Stealing Links from Graph Neural Networks

Xinlei He¹, Jinyuan Jia², Michael Backes¹, Neil Zhenqiang Gong², Yang Zhang¹

¹CISPA Helmholtz Center for Information Security
²Duke University
Era of Machine Learning
Modern machine learning excels at exploiting **grid-structured data**
Many Data are Graphs

Graphs are combinatorial structures, have arbitrary sizes, and contain multi-modal information.
Graph Neural Networks

Image source: https://tkipf.github.io/graph-convolutional-networks/
News source: https://syncedreview.com/2020/09/04/deepmind-uses-gnns-to-boost-google-maps-eta-accuracy-by-up-to-50/, https://syncedreview.com/2020/11/04/cornell-facebook-ai-simplified-graph-learning-approach-outperforms-sota-gnns/, https://blog.twitter.com/engineering/en_us/topics/insights/2020/graph-ml-at-twitter.html
Research question: Given two nodes used to train a black-box GNN, can we predict whether they are linked?
Attack Taxonomy

- Attacker can have either of these 3 knowledge
- Totally 8 different attack models

Node Features
- [12, 32, 6, 0.3]
- [14, 10, 9, 1.2]
- [22, 78, 5, 9.1]
- [15, 32, 9, 4.1]

Partial Graph

Shadow Dataset
- [10, 5, 8]
- [5, 3, 12]
- [8, 5, 13]
- [12, 7, 8]
Attack 0

Node Features:
- [12, 32, 6, 0.3]
- [14, 10, 9, 1.2]
- [22, 78, 5, 9.1]
- [15, 32, 9, 4.1]
- [61, 2, 13, 7.2]

Partial Graph:

Shadow Dataset:
- [10, 5, 8]
- [14, 6, 9]
- [5, 3, 12]
- [8, 5, 13]
- [12, 7, 8]

Posterior Difference

Unsupervised Attack
Correlation performs the best!

**Figure 1:** AUC for Attack-0 on all the 8 datasets with all the 8 distance metrics. The x-axis represents the dataset and the y-axis represents the AUC score.

**Table 15:** Prediction results for Attack-0 on all the 8 datasets with Correlation distance.

| Dataset       | Precision | Recall | F1-Score | AUC  |
|---------------|-----------|--------|----------|------|
| AIDS          | 0.524     | 0.996  | 0.687    | 0.691|
| COX2          | 0.523     | 0.987  | 0.684    | 0.867|
| DHFR          | 0.555     | 0.977  | 0.708    | 0.765|
| ENZYMES       | 0.501     | 1.000  | 0.667    | 0.630|
| PROTEINS_full | 0.540     | 0.998  | 0.701    | 0.815|
| Citeseer      | 0.788     | 0.991  | 0.878    | 0.959|
| Cora          | 0.777     | 0.966  | 0.861    | 0.929|
| Pubmed        | 0.691     | 0.965  | 0.806    | 0.874|

Use KMeans to give a concrete prediction.
Attack 1

Node Features
- [12, 32, 6, 0.3]
- [14, 10, 9, 1.2]
- [22, 78, 5, 9.1]
- [15, 32, 9, 4.1]
- [61, 2, 13, 7.2]

Partial Graph

Shadow Dataset
- [10, 5, 8]
- [14, 6, 9]
- [5, 3, 12]
- [8, 5, 13]
- [12, 7, 8]

Transfer Knowledge
Supervised Attack
Aditya Grover and Jure Leskovec. node2vec: Scalable Feature Learning for Networks. In KDD 2016.
**Attack 1**

- **GNN**
- **Distance (8)**
- **Entropy (4)**
- **MLP**

**Training**

- **GNN**
- **Distance (8)**
- **Entropy (4)**
- **MLP**

**Testing**

- **GNN**
- **Distance (8)**
- **Entropy (4)**
- **MLP**
### Attack 1

For all best performing shadow datasets, attack 1 is **better** than attack 0.

---

**Table 4: Average AUC with standard deviation for Attack-1 on all the 8 datasets. Best results are highlighted in bold.**

| Target Dataset | AIDS     | COX2     | DHFR     | ENZYMES   | PROTEINS_full | Citeseer | Cora     | Pubmed    |
|----------------|----------|----------|----------|-----------|---------------|----------|----------|-----------|
| AIDS           | -        | 0.720 ± 0.009 | 0.690 ± 0.005 | **0.730 ± 0.010** | 0.720 ± 0.005 | 0.689 ± 0.019 | 0.650 ± 0.025 | 0.667 ± 0.014 |
| COX2           | 0.755 ± 0.032 | -        | 0.831 ± 0.005 | 0.739 ± 0.116 | **0.832 ± 0.009** | 0.762 ± 0.009 | 0.773 ± 0.008 | 0.722 ± 0.024 |
| DHFR           | 0.689 ± 0.004 | **0.771 ± 0.004** | -        | 0.577 ± 0.044 | 0.701 ± 0.010 | 0.736 ± 0.005 | 0.740 ± 0.003 | 0.663 ± 0.010 |
| ENZYMES        | **0.747 ± 0.014** | 0.695 ± 0.023 | 0.514 ± 0.041 | -        | 0.691 ± 0.030 | 0.680 ± 0.012 | 0.663 ± 0.009 | 0.637 ± 0.018 |
| PROTEINS_full  | 0.775 ± 0.020 | 0.821 ± 0.016 | 0.528 ± 0.038 | 0.822 ± 0.020 | -        | **0.823 ± 0.004** | 0.809 ± 0.015 | 0.809 ± 0.013 |
| Citeseer       | 0.801 ± 0.040 | 0.920 ± 0.006 | 0.842 ± 0.036 | 0.846 ± 0.042 | 0.848 ± 0.015 | -        | **0.965 ± 0.001** | 0.942 ± 0.003 |
| Cora           | 0.791 ± 0.019 | 0.884 ± 0.005 | 0.811 ± 0.024 | 0.804 ± 0.048 | 0.869 ± 0.012 | **0.942 ± 0.001** | -        | 0.917 ± 0.002 |
| Pubmed         | 0.705 ± 0.039 | 0.796 ± 0.007 | 0.704 ± 0.042 | 0.708 ± 0.067 | 0.752 ± 0.014 | **0.883 ± 0.006** | **0.885 ± 0.005** | -        |
Figure 3: The last hidden layer’s output from the attack model of Attack-1 for 200 randomly sampled positive node pairs and 200 randomly sampled negative node pairs projected into a 2-dimension space using t-SNE. (a) Cora as the shadow dataset and Citeseer as the target dataset, (b) Cora as the shadow dataset and ENZYMES as the target dataset.
Evaluation of All Attacks

• More knowledge leads to better attack performance
• Partial graph contains the strongest signal
• Shadow dataset is the weakest
• Better performance than traditional link prediction, this means GNN indeed leaks graph information
Conclusion

- We are the first to propose link stealing attack against GNNs
- Our attacks can effectively steal the links from GNNs
- More information leads to better attack performance
- Transferring attack can achieve good performance

Questions?

Code is available at https://github.com/xinleihe/link_stealing_attack

Xinlei He
CISPA Helmholtz Center for Information Security
@AllenXinleiHe
http://www.xinlei.info/
Thanks!