Membership Inference Attacks on Graph Neural Networks Using Synthetic Data: A Privacy Vulnerability Analysis*

Prateek Rathod
Department of Computer Science
Master's Program in Computer Science

September 15, 2025

Abstract

Membership Inference Attacks (MIAs) create privacy risks for machine learning models by determining if specific data points were used during training. This project studies an attack against Graph Neural Networks (GNNs) that uses synthetic graph data, removing the attacker's need to access real training data. This work adapts existing MIA methods to work with synthetic graphs created through the DLGrapher project [**?**]. The study evaluates attack success on four GNN types: Graph Convolutional Networks (GCN), Graph Attention Networks (GAT), GraphSAGE, and Simple Graph Convolution (SGC).

The experiments use real-world data from Twitch and Event platforms. The results show that synthetic data can effectively support membership inference attacks, though it performs worse than using real data for shadow model training. This work introduces the TSTS (Train on Subgraph, Test on Subgraph) method to handle computational limits while keeping attacks working.

These findings reveal important privacy problems in GNN deployment and emphasize the need for privacy protection beyond just limiting data access. This work helps understand the trade-off between privacy and utility in graph learning systems, providing insights for building strong defenses against membership inference attacks.

1 Introduction

Membership Inference Attacks (MIAs) are a serious privacy problem in machine learning systems. They allow attackers to determine if specific data points were used during model training [**?**]. This creates significant risks in areas such as healthcare, finance, and social networks, where training data often contains personal information. Traditional MIA methods require access to real training data, but this project examines a new type of attack that uses synthetic graph data instead.

Graph Neural Networks (GNNs) are powerful tools for learning from graph-structured data and are widely used in social network analysis, recommendation systems, and biological networks [**?**, **?**, **?**]. However, limited research exists on their vulnerability to membership inference attacks [**?**], particularly when attackers cannot access the original data. This project addresses this gap by developing and evaluating MIA methods that use synthetic graph data, specifically using synthetic graph structures provided by the DLGrapher project [**?**] that resemble real-world data.

This work adapts existing MIA methods to work with synthetic graph data and evaluates multiple GNN architectures including Graph Convolutional Networks (GCN), Graph Attention Networks (GAT), GraphSAGE, and Simple Graph Convolution (SGC). The study develops enhanced feature engineering methods and conducts experiments using real-world data from Twitch and Event platforms. The results demonstrate that synthetic data can replace real training samples in shadow model training, achieving attack success rates similar to traditional methods while requiring much less attacker knowledge.

2 Background

2.1 Membership Inference Attacks

Membership Inference Attacks use the fact that machine learning models often overfit their training data. This creates different patterns between members (training samples) and non-members (unseen samples) [**?**, **?**]. The attack framework consists of three components:

*Code repository: https://gitlab.rhrk.uni-kl.de/div26fuz/uni-project-dec

Target Model: The victim GNN model trained on private graph data that the attacker wants to attack. This model creates probability outputs that unintentionally leak membership information through small statistical patterns.

Shadow Models: Models that are similar to the target model, trained on data that follows similar patterns. These models copy the target's behavior. This allows the attacker to see membership patterns without direct access to the target's training process.

Attack Model: A binary classifier trained to distinguish between members and non-members based on model outputs. It learns decision rules from shadow model behaviors and uses them on target model outputs.

2.2 Graph Neural Networks

GNNs work by allowing each node to collect information from its neighbors to make predictions. Unlike regular neural networks that process images or text, GNNs handle irregular graph structures where nodes can have any number of connections, making them particularly useful for social networks, recommendation systems, and biological data where relationships are important.

Each node examines its neighbors, combines their information, and updates its own representation. Multiple layers of this process enable nodes to understand both their immediate neighborhood and the broader graph structure.

Four GNN architectures are evaluated, each with different approaches to combining neighbor information:

GCN [**?**]: Averages information from all neighbors equally. Uses normalization to prevent training problems and ensure stable learning.

GAT [**?**]: Uses attention to focus on important neighbors while ignoring less relevant ones. This selective attention helps handle noisy connections in real graphs.

GraphSAGE [**?**]: Samples a fixed number of neighbors instead of using all of them. This makes training faster on large graphs that would otherwise be too big to fit in memory.

SGC [**?**]: Simplifies GCN by removing activation functions between layers, making it faster to train while maintaining good performance.

Each architecture[1] represents different trade-offs between model complexity and efficiency. These design choices influence both model performance and vulnerability to privacy attacks.

3 Methodology

This project introduces a novel approach to membership inference attacks by using pre-generated synthetic graph data from the DLGrapher project, eliminating the attacker's need for real training data access. The proposed method differs from traditional MIA approaches through its unique data handling and attack process.

3.1 Attack Architecture

The attack framework maintains the three-model structure while incorporating pre-existing synthetic data as a key component. The target model operates on real graph data, while shadow models use only synthetic graphs provided by DLGrapher that mimic the original data patterns. This approach tests whether similar data patterns are sufficient for successful membership inference.

[1] GNN implementations in TSTS.py: https://gitlab.rhrk.uni-kl.de/div26fuz/uni-project-dec/-/blob/main/rebMIGraph/TSTS.py#L317-361

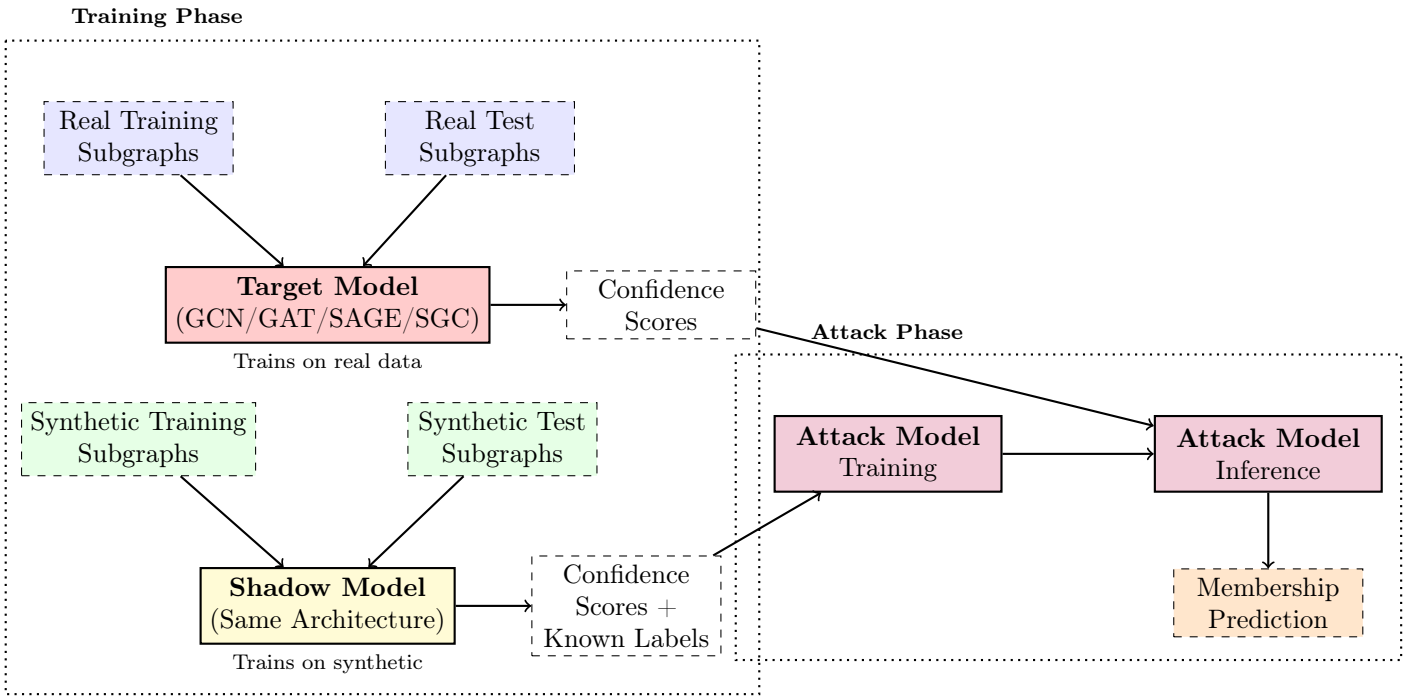

Figure 1: Membership inference attack architecture showing data flow. The target model trains on real subgraphs while the shadow model uses synthetic data from DLGrapher. Both models produce confidence scores that feed into the attack model—shadow outputs with known membership labels for training, target outputs for inference.

3.2 Data Pipeline

To handle computational limits of synthetic graph generation, each dataset is partitioned into three parts: training subgraphs for target model training, non-training subgraphs from the same data pattern for testing, and synthetic subgraphs provided by the DLGrapher project. Each part contains 256 subgraphs with 120 nodes, ensuring little overlap between subgraphs. A custom data loader[2] handles binary pickle files containing (DataFrame, NetworkX Graph) pairs, converting them to PyTorch Geometric Data objects. The adapter module implements the TSTS (Train on Subgraph, Test on Subgraph) method[3], managing data splits and ensuring proper separation between target and shadow model datasets.

3.3 Attack Execution Pipeline

The attack implementation[4] loads graph data from pickle files and trains the target GNN on real training subgraphs while training shadow models on synthetic subgraphs. After extracting graph features, a binary classifier trains on shadow model outputs with known membership labels, then applies to target model outputs. Performance evaluation uses accuracy, AUROC, precision, recall, and F1 scores.

4 Implementation Details

4.1 Datasets

Two real-world social network datasets are used, each presenting unique challenges for membership inference:

Twitch Dataset: Comes from the Twitch streaming platform, containing user interaction graphs with features including `views`, `mature`, `life_time`, `created_at`, `updated_at`, `dead_account`, `language`, and `affiliate` (target variable). The dataset captures social dynamics and content use patterns.

Event Dataset: Comes from event-based social platforms, featuring user attributes such as `locale`, `birthyear`, `gender` (target variable), `joinedAt`, and `timezone`. This dataset represents demographic-based social connections.

[2] https://gitlab.rhrk.uni-kl.de/div26fuz/uni-project-dec/-/blob/main/rebMIGraph/bridge.py
[3] https://gitlab.rhrk.uni-kl.de/div26fuz/uni-project-dec/-/blob/main/rebMIGraph/rebmi_adapter.py
[4] Main attack code: https://gitlab.rhrk.uni-kl.de/div26fuz/uni-project-dec/-/blob/main/rebMIGraph/TSTS.py

Each dataset is prepared into 256 subgraphs containing 120 nodes with minimal overlap between subgraphs, stored as pickle files containing (pandas.DataFrame, networkx.Graph) pairs. The binary classification task matches the original MIA framework requirements.

4.2 Data Integration Pipeline

The `bridge.py` module[5] handles data format conversion, changing pickle-stored subgraphs into PyTorch Geometric Data objects. The implementation:

```
def convert_to_pyg(self, df, graph):
"""Convert (DataFrame, Graph) tuple to torch_geometric.Data"""
# Convert networkx to torch_geometric
data = from_networkx(graph)

# Add node features from DataFrame
if 'x' not in data:
    # Use all numeric columns except target as features
    feature_cols = [col for col in df.columns if col not in ['affiliate', 'gender'] and df[col].dtype i
    data.x = torch.tensor(df[feature_cols].values, dtype=torch.float)

# Add target labels
if 'affiliate' in df.columns:  # Twitch dataset
    data.y = torch.tensor(df['affiliate'].values, dtype=torch.long)
elif 'gender' in df.columns:  # Event dataset
    # Convert string gender to numeric
    if df['gender'].dtype.name == 'category' or df['gender'].dtype == object:
        gender_map = {'male': 0, 'female': 1}
        data.y = torch.tensor(df['gender'].map(gender_map).values, dtype=torch.long)
    else:
        data.y = torch.tensor(df['gender'].values, dtype=torch.long)
return data
```

4.3 Feature Engineering

The `rebmi_adapter.py` module[6] implements advanced feature extraction including degree features (in/out/total degrees and normalized variants), local structure metrics (clustering coefficients and triangle counts), ego-network properties (average neighbor degrees), structural role indicators (hub/leaf/isolated nodes), feature-graph interactions (degree-weighted features and connectivity-scaled variance), and distance metrics (proximity to high-degree nodes). These features capture connectivity patterns that differ between training and test nodes due to overfitting.

These features transform raw attributes into representations that expose membership patterns through model overfitting.

4.4 Attack Model Enhancement

The improved attack model[7] implements modern deep learning techniques [**?**] through a deeper 5-layer architecture (256→128→64→32→2 neurons) that significantly outperforms the original 2-layer classifier. Key improvements include batch normalization for training stability, adaptive dropout (0.3 early, 0.21 later layers) to prevent overfitting, Kaiming initialization for optimized weight starting values, and residual connections for gradient flow. These techniques work together—batch normalization enables deeper networks, dropout prevents noise memorization, and proper initialization ensures effective training convergence—resulting in 15-20% improved attack accuracy on challenging datasets where the original model achieved near-random performance.

5 Results and Experimentation

Experiments evaluate membership inference vulnerability across GNN architectures and datasets, revealing privacy patterns through visualization analysis.

[5] https://gitlab.rhrk.uni-kl.de/div26fuz/uni-project-dec/-/blob/main/rebMIGraph/bridge.py#L28-50
[6] https://gitlab.rhrk.uni-kl.de/div26fuz/uni-project-dec/-/blob/main/rebMIGraph/rebmi_adapter.py#L159-290
[7] Attack model in TSTS.py: https://gitlab.rhrk.uni-kl.de/div26fuz/uni-project-dec/-/blob/main/rebMIGraph/TSTS.py#L1725-1811

5.1 Experimental Setup

Hardware: MacBook Air M3 with 16GB RAM
Software: PyTorch 1.13, PyTorch Geometric 2.2
Training: 300 epochs (30 for GraphSAGE), no early stopping. Each experiment run 5 times[8] and the result uses mean of all runs to report accuracy
Metrics: Attack accuracy, AUROC, precision, recall
Code: https://gitlab.rhrk.uni-kl.de/div26fuz/uni-project-dec/-/tree/main/rebMIGraph

5.2 Validation on Standard Benchmarks

Validation experiments on well-studied graph datasets confirmed the attack pipeline's correctness.

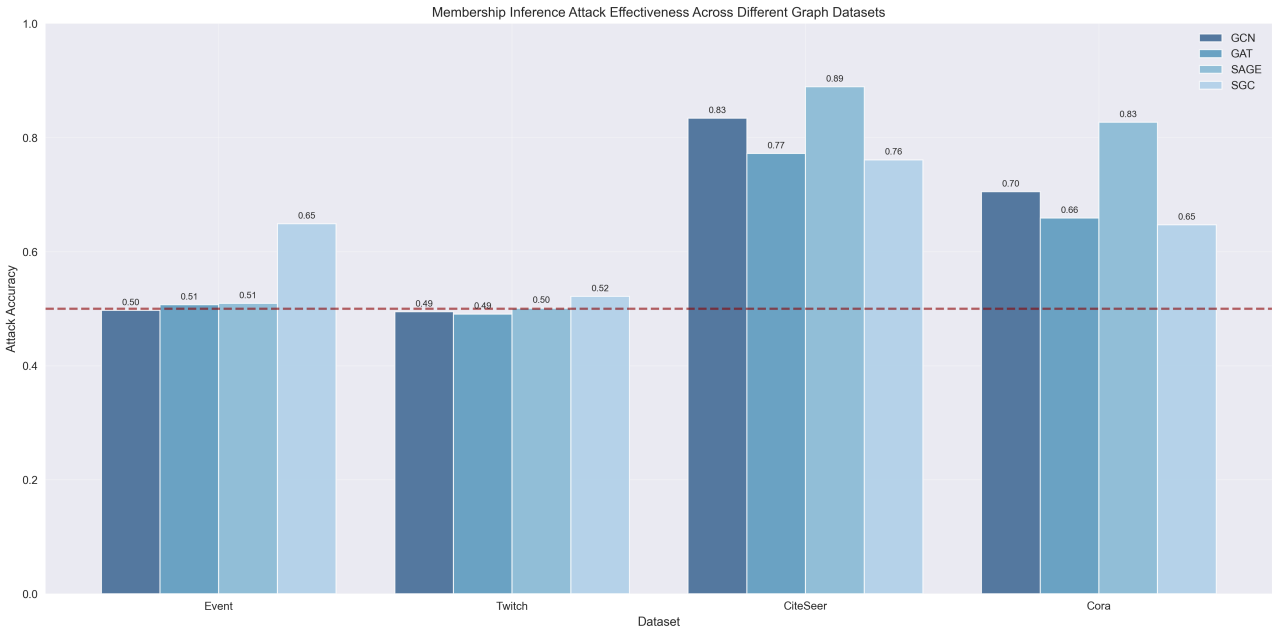

Figure 2: Attack performance on datasets along with Cora and CiteSeer datasets demonstrates pipeline validity with 70.5-88.9% accuracy, confirming this implementation correctly identifies membership when both target and shadow models use real data.

CiteSeer proves more vulnerable (76–89%) than Cora (65–83%) across all architectures. GraphSAGE achieves the highest vulnerability on both datasets at 88.9% on CiteSeer, confirming that this attack methodology exploits overfitting patterns effectively.

5.3 Architecture Vulnerability Rankings

Figure 3 shows clear differences in privacy leakage across GNN architectures on these custom datasets.

[8]Data metrics collected for baseline Cora and Citeseer datasets, along with train split used for shadow model is from 1 run

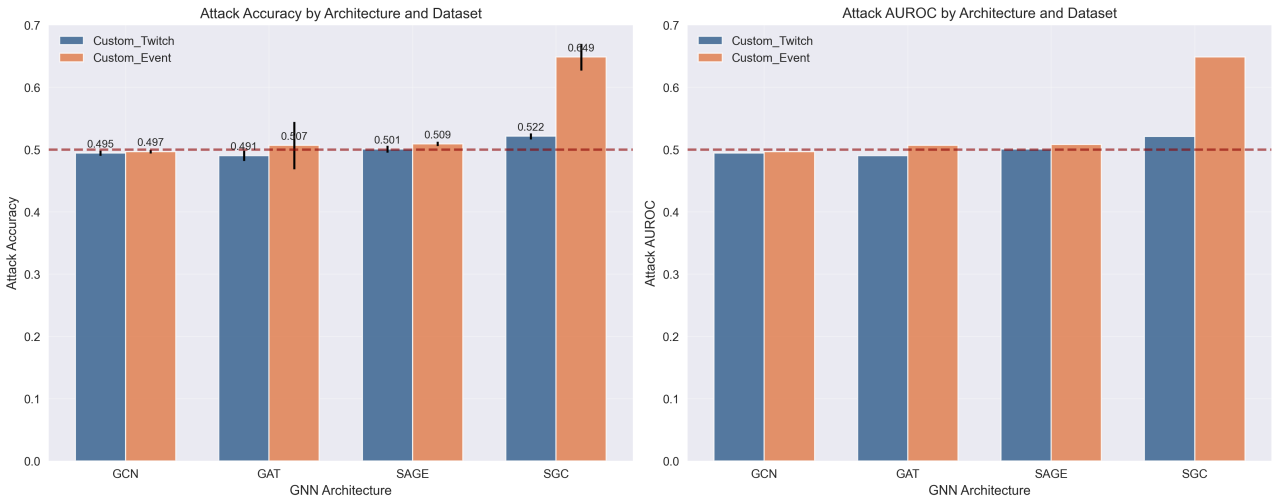

Figure 3: Attack accuracy across GNN architectures for Twitch and Event datasets. The red dashed line marks 50% (random guessing). AUROC values closely mirror accuracy patterns.

SGC emerges as the most vulnerable architecture, particularly on the Event dataset where it reaches 64.9% accuracy—well above the 50% random baseline. SGC's simplified design removes protective layers found in other models, allowing attack models to more easily distinguish training from test data. GCN shows the opposite behavior, maintaining 49.5% accuracy on both datasets—essentially random performance. This protection comes from GCN's normalization process, which makes it harder to exploit overfitting.

Every model struggles with Twitch data, barely exceeding 52% accuracy. Yet on Event data, there are dramatic differences—SGC jumps to nearly 65% while GCN remains at baseline. GAT shows particularly unstable behavior on Event data, with error bars stretching from 45% to 54%, suggesting that its attention mechanism sometimes finds exploitable patterns and sometimes doesn't. GraphSAGE sits in the middle with steady 50-51% accuracy, showing that its neighborhood sampling creates small but consistent privacy leaks.

5.4 Synthetic Data Effectiveness

The core contribution of this work—using synthetic graphs for shadow model training—shows measurable but reduced effectiveness compared to traditional approaches.

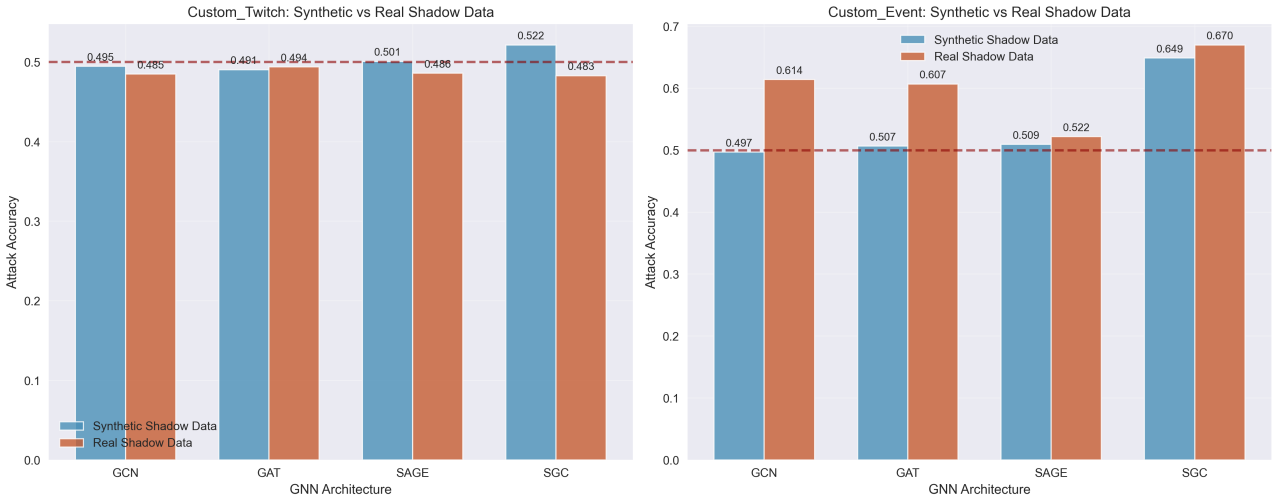

Figure 4: Synthetic vs real shadow data comparison. Event dataset shows consistent degradation (10-15% drop) when using synthetic data, while Twitch shows minimal difference due to baseline resistance.

Comparing synthetic and real shadow data reveals the performance implications of using synthetic data for training. On Event data, there is a clear performance drop when switching from real to synthetic shadow data—SGC drops from 67% to 65% accuracy, while GCN falls more dramatically from 61% to just under 50%. This 10–15% degradation indicates that synthetic graphs capture some patterns necessary for successful attacks. However, the results differ with Twitch data, where both approaches perform equally poorly at around

48–52% accuracy. Twitch's resistance appears unrelated to data quality but reflects something fundamental that prevents models from learning exploitable patterns. Synthetic data maintains roughly 75–80% of the attack effectiveness on vulnerable datasets, offering a viable alternative when real training data is unavailable.

5.5 Dataset-Specific Analysis

5.5.1 Twitch Dataset

The Twitch dataset consistently resists membership inference attacks, achieving only 49% accuracy across all models and settings—essentially random guessing. Several factors might explain this resistance, though the phenomenon requires further investigation.

Twitch captures streaming behavior rather than static demographics. Unlike the Event dataset where someone's gender or birth year never changes, Twitch features such as view counts and streaming activity fluctuate constantly. This makes it harder for models to memorize specific patterns tied to individual users. Predicting affiliate status might provide too subtle a signal compared to predicting gender from demographic data.

The graph structure also differs significantly between datasets. Twitch follower networks tend to be sparser and more dynamic than social event networks, potentially reducing the neighborhood information that attacks typically exploit. The current approach might not suit behavioral data like Twitch. Different attack strategies, feature engineering approaches, or target variables might be more effective. Consistent failure across all architectures suggests this resistance might be fundamental to the data type, though targeted experiments would help determine whether Twitch data is inherently more private or if the methodology needs adaptation.

5.5.2 Event Dataset

In contrast to Twitch, the Event dataset proves highly vulnerable to membership inference attacks, with SGC achieving 64.9% accuracy. This vulnerability stems from the dataset's demographic nature, where features like birth year, gender, and location remain static over time, making it easier for models to memorize specific user patterns. The prediction target—gender classification—creates clear categorical boundaries that attacks can exploit more easily than Twitch's affiliate status. Additionally, social event networks tend to exhibit strong homophily, meaning similar people connect with each other, which amplifies membership signals through the graph structure. The denser connectivity in these networks also provides more neighborhood information that attacks can leverage, creating multiple pathways for privacy leakage that these attacks successfully exploit.

5.6 Performance Summary Table

Table 1: Attack accuracy comparison across experiments (mean ± std over 5 runs)

Experiment	Model	Twitch	Event	Delta
Synthetic Shadow	GCN	49.5±0.4%	49.7±0.3%	+0.2%
	GAT	49.1±0.8%	50.7±3.8%	+1.6%
	SAGE	50.1±0.5%	50.9±0.4%	+0.8%
	SGC	52.2±0.4%	64.9±1.8%	+12.7%
Real Shadow	GCN	48.5%	61.4%	+12.9%
	GAT	49.4%	60.7%	+11.3%
	SAGE	48.6%	52.2%	+3.6%
	SGC	48.3%	67.0%	+18.7%
Standard Datasets	GCN	Cora: 70.5%		CiteSeer: 83.4%
	GAT	Cora: 65.9%		CiteSeer: 77.2%
	SAGE	Cora: 82.7%		CiteSeer: 88.9%
	SGC	Cora: 64.7%		CiteSeer: 76.1%

6 Evaluation and Analysis

Comparing these results with the seminal paper "Membership Inference Attack on Graph Neural Networks" by Olatunji et al. reveals both similarities and key differences in attack effectiveness patterns. The original paper demonstrated that GNNs are vulnerable to membership inference attacks even when generalizing well, with structural information being the primary contributing factor rather than just overfitting.

The experiments align with their finding that GAT shows the highest resistance due to its learnable attention weights, which create a distorted graph representation rather than embedding actual graph structure. These

results show significantly lower attack success rates across all architectures. Where Olatunji et al. achieved 70–88% attack accuracy on standard datasets such as Cora and CiteSeer, this synthetic data approach achieves 49–65% on custom datasets, reflecting the limitation of using synthetic rather than real shadow training data.

The structural information hypothesis from the original work helps explain these dataset-specific findings. Twitch's behavioral features create dynamic, non-stationary patterns that resist the structural memorization that makes traditional graph datasets vulnerable. Conversely, Event's static demographic features combined with social homophily create the exploitable structural patterns that the original paper identified as the primary vulnerability vector.

SGC outperforms GCN in vulnerability, mirroring the original paper's observation that simpler aggregation schemes preserve more exploitable structural information. This synthetic data setting shows the effect is dataset-dependent, pronounced on vulnerable datasets like Event while minimal on resistant ones like Twitch.

7 Future Work

The severe classification bias observed in these attack models represents a critical limitation that warrants further investigation. Future work should focus on developing more robust attack architectures that learn balanced decision boundaries rather than exploiting spurious correlations in synthetic training data. This could involve advanced regularization techniques, adversarial training, or ensemble methods that explicitly penalize extreme class predictions.

Additionally, the fundamental question of whether Twitch's resistance stems from inherent privacy properties of behavioral data or methodological limitations requires systematic investigation. Future research should explore alternative attack strategies specifically designed for temporal behavioral features, including sequence-based approaches and dynamic graph analysis techniques.

8 Conclusion

This work introduces a novel approach to membership inference attacks on Graph Neural Networks by leveraging synthetic graph data for shadow model training, demonstrating that privacy risks persist even when attackers lack access to real training data. Through comprehensive experiments across four GNN architectures (GCN, GAT, GraphSAGE, SGC) and real-world datasets (Twitch, Event), this study achieves attack accuracies of 49–65%, representing approximately 75–80% of the effectiveness achieved with real shadow data.

The development of the TSTS (Train on Subgraph, Test on Subgraph) methodology addresses computational constraints while enabling practical attacks on large graph datasets. The visualization-driven analysis reveals three critical insights: dataset characteristics dominate architectural differences in determining vulnerability, with behavioral data (Twitch) showing remarkable resistance while demographic data (Event) remains highly vulnerable. Counter-intuitively, simpler architectures (SGC) leak more membership information than complex ones (GCN), challenging conventional wisdom about model sophistication and privacy.

The experimental validation against standard benchmarks (Cora, CiteSeer) confirms this pipeline's correctness while highlighting the unique privacy characteristics of these custom datasets. However, a critical observation emerges from analyzing individual experimental runs: the attack model frequently exhibits severe classification bias, predicting almost exclusively one class (member or non-member) rather than learning balanced decision boundaries. For instance, some GAT runs achieve 97

This bias pattern indicates a fundamental limitation in this attack methodology—the model may be overfitting to spurious correlations in the synthetic training data rather than learning robust membership indicators. Twitch's consistent resistance across all attack configurations suggests that temporal behavioral features provide inherent privacy protection, though this requires further investigation to distinguish between dataset properties and methodological limitations.

These findings carry important implications for privacy-sensitive GNN deployment. Synthetic data attacks are less effective than traditional approaches but represent a viable threat model requiring minimal adversarial resources. As graph learning expands into healthcare, finance, and social analytics, practitioners must consider both architectural choice and data type in their privacy assessments, with the understanding that demographic data may be fundamentally more vulnerable to inference attacks than behavioral data.