I. SUPPLEMENTARY MATERIALS

A. Proof for Theorem 1

In this subsection, we first prove the cases in Theorem 1 where our optimization objective in Eq. (8) is both submodular and non-decreasing monotone. Then, we prove other cases in Tab. I and Tab. II where our objective does not satisfy the submodular and non-decreasing monotone properties one by one.

Given the clean graph $G = \{\mathcal{V}, \mathbf{A}, \mathbf{X}\}$ where $\mathbf{A} \in \{0,1\}^{\mathcal{V}\times\mathcal{V}}$ and $\mathbf{X} \in \{0,1\}^{\mathcal{V}\times d_x}$, the feature perturbation a_i has two types: (1) if $\mathbf{X}[v][j] = 0$, we can flip the *j*-th feature value of node v from 0 to 1, i.e., $(v, j, 1, \mathcal{F})$; (2) if $\mathbf{X}[v][j] = 1$, we can flip the *j*-th feature value of node v from 1 to 0, i.e., $(v, j, -1, \mathcal{F})$. Similarly, a topology modification t_j also has two types: (1) if $\mathbf{A}[v][u] = 0$, we can add one new edge between v and u, i.e., $(v, u, 1, \mathcal{T})$; (2) if $\mathbf{A}[v][u] = 1$, we can remove the existing edge between v and u, i.e., $(v, u, -1, \mathcal{T})$. As a result, given an attack set $S = \{a_i, t_j, ...\}$ that consists of feature perturbations and topology modifications, adjacency matrix $\mathbf{A}$, and node features $\mathbf{X}$, we can get the modified adjacency matrix $\hat{\mathbf{A}}_S$ and node features $\hat{\mathbf{X}}_S$ regarding the attack set S accordingly. Based on the above definition, we can change notations in Eq. (8) in the main text and re-state the optimization objective as

$$
\begin{aligned}
f(S) = &\sum_{v\in\mathcal{V}} \left\| \hat{\mathbf{A}}^2_{S,n}[v]\hat{\mathbf{X}} - \mathbf{A}^2_n[v]\mathbf{X} \right\|_p \\
&+ \lambda \sum_{v\in\mathcal{V}} \sum_{u\in\mathcal{N}_v} \left\| \hat{\mathbf{A}}^2_{S,n}[v]\hat{\mathbf{X}} - \mathbf{A}^2_n[u]\mathbf{X} \right\|_p \qquad (1)
\end{aligned}
$$

Then, we first give four lemmas that will be used to prove Theorem 1.

Lemma 1. *Given the constant $p \geq 1$ and a real number x, we can get $||[x]||_p = \sqrt[p]{|x|^p} = |x|$.*

Lemma 2. *Assuming that constants $p = 1$, $c > 0$, $x_1 > 0$, $x_2 > 0$, $x_3 > 0$ and $x_4 > 0$. $\forall x_1 \geq x_3$ and $x_2 \geq x_4$, $f(x_1, x_2, x_3, x_4) = (\sqrt[p]{(x_1+c)^p + x_2^p} - \sqrt[p]{x_1^p + x_2^p}) - (\sqrt[p]{(x_3+c)^p + x_4^p} - \sqrt[p]{x_3^p + x_4^p}) = 0$*

Proof. When $p = 1$, $c > 0$, $x_1 > 0$, $x_2 > 0$, $x_3 > 0$ and $x_4 > 0$, we can get $f(x_1, x_2, x_3, x_4) = c - c = 0$. □

Lemma 3. *Assuming that constants $p \geq 2$, $c > 0$, $x_1 > 0$, $x_2 > 0$, $x_3 > 0$ and $x_4 > 0$. $\exists x_1 \geq x_3$ and $x_2 \geq x_4$, $f(x_1, x_2, x_3, x_4) = (\sqrt[p]{(x_1+c)^p + x_2^p} - \sqrt[p]{x_1^p + x_2^p}) - (\sqrt[p]{(x_3+c)^p + x_4^p} - \sqrt[p]{x_3^p + x_4^p}) < 0$*

Proof. Let $x_1 = 2c, x_2 = 2c, x_3 = c$, and $x_4 = c$:

$$
\begin{aligned}
f(2c, 2c, c, c) &= (\sqrt[p]{3^p + 2^p} - \sqrt[p]{2^{p+1}})c - (\sqrt[p]{2^p + 1} - \sqrt[p]{2})c \\
&= (\sqrt[p]{3^p + 2^p} + \sqrt[p]{2} - \sqrt[p]{2^{p+1}} - \sqrt[p]{2^p + 1})c
\end{aligned}
$$

According to the visualization result[1], it is easy to find that $f(2c, 2c, c, c) < 0$ when $p \geq 2$ and $c > 0$.

□

[1] https://www.desmos.com/calculator/o0vfglvibt

Lemma 4. *Given constants $p \geq 1$, $c_1 > 0$, $c_2 > 0$, and the variable $x > 0$, and $f(x) = \sqrt[p]{(x+c_1)^p + c_2^p} - \sqrt[p]{x^p + c_2^p}$ is a non-decreasing monotone function.*

Proof. We can first get the derivative of $f(x)$ as follows:

$$
\begin{aligned}
f'(x) &= \frac{(x+c_1)^{p-1}}{((x+c_1)^p + c_2^p)^{\frac{p-1}{p}}} - \frac{x^{p-1}}{(x^p + c_2^p)^{\frac{p-1}{p}}} \\
&= (\frac{(x+c_1)^p}{(x+c_1)^p + c_2^p})^{\frac{p-1}{p}} - (\frac{x^p}{x^p + c_2^p})^{\frac{p-1}{p}}
\end{aligned}
$$

Let $g(x) = \frac{x}{x+c_2^p}$, and $g(x)$ is non-decreasing function when $x > 0$ and $c_2 > 0$. It indicates $\frac{(x+c_1)^p}{(x+c_1)^p+c_2^p} > \frac{x^p}{x^p+c_2^p}$, and $f'(x) > 0$. Therefore, $f(x)$ is a non-decreasing function. □

We then given the proof for Theorem 1 as follows.

Proof. For clarification, we prove the cases in Theorem 1 according to attack types.

1) Assuming that we can only flip feature value from 0 to 1. Given two feature perturbation sets S_1 and S_2 where $S_2 \subseteq S_1$, and one new feature perturbation $a_i = (u, j, 1, \mathcal{F})$ where $a_i \notin S_1$ and $a_i \notin S_2$. We then prove $f(S)$ is submodular by showing $f(S_1 \cup \{a_i\}) - f(S_1) \leq f(S_2 \cup \{a_i\}) - f(S_2)$. For clarification, we denote $S_3 = S_1 \cup \{a_i\}$, and $S_4 = S_2 \cup \{a_i\}$. Specifically, we can compute $\triangle f(a_i|S_1) = f(S_1 \cup \{a_i\}) - f(S_1)$ as follows:

$$
\begin{aligned}
\triangle f(a_i|S_1) &= \sum_{v\in\mathcal{V}} (\left\| \mathbf{A}^2_n[v](\hat{\mathbf{X}}_{S_1\cup\{a_i\}} - \mathbf{X}) \right\|_p - \left\| \mathbf{A}^2_n[v](\hat{\mathbf{X}}_{S_1} - \mathbf{X}) \right\|_p) \\
&= \sum_{v\in\mathcal{V}} (\left\| \mathbf{A}^2_n[v]\triangle\mathbf{X}_{S_1\cup\{a_i\}} \right\|_p - \left\| \mathbf{A}^2_n[v]\triangle\mathbf{X}_{S_1} \right\|_p) \\
&= \sum_{v\in\mathcal{V}} (\left\| [\triangle\mathbf{H}_{S_3}[v][0], \cdots, \triangle\mathbf{H}_{S_3}[v][j], \cdots, \triangle\mathbf{H}_{S_3}[v][d_x]] \right\|_p \\
&\quad - \left\| [\triangle\mathbf{H}_{S_1}[v][0], \cdots, \triangle\mathbf{H}_{S_1}[v][j], \cdots, \triangle\mathbf{H}_{S_1}[v][d_x]] \right\|_p) \\
&= \sum_{v\in\mathcal{V}} (\sqrt[p]{|\triangle\mathbf{H}_{S_3}[v][j]|^p + \sum_{i=0, i\neq j}^{d_x} |\triangle\mathbf{H}_{S_3}[v][i]|^p} \\
&\quad - \sqrt[p]{|\triangle\mathbf{H}_{S_1}[v][j]|^p + \sum_{i=0, i\neq j}^{d_x} |\triangle\mathbf{H}_{S_1}[v][i]|^p}) \\
&= \sum_{v\in\mathcal{V}} (\sqrt[p]{|\triangle\mathbf{H}_{S_3}[v][j]|^p + c_v^p} - \sqrt[p]{|\triangle\mathbf{H}_{S_1}[v][j]|^p + c_v^p}),
\end{aligned} \qquad (2)
$$

Where we can get $c_v^p = \sum_{i=0,i\neq j}^{d_x} |\triangle\mathbf{H}_{S_3}[v][i]|^p = \sum_{i=0,i\neq j}^{d_x} |\triangle\mathbf{H}_{S_1}[v][i]|^p$, because $a_i = (u, j, 1, \mathcal{F})$ only influence the *j*-th representation value of each node and thus its representation values in other dimensions are the same. Similarly, we can compute $\triangle f(a_i|S_2) = f(S_2 \cup \{a_i\}) - f(S_2)$ as follows:

$$
\begin{aligned}
\triangle f(a_i|S_2) &= \sum_{v\in\mathcal{V}} (\left\| \mathbf{A}^2_n[v]\triangle\mathbf{X}_{S_2\cup\{a_i\}} \right\|_p - \left\| \mathbf{A}^2_n[v]\triangle\mathbf{X}_{S_2} \right\|_p) \\
&= \sum_{v\in\mathcal{V}} (\sqrt[p]{|\triangle\mathbf{H}_{S_4}[v][j]|^p + d_v^p} - \sqrt[p]{|\triangle\mathbf{H}_{S_2}[v][j]|^p + d_v^p})
\end{aligned}
$$

TABLE I: $\lambda = 0$. $\checkmark^*$ denotes that our objective function is submodular when node feature dimension $d_x = 1$ or the norm distance $p = 1$. Also, $\times$, 1, and 0 denote that our objective function is not submodular, is non-decreasing monotone, and is not non-decreasing monotone, respectively.

	None	**Add Edge**	**Remove Edge**	**Both Topology Modifications**
None	-	$\times, 0$	$\times, 0$	$\times, 0$
Flip Feature from 0 to 1	$\checkmark^*, 1$	$\times, 0$	$\times, 0$	$\times, 0$
Flip Feature from 1 to 0	$\checkmark^*, 1$	$\times, 0$	$\times, 0$	$\times, 0$
Both Feature Perturbations	$\times, 0$	$\times, 0$	$\times, 0$	$\times, 0$

TABLE II: $\lambda > 0$. $\times$ and 0 denote that our objective function is not submodular and non-decreasing monotone, respectively.

	None	**Add Edge**	**Remove Edge**	**Both Topology Modifications**
None	-	$\times, 0$	$\times, 0$	$\times, 0$
Flip Feature from 0 to 1	$\times, 0$	$\times, 0$	$\times, 0$	$\times, 0$
Flip Feature from 1 to 0	$\times, 0$	$\times, 0$	$\times, 0$	$\times, 0$
Both Feature Perturbations	$\times, 0$	$\times, 0$	$\times, 0$	$\times, 0$

Similarly, we can get $d_v^p = \sum_{i=0,i\neq j}^{d_x} |\triangle \mathbf{H}_{S_4}[v][i]|^p = \sum_{i=0,i\neq j}^{d_x} |\triangle \mathbf{H}_{S_2}[v][i]|^p$. Thus, we can obtain $\triangle f(a_i|S_1) - \triangle f(a_i|S_2)$ as follows:

$$
\begin{aligned}
&\triangle f(a_i|S_1) - \triangle f(a_i|S_2) \\
&= \sum_{v\in\mathcal{V}} [(\sqrt[p]{|\triangle \mathbf{H}_{S_3}[v][j]|^p + c_v^p} - \sqrt[p]{|\triangle \mathbf{H}_{S_1}[v][j]|^p + c_v^p}) \\
&- (\sqrt[p]{|\triangle \mathbf{H}_{S_4}[v][j]|^p + d_v^p} - \sqrt[p]{|\triangle \mathbf{H}_{S_2}[v][j]|^p + d_v^p})]
\end{aligned} \tag{3}
$$

a) When $p = 1$ and $d_x \geq 1$, following Lemma 2, we can get

$$
\begin{aligned}
&\triangle f(a_i|S_1) - \triangle f(a_i|S_2) \\
&= \sum_{v\in\mathcal{V}} [(|\triangle \mathbf{H}_{S_3}[v][j]| + c_v - |\triangle \mathbf{H}_{S_1}[v][j]| - c_v) \\
&\qquad - (|\triangle \mathbf{H}_{S_4}[v][j]| + d_v - |\triangle \mathbf{H}_{S_2}[v][j]| - d_v)] \\
&= \sum_{v\in\mathcal{V}} [(|\triangle \mathbf{H}_{S_3}[v][j]| - |\triangle \mathbf{H}_{S_1}[v][j]|) \\
&\qquad - (|\triangle \mathbf{H}_{S_4}[v][j]| - |\triangle \mathbf{H}_{S_2}[v][j]|)]
\end{aligned} \tag{4}
$$

Since we only flip the value of features from 0 to 1 and $a_i = (u, j, 1, \mathcal{F})$, the difference between $|\triangle \mathbf{H}_{S_3}[v][j]|$ (resp., $|\triangle \mathbf{H}_{S_4}[v][j]|$) and $|\triangle \mathbf{H}_{S_1}[v][j]|$ (resp., $|\triangle \mathbf{H}_{S_2}[v][j]|$) is only brought by a_i. Therefore, we can get $|\triangle \mathbf{H}_{S_3}[v][j]| - |\triangle \mathbf{H}_{S_1}[v][j]| = |\triangle \mathbf{H}_{S_4}[v][j]| - |\triangle \mathbf{H}_{S_2}[v][j]|$. Thus, we find that $\triangle f(a_i|S_1) - \triangle f(a_i|S_2) = 0$ and so the function $f(S)$ is submodular.

Also, we can obtain $\triangle f(a_i|S_1) = f(S_1 \cup \{a_i\}) - f(S_1)$ as follows:

$$
\triangle f(a_i|S_1) = \sum_{v\in\mathcal{V}} [(|\triangle \mathbf{H}_{S_3}[v][j]| - |\triangle \mathbf{H}_{S_1}[v][j]|)] \tag{5}
$$

Since $|\triangle \mathbf{H}_{S_3}[v][j]| \geq |\triangle \mathbf{H}_{S_1}[v][j]|$ due to a_i, we can get $f(S_1 \cup \{a_i\}) \geq f(S_1)$. Therefore, the function $f(S)$ is non-decreasing monotone.

b) when $p \geq 2$ and $d_x = 1$, we regard that the node feature only has the j-th dimension. Given $a_i = (u, j, 1, \mathcal{F})$, following Lemma 1 we can get $\triangle f(a_i|S_1) - \triangle f(a_i|S_2)$ as follows:

$$
\begin{aligned}
&\triangle f(a_i|S_1) - \triangle f(a_i|S_2) \\
&= \sum_{v\in\mathcal{V}} [(|\triangle \mathbf{H}_{S_3}[v][j]| - |\triangle \mathbf{H}_{S_1}[v][j]|) \\
&\qquad - (|\triangle \mathbf{H}_{S_4}[v][j]| - |\triangle \mathbf{H}_{S_2}[v][j]|)] = 0
\end{aligned} \tag{6}
$$

Thus, the function $f(S)$ is submodular.

Also, similarly, we can compute $\triangle f(a_i|S_1) = f(S_1 \cup \{a_i\}) - f(S_1)$ as follows

$$
\begin{aligned}
\triangle f(a_i|S_1) &= \sum_{v\in\mathcal{V}} (\sqrt[p]{|\triangle \mathbf{H}_{S_3}[v][j]|^p} - \sqrt[p]{|\triangle \mathbf{H}_{S_1}[v][j]|^p}) \\
&= \sum_{v\in\mathcal{V}} |\triangle \mathbf{H}_{S_3}[v][j]| - |\triangle \mathbf{H}_{S_1}[v][j]|
\end{aligned} \tag{7}
$$

Since for each node $v \in \mathcal{V}$, $|\triangle \mathbf{H}_{S_3}[v][j]| \geq |\triangle \mathbf{H}_{S_1}[v][j]|$, and thus we can get $f(S_1 \cup \{a_i\}) \geq f(S_1)$. Therefore, the function $f(S)$ is non-decreasing monotone.

c) When $p \geq 2$ and $d_x \geq 2$, since S_1 (resp., S_3) contains more feature perturbations that only flip the value of node features from 0 to 1 than S_2 (resp., S_4), we can get that $|\triangle \mathbf{H}_{S_3}[v][j]| \geq |\triangle \mathbf{H}_{S_4}[v][j]|$ and $c_v \geq d_v$ in Eq. (3). Thus, based on Lemma 3, we cannot guarantee $\triangle f(a_i|S_1) - \triangle f(a_i|S_2) \leq 0$ in all cases, and so $f(S)$ is not submodular.

On the other hand, following Lemma 4, we can obtain $f(S_1 \cup \{a_i\}) - f(S_1) = \triangle f(a_i|S_1) \geq 0$ in Eq. (2). Therefore, the function $f(S)$ is non-decreasing monotone. Note that in this sub-case, $f(S)$ cannot satisfy **both** submodular **and** non-decreasing monotone properties.

The above proof in a), b), and c) has proved the cases where our objective $f(S)$ is both submodular and non-decreasing monotone when only flipping feature value from 0 to 1 under $p = 1$ or $d_x = 1$.

2) Similarly to case 1, assuming that we can only flip feature value from 1 to 0. Given two attack sets of feature perturbations S_1 and S_2 where $S_2 \subseteq S_1$, and one new feature perturbation $a_i = (u, j, -1, \mathcal{F})$ where $a_i \notin S_1$ and $a_i \notin S_2$. We then prove $f(S)$ is submodular by showing

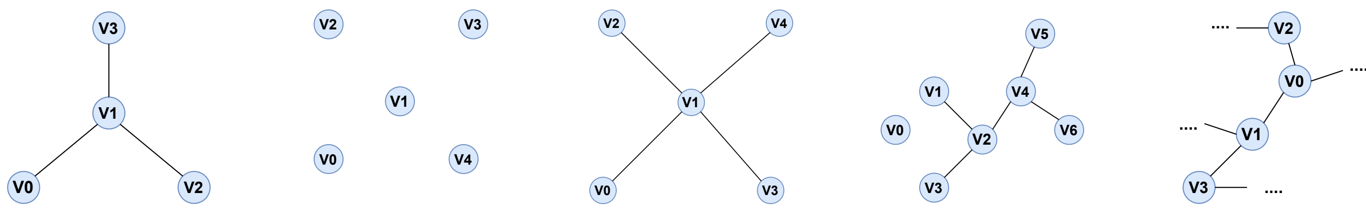

Fig. 1: Counter examples for proving that our function $f(S)$ is not submodular and non-decreasing monotone.

$f(S_1 \cup \{a_i\}) - f(S_1) \leq f(S_2 \cup \{a_i\}) - f(S_2)$. For clarification, we denote $S_3 = S_1 \cup \{a_i\}$, and $S_4 = S_2 \cup \{a_i\}$, and we can compute $\triangle f(a_i|S_1) - \triangle f(a_i|S_2)$ following Eq. (3).

a) When $p = 1$ and $d_x \geq 1$, following Eq. (4), we can get $|\triangle \mathbf{H}_{S_3}[v][j]| - |\triangle \mathbf{H}_{S_1}[v][j]| = |\triangle \mathbf{H}_{S_4}[v][j]| - |\triangle \mathbf{H}_{S_2}[v][j]|$. Thus, we find that $\triangle f(a_i|S_1) - \triangle f(a_i|S_2) = 0$ and so the function $f(S)$ is submodular. Similarly, we can obtain $\triangle f(a_i|S_1) = f(S_1 \cup \{a_i\}) - f(S_1) = \sum_{v \in \mathcal{V}} [(|\triangle \mathbf{H}_{S_3}[v][j]| - |\triangle \mathbf{H}_{S_1}[v][j]|)]$. Since $|\triangle \mathbf{H}_{S_3}[v][j]| \geq |\triangle \mathbf{H}_{S_1}[v][j]|$ due to a_i, we can get $f(S_1 \cup \{a_i\}) \geq f(S_1)$. Therefore, the function $f(S)$ is non-decreasing monotone.

b) when $p \geq 2$ and $d_x = 1$, we regard that the node feature only has the j-th dimension. Given $a_i = (u, j, -1, \mathcal{F})$, following Eq. (6), we can get $\triangle f(a_i|S_1) - \triangle f(a_i|S_2) = 0$. Thus, the function $f(S)$ is submodular.
Also, following Eq. (7), we can compute $f(S_1 \cup \{a_i\}) - f(S_1) = \sum_{v \in \mathcal{V}} |\triangle \mathbf{H}_{S_3}[v][j]| - |\triangle \mathbf{H}_{S_1}[v][j]|$. Since for each node $v \in \mathcal{V}$, the absolute value $|\triangle \mathbf{H}_{S_3}[v][j]| \geq |\triangle \mathbf{H}_{S_1}[v][j]|$, and thus we can get $f(S_1 \cup \{a_i\}) \geq f(S_1)$. Therefore, the function $f(S)$ is non-decreasing monotone.

c) When $p \geq 2$ and $d_x \geq 2$, similar to case 1 (c), we can get $|\triangle \mathbf{H}_{S_3}[v][j]| \geq |\triangle \mathbf{H}_{S_4}[v][j]|$ and $c_v \geq d_v$ in Eq. (4). Thus, based on Lemma 3, we cannot guarantee $\triangle f(a_i|S_1) - \triangle f(a_i|S_2) \leq 0$ in all cases, and so $f(S)$ is not submodular.
On the other hand, following Lemma 4, we can obtain $f(S_1 \cup \{a_i\}) - f(S_1) = \triangle f(a_i|S_1) \geq 0$ in Eq. (2). Therefore, the function $f(S)$ is non-decreasing monotone. But in this sub-case, $f(S)$ cannot satisfy **both** submodular **and** non-decreasing monotone properties.

The above proof in a), b), and c) has proved the cases where our objective $f(S)$ is **both** submodular **and** non-decreasing monotone when we only flip feature value from 1 to 0 under $p = 1$ or $d_x = 1$.

Thus, we have complete the proof for all cases in Theorem 1. □

Without loss of generality, we use counter examples to show our function $f(S)$ is not non-decreasing monotone and submodular in other cases in Tab. I and II. Specifically, we take GCN-mean [1] with one layer for proving, which can be easily extended to other variants of GCN with more layers. The normalized adjacency matrix $\mathbf{A}_n$ of GCN-mean is defined as follows:

$$\mathbf{A}_n = \mathbf{D}^{-1}(\mathbf{A} + \mathbf{I}), \tag{8}$$

where $\mathbf{D}$ is the degree matrix of $\mathbf{A} + \mathbf{I}$ and $\mathbf{I}$ is an identity matrix. Additionally, we set node feature dimension $d_x = 1$, which is adaptive to multi-dimensions, i.e., if $f(S)$ is not non-decreasing monotone and submodular under $d_x = 1$, $f(S)$ must not be non-decreasing monotone and submodular when $d_x \geq 1$.

Theorem 3. *Given graph $G(\mathcal{V}, \mathbf{A}, \mathbf{X})$ where $\mathbf{A} \in \{0,1\}^{|\mathcal{V}| \times |\mathcal{V}|}$ and $\mathbf{X} \in \{0,1\}^{|\mathcal{V}| \times d_x}$, and the hyper-parameter $\lambda = 0$, the optimization objective in Eq.* (1) *is not submodular or non-decreasing monotone when attacker can conduct feature perturbations by flipping feature from 0 to 1 and from 1 to 0.*

Proof. Assuming that the graph consists of four nodes v_0, v_1, v_2, and v_3 as shown in Fig. 1 (a), and the node feature $\mathbf{X}$ is :

$$\mathbf{X} = [[0], [0], [1], [1]]_{4 \times 1}.$$

Also, the adjacency matrix $\mathbf{A}$ and the normalized adjacency matrix $\mathbf{A}_n$ computed by Eq. (8) are listed as follows:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \quad \mathbf{A}_n = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}$$

Thus, the hidden representations $\mathbf{H}$ of these nodes are:

$$\mathbf{H} = \mathbf{A}_n \mathbf{X} = [[0], [\frac{1}{2}], [\frac{1}{2}], [\frac{1}{2}]]_{4 \times 1}.$$

Given two feature perturbation sets $S_1 = \{(v_0, 0, 1, \mathcal{F}), (v_2, 0, -1, \mathcal{F})\}$ and $S_2 = \{(v_0, 0, 1, \mathcal{F})\}$ where $S_2 \subseteq S_1$, and a feature perturbations $a_i = (v_3, 0, -1, \mathcal{F})$. Following Eq. (2), we can obtain $\triangle f(a_i|S_1) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$ and $\triangle f(a_i|S_2) = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$. Thus, $\triangle f(a_i|S_1) > \triangle f(a_i|S_2)$, and so the function $f(S)$ is not submodular.

Also, given a feature perturbation set $S_3 = \{(v_2, 0, -1, \mathcal{F}), (v_3, 0, -1, \mathcal{F})\}$ and a feature perturbation $a_j = (v_1, 0, 1, \mathcal{F})$, we can get that $f(S_3) = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = \frac{3}{2}$, and $f(S_3 \cup \{a_j\}) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$, and thus we find that $\triangle f(a_j|S_3) = f(S_3 \cup \{a_j\}) - f(S_3) < 0$. Therefore $f(S)$ is not non-decreasing monotone. □

Theorem 4. *Given graph $G(\mathcal{V}, \mathbf{A}, \mathbf{X})$ where $\mathbf{A} \in \{0,1\}^{|\mathcal{V}|\times|\mathcal{V}|}$ and $\mathbf{X} \in \{0,1\}^{|\mathcal{V}|\times d_x}$, and the hyper-parameter $\lambda = 0$, the optimization objective in Eq. (1) is not submodular or non-decreasing monotone in the following three cases.*

1) *We only conduct topology modification by adding new edges.*
2) *We only conduct topology modification by removing existing edges.*
3) *We conduct both topology modifications by adding new and removing existing edges.*

Proof. We prove $f(S)$ is not non-decreasing monotone and submodular one by one as follows:

1) Assuming that the graph consists of five singleton nodes $\{v_i\}_{i=0}^4$ as shown in Fig. 1 (b), and the node features are

$$\mathbf{X} = [[0],[1],[1],[1],[1]]_{5\times1}.$$

Specifically, the adjacency matrix $\mathbf{A}$ and the normalized adjacency matrix $\mathbf{A}_n$ computed by Eq. (8) are listed as follows:

$$\mathbf{A} = \begin{bmatrix} 0&0&0&0&0\\ 0&0&0&0&0\\ 0&0&0&0&0\\ 0&0&0&0&0\\ 0&0&0&0&0 \end{bmatrix} \quad \mathbf{A}_n = \begin{bmatrix} 1&0&0&0&0\\ 0&1&0&0&0\\ 0&0&1&0&0\\ 0&0&0&1&0\\ 0&0&0&0&1 \end{bmatrix}$$

Thus, the hidden representations $\mathbf{H}$ of these nodes are:

$$\mathbf{H} = \mathbf{A}_n\mathbf{X} = [[0],[1],[1],[1],[1]]_{5\times1}.$$

Given two sets of topology modification $S_1 = \{(v_0, v_1, 1, \mathcal{T}), (v_1, v_2, 1, \mathcal{T})\}$ and $S_2 = \{(v_0, v_1, 1, \mathcal{T})\}$ where $S_2 \subseteq S_1$, and a topology modification attack $t_i = (v_1, v_3, 1, \mathcal{T})$. We can obtain $\triangle f(t_i|S_1) = \frac{1}{4} - \frac{1}{3} = -\frac{1}{12}$ and $\triangle f(t_i|S_2) = \frac{1}{3} - \frac{1}{2} = -\frac{1}{6}$. Thus, $\triangle f(t_i|S_1) > \triangle f(t_i|S_2)$, and so the function $f(S)$ is not submodular. Also, since $\triangle f(t_i|S_1) < 0$, so the function $f(S)$ is not non-decreasing monotone.

2) To prove that $f(S)$ is not non-decreasing monotone and submodular in case 2, as shown in Fig. 1 (c), we assume that there are five nodes. Specifically, the node features matrix $\mathbf{X}$ is

$$\mathbf{X} = [[0],[1],[0],[1],[1]]_{5\times1}.$$

Also, the adjacency matrix $\mathbf{A}$ and the normalized adjacency matrix $\mathbf{A}_n$ computed by Eq. (8) are listed as follows:

$$\mathbf{A} = \begin{bmatrix} 0&1&0&0&0\\ 1&0&1&1&1\\ 0&1&0&0&0\\ 0&1&0&0&0\\ 0&1&0&0&0 \end{bmatrix} \quad \mathbf{A_n} = \begin{bmatrix} \frac{1}{2}&\frac{1}{2}&0&0&0\\ \frac{1}{5}&\frac{1}{5}&\frac{1}{5}&\frac{1}{5}&\frac{1}{5}\\ 0&\frac{1}{2}&\frac{1}{2}&0&0\\ 0&\frac{1}{2}&0&\frac{1}{2}&0\\ 0&\frac{1}{2}&0&0&\frac{1}{2} \end{bmatrix}$$

Similarly, we can obtain the final representation $\mathbf{H}$ based on GCN-mean with 1-layer as follows:

$$\mathbf{H} = \mathbf{A}_n\mathbf{X} = [[\frac{1}{2}],[\frac{3}{5}],[\frac{1}{2}],[1],[1]]_{5\times1}.$$

Then, given two sets of topology modifications $S_1 = \{(v_0, v_1, -1, \mathcal{T}), (v_1, v_2, -1, \mathcal{T})\}$, $S_2 = \{(v_0, v_1, -1, \mathcal{T})\}$, and a topology modification attack $t_i = (v_1, v_3, -1, \mathcal{T})$. Following Eq. (2), we can obtain $\triangle f(t_i|S_1) = 0$ and $\triangle f(t_i|S_2) = ||[\frac{1}{15}]||_p - ||[\frac{3}{2}]||_p = \frac{1}{15} - \frac{3}{2} < 0$. Thus, $\triangle f(t_i|S_1) > \triangle f(t_i|S_2)$, and so the function $f(S)$ is not submodular. Also, since $\triangle f(t_i|S_2) < 0$, so the function $f(S)$ is not non-decreasing monotone.

3) Since case 1 and case 2 are specific cases of case 3, thus $f(S)$ in case 3 is not non-decreasing monotone and submodular.

□

Corollary 1. *Given graph $G(\mathcal{V}, \mathbf{A}, \mathbf{X})$ where $\mathbf{A} \in \{0,1\}^{|\mathcal{V}|\times|\mathcal{V}|}$ and $\mathbf{X} \in \{0,1\}^{|\mathcal{V}|\times d_x}$, and the hyper-parameter $\lambda = 0$, the optimization objective in Eq. (1) is not submodular or non-decreasing monotone when covered attack types of attacker is any combination of feature perturbations and topology modifications. i.e., the other cases in Tab. I except the case in Theorem 1, Theorem 3, and Theorem 4.*

Proof. In each combination, we can find that at least one attack type does not satisfy the non-decreasing monotone and submodular properties. As a result, the function $f(S)$ is not the non-decreasing monotone and submodular under the combination of attacks. □

We have finish all proofs to support all cases shown in Tab. I. Then, we will prove $f(S)$ is not submodular and non-decreasing monotone in the cases under $\lambda > 0$ shown in Tab. II as follows.

Theorem 5. *Given graph $G(\mathcal{V}, \mathbf{A}, \mathbf{X})$ where $\mathbf{A} \in \{0,1\}^{|\mathcal{V}|\times|\mathcal{V}|}$ and $\mathbf{X} \in \{0,1\}^{|\mathcal{V}|\times d_x}$, and the hyper-parameter $\lambda > 0$, the objective function in Eq. (1) is not submodular or non-decreasing monotone in the following three cases.*

1) *We only conduct feature perturbations by flipping feature values from 0 to 1.*
2) *We only conduct feature perturbations by flipping feature values from 1 to 0.*
3) *We conduct both types of feature perturbations by flipping feature values from 0 to 1 and from 1 to 0.*

Proof. We prove $f(S)$ is not submodular and non-decreasing monotone one by one as follows:

1) Assuming we have 7 nodes $\{v_i\}_{i=0}^6$ as shown in Fig. 1 (d), and each node associates with 1-dimension feature. Specifically, the node features matrix $\mathbf{X}$ is

$$\mathbf{X} = [[0],[0],[0],[0],[1],[0],[1]]_{7\times1}.$$

Also, the adjacency matrix $\mathbf{A}$ and the normalized adjacency matrix $\mathbf{A}_n$ computed by Eq. (8) are listed as follows:

$$\mathbf{A} = \begin{bmatrix} 0&0&0&0&0&0&0\\ 0&0&1&0&0&0&0\\ 0&1&0&1&1&0&0\\ 0&0&1&0&0&0&0\\ 0&0&1&0&0&1&1\\ 0&0&0&0&1&0&0\\ 0&0&0&0&1&0&0 \end{bmatrix}$$

$$\mathbf{A}_n = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}$$

Thus, the hidden representations $\mathbf{H}$ of these nodes are:

$$\mathbf{H} = \mathbf{A}_n\mathbf{X} = [[0],[0],[\frac{1}{4}],[0],[\frac{1}{2}],[\frac{1}{2}],[1]]_{7\times1}.$$

Then, given two sets of feature perturbations $S_1 = \{(v_0,0,1,\mathcal{F}),(v_1,0,1,\mathcal{F})\}$, $S_2 = \{(v_0,0,1,\mathcal{F})\}$, and a feature perturbation attack $a_i = (v_3,0,1,\mathcal{F})$. Following Eq. (2), we can obtain $\triangle f(a_i|S_1) = \frac{3}{2}+\frac{3}{2}\lambda$ and $\triangle f(a_i|S_2) = \frac{3}{2}+\frac{1}{2}\lambda$. Thus, $\triangle f(a_i|S_1) > \triangle f(a_i|S_2)$, and so the function $f(S)$ is not submodular.

To prove $f(S)$ is not non-decreasing monotone in case 1, we use the graph in Fig. 1 (e) as an example. Assuming that node v_0 with feature [0] has $2n-1$ neighbors with feature [0] or [1], and the representation of node v_0 is $[\frac{n-1}{2n}]$. In particular, each neighbor of v_0 connects $2n-1$ nodes to guarantee that its hidden representation is $[\frac{1}{2}]$. Then, we take a v_0's neighbor v_1 with feature [0] as the perturbed node. Specifically, each neighbor v_3 of node v_1 also has $2n-1$ neighbors to guarantee its hidden representation is $[\frac{n+1}{2n}]$, and each neighbor of v_3 also has $2n-1$ neighbors whose final representation is $[\frac{n+2}{2n}]$.

When we flip the feature of node v_1 from 0 to 1, i.e., $a_1 = \{v_1,0,1,\mathcal{F}\}$, the representation of v_0 will become $[\frac{1}{2}]$ from $[\frac{2n-1}{2n}]$, and the final representation of v_1 will become $[\frac{n+1}{2n}]$ from $[\frac{1}{2}]$, and the representation of each neighbor v_3 of v_1 will become $[\frac{n+2}{2n}]$. According Eq. (1), $\triangle f(a_1|\{\}) = \frac{2n+1}{2n} - \frac{(2n-1)(2n-2)}{2n}\lambda$. We can find when $\lambda > \frac{2n+1}{(2n-1)(2n-2)}$, $\triangle f(a_1|\{\}) < 0$. Therefore, $f(S)$ is not non-decreasing monotone in case 1.

2) The same as case 1, assuming that we have 7 nodes $\{v_i\}_{i=0}^{6}$ and the graph topology is shown in Fig. 1 (d), and each node associates with 1-dimension feature. Unlike case 1, the node features matrix $\mathbf{X}$ is defined as:

$$\mathbf{X} = [[1],[1],[0],[1],[1],[0],[1]]_{7\times1}.$$

Then, given two sets of feature perturbations $S_1 = \{(v_0,0,-1,\mathcal{F}),(v_1,0,-1,\mathcal{F})\}$, $S_2 = \{(v_0,0,1,\mathcal{F})\}$, and a feature perturbation attack $a_i = (v_3,0,-1,\mathcal{F})$. Following Eq. (2), $\triangle f(a_i|S_1) = \frac{3}{4}+\frac{5}{4}\lambda$ and $\triangle f(a_i|S_1) = \frac{3}{4}-\frac{1}{4}\lambda$. Thus, $\triangle f(a_i|S_1) > \triangle f(a_i|S_2)$, and so the function $f(S)$ is not submodular.

Similar to case 1, it is easy to find a graph to show $f(S)$ is not non-decreasing monotone. Thus, we do not elaborate it in detail.

3) Since case 1 and case 2 are specific ones of case 3, thus $f(S)$ is not submodular or non-decreasing monotone in case 3.

□

Theorem 6. *Given graph $G(\mathcal{V},\mathbf{A},\mathbf{X})$ where $\mathbf{A} \in \{0,1\}^{|\mathcal{V}|\times|\mathcal{V}|}$ and $\mathbf{X} \in \{0,1\}^{|\mathcal{V}|\times d_x}$, and the hyper-parameter $\lambda > 0$, the optimization objective in Eq. (1) is not submodular or non-decreasing monotone in the following three cases.*

1) *We only conduct topology modifications by adding new edges.*
2) *We only conduct topology modifications by removing existing edges.*
3) *We conduct both types of topology modifications by adding new edges and removing existing edges.*

Proof. We prove $f(S)$ is not non-decreasing monotone and submodular one by one as follows:

1) The case 1 in Theorem 4 is a special case for this case. Therefore, $f(S)$ is not submodular and non-decreasing monotone.
2) Assuming that we have 4 nodes $\{v_i\}_{i=0}^{3}$ as shown in Fig. 1 (a), and each node associates with 1-dimension feature. Specifically, the node features matrix $\mathbf{X}$ is

$$\mathbf{X} = [[0],[1],[0],[1]]_{4\times1}.$$

Also, the adjacency matrix $\mathbf{A}$ and the normalized adjacency matrix $\mathbf{A}_n$ computed by Eq. (8) are listed as follows:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \quad \mathbf{A}_n = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}$$

Thus, the hidden representations $\mathbf{H}$ of these nodes are:

$$\mathbf{H} = \mathbf{A}_n\mathbf{X} = [[\frac{1}{2}],[\frac{1}{2}],[\frac{1}{2}],[1]]_{4\times1}.$$

Given two topology modification sets $S_1 = \{(v_0,v_1,-1,\mathcal{T})\}$ and $S_2 = \{\}$ where $S_2 \subseteq S_1$, and a topology modification attack $t_i = (v_1,v_2,-1,\mathcal{T})$. Following Eq. (2), we can obtain $\triangle f(t_i|S_1) = \frac{5}{6}+\frac{5}{6}\lambda$ and $\triangle f(t_i|S_2) = \frac{2}{3}+\frac{2}{3}\lambda$.Thus, $\triangle f(t_i|S_1) > \triangle f(t_i|S_2)$, and so the function $f(S)$ is not submodular.

To prove $f(S)$ is not non-decreasing monotone in case 2, we use the graph in Fig. 1 (e) as example. Assuming that node v_0 with feature [1] has $2n$ neighbors, and its representation is $[\frac{n}{2n+1}]$. Also, each neighbor of v_0 connects $2n+1$ nodes to guarantee that its hidden representation is $[\frac{1}{2}]$. Also, we take the edge $e(v_0,v_1)$ as the attacked edge where v_1's feature (a neighbor of v_0) is [0]. Also, each neighbor v_3 of v_1 also has $2n$ neighbors to guarantee its final representation is $[\frac{n}{2n+1}]$.

When we remove the edge $e(v_0,v_1)$, i.e., $t_1 = \{v_0,v_1,-1,\mathcal{T}\}$, the representation of v_0 will become $[\frac{1}{2}]$ from $[\frac{n}{2n+1}]$, and the final representation of v_1 will become $[\frac{n}{2n+1}]$ from $[\frac{1}{2}]$, and the representation of other neighbors of v_0 and v_1 will not change due to the 1-layer GCN. Therefore, we can obtain $\triangle f(t_1|\{\}) = \frac{1}{2n+1} - \frac{4n+1}{4n+2}\lambda$. We can find when $\lambda > \frac{2}{4n+1}$, $\triangle f(t_1|\{\}) < 0$. Therefore, $f(S)$ is not non-decreasing monotone in case 2.

3) Since case 1 and case 2 are specific ones to case 3, thus $f(S)$ is not submodular and non-decreasing monotone in case 3.

$\square$

Corollary 2. *Given graph $G(\mathcal{V}, \mathbf{A}, \mathbf{X})$ where $\mathbf{A} \in \{0,1\}^{|\mathcal{V}|\times|\mathcal{V}|}$ and $\mathbf{X} \in \{0,1\}^{|\mathcal{V}|\times d_x}$, and the hyper-parameter $\lambda > 0$, the objective function in Eq. (1) is not submodular or non-decreasing monotone when covered attack types of attacker is any combination of feature perturbations and topology modifications, i.e., the other cases in Tab. II except the cases in Theorem 5 and Theorem 6.*

Proof. Similar to Corollary 1, in each combination, we can find that at least one attack type in these combinations does not satisfy the non-decreasing monotone and submodular properties. As a result, the function $f(S)$ is not the non-decreasing monotone and submodular under the combination of attacks. $\square$

The above theorems and corollaries have proved Theorem 1 and all other conclusions in Tab. I and Tab. II.

B. Proof for Theorem 2

Proof. For simplicity and clarification, taking a 1-layer GCN $\mathcal{M}_\theta$ without non-linear functions, we can obtain the node label probability as $\hat{\mathbf{Z}} = \hat{\mathbf{A}}_n\hat{\mathbf{X}}\mathbf{W}$. Given $\mathcal{M}_\theta$ and the poison graph $\hat{G}(\mathcal{V}, \hat{\mathbf{A}}, \hat{\mathbf{X}}, \mathbf{Y})$, we can derive the $(\hat{\mathbf{A}}_n\hat{\mathbf{X}})[\mathcal{V}^{la}]$ as

$$(\hat{\mathbf{A}}_n\hat{\mathbf{X}})[\mathcal{V}^{la}] = \frac{1}{d|\mathcal{Y}|} \times \begin{bmatrix} d & \cdots & d \\ d & \cdots & d \\ d & \cdots & d \end{bmatrix}_{|\mathcal{V}^{la}|\times|\mathcal{Y}|},$$

where $\hat{\mathbf{A}}_n = \hat{\mathbf{D}}^{-\frac{1}{2}}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-\frac{1}{2}}$ and $\hat{\mathbf{D}}_{ii} = \sum_j \hat{\mathbf{A}}_{ij}$. We observe that each row of $(\hat{\mathbf{A}}_n\hat{\mathbf{X}})[\mathcal{V}^{la}]$ has the same elements due to the same context. Thus, the label probability predictions are $\hat{\mathbf{Z}}[\mathcal{V}^{la}] = (\hat{\mathbf{A}}_n\hat{\mathbf{X}})[\mathcal{V}^{la}]\mathbf{W}$, where all elements are the same. Since nodes in $\mathcal{V}^{la}$ are not in one category, and thus we can obtain $\mathcal{L}_{gnn}(\mathcal{M}_\theta, \hat{G}(\mathcal{V}, \hat{\mathbf{A}}, \hat{\mathbf{X}}, \mathbf{Y})) > 0$.

Given the augmented adjacency matrix $\hat{\mathbf{A}}'$ and GCN $\mathcal{M}_\theta$, since nodes in the same category have the same node representation, we first sample one node from each category to form a smaller node set $\mathcal{V}^s \in \mathcal{V}^{la}$. Therefore, we can obtain $(\hat{\mathbf{A}}'_n\hat{\mathbf{X}})[\mathcal{V}^s]$ as follows.

$$(\hat{\mathbf{A}}'_n\hat{\mathbf{X}})[\mathcal{V}^s] = \frac{1}{d|\mathcal{Y}| + \alpha} \times \begin{bmatrix} d+\alpha & \cdots & d \\ d & \cdots & d \\ d & \cdots & d+\alpha \end{bmatrix}_{|\mathcal{Y}|\times|\mathcal{Y}|},$$

According to Sherman–Morrison formula [2], the inverse matrix of $(\hat{\mathbf{A}}'_n\hat{\mathbf{X}})[\mathcal{V}^s]$ is derived as:

$$((\hat{\mathbf{A}}'_n\hat{\mathbf{X}})[\mathcal{V}^s])^{-1} = \mathbf{I} - \frac{(\hat{\mathbf{A}}'_n\hat{\mathbf{X}})[\mathcal{V}^s]}{1 - trace((\hat{\mathbf{A}}'_n\hat{\mathbf{X}})[\mathcal{V}^s]}.$$

Thus, we can find an optimal parameters and obtain the prediction for $\mathcal{V}^{la}$:

$$\mathbf{W}_* = ((\hat{\mathbf{A}}'_n\hat{\mathbf{X}})[\mathcal{V}^s])^{-1}\mathbf{Y}[\mathcal{V}^s],$$
$$\hat{\mathbf{Z}}'[\mathcal{V}^{\mathbf{la}}] = (\hat{\mathbf{A}}'_n\hat{\mathbf{X}})[\mathcal{V}^{la}]\mathbf{W}_* = \mathbf{Y}.$$

Therefore, the training loss $\mathcal{L}_{gnn}(\mathcal{M}_\theta, \hat{G}'(\mathcal{V}, \hat{\mathbf{A}}', \hat{\mathbf{X}}, \mathbf{Y})) = 0$. Thus, we have:

$$\mathcal{L}_{gnn}(\mathcal{M}_\theta, \hat{G}'(\mathcal{V}, \hat{\mathbf{A}}', \hat{\mathbf{X}}, \mathbf{Y})) < \mathcal{L}_{gnn}(\mathcal{M}_\theta, \hat{G}(\mathcal{V}, \hat{\mathbf{A}}, \hat{\mathbf{X}}, \mathbf{Y})).$$

$\square$

References

[1] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Advances in neural information processing systems*, 2017, pp. 1024–1034.

[2] Wikipedia contributors, "Sherman–morrison formula — Wikipedia, the free encyclopedia," 2021. [Online]. Available: https://en.wikipedia.org/wiki/Sherman%E2%80%93Morrison_formula