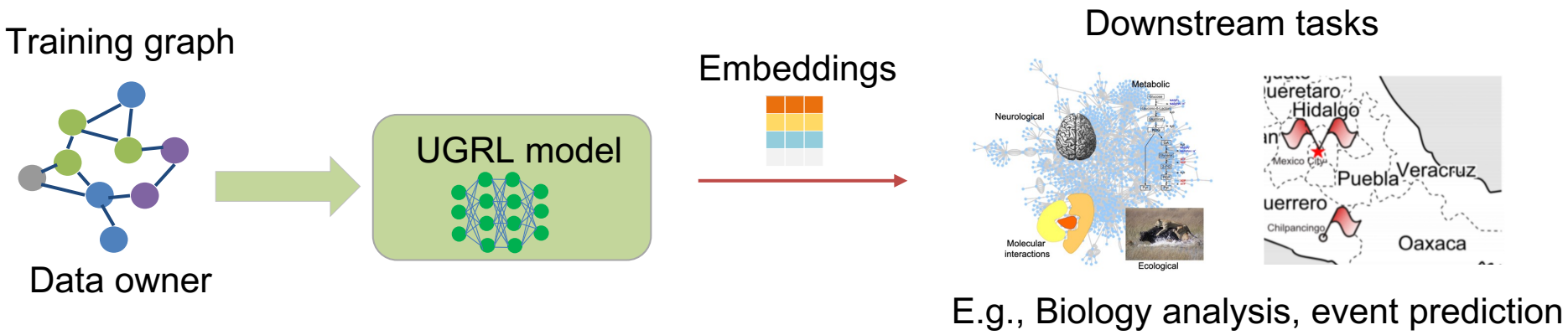
Link Membership Inference Attacks against Unsupervised Graph Representation Learning

Xiuling Wang (presenter), Stevens Institute of Technology, NJ, US
Wendy Hui Wang, Stevens Institute of Technology, NJ, US

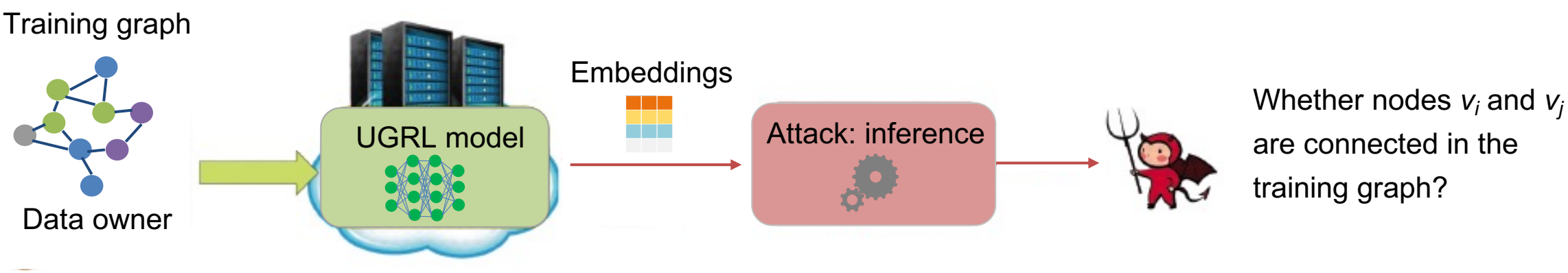
This work is funded by National Science Foundation (NSF)

Unsupervised Graph Representation Learning (UGRL)



- Goal: learn a mapping function Φ: $v \rightarrow Z$, 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 (LMIAs)
    - 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

	Graph Representation Learning		Adversary knowledge
	Supervised	**Unsupervised**	
[1]	GNN		Graph embedding
[2]	GNN		Node embedding + Auxiliary subgraph (mandatory)
[3]	GNN		Posteriors
Ours		DeepWalk, node2vec, LINE, GAE	Node embedding + Shadow graph (optional)

- Most of the works focus on GNN model (supervised model).
- None of the work investigates LMIA against UGRL models.
- None of the existing LMIAs 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.
  - [3] employs the posteriors output by GNN models for inference.

[1] Zhikun Zhang, Min Chen, Michael Backes, Yun Shen, and Yang Zhang. Inference attacks against graph neural networks. In USENIX Security Symposium, pages 4543–4560, 2022.
[2] Vasisht Duddu, Antoine Boutet, and Virat Shejwalkar. Quantifying privacy leakage in graph embedding. In MobiQuitous EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, pages 76–85, 2020.
[3] Xinlei He, Jinyuan Jia, Michael Backes, Neil Zhenqiang Gong, and Yang Zhang. Stealing links from graph neural networks. In Proceedings of the USENIX Security Symposium, pages 2669–2686, 2021.

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_{shadow} (optional)

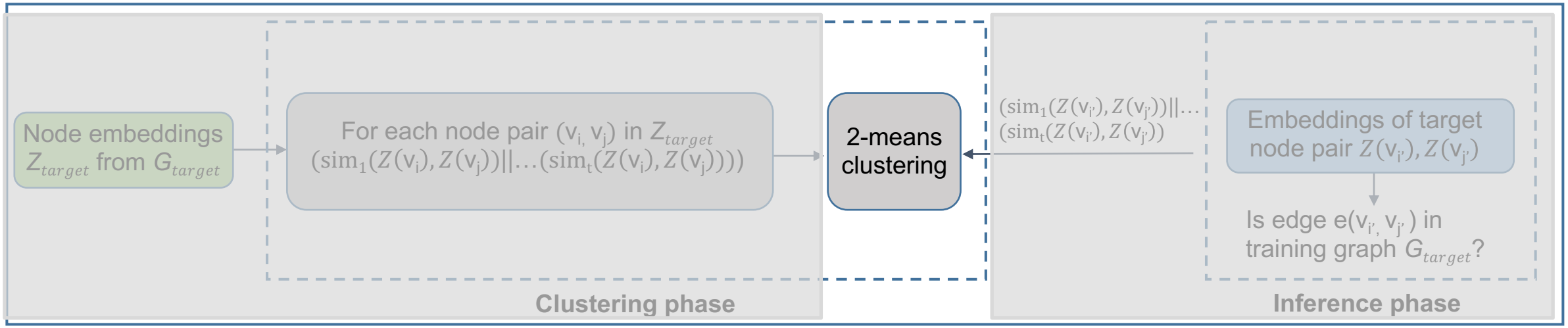
- LMIA is designed as a mapping function:

$$f: \boxed{\vec{z}_i, \vec{z}_j}, \boxed{K_{Aug}} \rightarrow \boxed{\{0, 1\}}$$

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

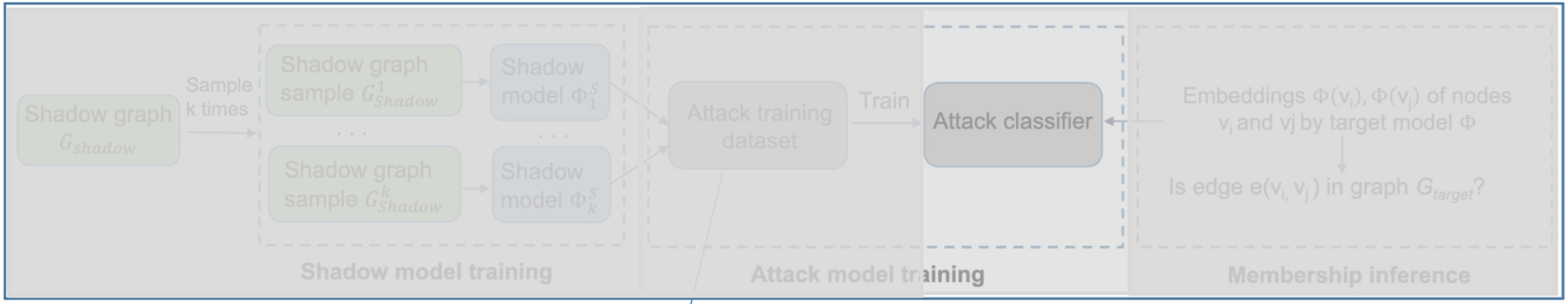
Without shadow graph: k-means clustering algorithm (k=2)



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^l_{Shadow}
$(\mathrm{sim}_1(\Phi^S_l(v_i), \Phi^S_l(v_j)) || \ldots (\mathrm{sim}_t(\Phi^S_l(v_i), \Phi^S_l(v_j))),$ "1"
. . .

Non-member edge $e(v_i, v_j)$ in G^l_{Shadow}
$(\mathrm{sim}_1(\Phi^S_l(v_i), \Phi^S_l(v_j)) || \ldots (\mathrm{sim}_t(\Phi^S_l(v_i), \Phi^S_l(v_j))),$ "0"
. . .

Roadmap

- Introduction
- Our attack
- Evaluation
- Defense
- Conclusion

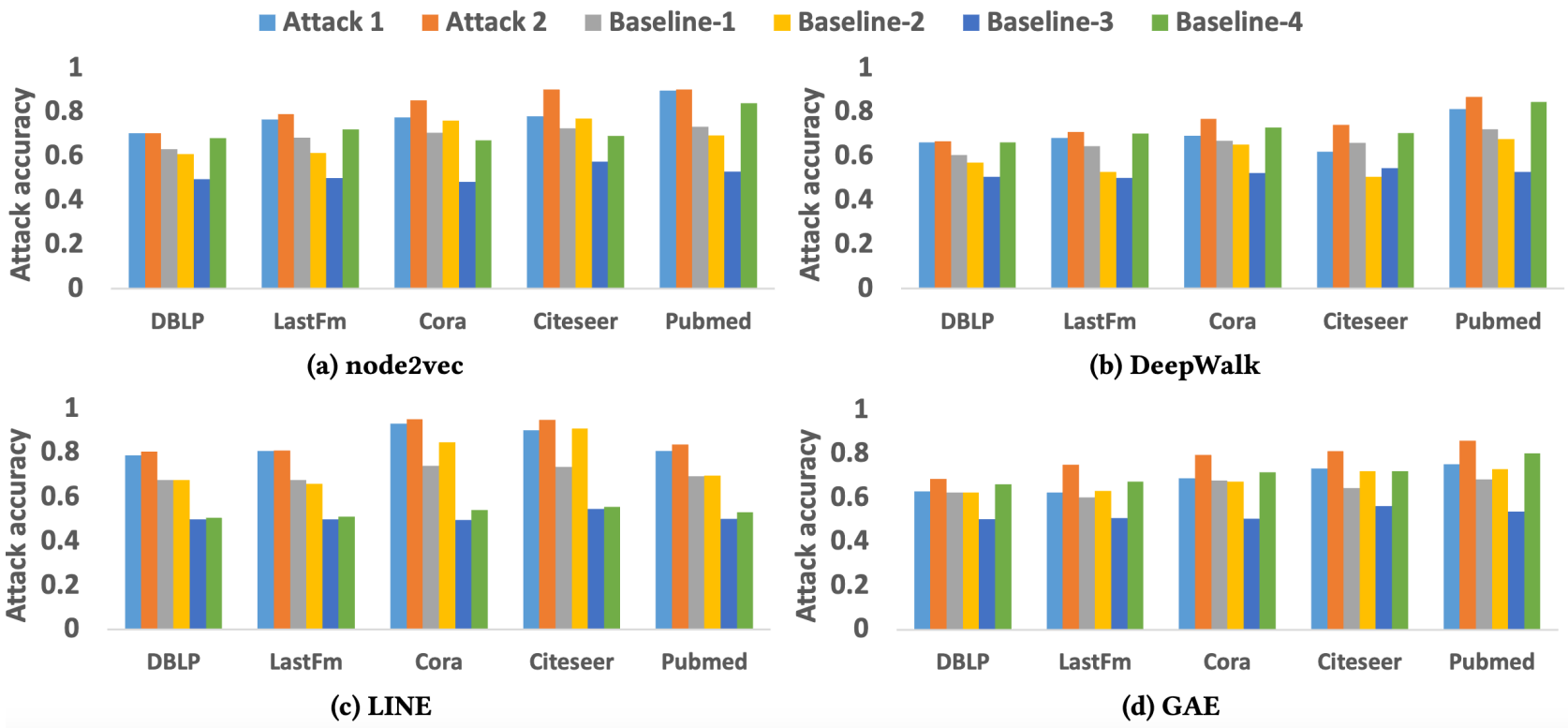
Setup

- 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

[1] Vasisht Duddu, Antoine Boutet, and Virat Shejwalkar. Quantifying privacy leakage in graph embedding. In MobiQuitous EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, pages 76–85, 2020.

Effectiveness of LMIA (1/2)

Setting 1 (Non-transfer setting)

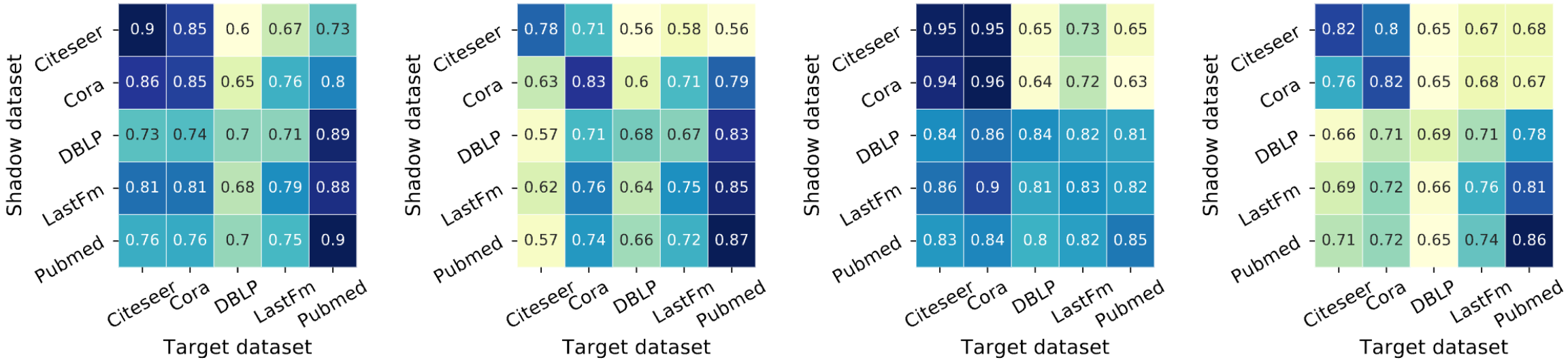


Observations:

- 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)

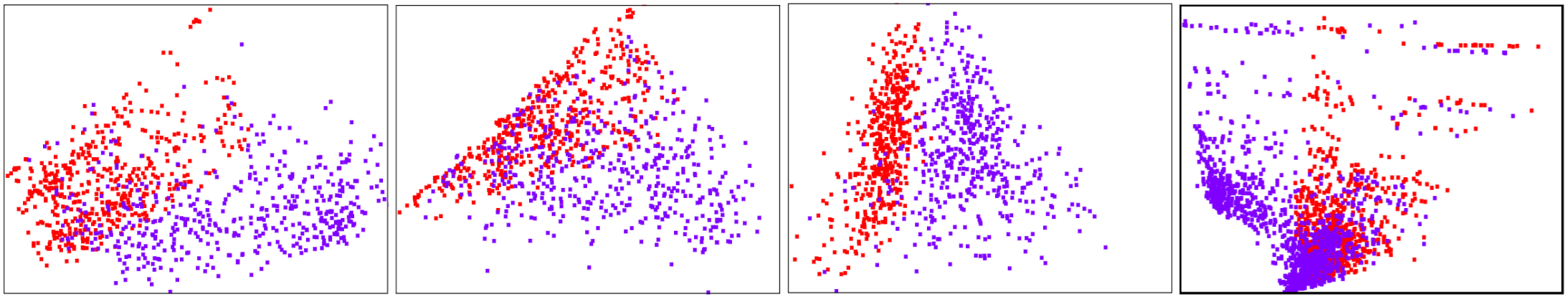


(a) node2vec **(b) DeepWalk** **(c) LINE** **(d) GAE**

Observation:

- The attack remains successful under the transfer setting, even when the target and shadow datasets are from different domains.

Why Our Attacks Work



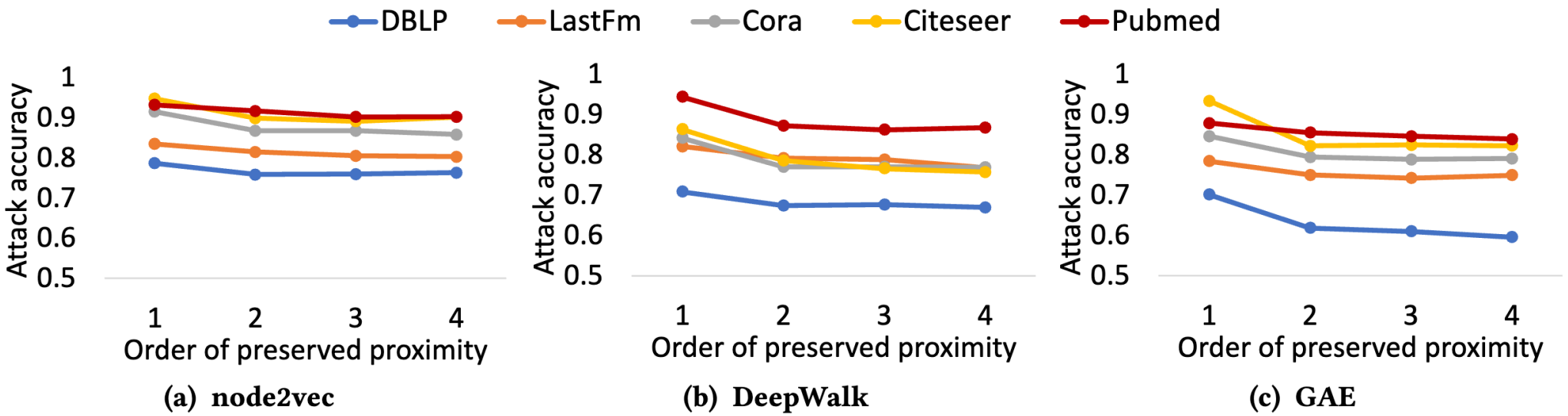
(a) Node2vec (b) DeepWalk (c) LINE (d) GAE

Distribution of LMIA input features for node pairs in both member and non-member node pairs on Cora dataset.
The blue dots are member node pairs, red dots are non-member node pairs.

Observation:

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

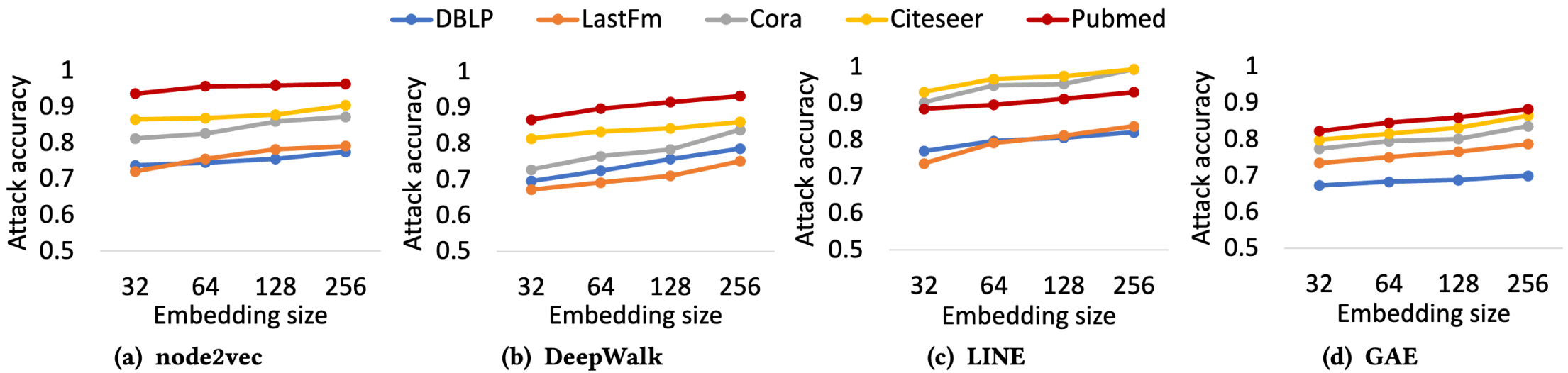
Impact of the Order of the Proximity Preserved in Embeddings on LMIA Performance



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: $\Delta = \frac{1}{2b} e^{-\frac{x-\mu}{b}}$

[1] André Altmann, Laura Tolocsi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. Bioinformatics, 26(10):1340–1347, 2010.
[2] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. Advances in neural information processing systems, 30, 2017.
[3] Leo Breiman. Random forests. Machine learning, 45:5–32, 2001.

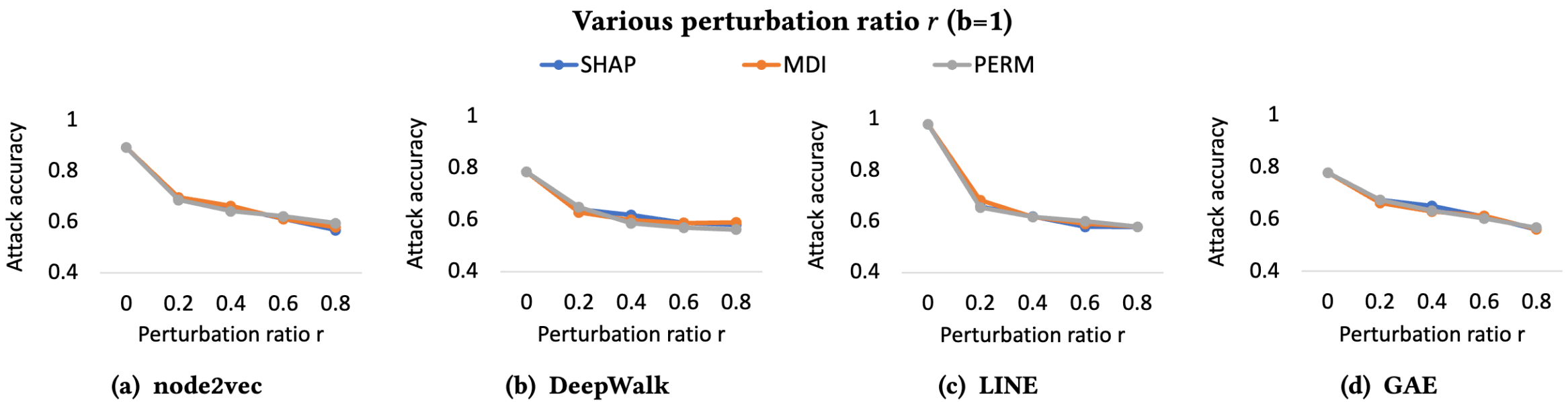
Setup

- 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 λ ={0.5, 1, 5, 10, 20}
- Evaluation metrics:
  - Attack effectiveness: attack accuracy
  - Target model performance: AUC of node classification

[1] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3-4), 2014.

[2] Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine learning with membership privacy using adversarial regularization. In Proceedings of the ACM SIGSAC conference on computer and communications security, pages 634–646, 2018.

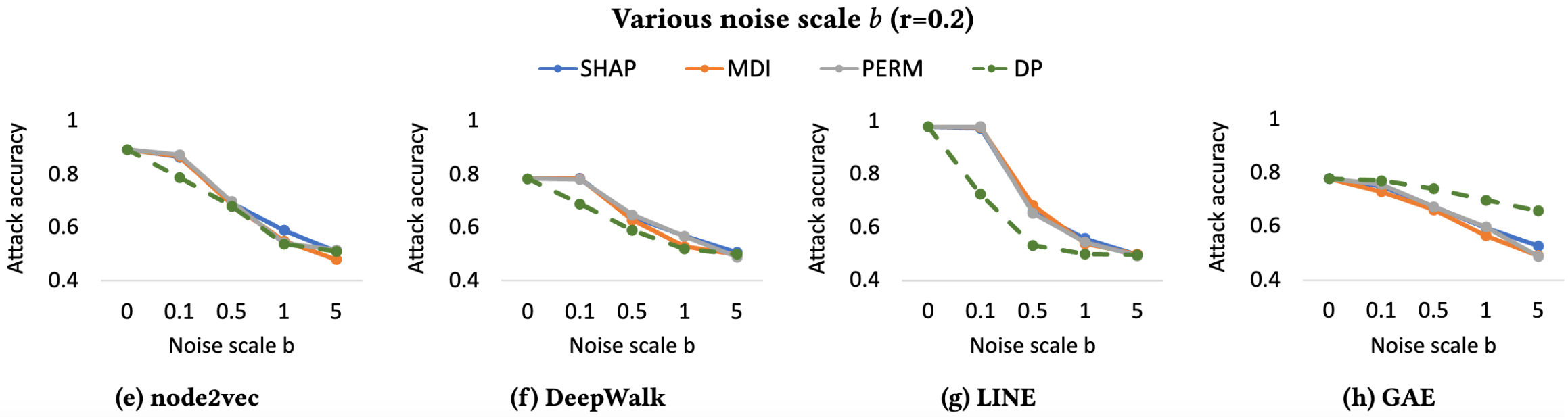
Defense Effectiveness (1/2)



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

Method	Cora				Citeseer				LastFm			
	node2vec	DeepWalk	LINE	GAE	node2vec	DeepWalk	LINE	GAE	node2vec	DeepWalk	LINE	GAE
PERM	0.253	0.326	0.365	0.247	0.303	0.229	0.274	0.243	0.153	0.207	0.204	0.162
MDI	0.258	0.341	0.366	0.252	0.401	0.258	0.282	0.258	0.171	0.207	0.205	0.182
SHAP	0.236	0.318	0.366	0.25	0.327	0.192	0.283	0.249	0.147	0.201	0.199	0.173
DP	0.211	0.271	0.361	0.042	0.257	0.155	0.265	0.04	0.129	0.186	0.177	0.008
AdvR	0.177	0.271	0.312	0.159	0.259	0.157	0.269	0.18	0.146	0.184	0.184	0.162

- 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 LMIAs 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