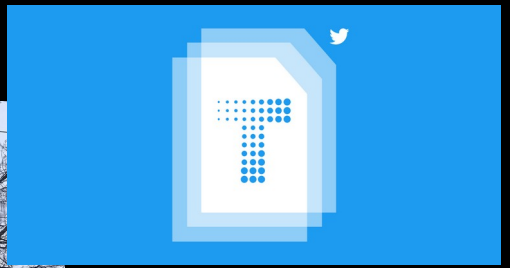
Twitter Influence Operations

Ron Thompson – CS 150 NS



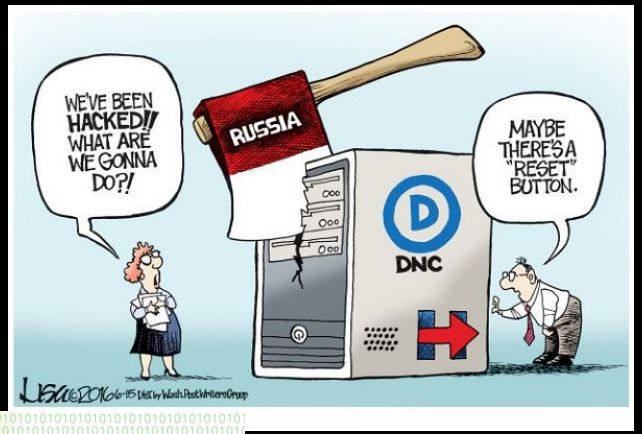
First, some definitions

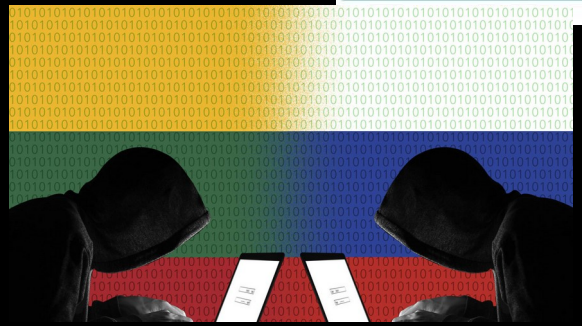
- Misinformation[1]
  - Communication of false information without the intent to deceive, manipulate, or otherwise obtain an outcome
- Disinformation[1]
  - Dissemination of explicitly false or misleading information
- Активные мероприятия[2]
  - Influence operations
  - Intelligence operations to subvert institutions of a foreign nation
  - Pioneered by the Cheka/KGB since 1920's
  - Technique adopted by the Revolutionary Guard and other actors that are hostile to US

"I just feel fortunate to live in a world with so much disinformation at my fingertips."

1. Benkler, Yochai, et al. Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics. Oxford University Press., 2018.
2. Bittman, Ladislav, and Roy Godson. The KGB and Soviet Disinformation: an Insider's View. Pergamon-Brassey's, 1985.

Social media has been a large threat vector for active measures since 2010’s





- Russian security services have leveraged active measures to influence politics in the Baltics and Ukraine
- Most notably came to ahead in the US in 2016
- Social media companies, such as Twitter, have been trying to identify disinformation networks and classify by the affiliated country
- Number of accounts have been flagged and removed

Identifying networks of foreign influence actors is important for foreign policy response

- Similarly to how it is important to classify and know what trademarks exist for hackers
- Can we identify the equivalent of Cozy Bear, Fancy Bear, Helix Kitten, or Rocket Kitten?
- Currently we are not classifying actions based on groups, but more generalized to countries
- How much are these groups focusing on domestic influence vs foreign influence

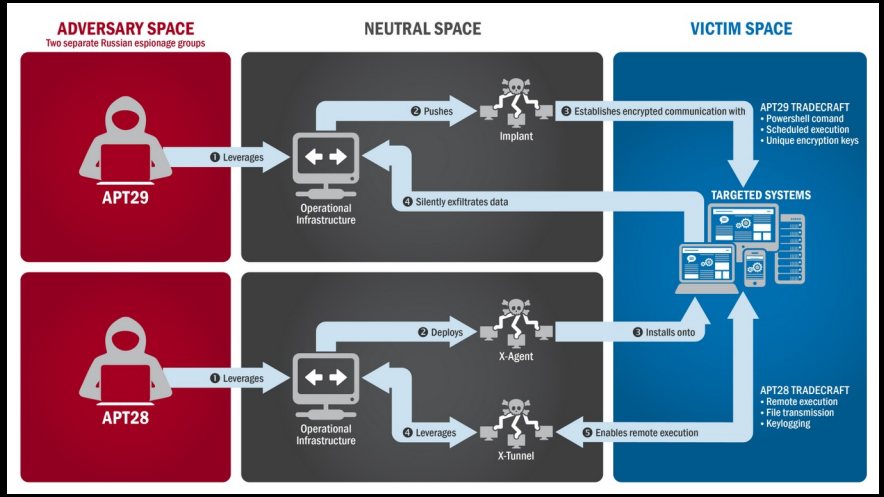
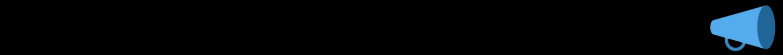


Diagram from US-CERT's report on APT28 and APT29 (aka GRIZZLY STEPPE) Description by source: "The tactics and techniques used by APT29 and APT 28 to conduct cyber intrusions against target systems"

Specific questions that this research will address

- How is the Russian disinformation network organized?
  - Can we differentiate between the Internet Research Agency and potentially other groups?
- What similarities exist between the Iranian and Russian disinformation networks?
  - Can we identify similar ways of spreading messages? How is the content similar?
- Are we able to identify suspicious accounts based on their activity and classify the associated group?



Data available through Twitter

Twitter Data

- Monthly data files with users/tweets that have been identified and removed
  - Split by country of origin
- Files have been generated since 2018
  - Actual tweets go further back
- Different data sets for tweets, users, and image content
- Usernames are hashed as is some other information as a means to protect user privacy
- Accounts were flagged by Twitter mainly with IP addresses
  - Other methods used are not made clear
  - Potentially more accounts out there that haven't been flagged or removed

Transformations Done to Date

- Subset of just Russian and Iranian affiliated accounts
- Combined data sets into six files
  - Iranian: Tweets, Users, Images
  - Russian: Tweets, Users, Images
- Started to build out the network based on account interactions
  - Unable to see follower/following information
  - Inferred connections based on tweets
- Using Python and packages from the semester, but may transition to PySpark given the size of the data and limits of personal machine

Methodology

- Measures related to the graphs
  - Do they have a strong central node?
  - How sparse are the graphs?
  - Are there interesting subgraphs?
  - Do these measures allow identification of different groups?
- Based on the information we have about the accounts and the tweets, can we identify accounts that should be removed?
  - Potential issue is that we do not have any counter examples so the model could overfit
- Are we able to identify Russian vs Iranian accounts?
  - What happens when we subset only to English language tweets?
- Do these networks overlap?