Game Theoretical Defenses against Poisoning Attack

-

October 3, 2019

1 Motivation

1.1 What harm does it cause?

1. The loss of revenue due to the decrease in user satisfaction. Fake ratings and reviews cause customers to purchase unwanted or lower quality items, which results in low satisfaction. Unsatisfied users will often switch to alternative platforms, and this creates a significant loss of future revenue from such users.

2. The loss of advertisement revenue. The sellers may choose alternative methods of advertising, for example, through fake purchases and hiring people to generate fake reviews, over the promotion service provided by the e-commerce platform. In this case, the platform will lose a significant amount of advertisement revenue.

2 Game-Theoretic Approach

2.1 Formal Model of the Game

We use a Stackelberg game with incomplete information to model the situation. In this game, given a clean dataset of n items, D_n, the attacker first chooses an adversarial dataset of size k or less, Δ_k, to add the items to D_n. The defender observes the combined (poisoned) dataset, $D_n \cup \Delta_k$, and sanitize it by selecting a subset $S\left(D_n \cup \Delta_k\right)$. Formally, the game can be specified by its extensive form $\Gamma = (N, H, P, I, u)$, in which,

- $N = \{A, D\}$ is the set of players A is the attacker, D is the defender.
- $H = \left\{\varnothing, D_n \in X^n, \left(D_n, \Delta_k\right) \in X^n \times X^{\leqslant k}\right\} \cup \left\{\left(D_n, \Delta_k, S\right) \in X^n \times X^{\leqslant k} \times \{0,1\}^{\left|D_n \cup \Delta_k\right|}\right\}$ is the set of histories including the initial history $\varnothing$, the non-terminal histories after nature's choice of D_n, and after the attacker's move Δ_k representing k or less elements from the set of possible data items in X, and the terminal histories after the defender's move S representing a selection of a subset of the poisoned dataset $D_n \cup \Delta_k$.
- $I = \{\varnothing\} \cup X^n \cup \left\{X^{\leqslant n+k} \backslash \sim_I\right\}$, where $\sim_I$ is the equivalence relation on the length-two histories $\left(D_n^{(1)} \cup \Delta_k^{(1)} \sim_I D_n^{(2)} \cup \Delta_k^{(2)}\right.$ if they are indistinguishable by the defender, for example, if these sets are just permutations of each other.
- $P : P(\varnothing) = \text{Nature}, P\left(D_n\right) = A \ \forall \ D_n \in X^n, P\left(D_n, \Delta_k\right) = D \ \forall \ \left(D_n, \Delta_k\right) \in X^n \times X^{\leqslant k}$, is the player function: nature chooses an attacker type D_n first according to some distribution $\mathcal{F}\left(D_n\right)$, the attacker

moves next and defender moves last after observing the information sets generated by nature and the attacker's action.

- $u : u_A(D_n, \Delta_k, S) = f(g(S(D_n \cup \Delta_k)) - c(\Delta_k), u_D(D_n, \Delta_k, S) = -d(g(D_n), g(S(D_n \cup \Delta_k)))$ is the payoff after terminal histories. In particular, the attacker gets a profit based on the parameter $\hat{\theta} = g(S(D_n \cup \Delta_k)$ learned from the poisoned data set. Here, g is the learning algorithm. The defender gets a profit based on some distance between $\hat{\theta}$ and the true parameter value $\theta^\star = g(D_n)$ learned from the clean data set.

2.2 Example

We use the following concrete example to describe the actions and payoffs of the game. Suppose the attacker is a seller on an e-commerce platform, and the defender is the platform. Then D_n is the set of true customer ratings of the seller's product. The possible set of ratings are X, for example, $X = \{1, 2, 3, 4, 5\}$. The seller then hires k or less fake customers to purchase the product and generate fake ratings $\Delta_k \in X^{\leqslant k}$. The platform observes $D_n \cup \Delta_k$ and removes suspected fake ratings using the filter function $S \in \{0,1\}^{|D_n \cup \Delta_k|}$, for example, $\{0,1,0,1,0\} \in S = \{0,1\}^5$ represents removing the first, third and fifth rating from a set of five ratings. Suppose the ratings are summarized by a single number, for example the average rating $\hat{\theta} = g(S(D_n \cup \Delta_k)), g(\cdot) =$ average $(\cdot)$, then payoff to the seller is the revenue generated from this average, $f\left(\hat{\theta}\right)$, minus the cost of hiring fake customers, $c(\Delta_k)$, and the payoff to the platform is the negative revenue or loss due to the discrepancy between the poisoned average $\hat{\theta}$ and the true average $\theta^\star = g(D_n), d\left(\hat{\theta}, \theta^\star\right)$, such as the loss of customers due to lower satisfaction, or the loss of reputation due to inaccurate ratings.

In the simplest numerical example, $X = \{0,1\}, n = k = 1$ and,

$$g(X) = \frac{1}{|X|} \sum_{x \in X} x,$$
$$d(x, y) = |x - y|.$$

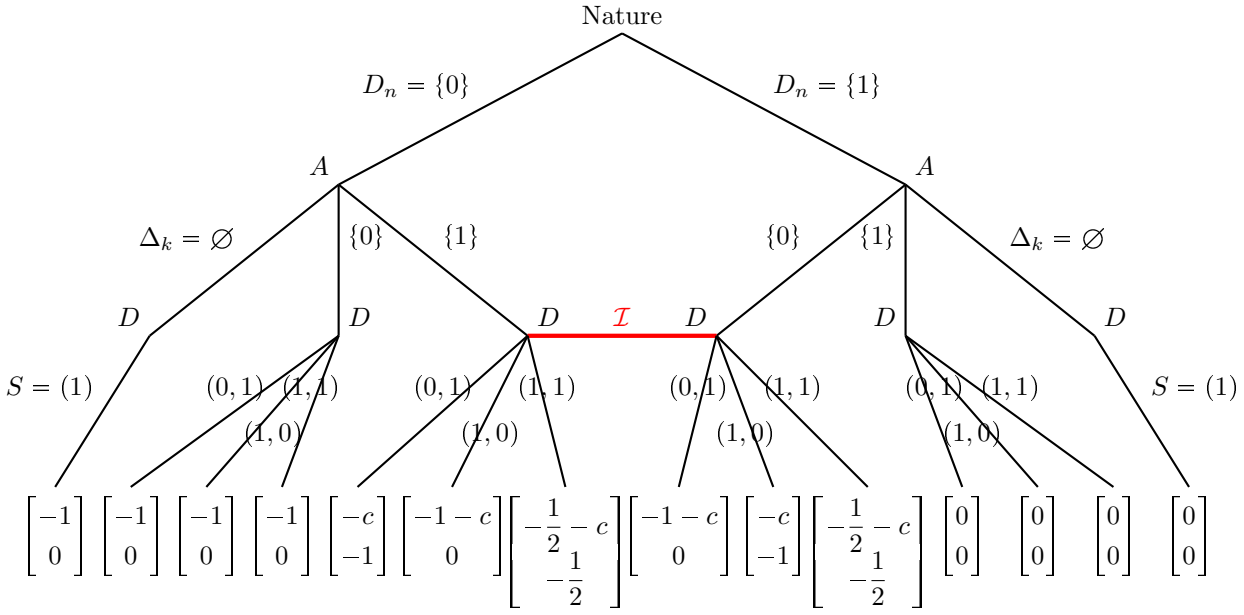
Here, the ratings are either 0 or 1, say "dislike" or "like". The platform displays the average sanitized rating $\hat{\theta} = g(S(D_1 \cup \Delta_1))$. The ideal average for the seller is $\theta^\dagger = 1$. The seller wants the rating to be close to $\theta^\dagger$ while the platform wants the rating to be truthful $\theta^\star = g(D_1)$. The prior belief of the true product rating is,

$$\mathbb{P}\{D_1 = \{0\}\} = p,$$
$$\mathbb{P}\{D_1 = \{1\}\} = 1 - p.$$

For simplicity, suppose that the cost of adding a rating that is different from the one in D_1 is c and the cost of adding a rating that is the same as the one in D_1 is 0, meaning,

$$c(\Delta_1 = \{0\} | D_1 = \{1\}) = c(\Delta_1 = \{1\} | D_1 = \{0\}) = c < \frac{1}{2},$$
$$c(\cdot) = 0, \text{ otherwise}.$$

The game Γ can then be represented by the following diagram.

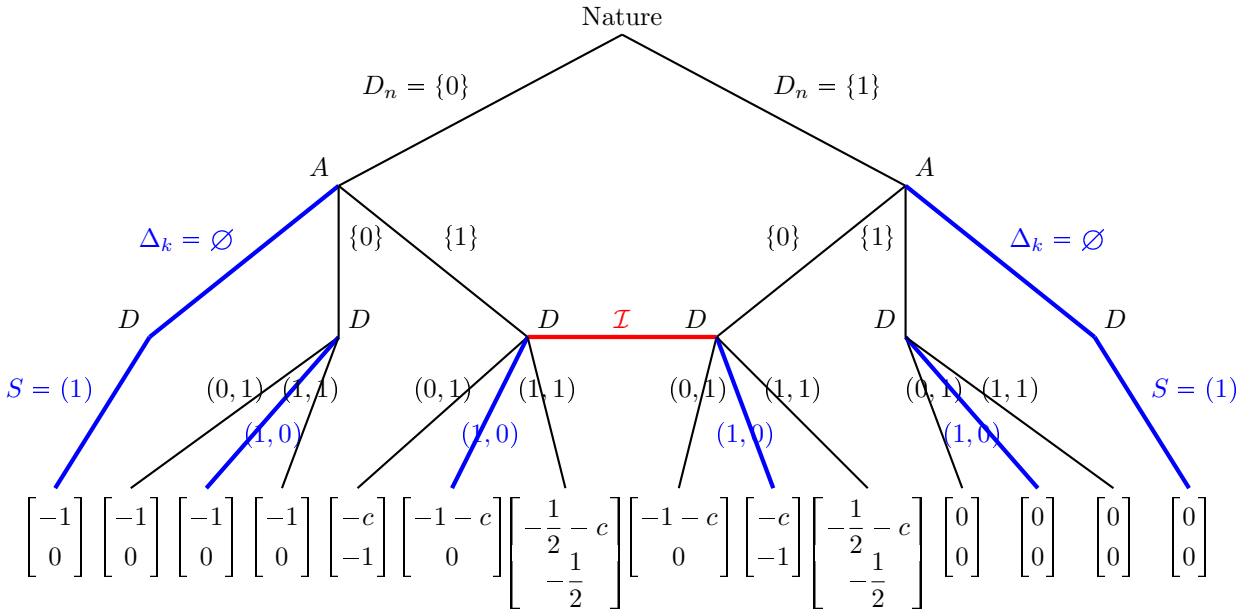


The only non-singleton information set $\mathcal{I}$ is highlighted in red. The training data the defender observe when $D_1 = \{0\}, \Delta_1 = \{1\}$ and $D_1 = \{1\}, \Delta_1 = \{0\}$ are indistinguishable because they are permutations of each other. Therefore, these two histories are put in the same information set.

The calculations of the payoffs for terminal histories, from left to right in the above diagram, are presented in the following table,

$$D' = D_1 \cup \Delta_1, S' = S\left(D'\right).$$

D'	$\{0\}$	$\{0,0\}$	$\{0,0\}$	$\{0,0\}$	$\{0,1\}$	$\{0,1\}$	$\{0,1\}$	$\{0,1\}$	$\{0,1\}$	$\{0,1\}$	$\{1,1\}$	$\{1,1\}$	$\{1,1\}$	$\{1\}$
S'	$\{0\}$	$\{0\}$	$\{0\}$	$\{0,0\}$	$\{1\}$	$\{0\}$	$\{0,1\}$	$\{1\}$	$\{0\}$	$\{0,1\}$	$\{1\}$	$\{1\}$	$\{1,1\}$	$\{1\}$
$\hat{\theta}$	0	0	0	0	1	0	$\frac{1}{2}$	1	0	$\frac{1}{2}$	1	1	1	1
$\theta^\star$	0	0	0	0	0	0	0	1	1	1	1	1	1	1
$\theta^\dagger$	1	1	1	1	1	1	1	1	1	1	1	1	1	1
π_A	-1	-1	-1	-1	$-c$	$-1-c$	$-\frac{1}{2}-c$	$-1-c$	$-c$	$-\frac{1}{2}-c$	0	0	0	0
π_D	0	0	0	0	-1	0	$-\frac{1}{2}$	0	-1	$-\frac{1}{2}$	0	0	0	0



One possible weak perfect Bayesian equilibrium is highlighted in blue. This equilibrium represents the case in which both types of attacker choose not to poison the dataset, meaning $\Delta_1 = \varnothing$, and the defender always removes the higher rating if the dataset contains two ratings,

$$\hat{\theta} = \min \{D_1 \cup \Delta_1\}.$$

Intuitively, the high type attacker (the case when nature chooses $D_1 = \{1\}$) would never poison the dataset because poisoning would not improve her payoff, and the reason the low type attacker (the case when nature chooses $D_1 = \{0\}$) would not poison the dataset is because the defender will remove the fake rating, so this type would prefer not to incur a wasteful cost. The defender will defend by removing the higher rating because she knows that only the low type attacker has the incentive to poison the dataset, so when she observes $D_1 \cup \Delta_1 = \{0, 1\}$, it must be generated by the low type attacker, so removing the 1 is the best response. Formally, the belief when reaching the information set $\mathcal{I}$ must be,

$$\mathbb{P}\{(D_1 = \{0\}, \Delta_1 = \{1\} | \mathcal{I}) = 1,$$
$$\mathbb{P}\{(D_1 = \{1\}, \Delta_1 = \{0\} | \mathcal{I}) = 0.$$

It can be shown that this equilibrium (with the above belief) is also a sequential equilibrium of the game, and it satisfies the Cho-Kreps' intuitive criterion, meaning it is robust to mistakes made by either type of attacker.

In general, with a more complex setting, we conjecture that there are similarly structured equilibria in which the defender can infer about the attacker's type at some information sets, and based on these posterior beliefs, deploy defense mechanisms with varying degrees. As a result, we should be able to identify a threshold $\bar{c}$ such that in equilibrium, no type of attacker would poison the dataset when the cost of generating fake ratings is higher than $\bar{c}$.

2.3 Interesting Questions

1. Characterize and analyze the equilibrium strategies of the attacker and the defender for simple parameterizations of the game. For example, the action S corresponding to removing extreme values from $D_n \cup \Delta_k$ could be the equilibrium action for some special class of parameterizations of the game.

2. Solve the mechanism design problem for the defender. For example, the defender can increase the (opportunity) cost of hiring fake reviewers by decreasing the advertising price on the platform. In this case, we attempt to find conditions on c such that $\Delta_k = \varnothing$ is on the equilibrium path, meaning that the attacker will choose not to attack in equilibrium.