A Certified Robustness

Although the GraphProt is effective in the empirical evaluation, it is unsure whether it can defense against adaptive attacks. We further propose a certified (provable) defense for our model GraphProt-R so that no further attacks can compromise the certified accuracy.

If the trigger involves injecting new nodes into the graph, we find the worst-case node number of the trigger graph. Let n_Δ denote the node numbers in poisoned subgraph $\mathcal{G}_\Delta$.

Theorem 1. *(Certified robustness for graph injection trigger). Given a testing graph G, a trained backdoored graph classifier f, and the ensemble classifier g defined in Eq. (8) with random subgraph sampling (GraphProt-R). Let $\mathcal{G}$ denote the subgraphs with s nodes sampled from G with replacement. Suppose y_A and y_B are the classes with the most votes and the second largest votes during the ensemble. We define $\underline{p_A}$ and $\overline{p_B}$ as the lower and upper bound of probability $\mathbb{P}(f(\mathcal{G}) = y_A)$ and $\mathbb{P}(f(\mathcal{G}) = y_B)$, respectively. We guarantee that the model still predicts class y_A for graphs G_Δ inserted with any trigger size smaller than r if:*

$$\max_{n_\Delta \leq n+r} (\frac{n_\Delta}{n})^s - 2(\frac{n_\Delta - r}{n})^s + 1 - (\underline{p_A} - \overline{p_B} - \delta_A - \delta_B) < 0, \tag{12}$$

where n and n_Δ are the node numbers in $\mathcal{G}$ and $\mathcal{G}_\Delta$, respectively, and $\delta_A = \underline{p_A} - (\lfloor \underline{p_A} \cdot n^s \rfloor)/n^s$, $\delta_B = (\lceil \overline{p_B} \cdot n^s \rceil)/n^s - \overline{p_B}$ are the residuals.

Proof. We begin by providing the necessary definitions and notations. We denote the testing graph G with n nodes as (V, E, X), and its structure can also be represented by adjacency matrix $A_{n\times n}$. Assume that the attacker attaches trigger Δ (e.g., specific subgraph) into the testing graph. We note the backdoored graph as $G_\Delta := (V_\Delta, E_\Delta, X_\Delta) = G + \Delta$. Assuming that the trigger size (the number of nodes involved in the trigger attachment) is r. If the trigger injection involves node injection, the node number of G_Δ might be increased by at most r: $n_\Delta := |V_\Delta| \leq n + r$. Let $I = V \cap V_\Delta$ denote the intersection nodes (with the same node features and one-hop neighbors) of the two graphs. Let $\mathcal{G} := \mathbb{S}(G)$ denote the subgraph induced from randomly selecting s nodes in G with replacement, and $\mathcal{G}_\Delta := \mathbb{S}(G_\Delta)$ denote the subgraph sampled from G_Δ. We note that the sample subgraph can be decided on the sampled nodes. As long as the $\mathcal{G}$ and $\mathcal{G}_\Delta$ have the same nodes among I, they have exactly the same adjacency matrix and node feature matrix. Equivalently, we represent the subgraph by a subset of nodes. Let $V_1 := \mathbb{S}(V)$ and $V_2 := \mathbb{S}(V_\Delta)$ denote the two node set sampled from V and V_Δ.

Next, we define an equivalent smoothed/ensemble model $g(G)$ in Eq. (8) as follows:

$$g(G) = \arg\max_{y \in \mathcal{Y}} p_y := \mathbb{P}(f(\mathcal{G}) = y), \tag{13}$$

where $\mathcal{G} = \mathbb{S}(G_\Delta)$. Given that $\mathbb{P}(f(\mathcal{G}) = y_A) \geq \underline{p_A}$ and $\mathbb{P}(f(\mathcal{G}) = y_B) \leq \overline{p_B}$, our goal is to show that $\mathbb{P}(f(\mathcal{G}_\Delta) = y_A) > \max_{y' \neq y} \mathbb{P}(f(\mathcal{G}_\Delta) = y')$, and equivalently, $\mathbb{P}(f(\mathcal{G}_\Delta) = y_A) > \mathbb{P}(f(\mathcal{G}_\Delta) = y_B)$.

We employ the Neyman-Pearson Lemma (Neyman and Pearson 1933) to establish the connection between $\mathbb{P}(f(\mathcal{G}))$ and $\mathbb{P}(f(\mathcal{G}_\Delta))$. According to the Variant of Neyman-Pearson Lemma provided in (Jia, Cao, and Gong 2021), we know that:

Let V_1 and V_2 denote two random variables in space Ω with probability densities μ_{V_1} and μ_{V_2}, $h : \Omega \to \{0, 1\}$ be any function. Then we have:

- If $S_1 = \{\omega \in \Omega : \frac{\mu_{V_1}(\omega)}{\mu_{V_2}(\omega)} \geq t\}$ for some $t > 0$, and $\mathbb{P}(h(V_1) = 1) \geq \mathbb{P}(V_1 \in S_1)$, then $\mathbb{P}(h(V_2) = 1) \geq \mathbb{P}(V_2 \in S_1)$.
- If $S_2 = \{\omega \in \Omega : \frac{\mu_{V_1}(\omega)}{\mu_{V_2}(\omega)} \leq t\}$ for some $t > 0$, and $\mathbb{P}(h(V_1) = 1) \leq \mathbb{P}(V_1 \in S_2)$, then $\mathbb{P}(h(V_2) = 1) \leq \mathbb{P}(V_2 \in S_2)$.

Let $h(\cdot)$ denote $\mathbb{I}(f(\cdot) = y_A)$, where $\mathbb{I}$ is an indication function. We can find a region S_1 such that $\mathbb{P}(f(V_1) = y_A) \geq \mathbb{P}(V_1 \in S_1) = \underline{p_A}$, then we have $\mathbb{P}(f(V_2) = y_A) \geq \mathbb{P}(V_2 \in S_1)$. Similarly, let $h(\cdot)$ denote $\mathbb{I}(f(\cdot) = y_B)$, where $\mathbb{I}$ is an indication function. We can find a region S_2 such that $\mathbb{P}(f(V_1) = y_B) \leq \mathbb{P}(V_1 \in S_2) = \overline{p_B}$, then we have $\mathbb{P}(f(V_2) = y_B) \leq \mathbb{P}(V_2 \in S_2)$. Note that we can conclude that $\mathbb{P}(f(V_2) = y_A) \geq \mathbb{P}(f(V_2) = y_B)$ if $\mathbb{P}(V_2 \in S_1) > \mathbb{P}(V_2 \in S_2)$.

Next, we define the regions S_1 and S_2 specifically. The space Ω can be divided into three subspaces:

$$\begin{aligned} \mathcal{O} &= \{\omega \in \Omega : \omega \subseteq V, \omega \nsubseteq I\}, \\ \mathcal{P} &= \{\omega \in \Omega : \omega \subseteq V_\Delta, \omega \nsubseteq I\}, \\ \mathcal{Q} &= \{\omega \in \Omega : \omega \subseteq I\}, \end{aligned} \tag{14}$$

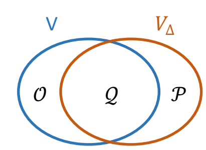

Figure 4: Space division

Because $V_1 := \mathbb{S}(V)$ and $V_2 := \mathbb{S}(V_\Delta)$ are the two node set of size s sampled from V and V_Δ with replacement, we have that:

$$\mathbb{P}(V_1 = \omega) = \begin{cases} \frac{1}{(n)^s}, & \text{if } \omega \in \mathcal{O} \cup \mathcal{Q}, \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

$$\mathbb{P}(V_2 = \omega) = \begin{cases} \frac{1}{(n_\Delta)^s}, & \text{if } \omega \in \mathcal{P} \cup \mathcal{Q}, \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

Let m denote the number of overlap nodes $|I|$, then we have $m = \max(n, n_\Delta) - r$. We have the probabilities:

$$\mathbb{P}(V_1 \in \mathcal{O}) = 1 - (\frac{m}{n})^s, \quad \mathbb{P}(V_2 \in \mathcal{O}) = 0, \tag{17}$$

$$\mathbb{P}(V_1 \in \mathcal{P}) = 0, \quad \mathbb{P}(V_2 \in \mathcal{P}) = 1 - (\frac{m}{n_\Delta})^s, \tag{18}$$

$$\mathbb{P}(V_1 \in \mathcal{Q}) = (\frac{m}{n})^s, \quad \mathbb{P}(V_2 \in \mathcal{Q}) = (\frac{m}{n_\Delta})^s, \tag{19}$$

We can find a subset $\mathcal{O}' \subseteq \mathcal{Q}$ such that:

$$\begin{aligned}\mathbb{P}(V_1 \in \mathcal{O}') &= \underline{p_A} - \delta_A - \mathbb{P}(V_1 \in \mathcal{O}), \\ &= \underline{p_A} - \delta_A - (1 - (\frac{m}{n})^s), \end{aligned} \quad (20)$$

where δ_A is an residuals such that $\underline{p_A} - \delta_A$ and integer multiple of $\frac{1}{n^s}$. Then, we define the region $S_1 = \mathcal{O} \cup \mathcal{O}'$. In this region, we know that $\frac{\mathbb{P}(V_1=\omega)}{\mathbb{P}(V_2=\omega)} \geq t$, where $t = (\frac{n_\Delta}{n})^s$. That is, the S_1 satisfies the requirement $\{\omega \in \Omega : \frac{\mu_{V_1}(\omega)}{\mu_{V_2}(\omega)} \geq t\}$.

Similarly, we can find the region S_2 by finding a subset $\mathcal{O} \subseteq \mathcal{Q}$ such that:

$$\mathbb{P}(V_1 \in \mathcal{P}') = \overline{p_B} + \delta_B, \quad (21)$$

where δ_B is an residuals such that $\overline{p_B} + \delta_B$ and integer multiple of $\frac{1}{n^s}$. Then, we define the region $S_2 = \mathcal{P} \cup \mathcal{P}'$. In this region, we know that $\frac{\mathbb{P}(V_1=\omega)}{\mathbb{P}(V_2=\omega)} \leq t$, where $t = (\frac{n_\Delta}{n})^s$. That is, the S_2 satisfies the requirement $\{\omega \in \Omega : \frac{\mu_{V_1}(\omega)}{\mu_{V_2}(\omega)} \leq t\}$.

Then, we calculate the probabilities $\mathbb{P}(V_2 \in S_1)$ and $\mathbb{P}(V_2 \in S_2)$:

$$\begin{aligned}\mathbb{P}(V_2 \in S_1) &= \mathbb{P}(V_2 \in \mathcal{O}) + \mathbb{P}(V_2 \in \mathcal{O}'), \\ &= \mathbb{P}(V_2 \in \mathcal{O}'), \\ &= [\underline{p_A} - \delta_A - (1 - (\frac{m}{n})^s)]/t. \end{aligned} \quad (22)$$

$$\begin{aligned}\mathbb{P}(V_2 \in S_2) &= \mathbb{P}(V_2 \in \mathcal{P}) + \mathbb{P}(V_2 \in \mathcal{P}'), \\ &= \mathbb{P}(V_2 \in \mathcal{O}'), \\ &= 1 - (\frac{m}{n_\Delta})^s + (\overline{p_B} + \delta_B)/t. \end{aligned} \quad (23)$$

We can conclude that $\mathbb{P}(f(V_2) = y_A) \geq \mathbb{P}(f(V_2) = y_B)$ if $\mathbb{P}(V_2 \in S_1) > \mathbb{P}(V_2 \in S_2)$. Finally, subtracting $\mathbb{P}(V_2 \in S_1)$ with $\mathbb{P}(V_2 \in S_2)$, we have the inequity (12). By definitions, we know that $g(G_\Delta) = y_A$ for all graphs G_Δ inserted with any trigger size smaller than r. □

If the trigger is attached to the existing nodes (involves node feature modification and edge modification among r nodes), we have the following simplified certifying condition:

Theorem 2. *(Certified robustness for in-graph trigger). Given a testing graph G, a trained backdoored graph classifier f, and the ensemble classifier g defined in Eq.* (8) *with subgraph random subgraph sampling (GraphProt-R). Let $\mathcal{G}$ denote the subgraphs with s nodes sampled from G with replacement. Suppose y_A and y_B are the classes with the most votes and the second largest votes during the ensemble. We define $\underline{p_A}$ and $\overline{p_B}$ as the lower and upper bound of probability $\mathbb{P}(f(\mathcal{G}) = y_A)$ and $\mathbb{P}(f(\mathcal{G}) = y_B)$, respectively. We guarantee that the model still predicts class y_A for graphs G_Δ inserted with any trigger size smaller than r if:*

$$2(\frac{n-r}{n})^s > 1 - (\underline{p_A} - \overline{p_B} - \delta_A - \delta_B), \quad (24)$$

where n is the node numbers in G, and $\delta_A = \underline{p_A} - (\lfloor \underline{p_A} \cdot n^s \rfloor)/n^s$, $\delta_B = (\lceil \overline{p_B} \cdot n^s \rceil)/n^s - \overline{p_B}$ are the residuals.

Proof. By setting $n_\Delta = n$ in (12) of Theorem 1, we have the simplified inequality (24). □

Note: In the main paper, $s = \lfloor p \cdot |V_G| \rfloor$.