Link Membership Inference Attacks against Unsupervised Graph Representation Learning

Xiuling Wang (presenter), Stevens Institute of Technology, NJ, US
Wendy Hui Wang, Stevens Institute of Technology, NJ, US
Unsupervised Graph Representation Learning (UGRL)

- Goal: learn a mapping function $\Phi: v \rightarrow Z$, while preserving the local structure information
  - The learning process does not require labeled data
  - Orders of preserved proximity: first-order, second-order, and high-order proximity
- Advantages
  - Independent from downstream learning tasks
  - No requirement for labeled data
Motivation

• **UGRL embeddings are not safe against privacy inference attacks**
  • Our focus: Link Membership Inference Attacks (LMIA$s$)
    • Attacker’s goal: infer whether any two nodes are connected in the target graph from the UGRL embeddings.

• Why LMIA is important?
  • Node connections in the input graph may contain sensitive information, such as social relations and mobility traces.
Insufficiency of Prior Works

<table>
<thead>
<tr>
<th>Graph Representation Learning</th>
<th>Adversary knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td>Unsupervised</td>
</tr>
<tr>
<td>[1] GNN</td>
<td></td>
</tr>
<tr>
<td>[3] GNN</td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>DeepWalk, node2vec, LINE, GAE</td>
</tr>
</tbody>
</table>

- Most of the works focus on GNN model (supervised model).
- None of the work investigates LMIA against UGRL models.
- None of the existing LMIA against graph representation models can be readily adapted to UGRL models.
  - [1] utilizes graph embedding for inference, as opposed to node embeddings.
  - [2] requires the auxiliary subgraph sampled from the target graph for inference.

Contributions

- Design of attacks and evaluation
  - We consider two distinct types of adversaries
    - The adversary has access solely to the node embeddings
    - The adversary is equipped with knowledge of a shadow graph
  - We devise two link membership inference attacks tailored to each type of adversary
  - We demonstrate the effectiveness of both attacks against four UGRL models

- Factor analysis
  - We investigate how varying degrees of preserved structural information in the embeddings impact the performance of LMIA.

- Design of defense mechanisms and evaluation
  - We design new defense mechanisms that add noise to the least important embedding dimensions only.
  - We demonstrate our defense mechanisms address the trade-off between defense and embedding quality.
Roadmap

- Introduction
- Our attack
- Evaluation
- Defense
- Conclusion
Problem Formulation

- Adversary Knowledge $K_{Aug}$
  - Node embeddings $Z$ of the target graph
  - Shadow graph $G_{\text{shadow}}$ (optional)

- LMIA is designed as a mapping function:

$$f: \overrightarrow{z_i}, \overrightarrow{z_j}, K_{Aug} \rightarrow \{0, 1\}$$

- Node embedding of $(v_i, v_j)$
- Adversary’s knowledge
  - 0: $v_i$ and $v_j$ are not linked in the target graph
  - 1: $v_i$ and $v_j$ are linked in the target graph
Unsupervised Attack

Without shadow graph: k-means clustering algorithm (k=2)

**Clustering phase**

- **Node embeddings** $Z_{target}$ from $G_{target}$
- For each node pair $(v_i, v_j)$ in $Z_{target}$:
  - Dot product similarity, Cosine similarity, and Euclidean distance

**Inference phase**

- 2-means clustering
- Embeddings of target node pair $Z(v_i), Z(v_j)$
- Is edge $e(v_i, v_j)$ in training graph $G_{target}$?

$\text{sim}()$: Dot product similarity, Cosine similarity, and Euclidean distance
Supervised Attack

With shadow graph: trains a binary classifier as the attack model

Member edge $e(v_i, v_j)$ in $G_{\text{Shadow}}$

$$\text{sim}_1(\Phi^*_I(v_i), \Phi^*_I(v_j)) || \text{sim}_t(\Phi^*_I(v_i), \Phi^*_I(v_j)), "1"$$

...\ 

Non-member edge $e(v_i, v_j)$ in $G_{\text{Shadow}}$

$$\text{sim}_1(\Phi^*_I(v_i), \Phi^*_I(v_j)) || \text{sim}_t(\Phi^*_I(v_i), \Phi^*_I(v_j)), "0"$$

...
Roadmap

- Introduction
- Our attack
- Evaluation
- Defense
- Conclusion
Setup

- Datasets: DBLP, LastFm, Cora, Citeseer, and Pubmed
- UGRL models: DeepWalk, node2vec, LINE, and GAE
- Two settings:
  - Non-transfer setting: both shadow and target graphs are sampled from the same dataset
  - Transfer setting: the shadow graph and the target graph are sampled from different datasets
- Evaluation metrics:
  - Attack effectiveness: attack accuracy, AUC, True-Positive Rate at False-Positive Rates (TPR@FPR)
  - Target model performance: AUC of node classification
- Baselines:
  - Baseline-1: an ensemble of three sub-attacks (threshold-based attack with a single similarity metric)
  - Baseline-2: an encoder-decoder network with the adversary knowledge of the auxiliary graph [1]
  - Baseline-3 (for unsupervised attack): replace the concatenation of embedding similarities with the concatenation of embeddings
  - Baseline-4 (for supervised attack): replace the concatenation of embedding similarities with the concatenation of embeddings

Effectiveness of LMIA (1/2)

Setting 1 (Non-transfer setting)

Observations:

- The attack accuracy of Attack 1 (unsupervised attack) & Attack 2 (supervised attack) is significantly higher than 0.5.
- Attack 2 (supervised attack) outperforms all four baselines in all the settings.
Effectiveness of LMIA (2/2)

Setting 2 (Transfer setting)

Observation:
- The attack remains successful under the transfer setting, even when the target and shadow datasets are from different domains.
Why Our Attacks Work

Observation:

- Members and non-members are distinguishable based on the attack features across all four models.
Observation:

- Embeddings that retain a lower level of proximity information are more susceptible to LMIA.
Impact of the Number of Embedding Dimensions on LMIA Performance

Observation:

- Embeddings of higher dimensions are more vulnerable to LMIA.
Roadmap

- Introduction
- Our attack
- Evaluations
- Defense
- Conclusion
Defense Mechanisms

- Key idea: add noise to the embeddings from UGRL models
- Challenge: add noise can cause significant quality loss on node embeddings
- Our solution: selectively add noise only to $[d \times r]$ dimensions of embeddings that are the least important ($r$: perturbation ratio)
  - **Step 1**: Estimate the importance of embedding dimensions
    - Permutation-based importance (PERM) [1]
    - SHAP value-based importance [2]
    - MDI-based importance [3]
  - **Step 2**: Rank the embedding dimensions based on their importance
  - **Step 3**: Add Laplace noise to $[d \times r]$ embedding dimensions of the lowest importance
    - Distribution of Laplace noise: $\Delta=\frac{1}{2b}e^{-\frac{x-\mu}{b}}$

---

Setup

- Baselines:
  - Baseline-1 (DP): differentially private deep learning method (DP-SGD) [1] that adds Laplace noise to the gradients during model training
  - Baseline-2 (AdvR): add an adversarial regularizer to the objective function as a min-max problem [2]

- Parameters:
  - Perturbation ratio $r = \{0.2, 0.4, 0.6, 0.8, 1\}$
  - Noise scale $b = \{0.1, 0.5, 1, 5, 10\}$
  - Regularization parameter $\lambda = \{0.5, 1, 5, 10, 20\}$

- Evaluation metrics:
  - Attack effectiveness: attack accuracy
  - Target model performance: AUC of node classification

Observations:

- Our perturbation methods are highly effective against LMIA.
- The defense strength increases with a higher perturbation ratio.
- The three importance-based methods demonstrate similar defense performance.
Defense Effectiveness (2/2)

Observations:

- The defense power of our perturbation methods increases at higher noise scales.
- Our methods provide defense capabilities comparable to DP.
Trade-off between Defense Effectiveness and Embedding Quality

- Trade-off measurement:
  - Step 1: we draw a defense-quality ROC curve
    - Defense effectiveness: 1 – attack accuracy
    - Embedding quality: AUC of node classification
  - Step 2: we measure the defense-quality trade-off as the Area Under the Curve (AUC) of the defense-quality ROC curve

<table>
<thead>
<tr>
<th>Method</th>
<th>Cora</th>
<th></th>
<th>Cora</th>
<th></th>
<th>Cora</th>
<th></th>
<th>Cora</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>node2vec</td>
<td>DeepWalk</td>
<td>LINE</td>
<td>GAE</td>
<td>node2vec</td>
<td>DeepWalk</td>
<td>LINE</td>
<td>GAE</td>
</tr>
<tr>
<td>PERM</td>
<td>0.253</td>
<td>0.326</td>
<td>0.365</td>
<td>0.247</td>
<td>0.303</td>
<td>0.229</td>
<td>0.274</td>
<td>0.243</td>
</tr>
<tr>
<td>MDI</td>
<td>0.258</td>
<td>0.341</td>
<td>0.366</td>
<td>0.252</td>
<td>0.401</td>
<td>0.258</td>
<td>0.282</td>
<td>0.258</td>
</tr>
<tr>
<td>SHAP</td>
<td>0.236</td>
<td>0.318</td>
<td>0.366</td>
<td>0.25</td>
<td>0.327</td>
<td>0.192</td>
<td>0.283</td>
<td>0.249</td>
</tr>
<tr>
<td>DP</td>
<td>0.211</td>
<td>0.271</td>
<td>0.361</td>
<td>0.042</td>
<td>0.257</td>
<td>0.155</td>
<td>0.265</td>
<td>0.04</td>
</tr>
<tr>
<td>AdvR</td>
<td>0.177</td>
<td>0.271</td>
<td>0.312</td>
<td>0.159</td>
<td>0.259</td>
<td>0.157</td>
<td>0.269</td>
<td>0.18</td>
</tr>
</tbody>
</table>

- Observations:
  - Our defense mechanisms outperform both baselines in terms of the defense-quality trade-off.
  - The MDI-based method exhibits the most favorable trade-off among the five defense methods.
Conclusion

• Our contributions:
  • We design LMIA{\textit{s}} for two different settings and evaluate their effectiveness against four state-of-the-art UGRL models.
  • We investigate how varying degrees of preserved structural information in the embeddings impact the performance of LMIA.
  • We propose effective defense mechanisms that introduce perturbation to the least important dimensions of embeddings.

• Future work:
  • We will investigate the vulnerability of UGRL models to other types of attacks (e.g., attribute inference attacks and model inversion attacks).
  • We will design new attacks that leverage knowledge transfer techniques for UGRL models.
Thank you for your attention!

Q&A

Contact: xwang193@stevens.edu
       Hui.wang@stevens.edu