Adversaries with Limited Information in the Friedkin-Johnsen Model

Sijing Tu, Stefan Neumann, Aristides Gionis

KTH Royal Institute of Technology

February 20, 2023

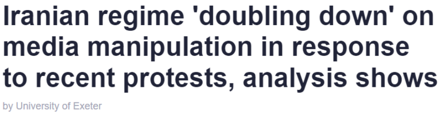
This work is under submission.

Malicious Actors are Attacking Social Networks

(The New York Times, 2016)

- 2016 Democratic National Committe email leak.
  - *Russian military and intelligence services have been using the Internet to sow discord and discredit legitimate political institutions* (TIME, 2016)

Malicious Actors are Attacking Social Networks



Iranian regime 'doubling down' on media manipulation in response to recent protests, analysis shows

by University of Exeter

(Phys.org, 2022)

- Mahsa Amini protests.
  - A recent analyis regarding the Iranian disinformation shows the main goal is to pit groups against each other. (The Washington Institute, 2022)

How Much Discord Can Attackers Sow on Social Networks, Given Limited Information?

How Do People Form Opinions

Friedkin-Johnsen Model (FJ model) (Friedkin, Johnsen; 1990)

- Each node i has innate and expressed opinions
- Innate opinion $s_i \in [0,1]$: fixed, kept private
- Expressed opinion $z_i^t \in [0,1]$: depends on time t, public

After convergence $\mathbf{z} = (I + L)^{-1}\mathbf{s}$, where L is the graph Laplacian.

- Expressed opinions $\mathbf{z}$ are determined by the network topology and innate opinions.

How to Use Opinions to Measure Societal Discord

In the literature, people use polarization and disagreement to indicate societal discord.

- Both polarization and disagreement can be measured by expressed opinoins.
  - Polarization: Variance of expressed opinions;
  - Disagreement: Tension of expressed opinions between neighbors.
- Expressed opinions are determined by network topology and innate opinions.
- ⇒polarization and disagreement are determined by innate opinions and network topology.

Assumptions

- The adversary can *only* access the network topology;
  - It is expensive for adversaries to obtain innate opinions
  - Network topology is accessible, e.g., by data crawling or and API.
  - Previous literature assumes both innate opinions and network topology are known.
- Small number of network users can be radicalized.
  - For Covid vaccination, the attacker may spread disinformation to make some net users to question its safety.
  - Strong assumption: The chosen users' opinions can be radicalized.

Problem Definition

Maximize discord with Limited Information

Problem (Full-information)

Given a social network's topology and its users' innate opinions, maximize the discord by radicalizing innate opinions of k users.

Problem (Limited-information)

Given a social network's topology, maximize the discord by radicalizing innate opinions of k users.

- We solve the problem under the limited-information setting.
- We compare our solutions with solutions obtained under full-information setting.

Our Results

Theoretical Contributions

- Algorithms that work well under the limited-information setting, also work well under full-information setting.
  - Under mild assumptions, i.e., the network initially has small discord.
- We propose an algorithm with provable guarantee under the limited-information setting.
- We show the hardness of an open problem under both settings.

Experimental Results

Regarding maximizing disagreement.

- The limited information algorithms outperform the baseline algorithms.
- Algorithms using limited information are within a factor of 2 compared to algorithms with full information.

Ethical Issues

Datasets

- Public datasets Twitter and Reddit have been used in various research papers.
  - Twitter contains ground-truth opinions.
  - Reddit contain synthetic opinoins generated using a power law distribution.
- karate, books, blogs are obtained from public data repository KONECT.
  - The datasets only contain network structure.
- The following datasets appeared in multiple published research papers. However, these datasets are not in the public domain.
  - Tweet:S5, Tweet:S2, Tweet:M5, G:Brexit, G:US-elect and Twitter100.

Ethical Reflection

- Our work investigates the power of a weak adversary and raises awareness to potential adversarial manipulations.
  - The adversary can sow great discord in the network given only network topology.
  - A simple, common, and scalable greedy algorithm works well for adversaries.
- Ways to mitigate the influence from adversaries:
  - Make attackers hard to obtain network topology.
  - Protect its users from adversary who tries to change their opinions.
    - Protect its users from dis- and mis-information.

Thank You