Chilling Effects: Online Surveillance and Twitter use

Vignoud Julien, Lepeytre Hugo, Benhaim Julien

***Abstract*—In this report we extend an analysis conducted in a paper titled "Chilling Effects: Online Surveillance and Wikipedia use" by Jonathon W. Penney. [1] While Wikipedia is analysed in the latter, we conduct the analysis on the social network Twitter. Our aim is to check if the Chilling Effects observed on Wikipedia page views is also observable on tweets related to privacy-sensitive subjects.**

I. Introduction

The original paper was interested in studying what is called "Chilling Effects": a tendency for users to change their internet consumption habits when they learn about government surveillance. The idea is that even if their actions are completely legal, they are afraid that the State may not protect their rights correctly and thus refrain from activities such as reading terrorism-related articles on Wikipedia even if they are otherwise completely legal. The paper studied the changes in Wikipedia page views on privacy-sensitive subjects after the NSA surveillance revelations by Edward Snowden [2], which made the general public aware of government surveillance on the Internet.

In this extension we decided to analyze a social network because we are looking for new insights as it may represent a different sample of users compared to Wikipedia. In addition, tweeting is a way for users to express their opinion and exercise their right to free speech, which is the kind of behavior that we expect to see changing due to Chilling Effects, especially after revelations of government monitoring. The reason we chose Twitter is because Tweets are a public form of Internet use and so users might think more carefully about what they express there. At the same time, it is completely legal to use those words in public so reservations that users may have on talking about these subjects are more likely to be due to Chilling Effects.

To conduct this extension, we propose to use a new dataset that we collected ourselves composed of every tweet containing any of the keywords listed in the original paper. While the original time-frame is from January 2012 to August 2014, we will extend ours to December 2014 in order to have a more reliable analysis of the long term effect.

II. Models and Methods

A. Data Scraping

We started by scraping Twitter using the Python library Twint [3]. We collected every tweet containing one of the selected keyword during the years 2012, 2013 and 2014. This represents 9 gigabytes of text data. Because of the RTT between web requests, and the delays introduced by the library, the scrapping was very long. Twint does not support parallelism out of the box. To speed up the process, we split the work into small parts by keyword and by time periods, going from one week to a month. We then launched hundreds of threads at the same time to make full use of our network bandwidth as well as our CPU computing power. Parallelism has been a key improvement in our data collection process, making the required scraping time go from weeks to a few days.

B. Collected data

There were a few adjustments we had to make during the data collection phase:

Firstly, while in the original paper all keywords translated pretty well to a Wikipedia page, it is not the case for tweets as a word can have different meanings and be used in different context. In the original paper, some of the pages selected are disambiguation pages or very general pages[1]. An improvement in the method could have been to select more precise pages for some keywords, or exclude them. But here we are scraping the whole of Twitter and so we cannot discriminate between relevant and irrelevant uses of the keyword. For this reason, we have to exclude some of them from the list. For example vague keywords such as 'recruitment' or 'attack' will be used on Twitter on a very wide range of subjects, most of which will have nothing to do with terrorism. After sorting those out, we were left with a smaller list of relevant keywords.

C. Data exploration

Once the keyword list was filtered and the scraping was finished, we were left with a big database of Tweets. For the count analysis we then created a new file containing the number of tweets each day for each keyword in a similar fashion as the original article. We started by performing an exploratory data analysis on the raw dataset, containing about 14 millions tweets. This part enabled us to get some useful insights about the data we hold. In Figure 1 we can visualize the most cited keywords in tweets of the terrorism-related dataset. For the domestic dataset, we can see the most represented keywords in the dataset in Figure 2.

[1]As an example, for the keyword "Recruitment" the authors selected the Wikipedia page https://en.wikipedia.org/wiki/Recruitment which is simply a definition of a general recruitment process. An alternative woul thed havye e could have taken it out, or selected something closer to terrorism such as this: https://en.wikipedia.org/wiki/Terrorism_and_social_media

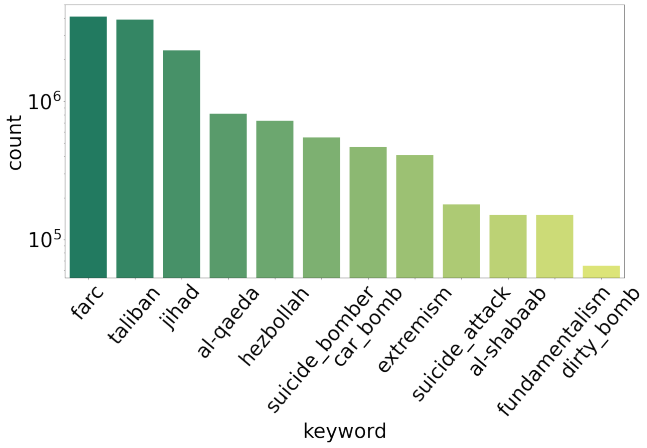

Figure 1. Tweet counts of Top 12 keywords for terrorism dataset

Figure 2. Cloudword of the most common words in the tweets of the domestic dataset.

Other interesting properties available in the dataset are related to what we call ”engagement” on social media. This word refers to users' interactions with publications. The metrics available are, for each tweet : the retweets, replies and likes count. We compute percentile of these values for the whole dataset and observe that the vast majority of tweet it contains generate no or almost no reaction. The figures are available in table I.

D. Tweet count analysis

Our analysis is an ITS (Interrupted Time Series) [4] where the interrupting event is the Snowden Revelations, like in the original paper. For both keyword list we count the number of tweets during each month, and then compare the results of the ITS analysis for both datasets. As in the original paper, the security-related dataset is assumed to be unaffected by Chilling Effects, it effectively acts as a quasi-control group. As for the Interrupted Time Analysis itself, it is a linear regression taking into account three explanatory parameters: the number of months since the beginning of the study period, whether or not the current month is after the interrupting event, and the number of months since the interrupting event (0 if we are before said event). In the end, the regression formula is:

$$Y_t = \beta_0 + \beta_1 \times time + \beta_2 \times intervention + \beta_3 \times postslope + \epsilon_1$$

Where the β are the regression parameters and ϵ_1 is the residuals.

Some obstacles and clarifications we have to make about our method are as follows:

First, Twitter usage follows a general trend which might be stable, increasing or decreasing. We have to make sure that what our analysis is going to capture is different from this general Twitter trend. This is what the quasi-control group with security-related keywords is here for. In the original paper, these keywords are chosen especially because they should not be affected by Chilling Effects. So the trend that we are going to capture in the group is unaffected by the Snowden Revelations and thus its variations will only reflect the global Twitter trend, especially in subjects close to those we actually want to study. If the trend in the terrorism-related group is different from the security-related group we will be able to know that there is something more than a simple global trend.

Secondly, what about Tweets that are not visible anymore in 2020? This includes deleted tweets as well as banned or deleted accounts. There are three different kinds of cases here: first if an account was deleted by the user or banned by Twitter for a reason unrelated to government surveillance. This is not a problem because it will be uniformly distributed over both the terrorism and security groups since there is not reason for an account who tweeted something legal about terrorism to get banned more often for unrelated reasons. Second: for tweets that were deleted by users out of fear of government surveillance. This is something that interests us because the drop in representation in those tweets in our dataset will be a direct consequence of Chilling Effects, so the bias introduced by those deleted Tweets is actually something we expect to observe and is analogous to users refraining from Tweeting out of the same fear of government surveillance. Third, tweets or account that were banned by Twitter because of illegal terrorism-related speech: those are not of interest to us because what we want is to study only the exercise of legal free speech. So it is a good thing that Twitter deletes all of the illegal speech [5] because it would only add noise to our dataset.

In the original article, another global trend dataset was used to as another form of control group: the Top Wikipedia

Table I
DISTRIBUTION OF ENGAGEMENT METRICS IN THE DATASET

Percentile	0.75	0.9	0.95	0.99	0.995	0.999	0.9999
Retweets	0	1	3	12	23	85	395
Replies	0	1	1	3	6	21	83
Likes	0	0	1	4	7	25	128

articles from the studied years. But the paper clearly mentions that it is "for illustrative purposes" [1] (p.41) and this is because a group with subjects closer to terrorism is much more likely to be affected by the same variations as our terrorism group, apart from the studied Chilling Effects

E. Sentiment analysis

After reproducing the ITS on the number tweets containing the keywords, we want to exploit the nature of our dataset since it is composed of text in which users express their opinions.

An idea that comes to mind with social media is Natural Language Processing. More precisely, we focused our study on sentiment analysis. An hypothesis is that we could observe a shift in the global sentiment tendency about privacy sensitive subject, potentially due to a chilling effect. For instance people would be less likely to talk positively or to appear enthusiastic about these topics if they know they are being monitored by their government.

To perform our analysis, we used the Vader library [6]. We labeled every tweet of the dataset with a proportion for each sentiment: positive, neutral, negative. From these values we can compute a compound value such than a value below -0.5 is negative, between -0.5 and 0.5 is neutral and above 0.5 is positive. With this compound value that takes into account all three sentiments we can analyse the evolution of the sentiments expressed in the tweets over time. For that, we first have to compute the compound value of every tweet, then we can compute the monthly average value for each keyword.

After this filtering, we can aggregate all articles to compute the overall trends. By performing this process on both the terrorism-related keywords and domestic security-related keywords we can compare the privacy-sensitive keywords and the control group. A stable trend for the domestic security keywords as well as a sudden change around the June 2013 revelations and a different slope after the intervention would show the presence of a chilling effect.

III. Results

In the Tweet count analysis, we found that for the terrorism-related group, there is a significant drop ($p = 0.033 < 0.05$) in the count after the Snowden revelations but no change in the trend (coefficient of -0.0257 and p-value of 0.551). Since in the control group the trend does not change either, the results point towards the absence of a long-term effect on the tweets. On the other hand, the drop detected in the terrorism group is absent in the security-related group, and can thus be attributed to the interrupting event. (For the control group, p-values are above 0.2)

For the sentiment analysis, we could find no significant difference in the trends of the two groups. Most sentiments have been found to be neutral and very steadily so.

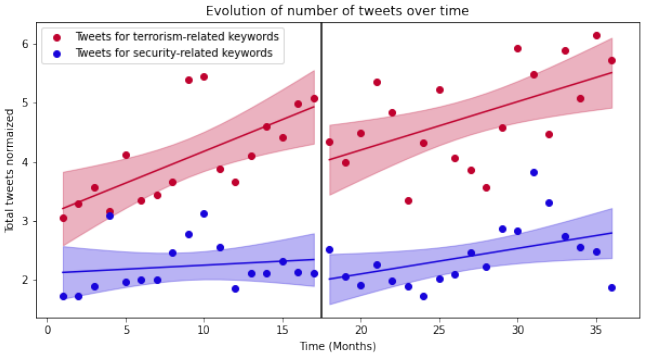

Figure 3. Tweets per month over time for both datasets

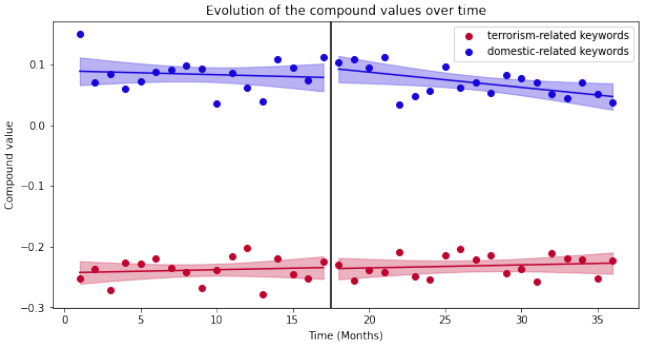

Figure 4. Sentiment over time for both datasets

IV. Conclusion

Unlike Penney's original paper, the data suggests an absence of long-term chilling effects both on the volume of tweets as well as on the general sentiments expressed about these subjects. On the other hand, the same kind of temporary volume drop was observed on Twitter and Wikipedia. The existence of such a drop has far-reaching implications both in terms of surveillance politics and economics. An example of this would be the lawsuit between Wikimedia Foundation and the NSA which was already mentioned in the original paper and as of today is still in appeal [7]. This suggests directions for further research both in terms of uncovering those political and economic consequences as well as in studying various other online services to find out if Chilling Effects can be observed in all kinds of Internet activity.

References

[1] J. W. Penney, "Chilling effects: Online surveillance and wikipedia use."

[2] B. Gellman, "Edward snowden, after months of nsa revelations, says his mission's accomplished," *The Washington Post*. [Online]. Available: https://www.washingtonpost.com/world/national-security/edward-snowden-after-months-of-nsa-revelations-says-his-missions-accomplished/2013/12/23/49fc36de-6c1c-11e3-a523-fe73f0ff6b8d_story.html

[3] T. Team, "Twint project." [Online]. Available: https://github.com/twintproject/twint

[4] R. McDowall, David; McCleary, *Interrupted Time Series Analysis*. SAGE Publications, Inc, 1980.

[5] Twitter, "Violent organizations policy." [Online]. Available: https://help.twitter.com/en/rules-and-policies/violent-groups

[6] E. Hutto, C.J.; Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text." [Online]. Available: https://pypi.org/project/vaderSentiment/

[7] A. Buatti, Jim; Palmer, "District court rules for government in wikimedia foundation's mass surveillance case against the nsa," *Wikimedia Foundation*. [Online]. Available: https://wikimediafoundation.org/news/2019/12/17/district-court-rules-for-government-in-wikimedia-foundations-mass-surveillance-case-against-the-nsa/

V. Appendix

As another example of NLP we can visualize how the keywords are related to each other in Figure 5: two keywords are linked if they appear in the same tweet and the more they appear together, the more intense the color of the link is.

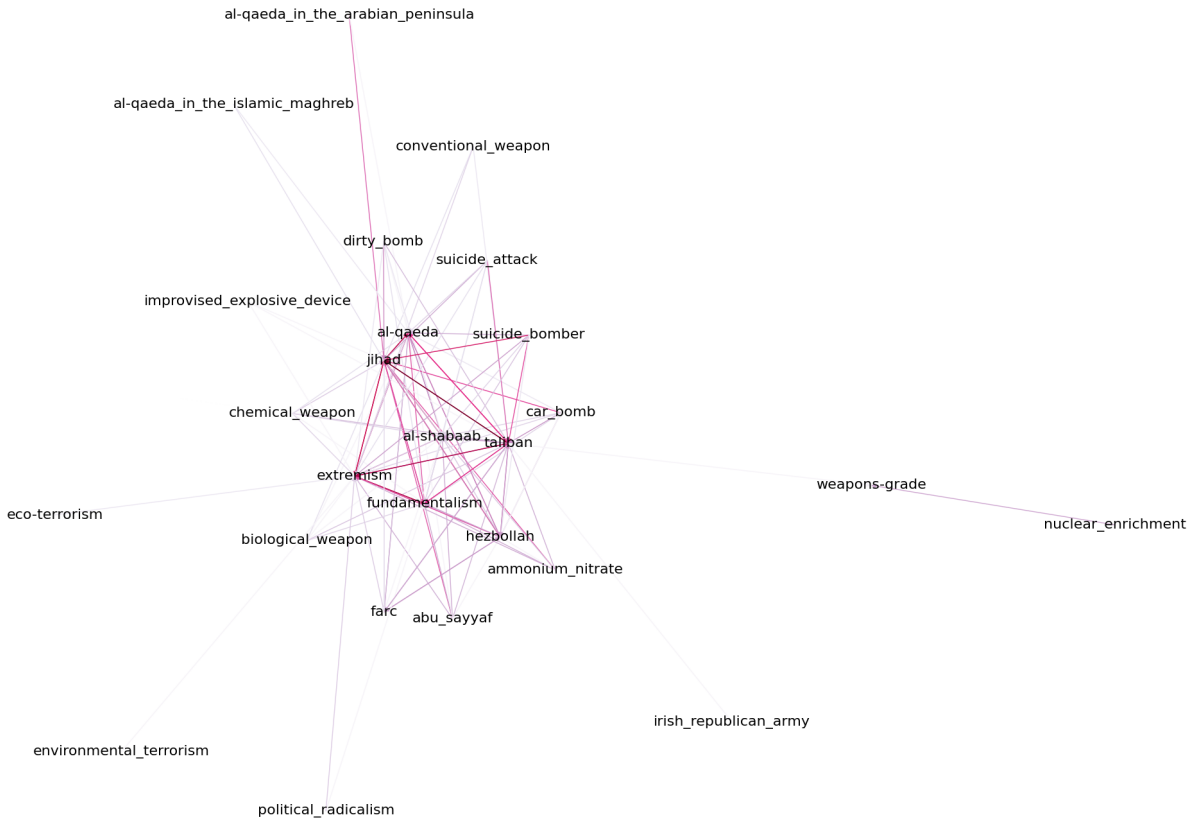

Figure 5. Dependency network between terrorism-related keywords. Two keywords are linked if they appear in the same tweet. The intensity of the edge color reflects such number of tweets. We can see in the center a cluster of words that often appear together and farther away keywords that are less commonly associated with others. It is interesting to notice how keywords are associated, such as ”nuclear enrichment” and ”weapons grade” or ”environmental terrorism” and ”biological weapon”.