Adversarial Classification on Social Networks

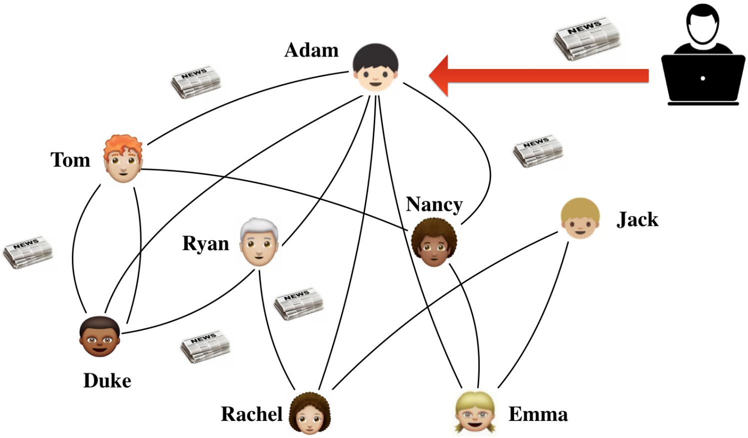
Sixie Yu[1] Yevgeniy Vorobeychik[1] Scott Alfeld[2]

[1]Electrical Engineering and Computer Science
Vanderbilt University

[2]Computer Science
Amherst College

AAMAS 2018

Problem Setting



Motivation

- Over 50% adults in the U.S. regard social media as primary sources for news. [**holcomb2013news**].
- Over 37 million news stories in 2016 U.S. Presidential election later proved fake. [**allcott2017social**]
- Anti-social posts/discussions are negatively affecting users and damage online communities. [**cheng2015antisocial**]
- Social network spams and phishing can defraud users and spread malwares.

Traditional Defense

- Train a "global" detector from past data and deploy it everywhere.
- Ignore network structures, propagation of messgaes, and adversarial behavior.

Not Adequate

- Adversaries can tune content to avoid being detected.
- Traditional learning approaches ignore network structures.
    - The impact of detection errors.
    - Being able to detect malicious content at multiple nodes creates a degree of redundancy.

Table of Contents

Continuous-Time Diffusion

- The propagation of a message depends on both the network structure and the features of the message (x).

Continuous-Time Diffusion

- 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.

Continuous-Time Diffusion

- 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 $f_e(t; \mathbf{w}_e, x)$.

Continuous-Time Diffusion

- 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.

Continuous-Time Diffusion

- 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.

Continuous-Time Diffusion

- 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 nodes over time window T.

Table of Contents

Defender Model

Innovations

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

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

- 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 the parameters of detectors (Θ), as well as manipulated messages ($z(x)$).
- $x \to z(x)$: adversarial manipulation.

Table of Contents

Attacker Model

Attacker's actions

- Find a node $s \in V$ to start propagation (reminiscent of the famous influence maximization problem).
- Transform $x \rightarrow z(x)$ in order to avoid detection.

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

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

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

Table of Contents

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.

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

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

Table of Contents

Solution Approach

Assumption

The defender *knows* the node being attacked.

- This assumption enables us to collapse the bi-level optimization into a single-level optimization.
- Assume the defender knows the node s will be attacked, by leveraging *Implicit Function Theorem*, we can solve the single-level optimization, which results in the optimal defense strategy Θ_s^*.

Relax the assumption

We relax the assumption that the defender *knows* the node being attacked, and introduce a heuristic algorithm to solve for $\big(\Theta, s(\Theta), z(x;\Theta)\big)$.

Heuristic algorithm:

- For each node $i \in V$ we solve for the Θ_i^*.
- $\Theta^* = \arg\max_{\Theta_i^*} U_d(\Theta_i^*)$

Table of Contents

Experiments

- In our experiments, we consider a specific detection model: logistic regression (LR)
- $\Theta = \{\theta_1, \theta_2, \cdots, \theta_{|V|}\}$: thresholds of detectors
- We compare our defense strategy against three others:
  - Baseline: simply learn a LR on training data and deploy it at all nodes
  - Re-training: iteratively augment the original training data with attacked instances, re-training the LR each time, until convergence
  - Personalized-single-threshold: this strategy is only allowed to tune a single node's threshold.

Experiments

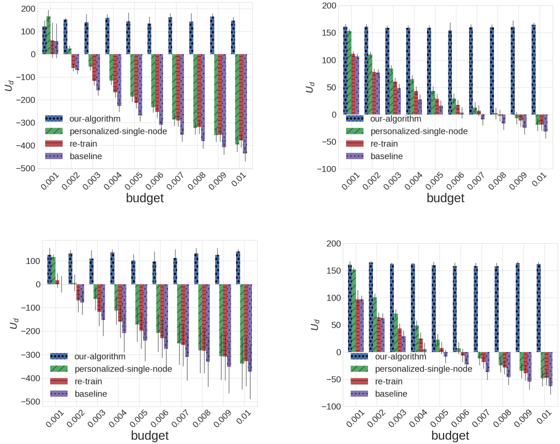


Figure: The performance of each defense strategy. Each bar is averaged over 10 random topologies. Left: BA. Right: Small-world

Our Contribution

- Explicitly model the propagation process of contents through networks as a *function* of the features of the contents.

Our Contribution

- Explicitly model the propagation 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, propagation of messages, and adversarial behavior into account.

Our Contribution

- Explicitly model the propagation 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, propagation of messages, and adversarial behavior into account.
- Fomalize the overall problem as a Stackelberg game between a defender and an attacker.

Our Contribution

- Explicitly model the propagation 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, propagation 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.

Thank you!

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

Table of Contents

References