Final Project: Group 5. Effect of Russian Trolls on the 2016 U.S. Presidential election.

Ryan Appel, Angel Arribas Lopez, Drew Breyer, Humza Kahn,
Devanshi Patel, Hanna Shin, Poonam Siyag

Impact of a cyberthreat on important outcomes

Existing Data

Describe the scale of the problem using existing data / opinions from expert reports or academic papers

> This problem was extremely concerning for the United States and governments around the world, and as a result, there are many academic sources and expert reports which looked into the impact that trolls had on the 2016 election. One example is out of Tennessee-Knoxville which analyzed 770,005 tweets in English from known Russian troll accounts, alongside data from FiveThirtyEight's archive of multiple polling outlets.
>
> Some of their findings were alarming, such as that for every 25,000 retweets "Russian accounts correlated to a 1% increase in Trump's poll numbers one week later" as well as seeing that "25,000 retweets would average around 10 retweets per tweet." (Ruck et al. 2019) Looking at Trump's opponent and similar social media activity, this kind of retweet volume did not have an impact on her numbers.
>
> The study additionally looked at how much of this social media activity could be attributed to bots and how much was genuine/organic social media reach. The study found that "91% of first retweeters of known Russian bots were non Russian bots, which suggests that propaganda spread into the networks of real U.S. citizens."

Who is the target of the cyberthreat?

> There were several "targets" of this disinformation campaign. Broadly speaking, the campaign was to sew discord and uncertainty into U.S. politics and the election. More specifically, the campaign clearly targeted Donald Trump and Hillary Clinton, to improve the likelihood of Donald Trump being elected as President of the United States. As we now know, this was successful we can see by the results of the election.

Do we know how many targets are affected in the population? What is the proportion of the target group in the overall population of interest?

> This is an extremely difficult task to know how many targets were affected. This is due to how social media works, as there are many links and ads and links/ads within those, which can then be shared with additional parties. Speaking to tweets coming from known Russian troll accounts, which would be far less than the true targeted population, one reputable study found there were 770,005 tweets from these accounts, which were then of course spread throughout the social media networks.
>
> Additionally, the more difficult task here is to know who was truly impacted by the social media campaign. That is, who was exposed to Russian bot activity AND changed their vote or interest in the election as a result of this material.

Ideal Conditions

Describe the ideal data you would need to obtain a precise estimate of scale of the problem (proportion of the targeted units in the overall population of interest)

> The ideal conditions would be to know exactly who was exposed to the Russian bot activity AND changed their vote because of having seen the propaganda. As a proportion of the larger population, this would be looking at everyone that voted in the U.S. election vs. who was exposed to troll activity and changed their vote, or decided to vote, as a result of having seen this material.

What are the major obstacles for collecting these data?

> Because this is all happening in cyberspace, there are logs and tweets that can tangibly be measured and dissected by a statistical approach. That said, there are enormous obstacles as some of this information is neither public or available at all. Collecting this data would mean looking into each account and associating the account with a known voter identity. Then that person would need to be polled or asked about their affinity for the election i.e. who they voted for and why.
>
> Making things more difficult is that it is hard to tell authentic reach from imitation reach. There are many people who have many twitter accounts, and even bots accounts are known to be bought and sold on various marketplaces. Discerning what is organic vs what is machine is a burden on this area of research.

Do we know how many targets are affected in the population? What is the proportion of the target group in the overall population of interest?

> By one measure, 770,000 messages sent from known Russian troll accounts but reach is difficult to tell because of retweets and 43 million other election related tweets sent over same timeline could be difficult to attribute to Russia.

Existing Studies

Describe the major results from existing studies (academic papers or reports) regarding the impact of your cyber threat on your outcomes of interest

- Cross-Platform State Propaganda: Russian Trolls on Twitter and Youtube during the 2016 U.S. Presidential Election
  - Troll accounts were primarily trying to help increase support for Donald Trump and conservative candidates.
  - Some accounts were "agnostic" trying to inflame partisan divisions by supporting either side.
- Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign
  - Conservatives retweeted Russian troll more often producing 36x more tweets
  - Only about 4.9% of conservative users retweeting content were "bots"
- Describe the data used in these studies
  - 1 - 770,005 tweets from known Russian troll accounts
  - 2 - 43 million elections-related posts shared on Twitter by 5.7 million users (Sept 16-Nov 9 2016)
- How do authors justify that the relationship between a cyber threat and their outcomes of interest are causal? In other words, describe their research design (DiD, RD, IV, matching, naïve regression, etc)
  - Observational: Sorted tweets by their "mean" ideologies using latent semantic analysis
  - DiD through comparing Members of Congress "ideology" to discover categories of the tweets
- Describe in detail the credibility of these results: potential concerns with data (e.g., selection bias) and the research design (e.g., lack/improper control group, violation of parallel trends in DiD, self-sorting in RD, exclusion restriction in IV, etc)
  - The findings are credible given they are looking at a subset of the larger population of tweets that they've used strong methods to detect "troll-ness" but there is possibility of violation of parallel trends is certainly a difficult bar to meet. This requires the absense of treatment in the diffrences

over-time which is something the authors did not address. This could potentially lead to biased estimation of the causal effect.

Ideal Experiment

- The observation can be voters who are in favor of either of Democratic or Republican party and do Twitter
- Treatment group will be exposed to foreign trolls through Twitter (retweet the foreign trolls)
- We can code treatment group as 1 if they retweet the foreign trolls
- Control group can be coded as 0 which means that they are not exposed to foreign trolls (do not retweet the foreign trolls)
- Our outcome of interest will be the change in voters' behavior after they are exposed to foreign trolls
- After the exposure to foreign trolls, voters' preferences change
- It is feasible. We can comprehend voters' preference based on how many the users have tweeted on the democratic party or the republican party. However, we need to know what twitter users are foreign trolls.
- We can do Difference in Difference design based on before and after exposure to foreign trolls

Existing Data Sources

- We can use dataset which Twitter provides. It has retweet information. We can find that who retweet the foreign trolls if we know which users are foreign trolls.
- Also, we can use Twitter dataset to know potential outcomes of interest. We can find the change in their preference based on post and retweet information such as what they post and what they retweet.

Structure of the Dataset

- Our structure of a dataset to estimate the impact of the foreign trolls
  - Vote preference 2012 election
  - Vote preference 2016 election
  - Exposure to foreign trolls
- Treatment variable
  - exposure to foreign trolls
- Outcomes of interest
  - Vote preference in 2016 election

Research Design

- Difference in Difference is part of "Design based inference" quasi experimental models.
- Widely used technique amongst economists, social scientists, and researchers.
- Top model to use for research better than observational studies
- Difference in difference method does not require randomization.

	2012	**2016**	**Difference**
Exposed to tweets (treatment)	A (not yet treated)	B (treated)	B - A
Not Exposed (control)	C (never treated)	D (never treated)	D - C
Difference	A - C	B - D	(B - A) - (D - C)

Code Analysis

```
path <- 'https://raw.githubusercontent.com/RappelBerryPi/PSCI6303FinalProject/main/0-data/data.csv'
d <- fread(path)
```

```
# Code Omitted for brevity
pander(d.display.head, style = "rmarkdown")
```

voting preference	pro democratic	anti democratic	pro republican	anti republican	vote 2012	vote 2016	main information sorce	vote changed
R	14	64	85	18	R	O	Online	1
R	21	72	90	23	R	R	Online	0
R	25	62	89	22	R	R	Social Media	0
R	8	73	96	29	R	R	Television	0
R	22	74	82	28	R	D	Online	1
R	23	60	85	18	R	R	Social Media	0

```
# Code Omitted for brevity
pander(head(d.condensed), style = "rmarkdown")
```

Total Tweets	vote 2012	vote 2016	vote changed	exposure to troll tweets
110	R	O	1	FALSE
137	R	R	0	FALSE
134	R	R	0	FALSE
118	R	R	0	FALSE
146	R	D	1	FALSE
124	R	R	0	FALSE

Regression Table Analysis

```
# Code Omitted for brevity
model.did <- lm(data = d.affected.total, voted_preference.2016 ~ affected +
                  voted_preference.2012 + voted_preference.2012:affected)
stargazer(model.did, type="text")
```

```
##
## ===============================================================
##                                        Dependent variable:
##                                    ---------------------------
##                                       voted_preference.2016
## ---------------------------------------------------------------
## affected                                      0.386***
##                                                (0.071)
##
## voted_preference.2012                         0.381***
##                                                (0.031)
##
## affectedTRUE:voted_preference.2012           -0.425***
##                                                (0.085)
##
## Constant                                      0.392***
##                                                (0.024)
##
## ---------------------------------------------------------------
## Observations                                    1,000
```

```
## R2                                                  0.137
## Adjusted R2                                         0.134
## Residual Std. Error                           0.446 (df = 996)
## F Statistic                                 52.603*** (df = 3; 996)
## ============================================================
## Note:                                     *p<0.1; **p<0.05; ***p<0.01
```

Credibility

```
p.1 <- get.percentage(d.affected.total[voted_preference.2012 == TRUE & affected == TRUE,],
                      d.affected.total[voted_preference.2012 == TRUE,])
p.2 <- get.percentage(d.affected.total[voted_preference.2012 == TRUE & affected == FALSE,],
                      d.affected.total[voted_preference.2012 == TRUE,])
p.3 <- get.percentage(d.affected.total[voted_preference.2012 == FALSE & affected == TRUE,],
                      d.affected.total[voted_preference.2012 == FALSE,])
p.4 <- get.percentage(d.affected.total[voted_preference.2012 == FALSE & affected == FALSE,],
                      d.affected.total[voted_preference.2012 == FALSE,])
cat(paste(p.1,p.2,p.3,p.4, sep = "\n"))
```

```
## 17.552%
## 82.448%
## 11.873%
## 88.127%
```

As we can see, both relative populations between those who were affected and those who weren't affected for both our control and treatment populations are nearly identical with both affected participants having nearly identical percentages at `17.552%` and `11.873%`. Similarly both groups of non-affected participants have percentages at `82.448%` and `88.127%`.

Conclusions

The size of the effect between voting behavior and those affected by troll tweet exposure is significant and shows evidence for our original hypothesis. Challenges in collecting actual data: the preference survey is a sensitive issue because it is political in nature which can cause a problems for data collection.

Policy to Mitigate a Cyber Threat

Existing Studies

Describe the major results from the existing studies (academic papers or reports) regarding the mitigation of your cyber threat

The widespread transmission of digital misinformation has been seen as a major threat to representative democracy. Communication, cognitive, social, and computer scientists are researching the complex causes of disinformation spreading virally, while online platforms are starting to implement countermeasures. To guide these efforts, little systematic, data-based evidence has been released. During the ten months of 2016 and 2017, researchgate article analyzed 14 million tweets and 400 thousand articles on Twitter. They discovered evidence that social bots had a disproportionate role in the distribution of low-credibility content. Bots boost such material in the early stages of its distribution, before it becomes viral. They also use replies and mentions to target users with a large number of followers. Users are vulnerable to this manipulation because they reshare bot-posted content. Social bots play an important role in the success of low-credibility sources. These findings show that limiting the use of social bots could be a useful method for combating the spread of online disinformation. Partnerships with social media platforms for Curbing social bots may be an effective strategy for mitigating the spread of online misinformation.

Describe the data used in these studies, Justify that the relationship between a policy and the scale of a cyber threat are causal, potential concerns with data

There are no existing studies that can help us specify data for the mitigation of cyber threat.

Ideal Experiment

We would partner with twitter to run an experiment on a subset of our current population divided into two groups. In this experiment, one group of people would be exposed to all the tweets which include both troll and legitimate tweets while other group will not be exposed to any troll tweets.

And we would collect the following columns: + Number of fake troll tweets seen + Number of fake troll tweets reported + Number of fake troll tweets interacted with + Plan to vote for in 2024 (after the experiment period ends)

In this experiment, our control group will be the group of people who would we able to see both, the troll tweets and the legitimate tweets. And our treatment group will be the group of people who do not see any troll tweets but they are exposed to regular tweets.

We are using an instrumental variable approach on, How the seeing and interacting with troll tweets affects people's patterns/preferences. For instance, whether they changed their vote or not. Finally, we should be able to tell whether the people who saw troll tweets (Control Group) are more likely to change their voting preferences than those who are not exposed to troll tweets at all (Treatment Group).

In order to be able to successfully conduct this experiment, we will need support from Twitter and our biggest concern would be Twitter's agreement. Everyone that is enrolled knows that they are part of an experiment. We would need to employ people to somehow generate the troll tweets or create a program that does it for us. It is probably unfeasible for Twitter to allow a 3rd party to conduct research about how their platform operates.

Existing Data Sources

- There are no existing sources to collect a dataset that could help us establish the impact of our policy since it is a new experiment that would need to be run first in order to collect and analyze the data and the impact of the policy on the cyber threat.
- The structure of our dataset will consist of these eight columns:
  - general_voting_preference
  - number_of_troll_tweets_seen
  - number_of_troll_tweets_interacted_with
  - number_of_troll_tweets_reported
  - voting_plan_2024
  - treated
  - voting_plan_different
  - affected
- Anyone who has the treated column as true, is part of the treatment group. This is regardless as to whether or not they reported any tweets or interacted with any tweets.
- The outcome of interest would be a DID model comparing whether or not someone was affected by the troll tweets seen (number_of_troll_tweets_seen > mean(troll tweets seen)) combined with whether or not the user is part of the treatment group or not.
- These two variables would describe as to whether the voting plan was different or not.

Research Design

- We used python to randomly create the control and treatment group as well as the results of the experiment. We used the DiD model to show that there is a difference between those who are part of the treatment and those who are not part of it.

- We randomly selected people from the population, and that we randomly showed them legitimate and fake troll tweets. The control group was always given the same chance to interact with them without any knowledge of the tweets being fake or real.

Regression Table

Skipped as per instructions

Credibility

Skipped as per instructions

Conclusions

- Although this data was generated with the intent to show that this experiment would succeed, we believe the experiment could work and a real research experiment and real data are needed in order to come up to a solid conclusion.
- The main challenge we might face in collecting this data is Twitter support to allow us to run the experiment.

Bibliography

Ruck, Damian J., Natalie M. Rice, Joshua Borycz, and R. Alexander Bentley. 2019. "Internet Research Agency Twitter Activity Predicted 2016 U.S. Election Polls." *First Monday*, June. https://doi.org/10.5210/fm.v24i7.10107.