Adversarial Classification on Social Networks

Sixie Yu, Yevgeniy Vorobeychik, Scott Alfeld

Vanderbilt University, Amherst College

Motivation

- Over 50% adults in the U.S. regard social media as primary sources for news. [1].
- Over 37 million news stories in 2016 U.S. Presidential election later proved fake. [2]
- Anti-social posts/discussions are negatively affecting users and damage online communities. [3]
- Social network spams and phishing can defraud users and spread malwares.

Introduction

A large social network enables unprecedented levels of social interaction in the digital space, as well as sharing of valuable information among individuals. It is also a treasure trove of potentially vulnerable individuals to exploit for unscrupulous parties who wish to gain an economic, social, or political advantage. We consider a problem where there is a defender and an attacker. The attacker spreads malicious contents (i.e., malwares, fake news, etc.) through social networks. The defender limits the propagation of malicious contents, while facilitating the propagation of benign or normal contents.

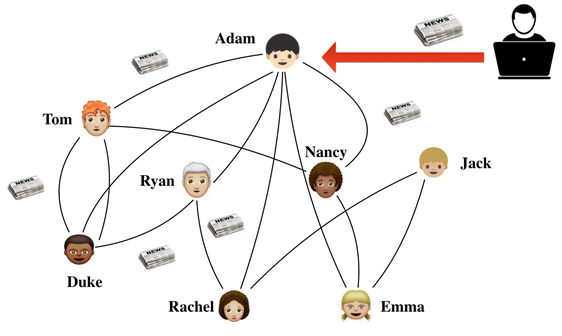

Figure 1: An example of the propagation of malicious contents.

Our Contributions

Our contributions are as follow:

- Explicitly model the diffusion process of contents through networks as a *function* of the features of the contents.
- Instead of deploying a "global" detector, we learn and deploy a collection of *heterogeneous* detectors, which takes network structures, diffusion of messages, and adversarial behavior into account.
- Fomalize the overall problem as a Stackelberg game between a defender and an attacker.
- Utilize *Implicit Function Theorem* to design a novel approach for solving the resulted Stackelberg game.

Defender Model

Innovations

- Learn and deploy *heterogeneous* detectors at different nodes.
- Explicitly considering both *diffusion* and *adversarial manipulation* during learning.

$$U_d = \alpha \sum_{x\in D^-} \sum_{i\in V} \sigma(i,\Theta,x) - (1-\alpha) \sum_{x\in D^+} \sigma(s,\Theta,z(x)) \tag{1}$$

- D^-, D^+ are benign and malicious data, respectively.
- $\Theta = \{\theta_1, \theta_2, \cdots, \theta_{|V|}\}$ being parameters of detectors at different nodes.
- The expected influence is now a function of detector parameters and manipulated messages.
- $x \to z(x)$: adversarial manipulation.

Continuous-Time Diffusion Model

This model is to simulate the propagatioin of contents through networks. It has several key properties:

- The propagation of a message depends on both the network structure and the features of the message (x).
- A message started from a node s propagates to other nodes in a breadth-first search fashion.
- The propagation time through an edge e is sampled from a distribution $f_e(t; \mathbf{w}_e, x)$.
- The time taken to affect a node i is the shortest path between s and i, where the weights of edges are propagation times associated with these edges.
- A node is affected if its shortest path to s is above T, which is externally supplied.
- The influence of a message initially affecting a node s is defined as $\sigma(s, x)$, which is the expected number of affected agents over time window T.

Attacker Model

Attacker's actions

- Find a node $s \in V$ to start diffusion.
- Transform $x \to z(x)$ in order to avoid detection.

For any original malicious instance $x \in D^+$:

$$\begin{aligned} &\max_{i,z} \ \sigma(i,\Theta,z) \\ &s.t \quad ||z-x||_p \le \epsilon \\ &\quad\ \ \mathbb{1}[\theta_j(z)=1] = 0, \forall j \in V \end{aligned} \tag{2}$$

- ϵ: the attacker's budget.
- $\theta_j(z) = 1$: the manipulated message is detected at node j.

Stackelberg Game

The interaction between the defender and the attacker is modeled as a Stackelberg game. which proceeds as follow:

- The defender first learns Θ (the parameters of detectors at different nodes).
- The attacker observes Θ and construct its optimal attack against the defender.

$$\begin{aligned} &\max_{\Theta} \ \alpha \sum_{x\in D^-} \sum_{i} \sigma(i,\Theta,x) - (1-\alpha) \sum_{x\in D^+} \sigma(s,\Theta,z(x)) \\ &s.t. : \forall x \in D^+ : \quad (s, z(x)) \in \arg\max_{j,z} \sigma(j,\Theta,z) \\ &\qquad\quad \forall x \in D^+ : \quad ||z(x)-x||_p \le \epsilon \\ &\qquad\quad \forall x \in D^+ : \quad \mathbb{1}[\theta_k(x)=1] = 0, \forall k \in V \end{aligned} \tag{3}$$

The equilibrium of this game: $(\Theta, s(\Theta), z(x;\Theta))$.

Experimental Results

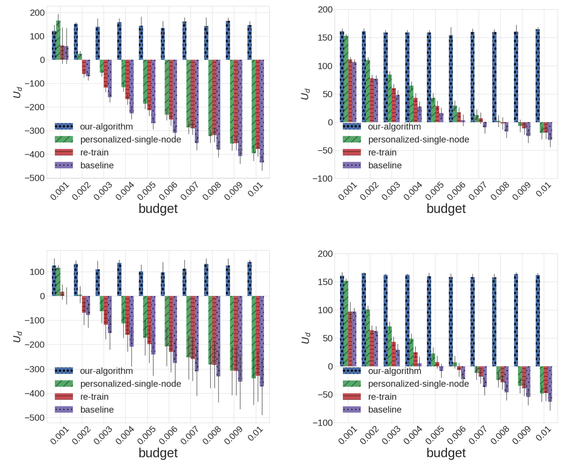

Figure 2: The performance of each defense strategy. Each bar is averaged over 10 random topologies. Left: BA. Right: Small-world

Contact Information

- Web: sixie-yu.org
- Email: sixie.yu@vanderbilt.edu