Supplementary Material for Attacking Social Media via Influential Interactions Poisoning

Anonymous Author(s)

CCS CONCEPTS

• **Security and privacy** → **Web application security**.

KEYWORDS

Social Media, Social Interaction, Poisoning Attacks.

ACM Reference Format:

Anonymous Author(s). 2018. Supplementary Material for Attacking Social Media via Influential Interactions Poisoning. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/1122445.1122456

1 DETAILED PROOFS

1.1 Proof of Proposition 4.1 in main body

Proposition 4.1. *Assume that the simulator $f(x, \theta)$ maps any sample x in dataset $\mathcal{D}$ to $[0, 1]$, and θ is the simulator's parameters. Suppose $x_{fix} = \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D},y=1} x$. Let $\mathcal{L}_{trn}(x, \theta)$ be the training loss, and H be the Hessian matrix of $\mathcal{L}_{trn}(x, \theta)$. If the attacker performs the interaction of user $u \in \mathcal{U}_{ctl}$ retweeting the user v's tweet t, then this poisoning action's influence $\mathcal{I}_{inject}(x_{u,v,t})$ on the attack objective $\mathcal{L}_{atk}$ is*

$$\begin{aligned} &\mathcal{I}_{inject}(x_{u,v,t}) := \mathcal{I}^T x_{u,v,t}, \\ &\textit{where } \mathcal{I} = (-\nabla_\theta \mathcal{L}_{atk}^T H_{\hat{\theta}}^{-1} \nabla_x \nabla_\theta \mathcal{L}_{trn}(x_{fix}, \hat{\theta}))^T. \end{aligned} \tag{1}$$

Proof. We first provide two related lemmas provides from [1].

Lemma 1.1. *For a model f with parameters θ, let the training loss of any sample $x \in \mathcal{D}$ be $\mathcal{L}_{trn}(x, \theta)$. Suppose that $\mathcal{L}(x_{text}, \theta)$ is the prediction loss of test sample x_{test}. If doubling a sample x, then its influence on the prediction of x_{test} can be linearly approximated as*

$$\begin{aligned} \mathcal{I}_{dbl}(x) &= \left.\frac{d\mathcal{L}(x_{test}, \hat{\theta}_{\epsilon,x})}{d\epsilon}\right|_{\epsilon=0} = \nabla_\theta \mathcal{L}(x_{test}, \hat{\theta})^T \frac{d\hat{\theta}_{\epsilon,x}}{d\epsilon} \\ &= -\nabla_\theta \mathcal{L}(x_{test}, \hat{\theta})^T H_{\hat{\theta}}^{-1} \nabla_\theta \mathcal{L}_{train}(x, \hat{\theta}). \end{aligned}$$

Lemma 1.2. *For a model f under the assumptions of Lemma 1.1. if perturbing a sample x with a noise δ, then the influence of sample from x to $x' = x + \delta$ on the prediction of x_{test} is*

$$\mathcal{I}_{mod}(x, x') = -\nabla_\theta \mathcal{L}(x_{test}, \hat{\theta})^T H_{\hat{\theta}}^{-1} \nabla_x \nabla_\theta \mathcal{L}_{trn}(x, \hat{\theta})(x' - x).$$

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

The two lemmas reveal the influence on the test sample when any sample is doubled and perturbed. It is natural to treat the attack loss $\mathcal{L}_{atk}$ defined in Eq. 3 of main body as $\mathcal{L}(x_{test}, \hat{\theta})$, to get the influence on the attack task when the sample x is doubled or perturbed.

Now we study the attack influence of retweeting. Section 3.2 reveals that user u retweeting user v's tweet t is equivalent to poisoning a sample $x_{u,v,t}$ to the dataset. Unfortunately, the above two lemmas are not suitable for the poisoning case. Because the poisoning sample $x_{u,v,t}$ does not exist in the original dataset, there is no doubling or perturbing.

Let us re-examine poisoning a new sample $(x_{u,v,t}, 1)$. It is equivalent to doubling a fixed sample $(x_{fix}, 1)$ and perturbing x_{fix} to x, where the perturbation $\delta = x - x_{fix}$. Therefore, the attack influence $\mathcal{I}_{inject}(x)$ of injecting sample $x_{u,v,t}$ approximates to the sum of doubling influence and perturbing influence, that is,

$$\mathcal{I}_{inject}(x_{u,v,t}) = \mathcal{I}_{dbl}(x_{fix}) + \mathcal{I}_{mod}(x_{fix}, x_{u,v,t}).$$

Inspired by [2], we set $x_{fix} = \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in D,y=1} x$, as this setting has the smallest influence estimation error on mathematical expectations. For a given dataset, x_{fix} is determined. Then even though x_{fix} is not in the dataset, we can inject it in advance. $\mathcal{I}_{dbl}(x_{fix})$ is a constant, ignoring this term does not affect the influence comparison of data. Replacing $\mathcal{L}(x_{test}, \hat{\theta})$ in Lemma 1.2 with $\mathcal{L}_{atk}$, then the influence of poisoning sample $x_{u,v,t}$ can be defined as

$$\begin{aligned} &\mathcal{I}_{inject}(x_{u,v,t}) := \mathcal{I}_{mod}(x_{fix}, x_{u,v,t}) \\ &= -\nabla_\theta \mathcal{L}_{atk}^T H_{\hat{\theta}}^{-1} \nabla_x \nabla_\theta \mathcal{L}_{trn}(x_{fix}, \hat{\theta})(x_{u,v,t} - x_{fix}) \\ &= -\nabla_\theta \mathcal{L}_{atk}^T H_{\hat{\theta}}^{-1} \nabla_x \nabla_\theta \mathcal{L}_{trn}(x_{fix}, \hat{\theta}) x_{u,v,t} + const, \end{aligned}$$

where $const = -\nabla_\theta L_{atk}^T H_{\hat{\theta}}^{-1} \nabla_x \nabla_\theta \mathcal{L}_{train}(x, \hat{\theta}) x_{fix}$. Similarly, it can also be ignored. Finally,

$$\mathcal{I}_{inject}(x_{u,v,t}) := -\nabla_\theta \mathcal{L}_{atk}^T H_{\hat{\theta}}^{-1} \nabla_x \nabla_\theta \mathcal{L}_{trn}(x_{fix}, \hat{\theta}) x_{u,v,t}.$$

1.2 Proof of Proposition 4.2 in main body

Proposition 4.2. *Assume a simulator $f(x, \theta)$ under the assumptions of Proposition 1.1. If the attacker modifies the profile x_i of user $i \in \mathcal{U}_{ctl}$, any associated sample $x_{u,v,t}$ of user i will change, where $u = i$ or $v = i$, and set it to $x'_{u,v,t}$ after modification. Then, the influence of changing $x_{u,v,t}$ to $x'_{u,v,t}$ on the attack objective $\mathcal{L}_{atk}$ is*

$$\begin{aligned} &\mathcal{I}_{mod}(x_{u,v,t}, x'_{u,v,t}) := \mathcal{I}^T (x'_{u,v,t} - x_{u,v,t}), \\ &\textit{where } \mathcal{I} = (-\nabla_\theta \mathcal{L}_{atk}^T H_{\hat{\theta}}^{-1} \nabla_x \nabla_\theta \mathcal{L}_{trn}(x_{fix}, \hat{\theta}))^T. \end{aligned} \tag{2}$$

Proof. Although Lemma 1.2 can be directly applied, the time-consuming $\nabla_x \nabla_\theta \mathcal{L}_{trn}(x, \hat{\theta})$ needs to be calculated for each sample. Reconsider the modification case, it can be considered as deleting the sample x and injecting a new sample x', so the influence of data modification is

$$\mathcal{I}_{mod}(x, x') = -\mathcal{I}_{inject}(x) + \mathcal{I}_{inject}(x') = \mathcal{I}^T(x' - x).$$

REFERENCES

[1] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *ICML'17*. PMLR, 1885–1894.

[2] Chenwang Wu, Defu Lian, Yong Ge, Zhihao Zhu, and Enhong Chen. 2021. Triple Adversarial Learning for Influence based Poisoning Attack in Recommender Systems. In *Proceedings of KDD'21*. 1830–1840.