On Strengthening and Defending Graph Reconstruction Attack with Markov Chain Approximation

Zhanke Zhou, Chenyu Zhou, Xuan Li, Jiangchao Yao, Quanming Yao, Bo Han

Paper

Code

Abstract: *We perform the first comprehensive study of graph reconstruction attack (GRA). By taking GNN as a Markov chain and attacking GNN via a flexible chain approximation, we systematically explore the underneath principles of GRA, and propose two information theory-guided mechanisms for attack and defense, respectively.*

Research Problem: Graph Reconstruction Attack (GRA)

Although powerful graph neural networks (GNNs) have boosted numerous real-world applications, the **privacy concerns** behind these applications raise with the Graph Reconstruction Attacks (GRA). Nevertheless, The **general principles for strengthening and defending** GRA are under-explored.

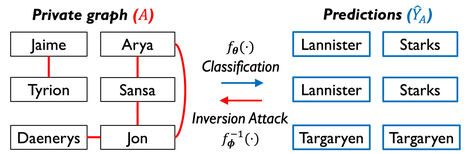


Figure 1. An illustration of Graph Reconstruction Attack.

The **forward** (blue arrow) is to predict the node category $\hat{Y}$, i.e., the family name; the **backward** (red arrow) is to recover the original adjacency A, i.e., the kinship among characters (red edges).

Given a set of prior knowledge K and a trained GNN $f_{\theta^*}(\cdot)$, the graph reconstruction attack aims to recover the original linking relations A^* of the training graph $G_{train} = (A, X)$, namely,

$$\text{GRA:} \quad \hat{A}^* = \arg\max_{\hat{A}} \mathbb{P}(\hat{A} | f_{\theta^*}, \mathcal{K}).$$

Here, $\mathbf{P}(\cdot)$ is the attack method to generate $\hat{A}$, and K can be any subset of $\{X, Y, H_A, \hat{Y}_A\}$ with node features X, node label Y, node embeddings H_A, and classification outputs $\hat{Y}_A$. Note that GRA is conducted in a post-hoc manner. That is, after training the target GNN $f_{\theta^*}(\cdot)$, the GRA is conducted.

Modeling GRA as Markov Chain Approximation

To adaptively support the white-box GRA that leverages the target model and any prior knowledge, and to properly utilize the interplaying variables of GNN forward in a generic manner, we cast the GRA problem as **approximating the original Markov chain by the attack chain.**

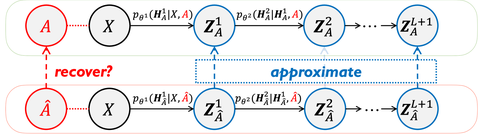


Figure 2. Modeling the GRA problem as approximating the **original chain** (upper chain with the original adjacency A) by the **attack chain** (lower chain with the recovered adjacency $\hat{A}$).

A Comprehensive Study of GRA

$$\text{ORI-chain: } Z^0 \xrightarrow[\theta^1]{A} Z_A^1 \xrightarrow[\theta^2]{A} Z_A^2 \to \cdots \xrightarrow[\theta^{L+1}]{A} Z_A^{L+1}$$

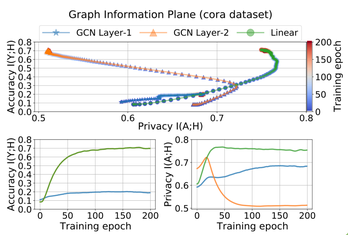
$$\text{GRA-chain: } Z^0 \xrightarrow[\theta^1]{\hat{A}} Z_{\hat{A}}^1 \xrightarrow[\theta^2]{\hat{A}} Z_{\hat{A}}^2 \to \cdots \xrightarrow[\theta^{L+1}]{\hat{A}} Z_{\hat{A}}^{L+1}$$

Observation 1: a single variable in ORI-chain can recover the original adjacency to some extent.

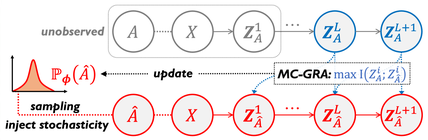
Observation 2: the linear combination of informative terms only brings marginal improvements in recovering.

Observation 3: the training procedure contains two main phases, i.e., fitting and compressing.

The information plane of $\big(I(A;H), I(Y;H)\big)$



MC-GRA: Markov Chain-based Graph Reconstruction Attack

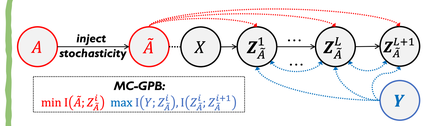


$$\text{MC-GRA: } \hat{A}^* = \arg\max_{\hat{A}} \underbrace{\alpha_p I(\boldsymbol{H}_A; \boldsymbol{H}_{\hat{A}}^i)}_{\text{propagation approximation}} + \underbrace{\alpha_o I(\boldsymbol{Y}_A; \boldsymbol{Y}_{\hat{A}}) + \alpha_s I(Y; \boldsymbol{Y}_{\hat{A}})}_{\text{outputs approximation}} - \underbrace{\alpha_c H(\hat{A})}_{\text{complexity}}.$$

Algorithm 1: A chain-based attack method MC-GRA to enhance attack

- Extract the knowledge stored in the target model while utilize all the prior knowledge simultaneous
- Promote $I(Z_{\hat{A}}^i; Z_A^i = K_i)$ as it provides supervision signals that can be tractably approximated
- Technically, parametrize of the recovered adjacency and optimize with injected stochasticity

MC-GPB: Markov Chain-based Graph Privacy Bottleneck

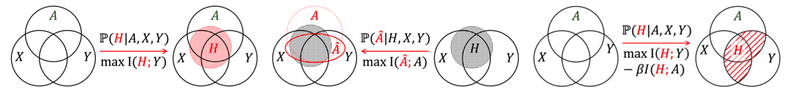


$$\text{MC-GPB: } \boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \sum_{i=1} \underbrace{-I(Y; \boldsymbol{H}_A^i)}_{\text{accuracy}} + \underbrace{\beta^i I(A; \boldsymbol{H}_A^i)}_{\text{privacy}} + \sum_{i=1}^{L-1} \underbrace{\beta_c I(\boldsymbol{H}_A^i; \boldsymbol{H}_A^{i+1})}_{\text{complexity}}.$$

Algorithm 2: A chain-based defense method MC-GPB to protect GNNs

- Make the graph representations H contain less information about adjacency A
- Balance the trade-off between accuracy $I(Y;H)$ and privacy $I(A;H)$
- Technically, apply differentiable similarity measurements and optimize with injected stochasticity

To Recover Better, You Must Extract More.
To Learn Safer, You Must Forget More.



(a) Standard training. (b) Reconstruction attack by MC-GRA. (c) Defensive training by MC-GPB.

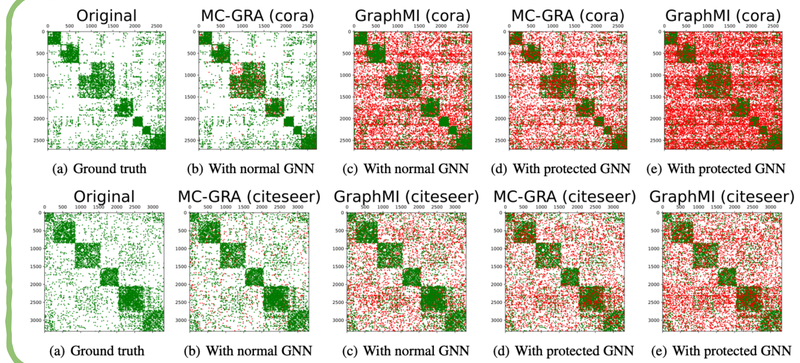
Empirical Study

Table 3: Results of MC-GRA with standard GNNs. Relative promotions (in %) are computed w.r.t. results in Tab. 2.

X	H_A	$\hat{Y}_A$	Y	Cora	Citeseer	Polblogs	USA	Brazil	AIDS
✓	✓			.864 (10.6% ↑)	.912 (3.5% ↑)	.831 (8.9% ↑)	.883 (3.8% ↑)	.771 (1.7% ↑)	.574 (10.1% ↑)
✓		✓		.839 (7.4% ↑)	.902 (2.3% ↑)	.836 (8.2% ↑)	.913 (10.5% ↑)	.800 (9.2% ↑)	.567 (8.8% ↑)
✓			✓	.896 (5.5% ↑)	.918 (1.2% ↑)	.837 (18.7% ↑)	.825 (13.3% ↑)	.753 (22.8% ↑)	.574 (9.9% ↑)
✓	✓	✓		.866 (10.8% ↑)	.921 (4.5% ↑)	.839 (9.9% ↑)	.878 (3.5% ↑)	.776 (2.6% ↑)	.572 (9.7% ↑)
✓	✓		✓	.905 (6.5% ↑)	.930 (2.5% ↑)	.832 (6.8% ↑)	.878 (3.5% ↑)	.758 (2.0% ↑)	.603 (15.5% ↑)
✓		✓	✓	.897 (5.6% ↑)	.928 (2.3% ↑)	.839 (6.8% ↑)	.870 (3.3% ↑)	.758 (3.7% ↑)	.567 (8.6% ↑)
✓	✓	✓	✓	.904 (6.4% ↑)	.931 (2.6% ↑)	.853 (9.2% ↑)	.870 (1.9% ↑)	.760 (5.9% ↑)	.588 (12.6% ↑)

Table 4: Results of GRA with MC-GPB protected GNNs. Relative reductions are computed w.r.t. results in Tab. 1. $I(A; \boldsymbol{H}_A), I(A; \hat{\boldsymbol{Y}}_A)$ are non-learnable GRA (He et al., 2021a). $I(A; \boldsymbol{H}_{\hat{A}}^1)$ is learnable, *i.e.*, GraphMI (Zhang et al., 2021).

MI	Cora	Citeseer	Polblogs	USA	Brazil	AIDS
$I(A; \boldsymbol{H}_A)$	.706 (7.8% ↓)	.750 (1.3% ↓)	.724 (5.1% ↓)	.716 (15.8% ↓)	.745 (1.7% ↓)	.564 (3.4% ↓)
$I(A; \hat{\boldsymbol{Y}}_A)$	.704 (0.1% ↓)	.730 (1.7% ↓)	.705 (8.7% ↓)	.587 (28.9% ↓)	.692 (5.5% ↓)	.559 (0.4% ↓)
$I(A; \boldsymbol{H}_{\hat{A}}^1)$	.625 (9.9% ↓)	.691 (9.8% ↓)	.506 (26.3% ↓)	.300 (64.5% ↓)	.609 (25.1% ↓)	.514 (10.6% ↓)
Acc.	.734 (3.0% ↓)	.602 (4.4% ↓)	.830 (1.1% ↓)	.391 (16.8% ↓)	.808 (5.1% ↑)	.668 (0.0% ↑)



(a) Ground truth (b) With normal GNN (c) With normal GNN (d) With protected GNN (e) With protected GNN

(a) Ground truth (b) With normal GNN (c) With normal GNN (d) With protected GNN (e) With protected GNN