Selected Papers of #AoIR2018:
The 19th Annual Conference of the
Association of Internet Researchers
Montréal, Canada / 10-13 October 2018

CYBERHATE ANONYMITY AND THE RISK OF BEING EXPOSED

Emma von Essen
Swedish Institute for Social Research, Stockholm University
Department of Economics and Business, Aarhus University

Joakim Jansson
Department of Economics, Stockholm University
Research Institute of Industrial Economics

Introduction

The digitalisation of our daily lives alters ways of civic discussions, which ultimately can affect our economic and social decision making. The traditional speakers' corners we now find in anonymous discussion forums online rather than on public squares. A crucial aspect of freedom of expression is anonymity, but anonymity is a contentious matter (e.g. Nagel and Frith 2015). On the one hand more anonymity enables individuals to discuss without fear of repercussions, but on the other hand, anonymity can lead individuals to involve in hateful writings threatening other's freedom, due to lack of accountability. Today online discussion forums are infamous for hate and harassment, and this disturbs how we take part in discussions. Also, much of the hate and harassment is directed against women and marginalized groups (e.g. Citron 2014). To find effective policies when combating hate online, understanding the empirical link between anonymity and hate is vital. This paper aims to quantify how anonymity links to writing hateful content in online.

The literature on the economics of transparency is diverse, theoretical and empirical research suggests that less transparency can give benefits as well as losses depending on the context (Acquisti et al. 2016). There is a lack of quantitative research joining anonymity and hate online. To our knowledge, there is only one previous paper and they investigate changes in offensive language caused by a real name policy in Korea (Cho et al. 2012). The main advantage of our study is that our design has a control and treatment group, which to partly avoids self-selection on non-observables.

Suggested Citation (APA): von Essen, E. and Jansson, J. (2018, October 10-13). *Cyberhate and the risk of being exposed*. Paper presented at AoIR 2018: The 19th Annual Conference of the Association of Internet Researchers. Montréal, Canada: AoIR. Retrieved from http://spir.aoir.org.

The context of our study is discussions on political topics at the Swedish discussion forum Flashback (similar to Reddit), where anonymity between end-users and between platform and user is an enforced rule. The forum is one of the most significant discussion platforms in Sweden. We elicited text from three sub-forums encompassing discussions about political topics: feminism, immigration and domestic policy. Participation in these discussions is likely to have an impact on individual decision making, about for example voting. Political topics are previously known to attract hate (Cheng et al. 2017; Citron 2014).

As a first step, we use a supervised machine-learning model to predict hateful content. Then we test how an unexpected exogenous change in anonymity affects hateful posting, using a so-called difference in difference model. In September 2014, the identity of one-third of the Flashback accounts registered before March 2007 was unexpectedly in the hands of journalists, and the journalists did publicly expose the identities of handful individuals. This event creates a quasi-experiment, using accounts that are registered before March 2007 as a treatment group, and as a control group, we use accounts registered after March 2007.

Data collection and Empirical strategy

We scraped the posts from all threads in the three sub-forums. The sample, we restricted to the period from January 1, 2012, until December 31, 2016. A research assistant manually classified posts from a random subset of the threads.

The entries were classified by whether it contained hateful writings and whether the hate was directed towards foreigners, females and feminists or others (classifications are inspired by Cohen et al. 2014). We use a simple bag of words model to make the text quantifiable and produce three Logistic Lasso prediction models. The three models is used to predict hate in the full data set.

To quantify the link between anonymity and hateful speech we used the predicted hate from the Lasso models to estimate the differences in changes of hateful writing before and after the event comparing the experimental and control group (see Hansen et al. 2014 for a similar methodology).

Results

Our full sample comprises roughly 48 000 users and 1 980 000 entries. The share of hateful content is 7 %, whereas the percentages of entries with hate against foreigners and females is 16 % and 14 % respectively. These shares are similar for the treatment and control group. Turning to the event, we find that an unexpected decrease in anonymity leads to a lower share of hateful posts in the discussions. The treatment group has about 1.5 percentage point lower probability of writing hateful comments in the post period compared to the control group. However, when we look at the hate directed towards specific groups, we find the share of hate against foreigners to decrease, whereas the share of misogyny increases. The changes are both in the number of hateful entries and the number of total entries. Restricting the sample to

users registered before the event, we see confirm that results are not driven by inflow or outflow of users.

To probe further into the findings, we look at the users we can follow before and after the event. Here, we find no effect in changes of the amount of misogynistic posts, but changes in the number of hateful entries against foreigners. The shares are however unaltered. To understand why we find an increase in the share of misogyny as well as no decrease in the amount of misogynous entries when looking at individuals we can follow, we also assess if users substitute hate. Indeed, we find that decreased hateful entry towards foreigners is partly substituted with increased hate against females.

Discussion

Our study contributes to the current policy debates on how to combat online hate. We quantify the link between anonymity and hateful content and find that reduced anonymity leads to reduced hate in online discussions about political topics. The effect seems to be driven by a combination of a decline in writing hateful posts, a decrease in the number of non-hateful posts and a substituting hate against foreigners with hate against females. One possible explanation for the substitution is that hate against foreigners is associated with previous convictions of hate speech in Sweden. Our results open up for an exciting avenue of research on how to understand different types of hate online.

Our results support the conclusion that a real name policies will not be effective in combating hateful content online, since it may create adverse effects for example substitution of hate from one target group to another. Also, a real name policy may decrease the general discussion activity. Fewer individuals that discuss politics online could lower political accountability (Strömberg 2015).

References

Acquisti, A., Taylor, C., & Wagman, L. (2016). The economics of privacy. *Journal of Economic Literature*, *54*(2), 442-92.

Citron, D. K. (2014). *Hate crimes in cyberspace*. Harvard University Press.

Cho, D., S. Kim, and A. Acquisti (2012). *Empirical analysis of online anonymity and user behaviors: the impact of real name policy*. In System Science (HICSS), 2012 45th Hawaii International Conference on, pp. 3041:3050. IEEE.

Cohen, K., Johansson, F., Kaati, L., & Mork, J. C. (2014). Detecting linguistic markers for radical violence in social media. *Terrorism and Political Violence*, *26*(1), 246-256.

Hansen, S., McMahon, M., & Prat, A. (2017). Transparency and deliberation within the FOMC: a computational linguistics approach. *The Quarterly Journal of Economics*.

Van der Nagel, E., & Frith, J. (2015). Anonymity, pseudonymity, and the agency of online identity: Examining the social practices of r/Gonewild. *First Monday*, *20*(3).

Strömberg, D. (2015). Media and politics. *economics*, *7*(1), 173-205.