size of the first column in all three tables, indicating that we are, if anything, overestimating the standard error in our baseline model. In 2014, there was a parliamentary election in Sweden. To ensure that the week of the election does not influence our results, we exclude it from the analysis in column 3. This leaves the coefficients in all tables qualitatively unchanged. Column 4 includes half-year dummies as a flexible control for the general temporal trend. This decreases the estimated effect slightly—qualitatively, the results remain the same. Column 5 introduces the registration month of the user as a control variable, since both the treatment assignment and the pre- and post-treatment periods are a function of time. Controlling for the registration date is then equivalent, in terms of the difference in the discontinuity before and after the revelation date between the control and treatment groups, to using the same slope before and after the revelation date and for all data on both sides of the cutoff. This decreases the coefficients slightly for hate and hate against foreign residents, but we also see an increase in the coefficient for misogyny.

In column 6, we drop all users registered after the event took place in September 2014. Flashback users can create multiple accounts and then become passive on one account and continue to discuss and produce hate using another account. However, there are strict forum rules to guide the discussions and accounts. Moderators on Flashback can suspend users from all forums if they, for example, use multiple accounts to support their arguments in the same subforum. Suspensions are automatically displayed for all other users on all previous posts that the suspended user has made on Flashback. A suspended user cannot use his or her current account and cannot create a new Flashback account. Thus, it is unlikely that there are individuals with multiple accounts producing hateful content in the same subforum. Nevertheless, users who worry about their identity being obtained by journalists could create a new account and continue to write in a hateful way using that new account. However, this is not without its cost to regular users, as they might have built up a certain reputation for their pseudonym. Additionally, there was no sharp increase in newly registered users after the event took place, and there is no trend in the number of users that post (see Figures A.5 and A.4 in the appendix). From column 1, we can also conclude that this should not be a major concern, since if anything, those

in the control group also decreased the amount of hate they produced. Nevertheless, column six restricts the sample to everyone registered before the event, that is, before September 2014. Thus, this column drops all post-event registered users. Overall, the main coefficients are largely unchanged. Thus, sorting out of the treatment group by creating new user accounts does not seem to be a major issue. It could, however, still be the case that some users moved to other forums or other media where they could write hate without fear of repercussion, though as we argue in Section II, there is no clear substitute in Swedish. To rule out possible temporal trends in the data, we also look at the number of writers per week and the number of posts per week. Except for the election week in 2014, which is marked by a spike in activity, there is no clear temporal trend or pattern (see Figures A.4 and A.3 in the appendix). Restricting the sample to the number of users who write at least one post per registration month, we see that the majority of users are from the control group and that there is no large influx of new users in the post-treatment period (see Figure A.5 in the appendix). In general, users do not seem to create new accounts to continue their (hateful) writings under a different username.

Column 7 also restricts the sample to those that were registered before the event and adds the control for the user start date, while column 8 interacts the variable of the user start date with the post variable to let the slope of the start date differ between the pre- and post-period. The last specification is thus a global version of the difference-in-discontinuity approach, as pioneered by Grembi, Nannicini, and Troiano (2016). If anything, the coefficients actually increase slightly, although the standard errors also become much larger. Estimates from difference-in-discontinuity regressions where the slope is also allowed to differ by treatment group are presented in Table A.5 along with a depiction of the discontinuity in Figure A.7. Needless to say, the power is even lower for these regressions. Nevertheless, the coefficients are more or less the same for hate against anyone and misogyny, although the estimate for hate against foreign residents decreases slightly.

Overall, the regressions confirm our conclusion that a threat of decreased anonymity leads to a decrease in hate, comprising a decrease in xenophobia and an increase in misogyny. In addition, we obtain similar results from a basic DD model using only the random sample assigned to our RA (see the online appendix Section D.C.).

VII. Possible Mechanisms

A decrease in the share of hateful content can be caused by both changes in the number of hateful entries and changes in the number of non-hateful entries. We start by standardizing the total number of entries and the number of hateful entries of the three different types by treatment group and then collapse these series at a weekly level (following equations 3 and 4). The decrease in the share of hateful entries is a result of both a decrease in the number of hateful entries and a decrease in the number of total entries (see Table A.1 in the appendix). Both the number of standardized entries and the number of standardized hateful entries against foreign residents and against anyone decreases, while the standardized number of misogynistic entries hardly changes at all.[17]

VII.A. Changes in Activity

A decrease in anonymity might lead individual users to stop writing hateful content and proceed to write non-hateful content. To this end, we change the unit of analysis to the individual and take a closer look at the individual behavior. We adjust the sample to comprise individual users who have at least one entry in both the pre- and the post-treatment periods, and we collapse the data into a pre- and post-observation for each individual. With these collapsed data, we run DD models for the standardized sum of entries and hateful entries by treatment group as well as the share of hateful entries. The collapsed data consist of 11159 individual users, with two observations per user. Table VII shows the results.

The first column shows that the treated group decreased their amount of hateful entries against anyone by 0.07 standard deviations from the pre- to the post-period compared to the control group. However, column two shows that the same is not true for the share of individual hateful comments against anyone, as the coefficient is close to zero. Column three highlights that the treated group decreased their hateful content against foreign residents by 0.056 standard deviations compared to the control group. Once again, however, the share of individual user hate against foreign residents remains unaltered, as shown in column four. In the fifth column, however, we note that there is no real change in the standardized number of misogynistic posts, which is in line with the results in Table VI. The individual share of misogyny remains at a similar level as in the pre-period. Last, column seven shows the estimated effect on the total number of user entries. We note that decreased anonymity decreases the level of overall activity,

[17] Figures A.1 and A.2 show graphs of the standardized differences in the numbers of entries in the appendix.

and individuals in the treatment group write approximately 0.06 standard deviations fewer entries in the post-period compared to the change in the control group. This is in line with individuals not changing what they write but rather simply decreasing their writing activity.

We also suspect that those most likely to decrease their activity are those who wrote the most hateful entries before the event, as they would be more likely to be identified by the journalists as haters. Table VIII first restricts the sample to only users in the bottom 2/3 of the distribution of the share of hate against anyone at the individual level in the pre-period. Compared to Table VIII, the coefficient is essentially zero and is insignificant. The second column gives this result for only the individuals in the top 1/3, i.e., the individuals who have the highest share of hateful posts against anyone in the pretreatment period. They decreased the number of posts they made by 0.18 standard deviations compared to the control group. The third and fourth columns use the bottom 2/3 and the top 1/3 of users in terms of pretreatment hate against foreign residents, providing a similar pattern to that for hate against anyone: there is no change in the number of posts among the bottom 2/3 of the pretreatment distribution, but a 0.13 standard deviation decrease among the top 1/3. Columns five and six explore misogyny in the same way. Interestingly, we find almost identical coefficients to those in column seven in Table VII for both high and low misogynistic individuals. Thus, it appears that individuals with a high rate of pre-period hate against anyone or against foreign residents altered their behavior, whereas those with a high proportion of misogyny did not appear to change more than the average. One possible explanation for the difference in response depending on prior hate is that journalists had a history of publicly exposing those expressing hatred towards foreign residents in other settings. Thus, the threat of being exposed was arguably more credible for those who had previously written much hateful content against foreign residents. It could, however, also be the case that it is more socially acceptable to write hateful comments against women. Gender, in contrast to race, is not a ground for hate crime in Swedish law. Most examples of convictions under hate crime laws in Sweden concern race.

In sum, a decrease in anonymity, in the form of a threat to expose users' identities, can lead to a decrease in hateful content in online discussions, but it can also lead to a decrease in non-hateful entries. Individuals leave the forum or decrease their activity level, particularly individuals who know they have misbehaved in the past.[18]

[18]The results indicate that one possible mechanism behind the decrease in the share of hateful entries is that the decrease in anonymity makes users leave the forum or decrease the frequency of writing both hateful and non-hateful posts in the forum. Looking at only the number of active users pre- and

VII.B. Changes in the Target Group of Hate

We find that less anonymity reduces the share of hate against foreign residents, while the share of hate against females rises. Part of these results can be explained by a decrease in activity by the individuals who had a high share of hate against foreign residents in the pre-period.

Another possible mechanism is that a diminished anonymity can make users substitute to some degree which group they direct their hate towards. Previous research suggests that anonymity can lower the perception of the possible repercussions of breaking a social norm (Suler 2004, Moore et al. 2012, Van Royen et al. 2017). If anonymity is the default, more transparency might induce individuals to be aware of the repercussions of breaking a social norm. If hate against foreign residents is less acceptable than hate against females and feminists, users might decrease their hate against foreign residents and increase their hate against females and feminists. To explore this possibility further, we estimate a regression corresponding to equation 5, where ΔY_i is the change in the individual number of misogynistic entries, ΔX_i is the change in the individual number of hateful entries against foreign residents and $Treated$ is the same treatment group dummy that we used in the regression above.

$$\Delta Y_i = \rho Treated_g * \Delta X_i + \pi \Delta X_i + \sigma Treated_g + \kappa_{ig} \tag{5}$$

Under the assumption that there are similar trends between the treated and non-treated individuals, ρ measures the degree to which the treated individuals substitute hate against foreign residents with misogyny. Table IX investigates the substitution effect using a regression corresponding to equation 5.

Column one in Table IX shows a negative coefficient for the interaction term in the first column, which implies that users in the treatment group who decrease their hateful entries against foreign residents by one between the pre- and post-period also increase their amount of misogyny by 0.57 entries. Thus, the treated seem to substitute one hateful entry against foreign residents with one-half entry containing misogyny. As expected from the main results, the second column indicates that there is no substitution between hate against foreign residents and hate against other groups (measured by hate

post-event, the number of users in the treatment group decreases 4.5 percentage points more than the number of users in the control group. It would be a natural step to study the probability that a user will drop out of the discussion. However, in our data, we only observe individuals if they are active in both the pre- and post-period. This implies that we cannot use the time dimension in the DD setting if our outcome is a dummy representing whether the user is active, since all users are, by definition, active in the pretreatment period.

against anyone).

VIII. Concluding Remarks

Hate, harassment, and threats in online discussions have become a growing democratic concern (Cheng et al. 2017). A vast majority of the hateful content online comes from users who disguise their identities, making their actions less open to repercussions (Citron 2014). Our study contributes to the current policy debates on how to combat online hate by providing descriptive evidence concerning the prominence of hate in online discussions of political topics and estimating the effect of anonymity on hateful content. Overall, we find that less anonymity leads to a reduced share of hate against foreign residents and an increased share of hate against females in online discussions of politics.

First, we predict hateful content in political discussion forums using a Swedish social media discussion forum. Here, we have an RA classify a random part of the entries as having hateful content or no hateful content. Using a supervised machine learning model, a logistic lasso, we predict hateful content in the full data set. Second, we use the predictions to quantify the causal effect of anonymity on hateful content using a DD design. The exogenous variation comes from an event where anonymity unexpectedly decreased—a threat of being exposed arose—for a well-defined subset of the users.

The effects of a decreased share of hate against foreign residents and an increased share of hate against females seem to be driven by a combination of two factors: i) treated users with a high share of hate against foreign residents before the event decrease their activity at Flashback after the event, and ii) treated users to some extent seemed to shift from writing hateful posts about foreign residents to writing hateful posts about females and feminists. However, individuals in the control group might also have believed that their general risk of being exposed increased after the event. These possible spill-over effects suggest that our estimate is a lower bound on how changes in anonymity affect hateful content. We also find evidence that suggests individuals substituted hate against foreign residents with hate against females. One possible explanation for this substitution is that hate against foreign residents was the main focus of the journalists when publicly exposing users. Our results open up an exciting avenue of research on understanding different types of hate online. Our results also point to the importance of high social costs or high costs with a low probability as deterrents for future criminal or anti-social behavior.

Previous research finds discussions on social media to be an important part of current political outcomes (Allcott and Gentzkow 2017, Qin, Strömberg, and Wu 2017). If fewer

individuals discuss politics online, this can have adverse consequences such as lower political accountability or less informed decisions (Strömberg 2015). The effects of there being less hateful content can include both changes in how individuals write their entries and that individuals stop discussing. In this paper, we see both effects. Our findings thus support the idea that making a policy combating online hate is not as simple as requiring users to expose their names, such as on Facebook and in comments on traditional media news articles.

References

Acemoglu, Daron, Tarek A. Hassan, and Ahmed Tahoun, "The Power of the Street: Evidence from Egypt's Arab Spring," *The Review of Financial Studies*, 31 (2017), 1–42.

Acquisti, Alessandro, Curtis R. Taylor, and Liad Wagman, "The Economics of Privacy," *Journal of Economic Literature*, 2 (2016), 442–492.

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

Ali, S. Nageeb, and Roland Bénabou, "Image Versus Information: Changing Societal Norms and Optimal Privacy," National Bureau of Economic Research, 2016.

Allcott, Hunt, and Matthew Gentzkow, "Social Media and Fake News in the 2016 Election," *Journal of Economic Perspectives*, 31 (2017), 211–236.

Angrist, Joshua D., and Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton: Princeton University Press, 2008).

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan, "How Much Should we Trust Differences-in-Differences Estimates?" *The Quarterly Journal of Economics*, 119 (2004), 249–275.

Bhuller, Manudeep, Tarjei Havnes, Edwin Leuven, and Magne Mogstad, "Broadband Internet: An Information Superhighway to Sex Crime?" *Review of Economic Studies*, 80 (2013), 1237–1266.

Bursztyn, Leonardo, Georgy Egorov, Ruben Enikolopov, and Maria Petrova, "Social Media and Xenophobia: Evidence from Russia," National Bureau of Economic Research Working Paper No. 26567, 2019.

Bursztyn, Leonardo, Georgy Egorov, and Stefano Fiorin, "From Extreme to Mainstream: How Social Norms Unravel," National Bureau of Economic Research Working Paper No. 23415, 2017.

Chan, Jason, Anindya Ghose, and Robert Seamans, "The Internet and Racial Hate Crime: Offline Spillovers from Online Access," *MIS Quarterly*, 40 (2016), 381–403.

Chandrasekharan, Eshwar, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert, "You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech," *Proceedings of the ACM on Human-Computer Interaction*, 1 (2017), 31.

Chen, Ying, Yilu Zhou, Sencun Zhu, and Heng Xu, "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety," in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)* (Amsterdam: IEEE, 2012), pp. 71–80.

Cheng, Justin, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec, "Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions," *arXiv preprint arXiv:1702.01119*, (2017).

Cho, Daegon, Soodong Kim, and Alessandro Acquisti, "Empirical Analysis of Online Anonymity and User Behaviors: The Impact of Real Name Policy," in *System Science (HICSS), 2012 45th Hawaii International Conference on* (Maui, HI: IEEE, 2012), pp. 3041–3050.

Citron, Danielle Keats, *Hate Crimes in Cyberspace* (Cambridge: Harvard University Press, 2014).

Cohen, Katie, Fredrik Johansson, Lisa Kaati, and Jonas Clausen Mork, "Detecting Linguistic Markers for Radical Violence in Social Media," *Terrorism and Political Violence*, 26 (2014), 246–256.

Conley, Timothy G., and Christopher R. Taber, "Inference with "Difference in Differences" with a Small Number of Policy Changes," *The Review of Economics and Statistics*, 93 (2011), 113–125.

Davidson, Thomas, Dana Warmsley, Michael Macy, and Ingmar Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," in *Eleventh International AAAI Conference on Web and Social Media* (2017).

Davidsson, Pamela, Matti Palm, and Asa Melin Mandre, *Svenskarna och Internet 2018* (IIS (Internetstiftelsen i Sverige), 2018).

Donald, Stephen G., and Kevin Lang, "Inference with Difference-in-Differences and Other Panel Data," *The Review of Economics and Statistics*, 89 (2007), 221–233.

Duggan, Maeve, *Online Harassment* (Washington, DC: Pew Research Center, 2014).

Engelberg, Joseph, and Pengjie Gao, "In Search of Attention," *The Journal of Finance*, 66 (2011), 1461–1499.

Enikolopov, Ruben, Alexey Makarin, and Maria Petrova, "Social Media and Protest Participation: Evidence from Russia," *Available at SSRN 2696236* (2019).

Enikolopov, Ruben, Maria Petrova, and Konstantin Sonin, "Social Media and Corruption," *American Economic Journal: Applied Economics*, 10 (2018), 150–174.

Froomkin, A Michael, "Lessons Learned Too Well: Anonymity in a Time of Surveillance," *Arizona Law Review*, 59 (2017), 95.

Glaeser, Edward L., "The Political Economy of Hatred," *The Quarterly Journal of Economics*, 120 (2005), 45–86.

Grembi, Veronica, Tommaso Nannicini, and Ugo Troiano, "Do Fiscal Rules Matter?" *American Economic Journal: Applied Economics*, 8 (2016), 1–30.

Hansen, Stephen, Michael McMahon, and Andrea Prat, "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach," *The Quarterly Journal of Economics*, 133 (2018), 801–870.

Hlavac, Marek, "stargazer: Well-Formatted Regression and Summary StatisticsTables," R package version 5.2.2. `https://CRAN.R-project.org/package=stargazer`, 2018.

Holmström, Bengt, "Moral Hazard and Observability," *The Bell Journal of Economics*, 10 (1979), 74–91.

——, "Managerial incentive problems: A dynamic perspective," *The review of Economic studies*, 66 (1999), 169–182.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, *An Introduction to Statistical Learning*, vol. 112 (New York: Springer, 2013).

Lööw, Helene, and Lotta Nilsson, "Hets mot folkgrupp," BRÅ Rapport, 17 (2001).

Moore, Michael J., Tadashi Nakano, Akihiro Enomoto, and Tatsuya Suda, "Anonymity and Roles Associated with Aggressive Posts in An Online Forum," *Computers in Human Behavior*, 28 (2012), 861–867.

Mullainathan, Sendhil, and Jann Spiess, "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, 31 (2017), 87–106.

Pettersson-Lidbom, Per, and Peter Skogman Thoursie, "Temporary Disability Insurance and Labor Supply: Evidence from a Natural Experiment," *The Scandinavian Journal of Economics*, 115 (2013), 485–507.

Postmes, Tom, Russell Spears, and Martin Lea, "Breaching or Building Social Boundaries? SIDE-Effects of Computer-Mediated Communication," *Communication Research*, 25 (1998), 689–715.

Prat, Andrea, "The Wrong Kind of Transparency," *The American Economic Review*, 95 (2005), 862–877.

Qin, Bei, David Strömberg, and Yanhui Wu, "Why Does China Allow Freer Social Media? Protests Versus Surveillance and Propaganda," *Journal of Economic Perspectives*, 31 (2017), 117–140.

Strömberg, David, "Media and Politics," *Economics*, 7 (2015), 173–205.

Suler, John, "The Online Disinhibition Effect," *Cyberpsychology & Behavior*, 7 (2004), 321–326.

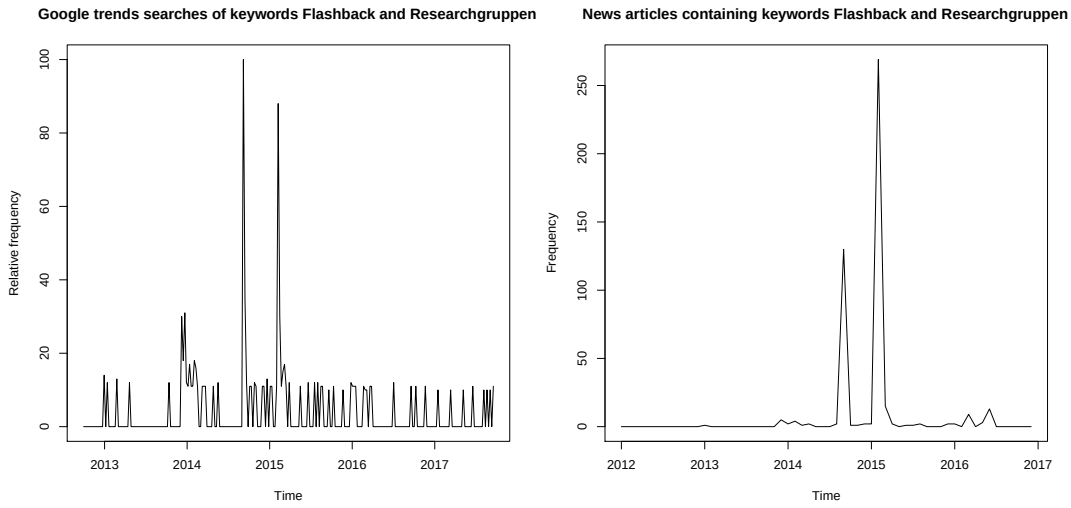
Van Royen, Kathleen, Karolien Poels, Heidi Vandebosch, and Philippe Adam, "'Thinking Before Posting?' Reducing Cyber Harassment on Social Networking Sites Through a Reflective Message," *Computers in Human Behavior*, 66 (2017), 345–352.

Wu, Alice H., "Gendered Language on the Economics Job Market Rumors Forum," in *AEA Papers and Proceedings*, vol. 108 (2018), pp. 175–79.

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

Zhuravskaya, Ekaterina, Maria Petrova and Ruben Enikolopov, "Political Effects of the Internet and Social Media," CEPR Discussion Paper No. DP13996, 2019.

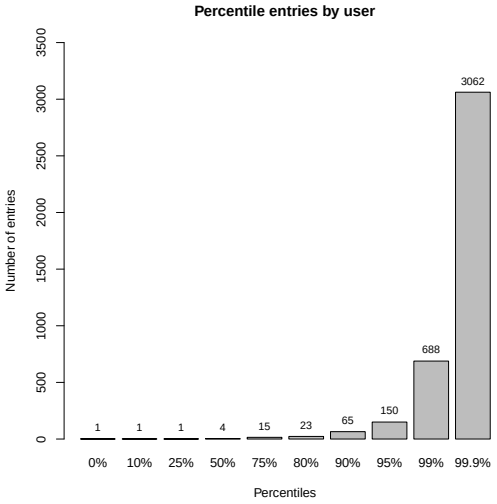
Figures and Tables



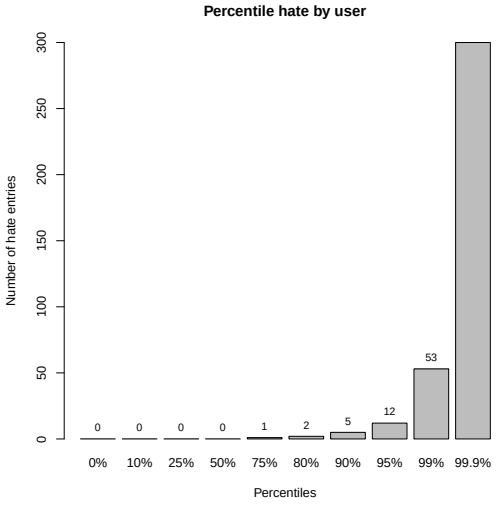
(A) (B)

FIGURE I
Reactions to the Event

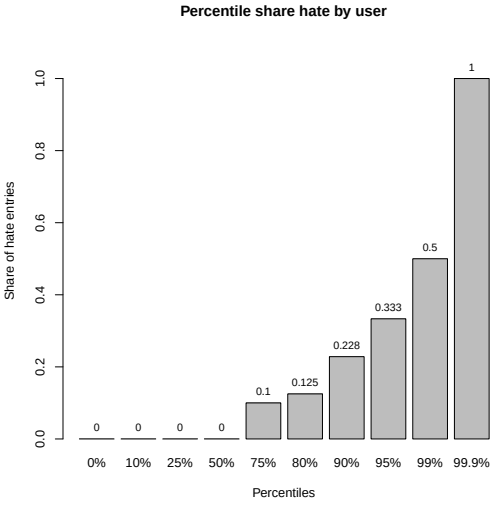
Panel Ia displays data from Google Trends for the weekly relative search frequencies of the terms Flashback and Researchgruppen between 2013 and 2017. The relative search frequency shows a spike if there are at least 50 searches, and all spikes are relative to the largest spike, which reaches 100 in the figure. Thus, a spike reaching 80 implies that the search terms reach 80 percent of the highest search frequency. Panel Ib displays a similar graph using the monthly number of news articles in Sweden containing the words "Flashback" and "Researchgruppen". The data on the number of news articles were obtained through searches in the database Mediearkivet (https://www.retriever.se/, accessed using a login from Stockholm University on the fourth of July 2018).



(A)

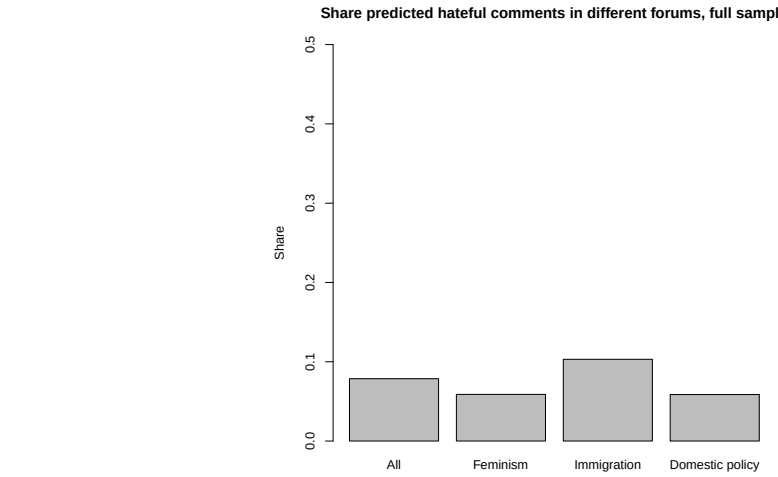


(B)

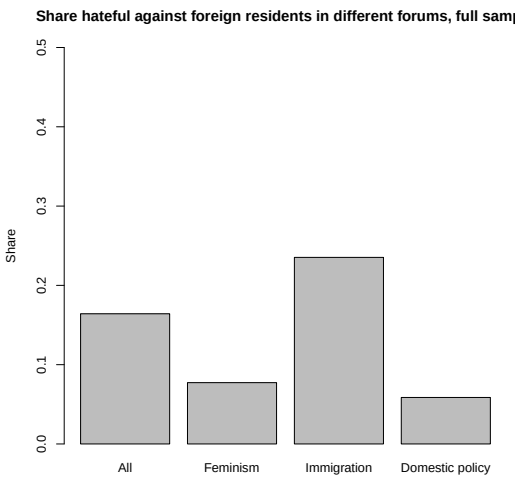


(C)

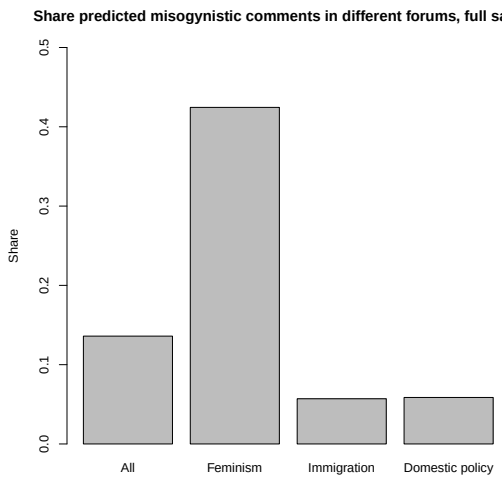
FIGURE II
Distribution of Entries and Hate



(A)

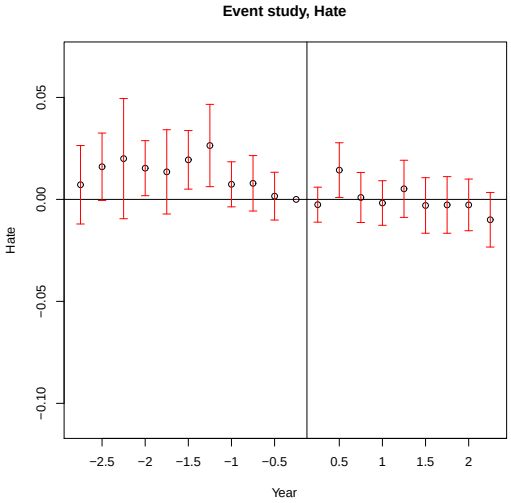


(B)

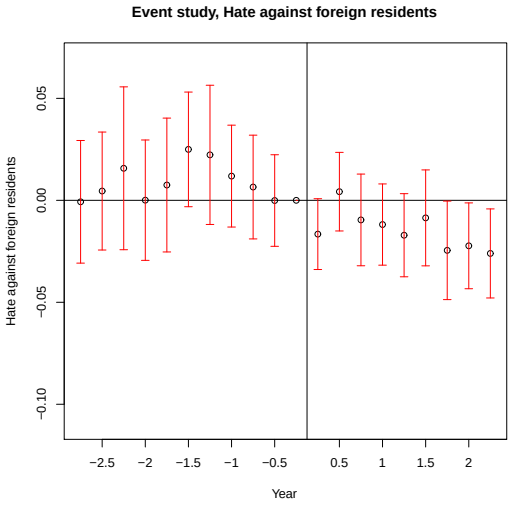


(C)

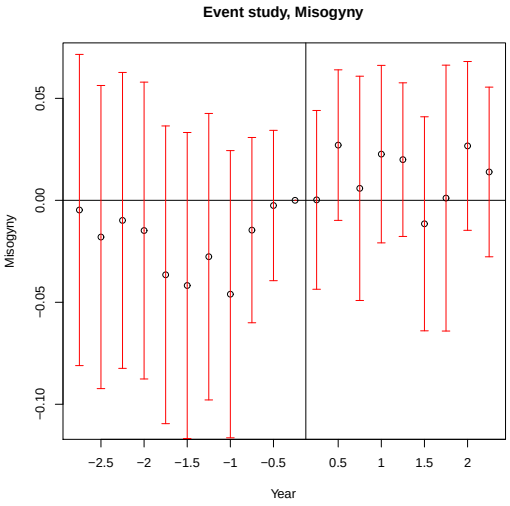
FIGURE III
Forum Shares



(A)



(B)



(C)

FIGURE IV
Event Study

The time period used in the regressions is from January 1, 2012, until December 31, 2016. Standard errors are clustered at the user level.

TABLE I
LOGISTIC LASSO MODEL FOR GENERAL HATE

name	coefficient
(Intercept)	-1.22
arab	0.39
blatt	1.07
dumm	0.01
hor	1.91
lill	1.74
miljon	1.56
muslimsk	3.87
parasit	0.06
patetisk	2.25
rån	3.30
what	0.23

TABLE II

Summary statistics, full sample

	Total, mean	Total, SD	Pre-event registered, mean	Pre-event registered, SD
No. entries	1984224.00		1754758.00	
No. users	48672.00		41350.00	
No. threads	29425.00		29312.00	
Hate against anyone	0.08	0.27	0.08	0.27
Misogyny	0.14	0.14	0.14	0.14
Hate against foreigners	0.16	0.37	0.16	0.37

The table presents the summary statistics for the lasso predicted dummy variables of hate, hate against foreigners and hate against females respectively.

TABLE III
SUMMARY STATISTICS BY TREATMENT/CONTROL, FULL SAMPLE

	Treated, pre	Treated, post	Control, pre	Control, post
No. entries	145385.00	98219.00	902416.00	838204.00
No. users	3588.00	2536.00	30872.00	22835.00
No. threads	13827.00	10892.00	16095.00	13274.00
Hate against anyone	0.09	0.07	0.08	0.08
Misogyny	0.11	0.12	0.15	0.13
Hate against foreigners	0.18	0.15	0.17	0.16

The table presents the summary statistics for the lasso predicted dummy variables of hate, hate against foreigners and misogyny respectively.

TABLE IV
DD RESULTS ON HATE

	Dependent variable:							
	Hate							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Post*Treated	−0.015		−0.015	−0.013	−0.012	−0.012	−0.011	−0.012
	(0.006)		(0.006)	(0.006)	(0.006)	(0.007)	(0.006)	(0.009)
Post reveal	−0.002	−0.013	−0.001	0.022	−0.005	−0.005	−0.006	0.334
	(0.002)	(0.003)	(0.002)	(0.004)	(0.002)	(0.002)	(0.002)	(1.952)
Treated	0.006		0.006	0.005	0.020	0.006	0.018	0.018
	(0.007)		(0.007)	(0.007)	(0.008)	(0.007)	(0.008)	(0.009)
Registration date					0.002		0.002	0.002
					(0.001)		(0.001)	(0.001)
Post*Registration date								−0.0002
								(0.001)
Constant	0.079	0.005	0.079	0.067	−4.794	0.079	−3.911	−4.063
	(0.002)	(0.002)	(0.002)	(0.002)	(1.105)	(0.002)	(1.500)	(1.713)
Exclude election week	No	No	Yes	No	No	No	No	No
Last registered	-	-	-	-	-	Aug 2014	Aug 2014	Aug 2014
Collapsed on weeks	No	Yes	No	No	No	No	No	No
Half year dummies	No	No	No	Yes	No	No	No	No
Observations	1984224	262	1961610	1984224	1984224	1754758	1754758	1754758

Note: The time period used in the regressions is from January 1st 2012 until December 31st 2016.
The standard errors are clustered at the user level except in the second column,
where Newey-West standad errorswith 4 lags on a collapsed time series is used.
The dependent variable is a dummy indicating if a post is predicted as hateful or not.

TABLE V
DD RESULTS ON HATE AGAINST FOREIGNERS

	Dependent variable:							
	Hate against foreigns							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Post*Treated	−0.029		−0.028	−0.025	−0.023	−0.021	−0.020	−0.020
	(0.010)		(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.017)
Post reveal	−0.008	−0.024	−0.006	0.037	−0.013	−0.016	−0.016	−0.044
	(0.004)	(0.004)	(0.004)	(0.005)	(0.004)	(0.004)	(0.004)	(3.972)
Treated	0.017		0.017	0.015	0.039	0.017	0.031	0.031
	(0.012)		(0.012)	(0.012)	(0.014)	(0.012)	(0.016)	(0.018)
Registration date					0.004		0.002	0.002
					(0.001)		(0.001)	(0.002)
Post*Registration date								0.00001
								(0.002)
Constant	0.167	0.015	0.167	0.136	−7.651	0.167	−4.639	−4.626
	(0.004)	(0.003)	(0.004)	(0.004)	(2.300)	(0.004)	(2.982)	(3.957)
Exclude election week	No	No	Yes	No	No	No	No	No
Last registered	-	-	-	-	-	Aug 2014	Aug 2014	Aug 2014
Collapsed on weeks	No	Yes	No	No	No	No	No	No
Half year dummies	No	No	No	Yes	No	No	No	No
Observations	1984224	262	1961610	1984224	1984224	1754758	1754758	1754758

Note: The time period used in the regressions is from January 1st 2012 until December 31st 2016. The standard errors are clustered at the user level except in the second column, where Newey-West standad errorswith 4 lags on a collapsed time series is used. The dependent variable is a dummy indicating if a post is predicted as hateful against foreigners or not.

TABLE VI
DD RESULTS ON HATE AGAINST FEMALES

	Dependent variable:							
	Misogyny							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Post*Treated	0.032		0.033	0.031	0.038	0.033	0.035	0.048
	(0.014)		(0.015)	(0.014)	(0.014)	(0.015)	(0.015)	(0.027)
Post reveal	−0.021	0.033	−0.019	−0.033	−0.026	−0.021	−0.023	−4.534
	(0.007)	(0.007)	(0.007)	(0.014)	(0.006)	(0.008)	(0.007)	(8.269)
Treated	−0.044		−0.044	−0.042	−0.022	−0.044	−0.013	−0.019
	(0.012)		(0.012)	(0.012)	(0.017)	(0.012)	(0.022)	(0.024)
Registration date					0.004		0.006	0.005
					(0.002)		(0.003)	(0.004)
Post*Registration date								0.002
								(0.004)
Constant	0.150	−0.043	0.150	0.165	−7.780	0.150	−10.958	−8.940
	(0.006)	(0.004)	(0.006)	(0.010)	(4.261)	(0.006)	(6.497)	(7.339)
Exclude election week	No	No	Yes	No	No	No	No	No
Last registered	-	-	-	-	-	Aug 2014	Aug 2014	Aug 2014
Collapsed on weeks	No	Yes	No	No	No	No	No	No
Half year dummies	No	No	No	Yes	No	No	No	No
Observations	1984224	262	1961610	1984224	1984224	1754758	1754758	1754758

Note: The time period used in the regressions is from January 1st 2012 until December 31st 2016. The standard errors are clustered at the user level except in the second column, where Newey-West standad errorswith 4 lags on a collapsed time series is used. The dependent variable is a dummy indicating if a post is predicted as hateful against females or not.

TABLE VII
INDIVIDUAL BEHAVIOR, FULL SAMPLE

	Dependent variable:						
	Std. No. hate	Share hate	Std. No. hate foreigners	Share hate foreigners	Std. No. misogyny	Share misogyny	Std. No. entries
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Post*Treated	−0.070	−0.002	−0.056	−0.007	−0.003	−0.0001	−0.063
	(0.030)	(0.004)	(0.027)	(0.006)	(0.017)	(0.005)	(0.024)
Post reveal	−0.015	−0.011	−0.027	−0.015	−0.021	−0.012	−0.013
	(0.009)	(0.002)	(0.010)	(0.003)	(0.010)	(0.002)	(0.010)
Treated	0.035	−0.006	0.028	−0.013	0.002	−0.021	0.031
	(0.034)	(0.004)	(0.034)	(0.005)	(0.024)	(0.004)	(0.031)
Constant	0.007	0.071	0.013	0.148	0.011	0.104	0.006
	(0.010)	(0.001)	(0.011)	(0.002)	(0.010)	(0.002)	(0.011)
Observations	22318	22318	22318	22318	22318	22318	22318

Note: The time period used in the regressions is from January 1st 2012 until December 31st 2016. Standard errors clustered at the user level. The dependent variable is either individual share hate, hate against foreigners or females in each period, or standardized number of individual hateful, hateful against foreigners or females entries in each period.

TABLE VIII
LOW AND HIGH HATERS

	Dependent variable:					
	Standardized No. entries					
	(1)	(2)	(3)	(4)	(5)	(6)
Post*Treated	0.0004	−0.183	−0.027	−0.133	−0.062	−0.073
	(0.018)	(0.059)	(0.022)	(0.056)	(0.023)	(0.054)
Post reveal	0.0002	−0.037	−0.005	−0.027	0.011	−0.055
	(0.012)	(0.018)	(0.013)	(0.015)	(0.009)	(0.022)
Treated	−0.031	0.157	0.0004	0.096	0.041	0.044
	(0.023)	(0.077)	(0.028)	(0.073)	(0.026)	(0.077)
Constant	−0.073	0.147	−0.028	0.068	−0.082	0.157
	(0.013)	(0.018)	(0.014)	(0.016)	(0.008)	(0.025)
Cut off	Hate< 0.051	Hate> 0.051	Hate foreign< 0.15	Hate foreign> 0.15	Misogyny< 0.063	Misogyny> 0.063
Observations	14274	8036	14296	7982	14236	8080

Note: The time period used in the regressions is from January 1st 2012 until December 31st 2016. Standard errors clustered at the user level. Each column uses the data in either the bottom 2/3 or top 1/3 of the distribution in individual share hate, share hate foreign and share hate females in the pre-treatment period, respectively. Outcome variable is standardized individual number of entries in the pre- and post-period

TABLE IX
INDIVIDUAL SUBSTITUTION OF HATE

	Dependent variable:	
	ΔNo. misogyny	ΔNo. hate
	(1)	(2)
Treated*ΔNo. hate foreign	−0.512	0.065
	(0.205)	(0.085)
ΔNo. hate foreign	0.753	0.428
	(0.204)	(0.042)
Treated	0.493	−0.025
	(0.988)	(0.264)
Constant	−0.644	0.135
	(0.688)	(0.137)
Observations	11159	11159

Note: The time period used in the regressions is from January 1st 2012 until December 31st 2016. Robust standard errors presented.

A. Appendix

TABLE A.1
Difference in number of standardized entries between treated and control

	Dependent variable:			
	Std. No. entries	Std. No. hateful entries	Std. No. hateful entries foreign	Std. No. misogyny entries
	(1)	(2)	(3)	(4)
Post reveal	−0.917	−1.193	−1.138	−0.148
	(0.165)	(0.185)	(0.176)	(0.223)
Constant	0.427	0.556	0.530	0.069
	(0.119)	(0.113)	(0.095)	(0.163)
Observations	262	262	262	262
Collapsed on weeks	Yes	Yes	Yes	Yes

Note: The time period used in the regressions is from January 1st 2012 until December 31st 2016. Data is collapsed on a weekly level and the standard errors are computed using the Newey-West estimator with 4 lags. The outcome variables are each a time series of the difference between the early and late adopters.

TABLE A.2
REGRESSIONS USING TWO-WAY CLUSTER

	Dependent variable:		
	Hate foreign	Misogyny	Hate
	(1)	(2)	(3)
Post*Treated	−0.029	0.032	−0.015
	(0.010)	(0.014)	(0.006)
Post reveal	−0.008	−0.021	−0.002
	(0.004)	(0.007)	(0.002)
Treated	0.017	−0.044	0.006
	(0.012)	(0.012)	(0.007)
Constant	0.167	0.150	0.079
	(0.004)	(0.006)	(0.002)
Observations	1984224	1984224	1984224

Note: Standard errors clustered at the user and thread level.

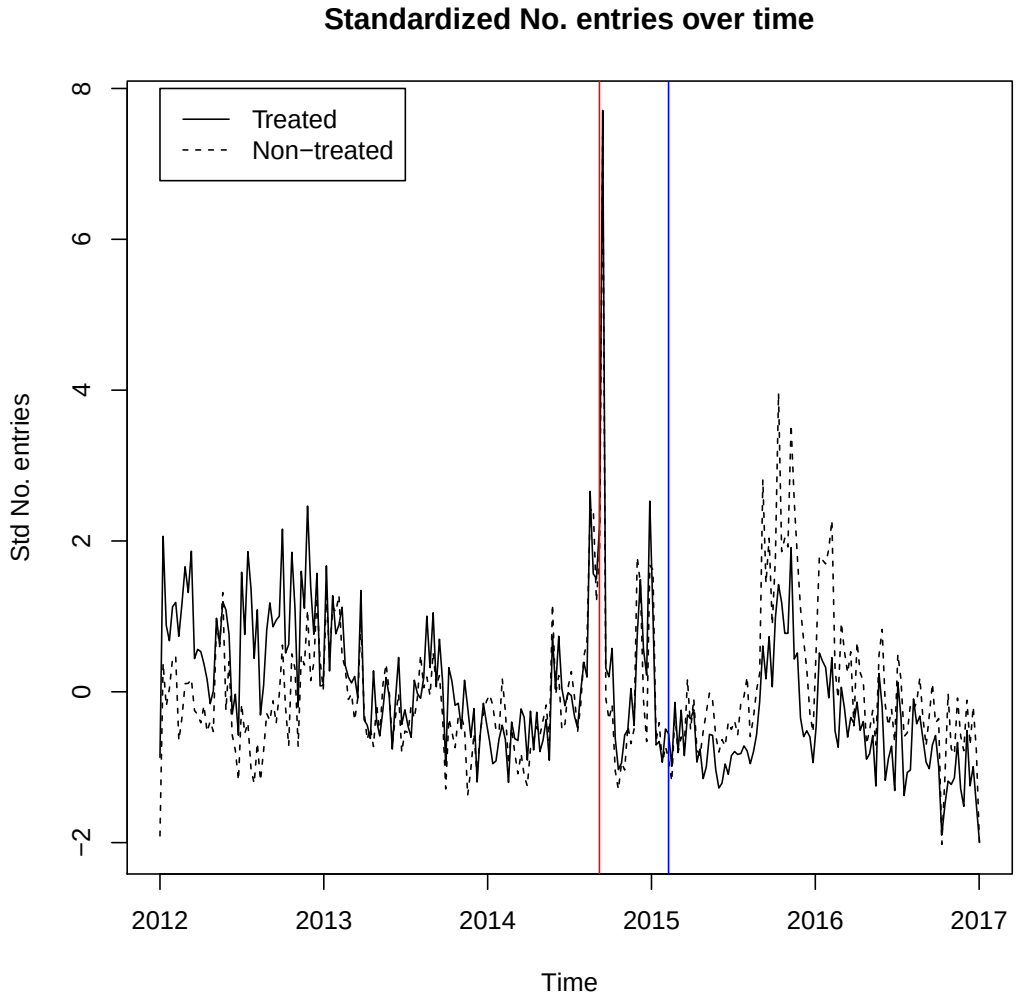


FIGURE A.1
Entries over time per treatment group

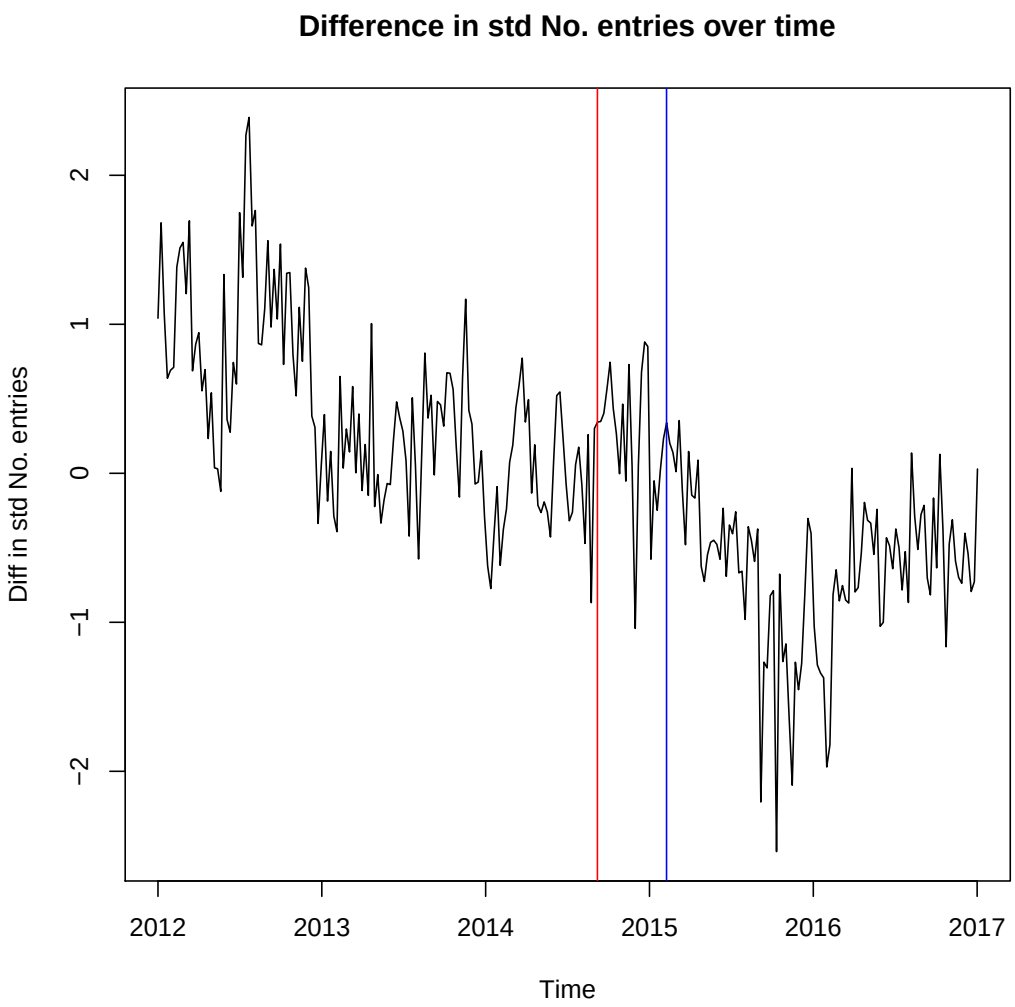


FIGURE A.2
Difference in entries over time between treatment and control

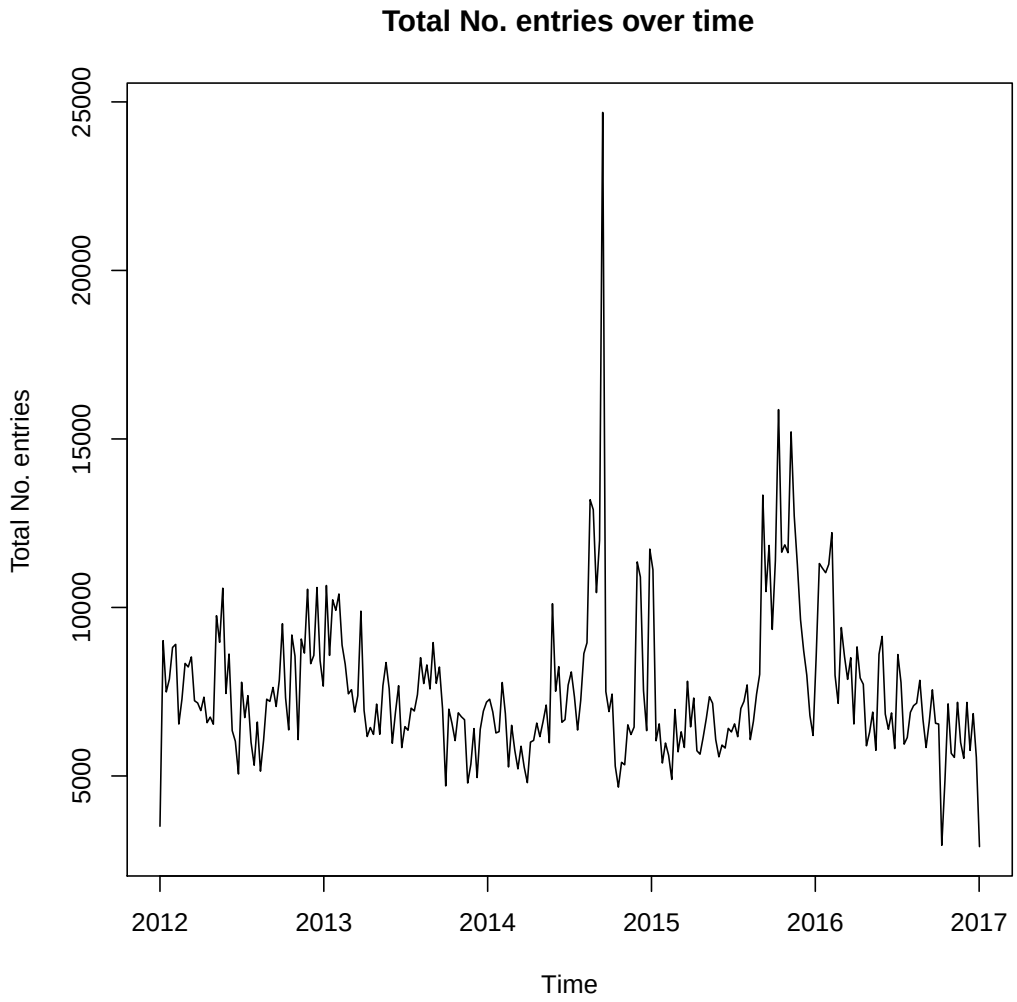


FIGURE A.3
Total entries over time

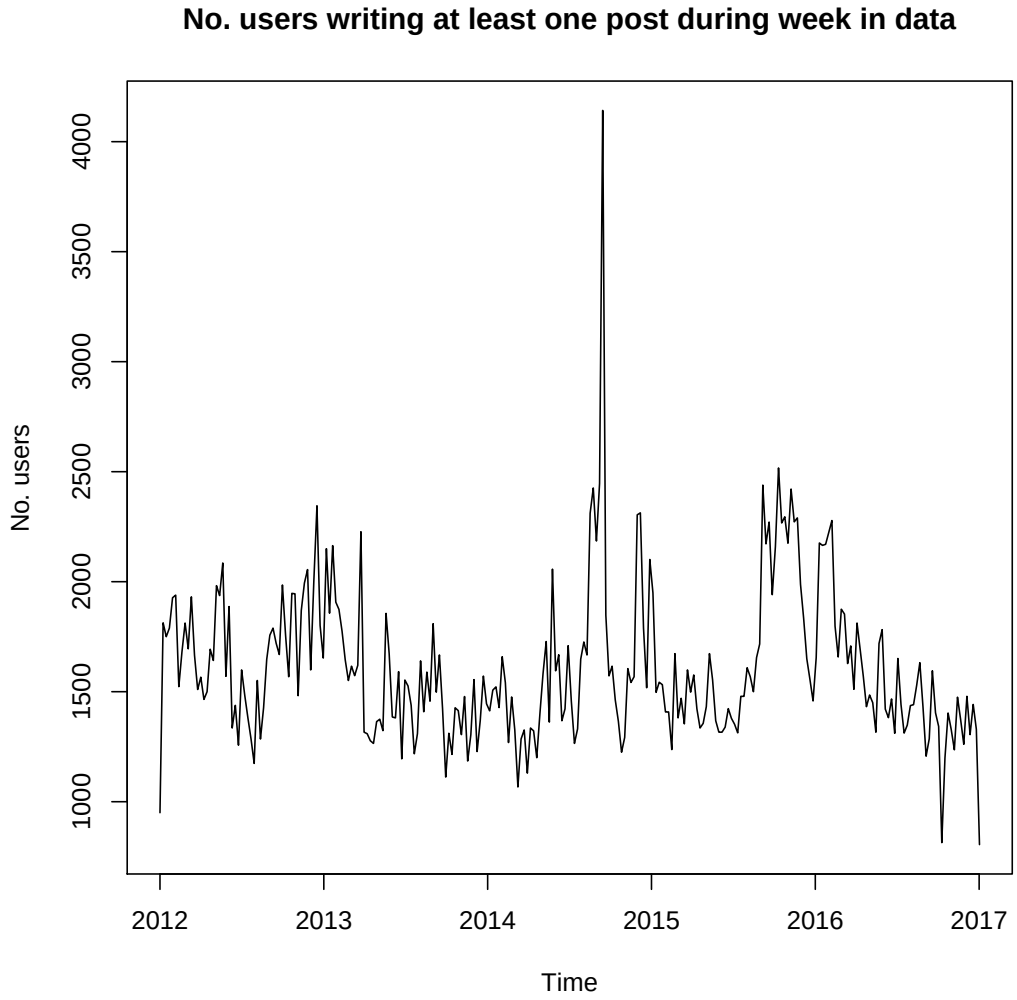


FIGURE A.4
Active users per week in data

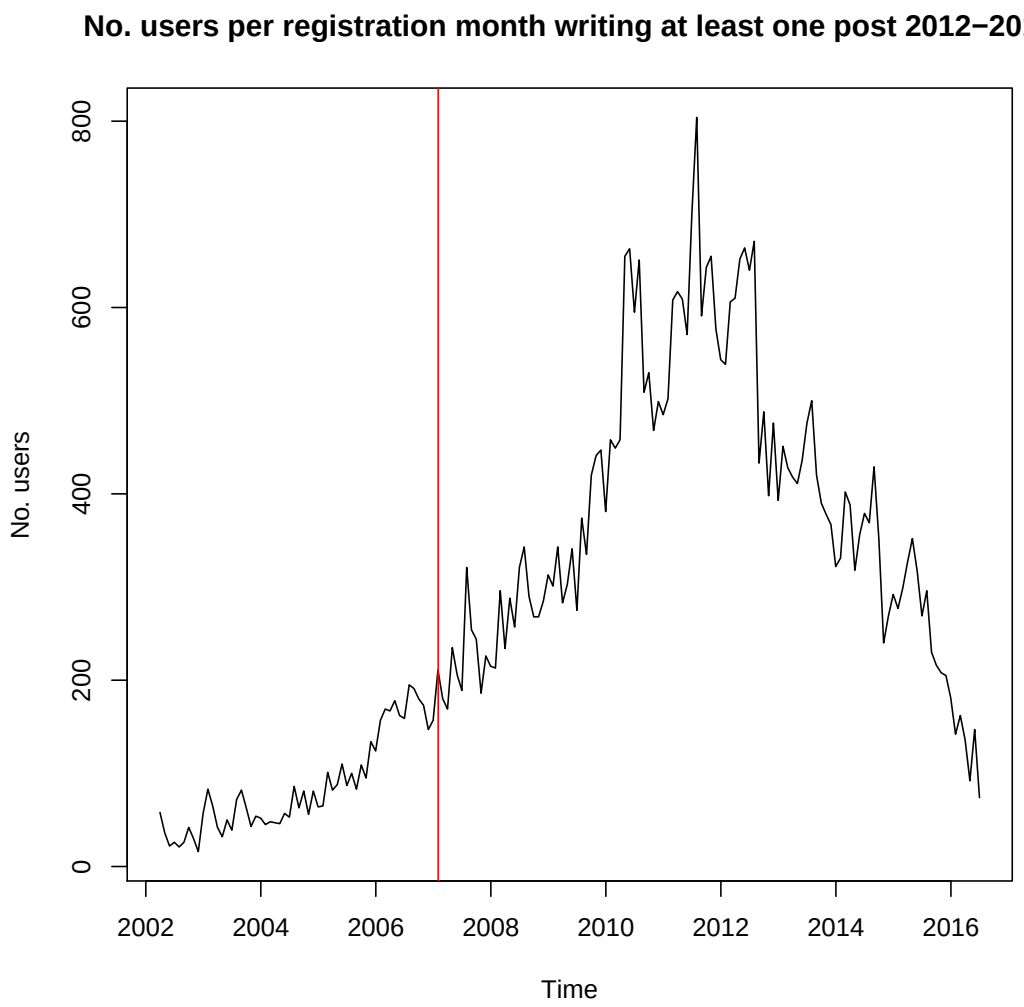


FIGURE A.5
Users per registration month in data

TABLE A.3
HATE AS RESPONSE TO HATE

	Dependent variable:								
	Hate resp.	Hate resp.	Hate resp.	Hate foreign resp.	Hate foreign resp.	Hate foreign resp.	Misogyny resp.	Misogyny resp.	Misogyny resp.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Initial hate	0.202			0.078			0.054		
	(0.022)			(0.020)			(0.014)		
Initial hate foreign		0.162			0.186			−0.020	
		(0.034)			(0.032)			(0.014)	
Initial misogyny			0.195			−0.066			0.206
			(0.036)			(0.016)			(0.033)
Constant	0.194	0.225	0.227	0.076	0.078	0.097	0.048	0.061	0.048
	(0.010)	(0.009)	(0.009)	(0.007)	(0.007)	(0.007)	(0.005)	(0.005)	(0.005)
Observations	2942	2942	2942	2942	2942	2942	2942	2942	2942

Note: Standard errors clustered at the initial post level.

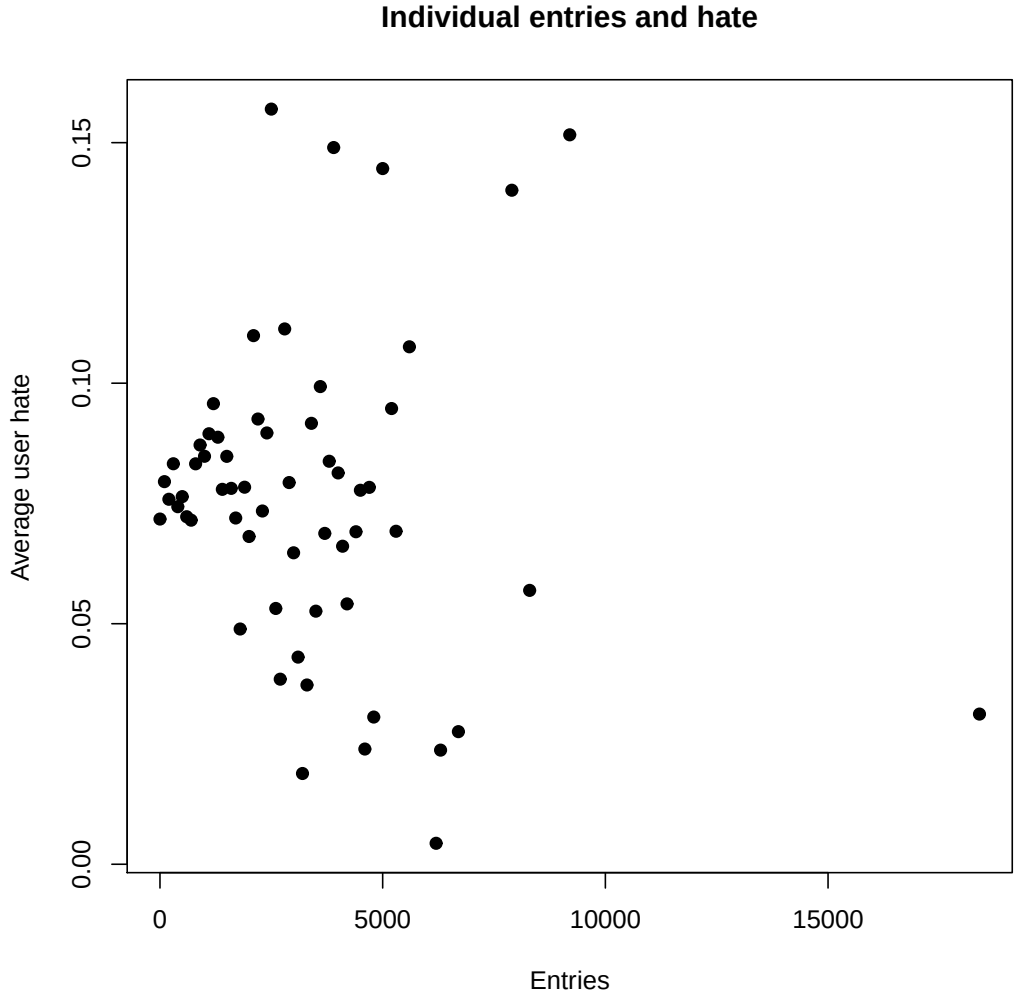


FIGURE A.6
Hate–entries relationship at user level

Binned scatter plot showing relationship between user number of entries and share hate.

TABLE A.4
SIMPLE CORRELATIONS

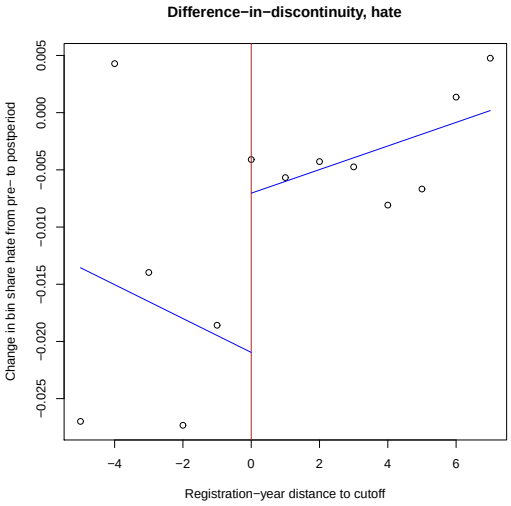
	Dependent variable:							
	Individual share hate	Hate		Misogyny	Individual share hate		Individual share misogyny	Hate
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
No. individual entries	0.00000							
	(0.00000)							
Misogyny		0.058						
		(0.001)						
Hate foreign			0.250	0.009				
			(0.0005)	(0.001)				
Individual misogyny					0.071			
					(0.004)			
Individual hate foreign						0.263	−0.008	
						(0.003)	(0.004)	
First entry is hateful								0.028
								(0.001)
Constant	0.072	0.071	0.038	0.135	0.065	0.033	0.107	0.074
	(0.001)	(0.0002)	(0.0002)	(0.0003)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	48672	1984224	1984224	1984224	48672	48672	48672	1958250

Note: Standard errors clustered at the thread level in the 8th column.

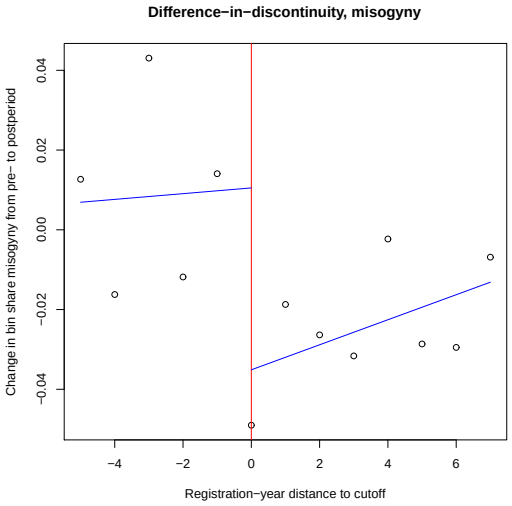
TABLE A.5
DIFFERENCE IN DISCONTINUITY ESTMATES

	Dependent variable:		
	Hate	Hate against foreigners	Misogyny
	(1)	(2)	(3)
Post*Treated	−0.014	−0.016	0.046
	(0.012)	(0.021)	(0.024)
Linear spline	Yes	Yes	Yes
Observations	1754758	1754758	1754758

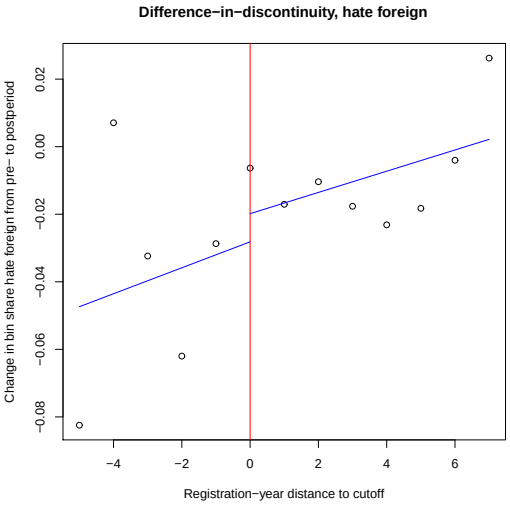
Note: The time period used in the regressions is from January 1st 2012 until December 31st 2016. Standard errors clustered at the user level.



(A)



(B)



(C)

FIGURE A.7
Difference-in-Discontinuity

TABLE A.6
Lasso Coefficients for Hate Against Gender

name	coefficient
(Intercept)	-2.95
alltm	2.95
avundsjukan	1.21
beatric	5.95
betrak	1.27
bottn	0.03
egotripp	0.06
feminism	0.14
feminist	1.32
fjortis	2.33
hor	8.81
klubb	3.29
kvinn	2.19
mental	1.66
oskyd	2.25

TABLE A.7
Lasso Coefficients for Hate Against Foreign

name	coefficient
(Intercept)	-2.39
arab	3.08
blatt	2.05
fruar	4.27
förankr	0.54
gruppvåldtäk	2.87
intel	0.46
kameran	0.59
komplet	0.24
koranskol	2.19
krasch	2.08
käk	1.82
landskron	1.86
muslimsk	6.21
neg	1.48
negr	0.22
parasit	3.67
rån	7.87
separat	0.32
serb	2.04
svennehor	2.28
svensk	0.13

B. Theoretical Framework

Ali and Bénabou (2016) explicitly models anonymity (transparency) and reputation in a public goods setting. Contribution to the public good entails writing in an honest and respectful manner, free from the pollution of hateful speech. For the purpose of our study, we will use the agents' equilibrium behavior to form empirical expectations. Below, we describe in brief the relevant parts of the model.

A single state or a general authority is concerned with all citizens having access to

civil political discussions in social media, i.e., discussions free from hateful content. A continuum of forum users (agents) $i \in [0,1]$ take part in political discussions on social media and can contribute to the public good by not resorting to hateful comments. Anonymity guides the degree to which the other agents as well as the principal can view individual contributions.

A user's contributions depend on 1) an intrinsic preference for having political debates that are free from hateful comments, v_i; 2) an individual signal or perception of the common value of a hate-free debate, θ_i; and 3) a concern for reputation μ_i (social image). Users are assumed to care about others' beliefs about them and thus wish to appear prosocial. The strength of reputational concern varies across individuals, communities and time periods. User $\jmath$ estimates other users' reputation by using his own signal and reputational concern as well as the aggregate contribution $\bar{a}$. User $\imath$ incorporates how she will be judged by others and makes contributions thereafter. The user's contribution decision a_i at a cost $C(a_i)$ depends on her nonreputational payoff and reputational payoff [19]

$$\max_{a_i \in \mathbb{R}} \{E[U_i(v_i\ \theta_i\ w; a_i\ \bar{a}\ a_p)|\theta_i] + x\mu_i[R(a_i\ \theta_i\ \mu_i) - \bar{v}]\} \tag{8}$$

The degree of anonymity x can change, for example, through an exogenous shock.[20] Anonymity affects utility only through reputational concerns: the risk of being exposed as a person producing hateful content. If a user $\jmath$ observes another user $\imath$'s increase in contributions relative to the aggregate contribution, all he or she knows is that this could have been motivated by a strong intrinsic motivation—a strong social signal of a preference for hate-free debates or a high concern for reputation. With linear strategies, there is a unique equilibrium ($x \geq 0$), and the expected returns for social image are the same for all users, despite them having different signals of the common value of hate-free discussions in social media and a different strength of reputational concern. When there is no variation in reputational concerns across individuals ($s_\mu^2 = 0$), the

[19]

$$U_i(v_i\ \theta_i\ w; a_i\ \bar{a}\ a_p) \equiv (v_i + \theta)a_i + (w + \theta)(\bar{a} + a_p) - C(a_i) \tag{6}$$

$$R(a_i\ \theta_i\ \mu_i) - \bar{v} \equiv E_{\bar{a},\theta\_i,\mu\_i}\Big[\int_0^1 E[v_i|a, \bar{a}, \theta_j, \mu_j]\mathrm{d}j\ |\theta_i, \mu_i\Big] \tag{7}$$

(a_p is the contribution by the principal.)

[20] A current debate concerns state regulation of anonymity on the Internet by, for example, requiring the retention of information (Froomkin 2017).

marginal return to reputation becomes a value, implying that a decrease in anonymity increases the aggregate contribution one to one. If individuals vary in their reputational concern ($s_\mu^2 > 0$), the signal of an individual writing in a less hateful way becomes less informative. The behavior could be due to reputational concerns. The marginal return to image concerns $\xi(x)$ decreases when anonymity decreases. This, in turn, will have less than a one-to-one impact on aggregate behavior.

C. R Packages Used

The regression tables in this document were created using (Hlavac 2018).

D. Online Appendix: Prediction of Hate

D.A. Prediction of Hate

The data used in the study come from text-based messages posted on Flashback. Using a custom-built script in Python, we scraped all posts (entries) in the three forums (feminism, domestic politics and immigration) from the time each respective forum started until January 2017.[21] Next, we randomly selected 100 threads from each forum and had a research assistant (RA) classify the first twelve and last five posts from the threads. The randomization was implemented at the thread level because we wanted to classify whether initial hateful content was followed by more or less hateful posts and whether a debate occurred criticizing previous posts. The RA received instructions from us with definitions of the classifications of content types—hateful content, threatening content, and aggressive content—and of the groups towards whom the user directed the hateful content—females and feminists, foreign residents, or others. The RA also classified each post according to whether the post confirmed or questioned the argument or topic discussed in the previous posts, whether the post expressed support for or against a specific political party, and whether the post contained the language of "us and them". The final data set contained 4018 classified posts, divided approximately equally across the three forums.[22] In this paper, we focus on the classification of hateful content. Please see Section D.F. in the online appendix for the full instructions to the RA.

[21] Specifically, we downloaded all posts (entries) in these forums from the start of each respective forum until the day each script ended (they ended sequentially between January 2, 2017, and February 9, 2017), except for domestic politics, from which we collected all posts after May 26, 2000. The feminism and immigration forums started later, on May 25, 2005, and July 4 2007.

[22] The data obtained from the RA contain 4040 observations; however, 22 of these are not used in the analysis because they contain only stopwords or numbers.

D.AA. Bag of Words and Logistic Lasso

To translate the text into a quantifiable measure, we used a so-called bag of words approach. First, we created a matrix containing the posts as rows and each word of the classified data as a column name. Second, we removed common stop words. Stop words are topic-neutral words such as articles and conjunctions. To reduce the dimensionality of the matrix, we also stemmed the data. Stemming is a common computer linguistic process that removes some ending characters of a word and groups similar words together. For example, words such as argues, arguing, and argue are reduced to argu. The discussions are all in Swedish, and thus, the processes of removing stop words and stemming were adapted to the Swedish language.[23] To fill the cells in the matrix with a statistic that reflects the importance of a word to a post in the data set, we estimated a weighting factor for each word in each post. The type of weighting scheme we used is called the *term frequency-inverse document frequency* (tf-idf). The value of tf-idf increases proportionally to the number of times a word appears in a post but is adjusted according to the frequency of the word in the entire data set. For example, tf-idf is a common weighting scheme in recommender systems in digital libraries. The weights in each cell are estimated using the following procedure:

$$tfidf(t_k, d_j) = \#(t_k, d_j) * log\frac{|T_r|}{\#_{Tr}(t_k)}, \tag{9}$$

where $\#(t_k, d_j)$ is the number of times the word t_k occurs in post d_j and $\#_{Tr}(t_k)$ is the number of posts in the entire data set T_r in which t_k occurs.

The tf-idf matrix comprises the right-hand-side variables in the prediction models. The left-hand-side variable is a dummy for hateful content against anyone, a dummy for hateful content against immigrants or a dummy for misogynistic content. In line with methodological practice in machine learning literature (James et al. 2013), we split the classified data set into one training set of 2812 posts (observations)—approximately 70 percent—and one test set of 1206 posts. Moreover, we removed the words that did not appear in the training data but only in the test set. We started by describing the data of the full manually classified set. Then, we ran a logistic lasso as the machine learning model using only the training set. We compared the predictions of this model with the

[23] To create the matrix and to remove the stop words as well as to perform stemming, we used the statistical software R. For stemming, we used the package *SnowballC*.

actual (true) classifications by the RA in the test data set. Then, we evaluated the model with so-called confusion matrices, representing the probabilities of correct and incorrect classifications.

We then ran a logistic lasso as the machine learning model using only the training set.[24] The lasso is a regression analysis method that performs variable selection and regularization to increase the prediction precision. The lasso reduces the coefficient estimates towards zero to balance the variance-bias trade-off, with some variable coefficients being reduced to zero. Formally, the logistic lasso computes a penalized maximization problem of the form given in equation 10.

$$\max_{\beta_0,\beta} \left\{ \sum_{i=1}^{N} \left[y_i(\beta_0 + \beta^T x_i) - log(1 + e^{\beta_0+\beta^T x_i}) \right] - \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{10}$$

Equation 10 is thus equivalent to the standard log-likelihood function for a logistic regression with an added penalty term $-\lambda \sum_{j=1}^{p} |\beta_j|$. The key parameter λ is chosen by tenfold cross-validation. Cross-validation is a method of evaluating models based on the idea of using the whole training data set during training. Using this method, we divide the training data into k subsets, and when each of the subsets constitutes a test set, the other k-1 subsets become the training set. Running the algorithm k times, each observation is in a test set once and in a training set k-1 times. Finally, the average error across all k trials is computed.[25] In this paper, we focus on the variables measuring hate in general, hate against foreign residents and misogyny as outcomes in three separate prediction models, and the tf-idf matrix comprises the regressors.

D.B. Evaluation of the Predictions of the Classified Data

When evaluating the performance of the predictions from the logistic lasso model, we focus on maximizing the sum of the true positive rate or sensitivity,

$$\frac{Truepositives}{Truepositives + Falsenegatives},$$

[24] We also ran a support vector machine model on the coded data. The lasso made better predictions, with fewer incorrect and more correct classifications. The result of this exercise is presented in Table D.11.

[25] We use R as our statistical software along with the package *glmnet* for the logistic lasso.

and the true negative rate or specificity,

$$\frac{Truenegatives}{Truenegatives + Falsepositives}.$$

Accuracy is another evaluation measure, defined as

$$\frac{True\ positives + True\ negatives}{Total\ cases}.$$

Since all our outcomes are heavily skewed towards zero, focusing on maximizing the accuracy would not yield any fruitful predictions, as the best accuracy will typically be obtained by predicting all posts as non-hateful. However, we report all three evaluation measures for each of the outcomes below. Figures D.8a, D.8b and D.8c show how we trade off the true positive rate against the true negative rate using receiver operating characteristic (ROC) curves. Intuitively, this is a trade-off between type I and II errors, where we strive to minimize the sum of these two.

Table D.8 presents the prediction results of the logistic lasso model for the hate split by the true classifications. Tables D.9 and D.10 do the same for hate against females and feminists and hate against foreign residents. Starting with general hate in Table D.8 shows that out of the 1206 cases, the lasso classifier predicted hateful content in 101 cases and no hateful content in 1105 cases. Posts that were classified by the RA as having hateful content constituted 290 cases, and 916 cases had no hateful content. The true positive rate in Table D.8 is $\frac{62}{62+228} \approx 0.214$, while the true negative rate is $\frac{877}{877+39} \approx 0.957$, implying that our prediction for general hate still makes less than 5 percent type I errors. The accuracy of the prediction is $\frac{62+877}{1206} \approx 0.779$. Proceeding to Table D.9, we see that the algorithm allows for a higher false positive rate, thus giving us a true negative rate of $\frac{969}{969+164} \approx 0.855$ and a true positive rate of $\frac{47}{26+47} \approx 0.644$. The accuracy rate is $\frac{47+969}{1206} \approx 0.843$, indicating that we are more successful in predicting hate against females compared to general hate. However, as noted above, misogyny is a more skewed variable than hate, and comparing the accuracy of the two is not very meaningful since we can obtain a higher accuracy for misogyny simply by classifying all data as non-misogynistic. In the appendix, we compare the two classifications according to the areas under the ROC curves in Figures D.8a and D.8b. The logistic lasso predicts misogyny better than hate in general; the area under the ROC curve (AUC) for general hate is 0.58, while it is 0.75 for misogyny. The AUC is a fairly standard quality-of-prediction measure in machine learning applications. A value of 0.5 implies

that we are not doing any better than chance.

Table D.10 provides the results of the prediction of the logistic lasso on hate against foreign residents. For this variable, we have a true positive rate of $\frac{59}{68+59} \approx 0.465$ and a true negative rate of $\frac{973}{973+106} \approx 0.902$. The accuracy, in turn, is $\frac{973+59}{1206} \approx 0.856$. Thus, the accuracy indicates that we can predict hate against foreign residents better than both general hate and misogyny. However, the area under the ROC curve is 0.69, implying that we can predict hate against foreign residents better than hate in general but not better than misogyny. In sum, we can conclude that general hate is our noisiest measure, and misogyny is the most precise.

TABLE D.8
Confusion Matrix from Logistic Lasso on Hate

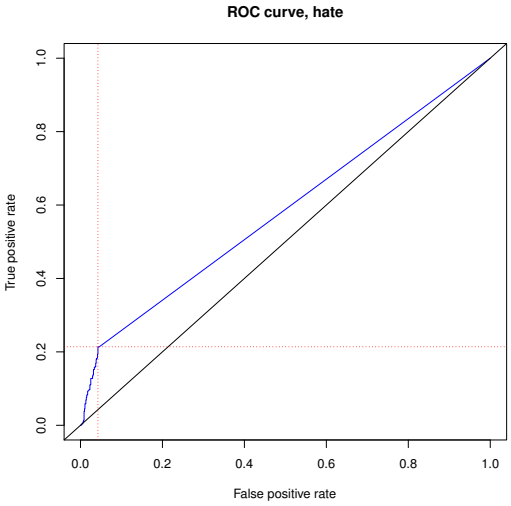
	Truth	
Predict	No hate	Hate
No hate	877	228
Hate	39	62

TABLE D.9
Confusion Matrix from Logistic Lasso on Misogyny

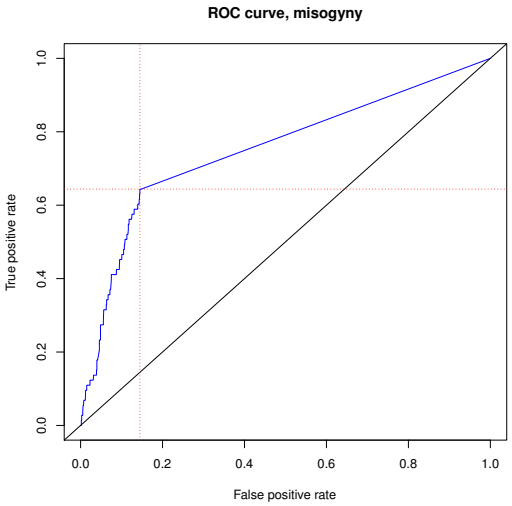
	Truth	
Predict	No misogyny	Misogyny
No misogyny	969	26
Misogyny	164	47

TABLE D.10
Confusion Matrix from Logistic Lasso on Hate Against Foreigners

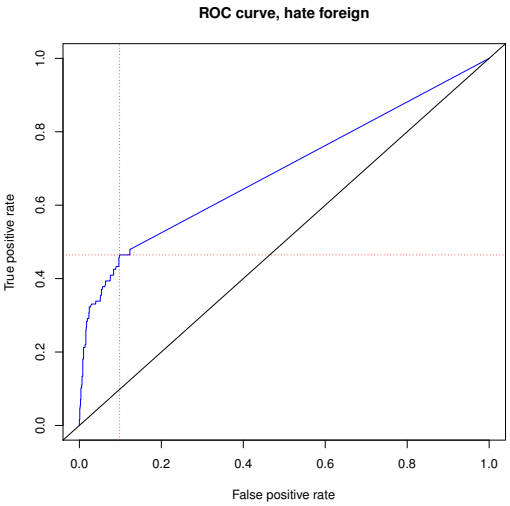
	Truth	
Predict	No hate foreign	Hate foreign
No hate foreign	973	68
Hate foreign	106	59



(A)



(B)



(C)

FIGURE D.8
ROC Curves

D.C. The Classified Subset of the Data

It is difficult to define hateful content and hate speech because it is not monolithic. The RA received instructions from us with definitions of the classifications of content types—hateful content, threatening content, and aggressive content—and the groups towards

TABLE D.11
CONFUSION MATRIX FROM SVM ON HATE

	Truth	
Predict	0	1
No hate	804	234
Hate	112	56

whom the user directed the hateful content—females and feminists, foreign residents, and others. The RA also classified each post according to whether the post confirmed or questioned the argument or topic discussed in the previous posts, whether the post expressed opinions for or against a specific political party, and whether the post contained the language of "us and them". Previous research on linguistic markers guided us in forming the classification definitions (Cohen et al. 2014). The final data set contains 4040 classified posts divided approximately equally across the three forums, but 22 of the posts are not included in the prediction model because they contain only stop words or numbers. In this online appendix, we focus on the classification of hateful content. The online appendix also includes the instructions to the RA.

Table D.12 presents the summary statistics of the data that were manually classified, presented for the full sample of classifications as well as for the samples before the event (September 2014). The vast majority of the entries and users come from the period before the event.

Among the classified data, approximately one in five posts has some hateful content; every tenth post has hateful content aimed at foreign residents, while every twentieth contains hate against females and feminists. Across the forums, the largest share of hate is found in the immigration forum, where every third post contains hateful content, and the lowest share is found in the domestic policy forum. Hate against foreign residents and hate against feminists and females are mainly found in separate forums. The results are presented in Figures D.9a–D.9c in the Appendix. Compared with the share of hateful posts, a higher share, 44 percent, of the posts were classified as aggressive, but a far lower share, one percent, contained an actual threat.

Discriminatory speech has been found to use the ideas of "them" and "us" to preserve differences between groups (Cohen et al. 2014). In our classified data set, 13 percent of the posts use "us and them" reasoning to mark differences between groups. Moreover, in regard to political content, 3 percent of the posts provide support for the right-wing populist party of the Sweden Democrats, while 7 percent are critical of all the other parties and 9 percent express support or critique of some political party in Sweden. The

data display variation in opinions; 27 percent of the posts dispute a claim in a previous post, while only 13 percent agree with a previous post.

TABLE D.12
Summary statistics, classified data

	Total, mean	Total, SD	Pre-event registered, mean	Pre-event registered, SD
No. entries	4040.00		3813.00	
No. users	2043.00		1897.00	
No. threads	300.00		299.00	
Hate	0.23	0.42	0.23	0.42
Hate against foreigners	0.09	0.29	0.09	0.28
Hate against females	0.06	0.23	0.05	0.23
Threat	0.01	0.09	0.01	0.09
Aggressive	0.44	0.50	0.44	0.50
We/them-reasoning	0.13	0.33	0.12	0.33
Disputing	0.27	0.44	0.27	0.44
Consenting	0.13	0.33	0.13	0.33
For rightwing poplusist	0.03	0.16	0.03	0.16
Against all other parties	0.07	0.25	0.07	0.25
Against left parties	0.04	0.20	0.04	0.20
Express political party opinion	0.09	0.28	0.09	0.28

All variables in the table are dummy variables.

The share of hateful entries against women and against foreigners do not sum to the total share of hateful entries due to the fact that there is also general hateful comments and hate towards particular individuals, such as politicians and celebrities, in the data as well.

Table D.13 presents the same summary statistics for the main variables in Table D.12 but breaks down the statistics for each treatment group and before and after the event. From this table, we can compute the simple difference-in-difference (DD) estimate for hate as $(0.14 - 0.21) - (0.28 - 0.23) = -0.12$, while the same estimate for hate against foreign residents is $(0.03 - 0.08) - (0.11 - 0.10) = -0.06$ and that for misogyny is $(0.04-0.03)-(0.09-0.06) = -0.02$. Furthermore, for threats, the estimate is $(0.00-0.01)-(0.01-0.01) = -0.01$, and for aggression, it is $(0.29-0.42)-(0.46-0.44) = -0.15$. Table D.14 in the Appendix provides the coefficients and standard errors for the corresponding regressions. Unsurprisingly, the estimated effects are all of the same magnitudes as the computations we just performed, though only the estimates for hate, hate against foreign residents and aggression are significant at the 5 percent level. It is worth noting that threats come remarkably close to being significant at the 5 percent level, though a single threatening post in the treatment group in the post-period would be enough to offset the negative coefficient completely.

TABLE D.13
Summary statistics by treatment/control, classified data

	Treated, pre	Treated, post	Control, pre	Control, post
No. entries	1065.00	96.00	1973.00	906.00
No. users	500.00	54.00	1053.00	504.00
No. threads	182.00	37.00	182.00	85.00
Hate	0.21	0.14	0.23	0.28
Hate against foreigners	0.08	0.03	0.10	0.11
Hate against females	0.03	0.04	0.06	0.09
Threat	0.01	0.00	0.01	0.01
Aggresive	0.42	0.29	0.44	0.46

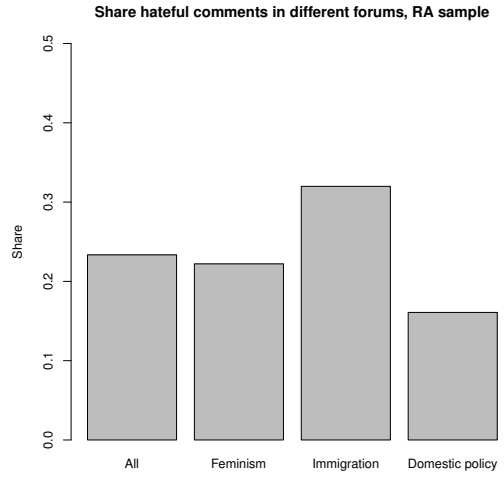
All variables in the table are dummy variables.

The share of hateful entries against women and against foreigners do not sum to the total share of hateful entries due to the fact that there is also general hateful comments and hate towards particular individuals, such as politicians and celebrities, in the data as well.

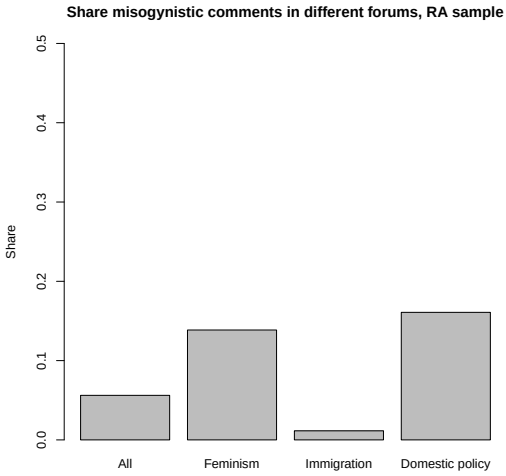
TABLE D.14
Regressions using only RA data

	Dependent variable:				
	Hate	Hate foreign	Misogyny	Threat	Aggression
	(1)	(2)	(3)	(4)	(5)
Post*Treated	−0.127	−0.064	−0.023	−0.010	−0.143
	(0.043)	(0.025)	(0.024)	(0.005)	(0.061)
Post reveal	0.050	0.011	0.034	0.0002	0.018
	(0.020)	(0.015)	(0.012)	(0.004)	(0.024)
Treated	−0.015	−0.012	−0.024	0.001	−0.028
	(0.019)	(0.013)	(0.009)	(0.004)	(0.023)
Constant	0.229	0.096	0.055	0.009	0.444
	(0.011)	(0.008)	(0.006)	(0.002)	(0.014)
Observations	4040	4040	4040	4040	4040

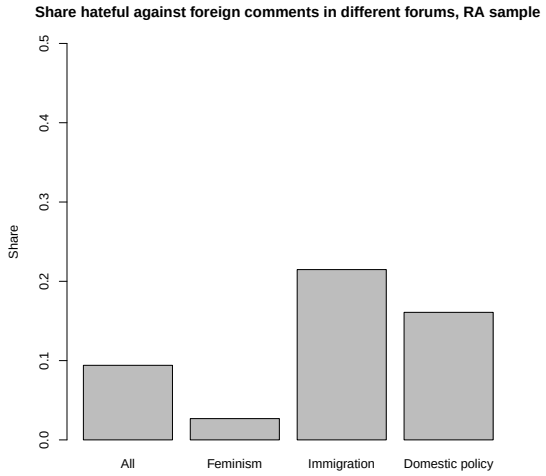
Note: Standard errors clustered at the user level.



(A)



(B)



(C)

FIGURE D.9
Forum Shares, RA Sample

D.D. Measurement Error

Here, we outline a simple framework for how we think about measurement error in our setting. Suppose there is some unobserved true amount of hate in a post, denoted by $Y^\star$. Furthermore, a post is deemed hateful ($Y = 1$) by a representative individual if $Y^\star$

exceeds some unobserved threshold c;

$$Y = \begin{cases} 1, & \text{if } Y^\star \geq c \\ 0, & \text{if } Y^\star < c \end{cases} \tag{11}$$

In addition, we are interested in the effect of a binary treatment indicator x taking the value 1 in case of treatment (no anonymity) and zero otherwise (anonymity). In other words, we have a standard binary variable as an outcome setting, and we are interested in the relationship

$$Y = \alpha + \beta x + \varepsilon. \tag{12}$$

Now, let us further suppose that we do not observe Y either, due to feasibility constraints, but rather the coding of our RA, $\tilde{Y}$, given by:

$$\tilde{Y} = \begin{cases} Y, & \text{with probability } 1 - q_{01} - q_{10} \\ 1, & \text{if } Y = 0 \text{ with probability } q_{01} \\ 0, & \text{if } Y = 1 \text{ with probability } q_{10} \end{cases} \tag{13}$$

In other words, with probability q_{10}, we classify a hateful entry as non-hateful, and with probability q_{01}, we classify a non-hateful entry as hateful. By construction, the relationship between Y and $\tilde{Y}$ is less than one, as either $\tilde{Y} = Y$, i.e., a one-to-one relationship, or $\tilde{Y} = 1 - Y$, i.e., a negative measurement error. A regression of $\tilde{Y}$ on Y of the form:

$$\tilde{Y} = \theta_0 + \theta_1 Y + \varphi \tag{14}$$

will thus give us $\theta_0 > 0$ and $\theta_1 < 1$. We hence have a mean reversion measurement error on the left-hand side in our regressions, which will attenuate our estimate towards zero.[26] The feasible population regression function in our setting is thus

$$\tilde{Y} = \tilde{\alpha} + \tilde{\beta} x + \tilde{\varepsilon}. \tag{15}$$

Furthermore, we also use the predicted values from the logistic lasso as our outcome variable in our main estimates. We consider this an additional misclassification problem for the machine learning model. We do not, however, focus on where this error stems

[26]This hinges on the assumption that $\theta_1 > 0$, which means that there is a positive, though imperfect, relationship between $\tilde{Y}$ and Y. In other words, our RA needs to outperform random chance. If, however, the RA does not outperform random chance, so that $\theta_1 < 0$, our estimates will have reversed signs. This would indeed be a great problem, but given the situation at hand, we do not consider it a potential issue.

from. Thus, we have:

$$\bar{Y} = \begin{cases} \tilde{Y}, & \text{with probability } 1 - p_{01} - p_{10} \\ 1, & \text{if } \tilde{Y} = 0 \text{ with probability } p_{01} \\ 0, & \text{if } \tilde{Y} = 1 \text{ with probability } p_{10} \end{cases} \tag{16}$$

implying that we will misclassify a greater share of entries. Hence, this will further attenuate our estimate towards zero. The actual regressions we run are thus of the form

$$\bar{Y} = \bar{\alpha} + \bar{\beta}x + \bar{\varepsilon}. \tag{17}$$

We can then, however, note that

$$\bar{\beta} = P[\bar{Y} = 1|x = 1] - P[\bar{Y} = 1|x = 0] \tag{18}$$

$$= (1 - p_{01} - p_{10})(P[\tilde{Y} = 1|x = 1] - P[\tilde{Y} = 1|x = 0])$$

$$= (1 - p_{01} - p_{10})\tilde{\beta} \Rightarrow$$

$$\tilde{\beta} = \frac{\bar{\beta}}{(1 - p_{01} - p_{10})} \tag{19}$$

Now, since we evaluate the performance of our prediction model with the test set, we actually obtain estimates of both p_{01} and p_{10}, which we can use to recover $\tilde{\beta}$ from our obtained $\bar{\beta}$. In other words, we can take the measurement error from the prediction part into account when we compute our estimates, but we cannot take into account any potential faulty classification by our RA. Thus, we cannot obtain β, only $\tilde{\beta}$. As an alternative approach, we then cross-check the classification of our RA against that of one other human for a subsample of the data classified by the RA. Finally, it is worth noting that equation 19 implies that since we minimize the sum of type I and II errors in the prediction part, we also minimize the attenuation bias in our treatment estimates.

D.E. Example of Adjusting Coefficients Using the False Positive and False Negative Rates

In Section D.B., we find a true positive rate of 0.214 for hate. This in turn gives us a false negative rate of $1 - 0.214 = 0.786$. We also have a true negative rate of 0.957, giving us a false positive rate of $1 - 0.957 = 0.043$. The coefficient in turn is -0.015. Thus, adjusting for the false classifications gives us an estimated effect of

$-0.015/(1-0.786-0.043) = -0.088$. For hate against foreign residents, we have a true positive rate of 0.465, giving us a false negative rate of $1 - 0.465 = 0.535$ and a true negative rate of 0.902 and thus a false positive rate of $1 - 0.902 = 0.098$. We thus obtain an estimate of $-0.027/(1 - 0.535 - 0.098) = -0.074$. Finally, for misogyny, we have a true positive rate of 0.644, giving us a false negative rate of $1 - 0.644 = 0.356$ and a true negative rate of 0.855 with a false positive rate of $1 - 0.855 = 0.145$. Thus, our estimate is $0.032/(1 - 0.356 - 0.145) = 0.064$.

D.F. Instructions to the RA

Instructions to Research Assistant Spring 2017.

The population we investigate is from the Internet forum Flashback. We have scraped text from the following three sub-forums; immigration, feminism and domestic policy. We have drawn a random sample of 100 threads from each forum. Each thread and each post has an id-number. We want you to code 12 posts starting from the beginning of the thread and 5 posts starting from the end of the thread.

The unit of coding is the post. Please read the full post. You will receive the threads and posts in an Excel sheet, where we want you to insert your classifications. Below you can find descriptions of how we want you to classify the posts.

Start with 2 threads and after this we can meet to discuss the progress before you proceed.

	Responds to Svarar på	
	0 = Doesn't seem to respond to any particular post 999 = Response to several posts from several authors [tomt] = Responds to a post that's not in the sample	*The value noted here is the id-number of the post to which the writer seems to respond.*
1	Questioning Ifrågasättande	
	0 = Neither nor 1 = Affirmative 2 = Nuancing 3 = Questioning	*- If the post **quotes another post** the coding relates to the quoted post.* *- If the post **doesn't contain a quote** the coding relates to the first post.* *- The **first post is always coded** as "neither-nor"*
2	Understanding Förståelse	
	0 = No 1 = Yes	If the writer shows understanding of the thoughts and intentions expressed in an earlier post, the coding should be "yes". **Regardless of whether the writer agrees or not.**
3	Party politics positive. *Does the post express a opinion in favor of any party or coalition of parties or Feminist Initiative?*	
	0 = No, not positive to any coalition of parties 1 = Yes, the red-green coalition 2 = Yes, the liberal-conservative coalition 3 = Yes, the Sweden Democrats 4 = Feminist Initiative 5 = Feminist initiative and the Left Party 6 = The seven traditional parties 7 = Sweden Democrats and the Right.	***Only to be coded "yes" if it is obvious,** e.g. when the parties or their representatives are mentioned, either explicitly or through paraphrases*
4	Party politics negative. *Does the post express a **negative** opinion of any party or coalition of parties or Feminist Initiative?*	
	0 = No, not positive to any coalition of parties 1 = Yes, the red-green coalition 2 = Yes, the liberal-conservative coalition 3 = Yes, the Sweden Democrats 4 = Feminist Initiative 5 = Feminist Initiative and the Left Party 6 = The seven traditional parties 7 = Sweden Democrats and the Right.	***Only to be coded "yes" if it is obvious,** e.g. when the parties or their representatives are mentioned, either explicitly or through paraphrases*

	Aggressiveness (tone)		
	0 = Not at all aggressive 1 = Partly aggressive 2 = Predominantly aggressive.	*In what **tone** is the post as a whole written?* *- If **some part** of the post is read as aggressive, it should be coded **partly aggressive**.* *- If the post contains **mostly aggressive** text, it should be coded **predominantly** aggressive.*	
	HATRED HAT		
6	Another Flashback user	0 = No 1 = Yes	*If the post contains **words or statements** that indicate persecution (in the broad sense of the term) of **a group or an individual,** it should be coded ”yes”, Possible examples are:* *- threat* *- expressions of disrespect* *- insults* *- verbal violations* ***Use the coding ”yes” also for isolated hateful statements. It doesn’t have to be blatant.***
7	Specific public person		
8	Persons with specific sex/gender		
9	Persons who were born abroad, or whose parents are born abroad		
10	Persons with a specific ethnicity		
11	Persons with a specific sexual inclination		
12	Persons with specific skin color		
13	Something else		
14	If the hatred is pointed toward something else, please specify	Text	
	HOT		
15	Another Flashback user	0 = No 1 = Yes	*Does the post contain words that are explicit threats or assault? Assault means that someone threatens to **harm an individual or his or her property.** The assault **can be directed to other persons, animals or objects** that are important to the individual.*
16	Specific public person		
17	Persons with specific sex/gender		
18	Persons who were born abroad, or whose parents are born abroad		
19	Persons with a specific ethnicity		
20	Persons with a specific sexual inclination		
21	Persons with specific skin color		
22	Something else		
23	If the threat is pointed toward something else, please specify	Text	
24	Male preference		
	0 = No 1 = Yes	If the post contains **words that in any way state the superiority of men** over women, it should be coded ”yes”	
25	Female preference		
	0 = No 1 = Yes	If the post contains **words that in any way state the superiority of women** over men, it should be coded ”yes”	
26	Gender equality preference		
	0 = Nej 1 = Ja	If the post contains **words of men and women being equal**, it should be coded ”yes”	
27	Attitudes towards foreigners		
	0 = No opinion 1 = Positive attitude 2 = Neutral attitude 3 = Negative attitude	By foreigners is meant people who are **born abroad or whose parents were born** abroad.	

28	Gender - disadvantaged	
	0 = No opinion 1 = Men are disadvantaged 2 = Women are disadvantaged	If the post contains words that express women as disadvantaged or men as disadvantaged, it shall be coded as 1 or 2 respectively.
29	Ethnicity - disadvantaged	
	0 = No opinion 1 = Swedes are disadvantaged 2 = Immigrants are disadvantaged	If the post contains words that express Swedes as disadvantaged or immigrants as disadvantaged it shall be coded as 1 or 2 respectively.
30	Us and Them	
	0 = No 1 = Yes	If the post explicitly contains a language of ”us and them” or clearly expresses an in-group out-group view the variable should be coded yes.
31	Sarcasm or irony	
	0 = No 1 = Yes partly 2 = Yes, fully	If the post contains sarcasm or irony in part or in full it should be coded yes.