

비트코인 거래 네트워크를 통한 GNN의 차등 프라이버시 강도에 대한 실험적 분석

우에다 마사토¹ · 콘도 코이치¹ · 황 석 형² · 정 영 애² · 황 세 웅^{2*}

¹선문대학교 소프트웨어학과 석사과정

²선문대학교 소프트웨어학과 교수

Experimental Analysis of Differential Privacy Strength in GNNs Over Bitcoin Transaction Networks

Masato Ueda¹ · Koichi Kondo¹ · Suk-Hyung Hwang² · Young-Ae Jung² · Se-Woong Hwang^{2*}

¹Master's Course, Department of Artificial Intelligence and Software Technology, Sunmoon University, Asan 31460, Korea

²Professor, Department of Artificial Intelligence and Software Technology, Sunmoon University, Asan 31460, Korea

[요 약]

본 연구는 그래프 신경망(GNN)의 프라이버시 보호 기법인 DPAR(Differentially Private Approximate Personalized PageRank)를 비트코인 거래 네트워크에 적용해, 프라이버시 수준(ϵ)에 따른 모델 유용성과 보안성의 상충 관계를 분석하였다. 유용성과 보안성은 분류 정확도와 속성 추론 공격 성공률로 평가되었다. 실험 결과, ϵ 이 낮아질수록 유용성은 저하되었으나 보안성은 전 구간에서 공격 성공률을 약 50%로 유지했다. 특히 $\epsilon=7,000\sim 8,000$ 구간에서 약 60%의 유용성을 확보하면서 공격 성공률을 55% 이하로 억제할 수 있었다. 이는 금융 네트워크에서도 유용성을 크게 훼손하지 않고 효과적인 프라이버시 보호가 가능함을 보여주며, 구체적 파라미터 가이드라인을 제시한다.

[Abstract]

This study applies Differentially Private Approximate Personalized PageRank—a privacy-preserving technique for graph neural networks—to a Bitcoin transaction network and empirically analyzes the tradeoff between model utility and privacy under different privacy budgets (ϵ). Model utility and privacy protection were evaluated using classification accuracy and the success rate of attribute inference attacks, respectively. Experimental results show that as ϵ decreases, model utility declines while security remains robust, thus maintaining attack success rates of approximately 50% across the entire range. Notably, we identified an optimal privacy parameter range of $\epsilon = 7,000\text{--}8,000$, where model utility is preserved at approximately 60% while attack success rates are suppressed below 55%. These findings indicate that privacy can be preserved meaningfully in real-world financial networks without severely compromising utility, thus providing practical guidelines for selecting effective privacy parameters.

색인어 : DPAR, 비트코인, 그래프 신경망, Differential Privacy, 모델 공격

Keyword : DPAR, Bitcoin, GNN, Differential Privacy, Model Attack

<http://dx.doi.org/10.9728/dcs.2025.26.11.3139>



This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received 23 September 2025; **Revised** 16 October 2025

Accepted 22 October 2025

***Corresponding Author, Se-Woong Hwang**

Tel: +82-41-530-2233

E-mail: indimoa@gmail.com

I. 서론

비트코인은 중앙 금융기관의 개입 없이 개인 간(P2P)의 가치 전송을 가능하게 하는 분산형 디지털 화폐로, 거래 효율성과 자산 가치 상승의 잠재력으로 인해 주목받고 있다. 그 개념은 사토시 나카모토의 백서에서 처음 제안되었다[1]. 비트코인 네트워크에서 발생하는 모든 거래는 블록체인이라는 공개 원장에 기록되어 투명하게 검증 가능하지만, 거래 주체의 신원을 직접적으로 드러내지 않는 공개키 기반의 거래 방식은 의사익명성(pseudo-anonymity)을 제공한다. 이러한 특성은 마약 조직의 자금 이동, 자금세탁, 금융 사기 등 각종 불법 행위에 악용될 소지를 낳았다. 그러나 거래 추적 연구들은 이러한 익명성이 완전하지 않음을 입증했으며, 실제로 불법 자금 흐름을 식별한 다수의 사례가 보고되었다[2]. 심지어 한 연구에서는 전체 비트코인 거래의 약 46%가 불법 활동과 연관되어 있다고 추정하기도 했다[3].

이러한 이상 거래를 탐지하기 위해 그래프 형태의 비유클리드 데이터를 효과적으로 처리하는 GNN(Graph Neural Network) 기반의 방법론이 활발히 제안되었지만, 기존 연구들은 모델의 성능과 개인정보 보호 간의 고질적인 상충 관계(trade-off)를 해결하지 못하는 한계를 보였다[4]. 이는 이상(불법) 노드를 탐지하는 과정에서 의료 기록이나 금융 정보와 같은 민감한 데이터를 사용할 경우, 그래프의 노드 속성이나 연결 구조를 통해 개인 식별이 가능해져 심각한 프라이버시 침해로 이어질 수 있기 때문이다. 즉, 이는 탐지 정확도를 높이기 위해 민감 정보 활용 수준을 높일 것인지, 아니면 정확도의 일부 손실을 감수하고 개인정보를 보호할 것인지에 대한 근본적인 딜레마라 할 수 있다.

이러한 상충 관계를 해결하기 위해, 탐지 성능을 최대한 보존하면서 프라이버시 보호 수준을 강화하는 다양한 모델이 제안되었다[5],[6]. 특히 GCN[7]이나 GAT[8]와 같은 GNN 아키텍처를 기반으로 한 연구가 활발히 진행되었으며, 이는 금융 사기 탐지 분야에서 높은 적용 가능성을 입증했다. 이때 개인정보 보호를 위한 대표적인 기술적 접근법인 차등 프라이버시(Differential Privacy)는 데이터에 통계적 노이즈를 주입하여, 특정 개인의 정보 포함 여부가 분석 결과에 미치는 영향을 통계적으로 구분할 수 없게 만드는 엄격한 프라이버시 보장 모델이다. 주입되는 노이즈의 양은 프라이버시 보장 수준을 정량화하는 파라미터인 엡실론(ϵ)에 의해 결정되며, 머신러닝 전반에서 ϵ 값과 모델 정확도 간의 상충 관계는 잘 알려진 사실이다[9]. 차등 프라이버시를 GNN에 접목하려는 연구는 다수 존재했지만, 이를 비트코인 거래 네트워크와 같은 실제 금융 환경에 적용한 사례는 부족한 실정이다. 이에 본 연구는 차등 프라이버시가 적용된 GNN 모델을 실제 금융 데이터에 적용하고 그 실효성을 검증하는 선도적인 사례를 제시하고자 한다.

이에 본 연구는 차등 프라이버시 GNN 모델인 DPAR을 비트코인 거래 네트워크에 적용하는 선구적인 연구를 제안한다.

본 연구의 핵심 목표는 프라이버시 보장 수준을 결정하는 핵심 파라미터 엡실론(ϵ) 값을 조정하며, 비트코인 거래 네트워크 환경에서 모델의 유용성과 프라이버시 보호 간의 최적 균형점을 탐색하는 것이다.

본 연구에서 활용하는 DPAR 모델은 구조 정보 학습과 특성 집계를 분리하여 노드 수준(node-level)의 차등 프라이버시를 효율적으로 구현하는 GNN 기법이다. 비트코인 데이터는 개별 노드가 특정 주체를 나타낼 수 있어 노드 수준의 보호가 필수적이다. DPAR은 Personalized PageRank를 기반으로 각 노드에 중요한 Top-K 이웃만을 선별하여 정보 전파 범위를 제한함으로써, 민감도(sensitivity)를 낮추고 우수한 프라이버시-유틸리티 균형을 달성한다[6]. 본 연구는 새로운 모델을 제안하기보다, 기존 DPAR 모델을 BitcoinTemporalGraph 데이터셋[10]에 적용하여 ϵ 값의 변화에 따라 프라이버시 보호 기제가 효과적으로 작동하는 범위와 단순 노이즈로 전락하는 임계점을 실증적으로 분석하는 데 초점을 맞춘다.

본 논문의 구성은 다음과 같다. 제2장에서는 GNN 기반 사기 탐지 및 관련 프라이버시 모델에 대한 선행 연구를 소개한다. 제3장에서는 데이터셋의 특성과 전처리 과정, 실험 절차 및 평가 방법을 상세히 기술하며, 제4장에서는 실험 결과를 바탕으로 ϵ 값의 변화에 따른 모델 성능과 프라이버시 강도를 심층적으로 논의한다.

II. 본론

2-1 차등 프라이버시(Differential Privacy)

머신러닝의 학습 과정에서 개인정보를 포함한 민감 데이터의 사용이 불가피한 경우가 많다. 차등 프라이버시(Differential Privacy)는 이러한 데이터 분석 과정에서 특정 개인의 정보가 포함되었는지 여부가 결과에 거의 영향을 미치지 않도록 보장함으로써, 데이터 제공자가 겪을 수 있는 잠재적 불이익을 원천적으로 차단하는 엄격한 수학적 프라이버시 모델이다[11].

기술적으로 이는 특정 개인의 데이터 유무에 따른 두 인접 데이터셋(adjacent datasets)에 대해, 동일한 분석 알고리즘을 적용했을 때 출력 결과의 확률 분포가 거의 동일하도록 강제하는 속성으로 정의된다. 이때 확률 분포의 유사성 정도는 프라이버시 예산(privacy budget)이라 불리는 엡실론(ϵ)에 의해 제어된다. ϵ 값이 작을수록 더 강력한 프라이버시 보호를 의미하지만, 분석 결과의 유용성(utility), 즉 정확성은 감소하는 상충 관계가 존재한다.

중요한 점은 차등 프라이버시는 특정 알고리즘이 아닌, 달성해야 할 프라이버시 보장의 수준을 정의하는 하나의 프레임워크라는 것이다. 실제로 이를 구현하기 위해 다양한 메커니즘이 사용되며, 본 연구에서 활용하는 DPAR 모델은 지수 메커니즘(The Exponential Mechanism)과 가우스 메커니즘(Gaussian Mechanism)을 채택하고 있다[12]-[14].

2-2 DPAR(Decoupled Graph Neural Networks with Node-Level Differential Privacy)

표준적인 GNN 모델은 다층 메시지 전파(Multi-layer Message Passing) 구조를 통해 각 노드가 이웃의 정보를 반복적으로 집계하여 자신의 특성 벡터를 갱신한다. 이러한 구조는 단일 노드의 변화가 다수의 이웃 노드에 연쇄적으로 영향을 미쳐, 차등 프라이버시에서 정의하는 민감도(sensitivity)를 급격히 높이는 원인이 된다. 민감도가 높을수록 프라이버시 보장을 위해 더 큰 노이즈를 추가해야 하므로, 모델의 학습 성능과 정확도가 크게 저하되는 문제가 발생한다.

DPAR은 이러한 문제를 해결하기 위해 구조 정보 학습과 특성 집계를 분리(decouple)하는 접근법을 사용한다. 핵심은 전통적인 메시지 전파 방식 대신 근사 개인화 페이지랭크(Approximate Personalized PageRank, APPR)를 활용하여 GNN 학습 이전에 각 노드에 가장 중요한 Top-K 이웃을 미리 식별하는 것이다. APPR은 구글의 페이지랭크 알고리즘 [15]에서 파생된 기법으로, 전체 그래프를 탐색하지 않고 특정 시작 노드(seed node)와 밀접한 국소적 클러스터(local cluster)를 효율적으로 발견하는 데 특화되어 있다[16]. 이렇게 영향력 있는 이웃의 범위를 Top-K로 한정함으로써 단일 노드 변화의 파급효과를 제어하여 민감도를 낮추고, 결과적으로 적은 노이즈만으로도 강력한 노드 수준의 차등 프라이버시 보장을 가능하게 한다.

이러한 DPAR의 접근법은 Cora-ML, Reddit 등 다양한 벤치마크 데이터셋에서 GraphSAGE와 같은 기존 모델보다 우수한 성능을 입증했다. 예를 들어, Reddit 데이터셋에서 DPAR은 0.934의 테스트 정확도를 기록하며 기존 연구(GAP)의 0.7047을 크게 상회하는 성과를 보였다[6]. DPAR은 이웃 선택 과정과 모델 학습 과정에 각각 차등 프라이버시를 적용하며, 이를 위해 선택 기반의 지수 메커니즘(Exponential Mechanism)과 연속적인 노이즈를 추가하는 가우스 메커니즘(Gaussian Mechanism)을 활용한다[14].

Exponential Mechanism (DP-APPR-EM)은 모델의 유용성을 최대한 확보하기 위해 품질 점수가 높은 이웃을 높은 확률로 선택하는 동시에, 품질 점수가 낮은 이웃도 0이 아닌 확률로 선택될 수 있도록 설계된 기법이다. 이처럼 모든 후보에게 선택의 기회를 부여하는 방식은 결과에 대한 그럴듯한 부인 가능성(plausible deniability)을 제공하여, 전체 선택 과정을 모호하게 만들어 프라이버시를 보호한다.

$$\Pr [EM_q(X) = y] \propto \exp \left(\frac{\epsilon}{\Delta(q)} q(X, y) \right) \quad (1)$$

- $\Delta(q)$: 민감도(인접한 데이터 간의 점수 함수의 최대 변화량)
- $q(X, y)$: 품질 점수(데이터 세트 X에서 후보 r의 '우수성'을 평가하기 위한 함수)
- ϵ : 프라이버시 감도

Gaussian Mechanism (DP-APPR-GM)은 각 노드의 점수에 연속적인 가우스 노이즈를 추가하여 프라이버시를 보호한다. 노드 집합의 선택과 관련된 수치에 직접 노이즈를 추가하기 때문에 DP-APPR-EM에 비해 구현이 용이하다. 따라서 대규모 그래프에 효율적이지만, 모델의 정확도가 노이즈의 양에 따라 달라지거나 노이즈의 양과 점수 분포에 따라 중요한 노드가 누락될 수 있다. 데이터 셋D에 대해 실행되는 차등 프라이버시를 만족하는 무작위화 알고리즘의 출력은 다음과 같다.

$$M(D) = f(D) + N(0, \sigma^2) \quad (2)$$

- $f(D)$: 데이터베이스 D에 대한 쿼리 결과
- $N(0, \sigma^2)$: 평균 0, 분산 σ^2 의 다변량 정규분포를 따르는 노이즈

노이즈의 크기 σ 는 다음의 조건을 충족한다.

$$\sigma \geq \frac{\sqrt{2 \ln(1.25/\delta)} \cdot \Delta f}{\epsilon} \quad (3)$$

- Δf : 함수 f 의 감도(한 사람의 데이터 변경에 따른 최대 출력 변화)
- ϵ : 프라이버시 감도
- δ : 프라이버시 침해 허용 확률

2-3 금융 사기 탐지를 위한 그래프 신경망

금융 사기 탐지 분야에서 GNN의 효용성은 Dawei Cheng 등의 연구에서 체계적으로 증명되었다. 이들은 현대 금융 거래가 IT 기술 발전으로 고도화되고 거래의 상호 연결성이 폭발적으로 증가함에 따라, 개별 거래 데이터에만 의존하는 전통적인 사기 탐지 방식이 한계에 직면했다고 지적했다. 규칙 기반 시스템이나 표준적인 머신러닝 모델은 고립된 특징(feature)을 분석하는 데는 유용하지만, 여러 계정을 거치는 자금 세탁이나 공모 관계와 같이 정교하게 조직된 사기 네트워크의 관계적 패턴을 포착하기 어렵기 때문이다.

이러한 한계를 극복하기 위한 대안으로 GNN이 주목받고 있다. GNN은 금융 거래 네트워크를 사용하는 노드(node), 거래는 엣지(edge)로 표현되는 그래프로 모델링한다. 이를 통해 각 노드(사용자)의 특징뿐만 아니라, 이웃 노드와의 상호작용 및 네트워크 내에서의 구조적 역할을 종합적으로 학습하여 풍부한 표현 벡터를 생성한다. 이렇게 학습된 표현 벡터는 개별 거래에서는 드러나지 않는 이상 징후나 비정상적인 자금 흐름의 패턴을 효과적으로 식별할 수 있게 한다.

Dawei Cheng 등은 이러한 GNN 기반 사기 탐지 연구를 GNN의 기법, 역할, 설계, 실제 활용, 당면 과제라는 다섯 가지 관점으로 종합하여 분석하는 프레임워크를 제시했다. 해당 연구에 따르면 GNN은 주로 GCN(Graph Convolutional Network)이나 GAT(Graph Attention Network)와 같은 아

키텍처로 분류된다. GCN은 이웃 노드의 정보를 균등하게 집계하여 패턴을 학습하는 반면, GAT는 어텐션 메커니즘을 통해 사기 행위와 더 관련이 깊은 이웃 노드의 정보에 가중치를 부여하여 학습함으로써 더 정교한 탐지가 가능하다. 결론적으로 GNN은 복잡하고 동적인 금융 네트워크 환경에서 높은 유연성과 적응력을 기반으로, 데이터에 내재된 관계를 스스로 학습하여 기존의 규칙 기반 방식보다 우수한 탐지 성능을 보임이 입증되었다[17].

2-4 모델 공격 및 프라이버시 평가

차등 프라이버시와 같은 보호 기법이 적용된 모델의 실질적인 프라이버시 강도를 평가하기 위해, 가상 공격 시나리오를 통한 정량적 측정이 요구된다. 본 논문의 실험에서는 대표적인 프라이버시 공격 기법인 멤버십 추론 공격(MIA)과 속성 추론 공격(Attribute Inference Attack)을 평가 척도로 활용한다.

1) 멤버십 추론 공격(MIA)

멤버십 추론 공격은 특정 데이터 레코드가 모델의 학습 데이터셋에 포함되었는지 여부를 추론하는 것을 목표로 하는 대표적인 프라이버시 공격 기법이다. 이 공격은 주로 모델의 내부 파라미터나 학습 데이터에 직접 접근할 수 없는 블랙박스(black-box) 환경을 가정하여, 모델의 예측 출력값만을 기반으로 수행된다. 따라서 실제 상용 서비스가 노출될 수 있는 프라이버시 위험을 측정하는 데 효과적이다.

공격은 일반적으로 ‘새도우 훈련(shadow training)’ 방법론을 따른다. 공격 과정은 다음과 같다[18].

- (1) 먼저 공격자는 타겟 모델(target model)의 동작을 모방하는 여러 개의 새도우 모델(Shadow model)을 생성한다. 이 새도우 모델들은 타겟 모델의 학습 데이터와 유사한 분포를 갖는 데이터로 학습된다.
- (2) 다음으로, 공격자는 새도우 모델 학습에 어떤 데이터가 사용되었는지 이미 알고 있다. 이 정답 정보를 활용하여, 새도우 모델이 학습에 사용한 데이터(member)와 사용하지 않은 데이터(non-member)에 대해 보이는 출력(예측 확률 벡터)의 차이를 학습하는 공격 모델(attack model)을 훈련시킨다. 이 공격 모델은 입력된 예측 벡터가 학습 데이터로부터 나온 것인지 아닌지를 판별하는 이진 분류기(binary classifier)로 작동한다.
- (3) 실제 공격 시에는 타겟 모델에 특정 데이터를 입력해 예측 벡터를 얻은 뒤, 이 벡터를 공격 모델에 입력하여 해당 데이터가 학습 데이터셋에 포함되었는지 추론한다.

공격 모델의 정확도가 무작위 추측 수준인 50%에 근접할수록 목표 모델이 멤버십 정보를 효과적으로 보호하고 있음을 의미한다.

2) 속성 추론 공격(Attribute Inference Attack)

속성 추론 공격은 학습 데이터의 멤버십 여부를 넘어, 데이터 레코드에 포함된 특정 민감 속성(sensitive attribute)을 직접 추론하는 것을 목표로 한다[19]. 이는 개인의 프라이버시에 더 직접적인 위협이 될 수 있다.

공격 시나리오는 공격자의 가정과 능력에 따라 다양하다. 한 가지 접근법은 공격자가 일부 데이터에 대해 공개 속성(예: 그래프 임베딩)과 그에 해당하는 민감 속성 쌍을 확보하고 있다고 가정하는 것이다. 공격자는 이 데이터를 사용하여, 공개 속성으로부터 민감 속성을 예측하는 공격 모델을 훈련시킨다. 이 방식은 특히 노드 간의 특징 및 관계가 풍부한 GNN에서 임베딩 벡터와 사용자 속성 간에 강한 상관관계가 존재할 경우 매우 유효한 공격이 될 수 있다.

다른 블랙박스 시나리오에서는 모델의 신뢰도 점수(confidence score)를 활용한다. 공격자는 대상 사용자의 알려지지 않은 특정 속성에 대해 가능한 모든 후보값을 모델에 순차적으로 질의한다. 이후 가장 높은 신뢰도 점수를 반환하는 후보값을 실제 사용자의 민감 정보로 판단하는 방식으로 공격을 수행한다.

표 1. 관련 문헌 연구

Table 1. Related works

Type	Description	Reference
Differential Privacy	A guarantee that an analysis is not significantly influenced by any single individual's data.	[11]
DPAR	A GNN is trained by identifying the top-k most influential neighbors of each node via Differentially Private PageRank.	[6]
Anomaly (Illicit) Node Detection	A Systematic Analysis of GNN Techniques for Evaluating their Usefulness in Financial Fraud Detection.	[17]
Model Attacks	Model attacks, such as Membership Inference (MIA) and Attribute Inference, pose significant privacy threats. MIA demonstrates the vulnerability of black-box services using shadow training, while Attribute Inference allows adversaries to deduce sensitive attributes of data subjects through model queries. Successful attacks risk the exposure of valuable assets, including sensitive personal data and intellectual property.	[18]
Research Dataset	A processed and extended version of the Elliptic Bitcoin transaction dataset from 2008 to 2020, adapted for the purposes of this study.	[10]

2-5 연구 데이터셋(Bitcoin Temporal Graph)

이 데이터셋은 Elliptic 사가 2019년에 발표한 비트코인 데이터셋 Elliptic2를 확장한 것이다. 이 데이터셋은 연구 목적에 적합한 공개 데이터셋이 매우 부족하다는 점, 그리고 이러한 점이 비트코인의 구조에 대한 이해와 리스크 분석에 장

애가 된다는 문제의식에서 출발하여 구축된 것이다. 해당 논문에서는 이 데이터의 활용 사례로 이상 탐지를 소개하고 있다. 자세한 내용은 제3장에서 서술한다.

III. 실험 구조 및 설계

3-1 모델 비교

본 연구는 DPAR의 효용성을 검증하기 위해, 프라이버시 보호 기술이 적용된 모델과 적용되지 않은 베이스라인 모델을 유용성(utility)과 프라이버시 견고성(privacy robustness)이라는 두 가지 핵심 관점에서 비교 평가한다.

모델 유용성 평가는 DPAR의 구조적 특성을 고려하여 설계되었다. DPAR은 차등 프라이버시를 보장하기 위해 데이터에 통계적 노이즈를 주입하므로, 그 성능이 베이스라인 모델보다 필연적으로 낮게 나타난다. 따라서 두 모델의 절대적인 성능을 직접 비교하기보다, DPAR 모델의 프라이버시 핵심 파라미터인 엡실론(ϵ) 값을 체계적으로 변화시키며 유용성과 프라이버시 간의 상충 관계(trade-off)를 관찰하는 데 초점을 맞춘다.

반면, 프라이버시 견고성 평가는 멤버십 추론 공격(MIA)을 통해 두 모델을 직접 비교하는 방식으로 수행된다. 베이스라인 모델 대비 DPAR이 적용된 모델에 대한 공격 성공률(attack success rate)이 유의미하게 감소하는 경우, 이는 DPAR의 프라이버시 보호 메커니즘이 효과적으로 작동함을 실증하는 결과로 해석한다.

3-2 데이터셋

본 연구에서 활용한 데이터셋은 Eliptic사가 수집한 비트코인 거래 데이터를 연구 목적에 맞게 가공, 확장한 그래프 구조의 데이터로, 총 10종의 엔티티 레이블을 포함하고 있다. 해당 데이터셋은 Figshare에 “BitcoinTemporalGraph” Ver3로 공개되었으며, 2025년 2월 6일에 배포되었다. 수집 기간은 제네시스 블록(2009년 1월 3일)부터 2022년 4월 17일까지이며, 총 노드 수는 약 2억 5200만개, 총 엣지 수는 약 7억 8500만개에 달한다. 이 중 34,098개 노드에는 엔티티 타입 레이블이 부여되어 있다. 라벨은 표 2와 같다.

대규모 그래프의 효율적인 처리 및 학습을 위해, 라벨이 부여된 노드를 대상으로 계층적 이웃 샘플링을 적용하였다. 구체적으로 각 시작 노드(seed node)에 대해 1-hop 이웃은 최대 10개, 2-hop 이웃은 최대 5개까지 무작위로 추출하였다. 이를 통해 원본 그래프를 처리 가능한 크기의 부그래프(subgraph)로 축소하였으며, 이는 계산 복잡도와 메모리 부하를 크게 낮춰 대규모 네트워크에서도 효율적인 학습을 가능하게 한다.

모델 학습에 앞서, 노드의 수치형 특징량에 대해 일련의 전처리 과정을 수행하였다. 먼저 기존 원본 특징에 더하여 총

표 2. 컬럼 설명

Table 2. Column of the data set

Feature	Description
Individual	A regular or individual user.
Mining	An individual or entity that performs the creation and validation of blocks.
Exchange	An online platform that facilitates the trading of cryptocurrencies and fiat currencies.
Marketplace	A platform for buying and selling goods or services using Bitcoin.
Gambling	A platform offering gambling services such as casinos, sports betting, and lotteries.
Bet	An address associated with a specific bet, typically issued by a gambling service.
Faucet	A service that distributes small amounts of Bitcoin in exchange for completing simple tasks.
Mixer	A service that enhances anonymity by obfuscating the trail of transactions, making them difficult to trace.
Ponzi	A fraudulent investment structure promising high returns.
Ransomware	An operator of malware that encrypts a victim's data and demands a ransom.
Bridge	A protocol that facilitates the transfer of value between Bitcoin and other blockchains.
Default	Unlabeled or not applicable.

표 3. 파생 변수 설명

Table 3. List of derived feature columns in the dataset

Feature	Description
avg_sent	Average amount per outgoing transaction.
avg_received	Average amount per incoming transaction.
Age	Time difference between the first and the last transaction.
degree_age_ratio	Ratio of the node's age (activity period) to its degree.
degree_in_age_ratio/degree_out_age_ratio	Ratio of the node's age to its in-degree and out-degree, respectively.
transactions_in/out_age_ratio	Ratio of the node's age to the number of incoming/outgoing transactions.
cluster_size_age_ratio/cluster_num_edges_age_ratio	Ratio of the node's age to its cluster's size and number of internal edges.
proportion_nodes_in_cc	Proportion of nodes in the main connected component.
time_before_first_transaction	Time difference between the first incoming and the first outgoing transaction.
degree_out_in_ratio	The ratio of out-degree to in-degree.
min/total/avg/max_sents_usd	Minimum, total, average, and maximum sent amounts in USD.
min/total/avg/max_received_usd	Minimum, total, average, and maximum received amounts in USD.

21종의 파생 변수를 계산하고, 이를 결합하여 특징 벡터를 확장하였다. 이후 각 특징에 대해 다음과 같은 전처리 절차를

순차적으로 적용하였다.

- (1) 로그 변환: 분포의 왜도(skewness)를 완화하기 위해 일부 수치형 특징에 로그 변환을 적용하였다.
- (2) 분위수 기반 정규화: 특징 그룹별로 분위수(quantile)를 기준으로 최소-최대 정규화를 차등 적용하였다. 차수 및 클러스터 관련 특징은 하위 0%-상위 95%를, 송수신 금액 관련 특징은 하위 5%-상위 95%를 기준으로 하였다.
- (3) 클리핑: 정규화된 값이 [0, 1] 범위를 벗어나는 경우, 해당 범위로 값을 제한하는 클리핑(clipping)을 수행하였다.
- (4) 결측치 처리: 대부분의 결측치는 0으로 대체하였다. 단, proportion_nodes_in_cc와 같이 비율 계산 과정에서 발생하는 특정 결측치(NaN)는 보수적 기준에 따라 1로 처리하였다.

본 연구는 라벨이 부여된 34,098개 노드를 대상으로, 층화 샘플링(stratified sampling)을 통해 학습, 검증, 테스트 세트를 약 8:1:1 비율로 분할하였다. 분할은 2단계 층화 추출 절차를 따랐다. 먼저 전체 데이터에서 클래스 비율을 유지하며 10%를 테스트 세트로 분리했다. 이후, 잔여 90% 데이터에 대해 다시 한번 층화 추출을 수행하여 검증 세트(잔여 데이터의 11.11%)를 분리함으로써 최종적으로 목표 비율을 구성했다. 분할 과정의 재현성을 확보하기 위해 난수 시드는 42로 고정하였다. 이 절차에 따라 표본 수는 학습 셋 27,278개, 검증 셋 3,410개, 테스트 셋 3,410개가 사용된다.

원본 데이터셋은 클래스 간 분포가 심각하게 불균형한 문제가 있어, 이로 인한 학습 불안정성을 완화하고자 분할된 각 세트(split) 내부에서 독립적으로 클래스별 샘플 수를 조정하는 리샘플링(resampling)을 수행했다. 구체적으로, 클래스별 샘플 수에 하한 700개와 상한 4,000개의 제약을 적용했다. 샘플 수가 700개 미만인 클래스는 복원 추출을 통해 700개로 증가하였고, 4,000개를 초과하는 클래스는 비복원 추출을 통해 4,000개로 축소하였다.

이러한 리샘플링 과정은 각 세트 내에서만 독립적으로 수행되므로 세트 간 데이터 누수(data leakage)는 발생하지 않는다. 다만, 복원 추출로 인해 동일 노드가 중복 포함될 수 있으므로, 그림 1에 제시된 샘플 수는 고유 노드 수가 아닌 유효 학습 표본 수를 의미한다.

추가적으로, 차등 프라이버시 실험에서는 ‘샘플링을 통한 증폭(amplification by sampling)’ 효과를 얻기 위해 학습 세트에 대해 추가적인 서브샘플링을 선택적으로 적용했다. 따라서 그림 1의 표본 수는 초기 8:1:1 분할 및 리샘플링뿐만 아니라, 학습 세트에 한해 이러한 추가 샘플링 효과까지 모두 반영된, 실제 실험에 사용된 최종 유효 표본 분포이다. 이러한 다단계 샘플링 설계는 극단적인 클래스 불균형으로 인한 학습 불안정성을 완화하는 동시에, 평가 세트의 독립성과 실험

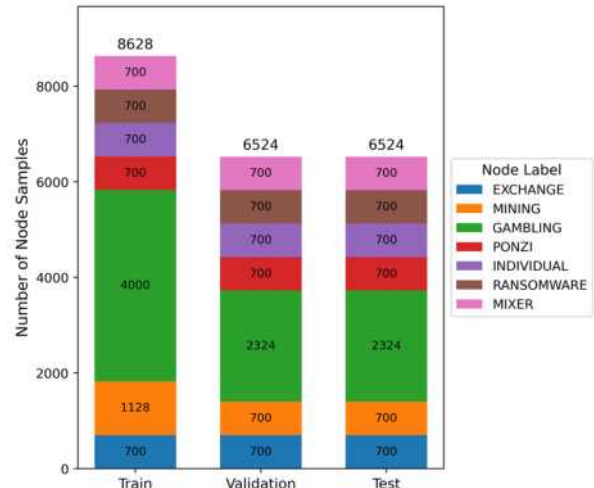


그림 1. 각 데이터 서브셋(Subset)의 클래스별 노드 샘플 분포
Fig. 1. Node distribution by class and subset

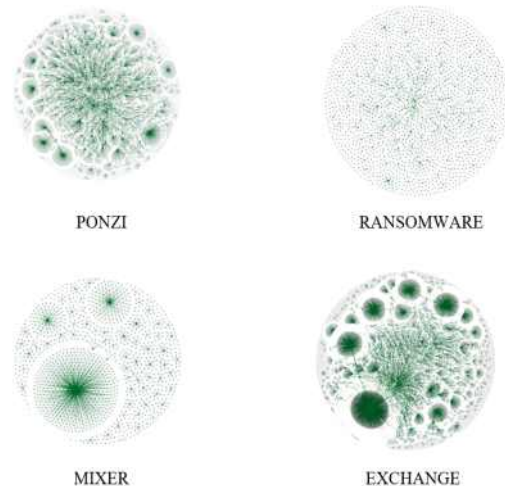


그림 2. 이상 노드 시각화
Fig. 2. Visualization of derived nodes

의 공정성을 보장하기 위함이다.

또한, 본 실험에서는 전체 12개 클래스 중 대표성을 갖는 7개 클래스(EXCHANGE, MINING, GAMBLING, PONZI, INDIVIDUAL, RANSOMWARE, MIXER)만을 분석 대상으로 선정했다. 이는 실제 금융 환경에서는 모든 유형의 불법 행위가 사전에 완벽히 정의되고 라벨링되어 있지 않다는 현실을 반영한 설정이다. 데이터셋의 원본 연구 역시 ‘일부 대표적인 이상 클래스로 모델을 학습하고, 학습되지 않은 미지의 클래스를 탐지하는 접근이 더 현실적’이라고 제안한 바 있다. 이러한 관점에 따라, 본 연구에서는 극단적으로 희소하거나 정의가 불명확한 클래스는 실험의 효율성과 분석의 명확성을 위해 제외하였다.

본 절에서는 네트워크의 구조적 특성을 탐색하기 위해, 주요 클래스에 대한 시각화 분석을 수행한다. 분석 대상은 대표적인 이상 행위 클래스인 ‘MIXER’, ‘RANSOMWARE’,

‘PONZI’와 정상 행위 클래스인 ‘EXCHANGE’이다. 각 클래스가 형성하는 부그래프(subgraph)의 구조적 차이를 시각적으로 비교함으로써, GNN 모델이 학습하게 될 유의미한 패턴을 직관적으로 확인하고자 한다.

시각화에는 오픈소스 네트워크 분석 도구인 Gephi를 활용하였으며, 레이아웃 알고리즘으로는 Fruchterman-Reingold를 적용했다(Area=10000.0, Gravity=10.0, Speed=1.0). 본 그래프는 송금 방향을 나타내는 방향성 그래프(directed graph)이며, 노드(node)는 차수(degree)에 따라, 엣지(edge)는 송금자 레이블 및 거래 가중치(weight)에 따라 시각적으로 구분되도록 표현했다. 이하에서는 각 클래스가 형성하는 부그래프(subgraph)의 구조적 차이를 시각적으로 비교함으로써, GNN 모델이 학습하게 될 유의미한 패턴을 직관적으로 확인하고자, 각 클래스별로 관찰된 시각적 패턴과 그에 대한 해석을 기술한다.

EXCHANGE 클래스의 시각화 결과, 네트워크 중앙에 극히 고밀도의 거대 연결 성분이 형성되고 그 주변으로 다수의 허브가 분포하는 중심-허브-주변부(core-hub-periphery) 구조가 뚜렷하게 나타났다. 이 거대 연결 성분(core)은 높은 차수의 노드를 중심으로 하는 여러 별 모양의 성분(hub)들과 소수의 ‘브리지’ 엣지로 연결되며, 허브들은 다시 외부의 소규모 노드 집단(periphery)으로 방사형으로 확장되는 형태를 보였다. 이러한 계층적 구조는 비트코인 생태계에서 대규모 거래 플랫폼(core)과 개별 거래소 클러스터(hubs)가 상호 연결되며 일반 사용자 집단(periphery)으로 확장되는 실제 상황을 반영하는 것으로 해석할 수 있다. 이는 소수의 허브가 대부분의 연결을 차지하는 ‘무척도 네트워크(scale-free network)’의 전형적인 특징으로, 정상적인 금융 활동에서 나타나는 복합적이고 다방향적인 자금 흐름을 보여준다.

MIXER 클래스의 시각화에서는 자금이 하나의 노드에서 다수의 노드로 병렬적으로 분산되는 팬아웃(fan-out) 패턴이 주로 나타났다. 짧은 거리에서 여러 목적지로 분기하는 별 모양 구조가 반복적으로 관찰되었으며, 네트워크 외곽에는 추가 거래를 수행하지 않는 1회성 터미널 노드가 다수 분포했다. 이는 거래 추적을 어렵게 하기 위해 자금을 수많은 주소로 분산시켜 흔적을 희석시키려는 믹싱 서비스의 전형적인 전략과 부합한다. 특히, 1회성 노드의 존재는 추적을 회피하기 위해 일회용 주소를 대량 생성하는 의도를 시사하며, 간헐적으로 관찰되는 중계 노드는 추적의 복잡성을 한층 더 높이는 장치로 볼 수 있다.

PONZI 클래스는 다수로부터 자금을 모으는 팬인(fan-in)과 일부에게 분배하는 팬아웃 패턴이 뚜렷하게 공존하는 양상을 보였다. 네트워크 외곽부에는 단발성 입출금 트랜잭션으로만 연결된 소규모 노드들이 다수 분포했으며, MIXER 클래스와 같은 대규모의 일방향적 팬아웃 패턴은 관찰되지 않았다. 이러한 양방향 자금 흐름은 다수의 참여자로부터 투자금을 유치(fan-in)하고 일부에게 수익금을 지급(fan-out)하는 폰지 사기의 재분배형 자금 순환 방식과 일치하는 것으로 해

석된다.

RANSOMWARE 클래스는 여러 피해자로부터 특정 주소로 자금이 집중되는 강력한 팬인으로 시작하여, 이후 자금이 소규모 단위로 연속 재송금되는 직선 형태의 긴 엣지 체인(edge chain)이 다수 관찰되었다. 이러한 직선적인 자금 이체 흐름은 대표적인 자금 세탁 기법인 필 체인(peel chain)의 시각적 특징과 정확히 부합한다. 이는 자금의 주된 흐름에서 소액을 ‘벗겨내(peel off)’ 여러 주소를 거치게 함으로써 추적을 어렵게 만드는 방식이다. MIXER가 자금을 병렬적으로 넓게 확산시키는 것과 대조적으로, RANSOMWARE는 자금을 직렬적으로 길게 이동시켜 흔적을 희석시키는 경향이 있음을 보여준다.

종합적으로, 본 시각화 분석은 각 클래스가 네트워크상에서 뚜렷하게 구별되는 고유한 구조적 특징을 가짐을 명확히 보여준다. 이는 GNN과 같은 그래프 기반 모델이 개별 거래의 특징만으로는 파악하기 어려운 네트워크의 거시적 구조 정보를 학습함으로써 이상 거래를 효과적으로 탐지할 수 있음을 시사한다.

3-3 DPAR

DPAR의 전체적인 흐름은 APPR행렬 계산과 특징 정보 학습이라는 크게 두 단계로 구성된다. 첫 번째 APPR 행렬 계산 단계에서는 차등 프라이버시가 적용된 PageRank를 이용하여 그래프의 구조를 분석한다. 이를 통해 각 노드에 대해 가장 중요한 Top-K의 이웃 노드를 특정한다. 다음으로 특징 정보 학습 단계에서는 앞서 특정된 Top-K의 이웃으로부터만 특징을 집계(aggregate)하고 DP-SGD(Differentially Private Stochastic Gradient Descent)를 적용하여 GNN 모델을 학습시킨다.

DPAR에서는 APPR을 계산하기 위해 ISTA(Iterative Shrinkage-Thresholding Algorithm)를 사용한다. 이는 APPR이 최소화하려는 목적 함수가 무엇인지 분석한 결과, 해당 문제가 ‘L1정규화 PageRank 문제(L1 regularized PageRank problem)’라는 최적화 문제에 해당한다는 결론에 따른 것이다. 따라서 이 문제를 해결하기 위한 효율적인 기법으로 ISTA가 채택되었다[18].

L1정규화 PageRank 문제는 아래와 같이 나타난다.

$$\psi(q) := \rho\alpha \|D^{1/2}q\|_1 + f(q) \quad (4)$$

이 식은 두 가지 주요 부분으로 구성되어 있다. 먼저 PageRank 계산 부분이다. 시드 벡터 s 를 기반으로 그래프 구조 Q 를 고려하여 어떤 노드가 중요한지를 평가한다. 이 항을 최소화하는 것은 시드에 대해 관련성이 높은 노드 그룹을 찾는 것에 해당한다.

$$f(q) := \frac{1}{2} \langle q, Qq \rangle - \alpha \langle s, D^{-1/2}q \rangle \quad (5)$$

$\rho\alpha\|D^{1/2}q_k\|_1$ 는 L1정규화에 해당한다. 이 항을 목적 함수에 추가함으로써 그래프 전체가 아닌 시드 주변의 국소적인 클러스터(local cluster)를 찾는 것이 가능해진다. 그러면 이 문제를 풀기위한 ISTA의 기본적인 식은 다음과 같이 표현된다[20].

$$q_{k+1}(i) = \underset{\rho\alpha d_i^2\| \cdot \|_1}{\text{prox}} (q_k(i) - \nabla_i f(q_k)) \quad (6)$$

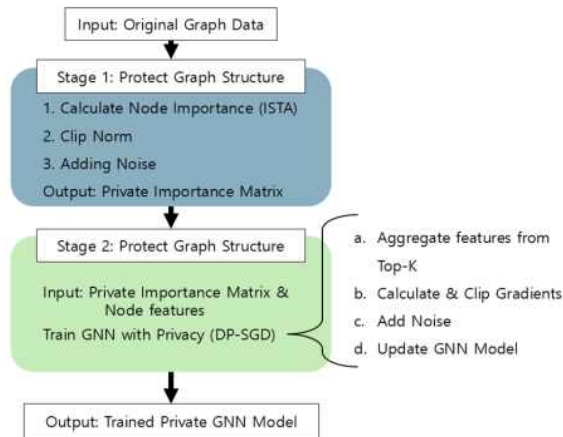


그림 3. DPAR프레임워크
Fig. 3. DPAR framework

ISTA를 통해 중요한 노드를 계산한 후 차등 프라이버시에서의 민감도(Sensitivity)를 억제하기 위해 Clip Norm을 수행한다. 간단히 말하면 민감도란 개별 데이터 하나가 결과에 미치는 영향의 크기이다. 추가해야 할 노이즈는 이 민감도에 비례한다[6]. 민감도가 크면 노이즈가 무한대가 되어 데이터가 쓸모없게 된다. 클리핑(clipping)은 이 민감도에 상한을 설정하기 위한 준비 단계이다. 구체적으로는 계산된 벡터의 놈(norm)을 측정하고, 그 크기가 사전에 정한 상한값을 초과하지 않는지 확인하며, 초과했을 경우 상한값까지 축소한다.

다음으로 노이즈를 추가하는 처리인데, 여기서는 좋은 결과를 보이는 가우시안 메커니즘을 사용한다[6].

IV. 실험 결과 및 논의

4-1 모델 결과

1) 베이스라인 모델 성능(No Differential Privacy)

베이스라인(baseline) 모델은 차등 프라이버시를 적용하지 않은 No-DP(No Differential Privacy) 설정이다. 이는 DPAR 모델과 동일한 아키텍처를 사용하되 프라이버시 보호를 위한 노이즈 주입 절차만 생략한 모델로, 달성 가능한 성능의 상한선(upper bound)을 제시하는 역할을 한다. 공정한 비교를 위해 데이터 분할(8:1:1 층화추출, seed=42), 전처리, 모델 구조, 하이퍼파라미터 등 모든 실험 조건은 DP 적용 모델과 동일하게 유지했다. 평가 지표로는 Test Accuracy와

Macro-F1 Score를 사용했으며, 베이스라인 모델의 최종 성능은 다음과 같다.

- Test Accuracy : 0.6312
- Test Macro-F1 : 0.6167

2) DP-PPR(전처리 단계) 성능

본 절에서는 DPAR의 전처리 단계, 즉 개인화 PageRank (PPR) 기반의 이웃 노드 중요도 계산 과정에 가우스 메커니즘을 적용했을 때의 효과를 분석한다. 실험은 프라이버시 예산(ϵ)을 1부터 8까지 변화시키며 수행했으며(seed=42), 각 ϵ 값은 노이즈의 크기를 결정하는 sigma_ista 파라미터를 조정하여 달성했다. 이때 delta 및 clip_bound_ista와 같은 다른 DP 관련 파라미터는 고정했으며, 그 외 모든 실험 조건은 베이스라인 모델과 동일하게 유지했다. 그림 4의 DP-PPR 곡선에서 확인할 수 있듯이, ϵ 값이 감소함에 따라(프라이버시 보호 강도가 높아짐에 따라) 모델의 정확도가 급격히 하락하는 전형적인 프라이버시-유용성 상충 관계가 관찰되었다. 특히, $\epsilon=6$ 부근에서는 정확도가 일시적으로 하락하는 비단조적 변동(non-monotonic fluctuation)이 나타나기도 했다. 각 ϵ 값에 해당하는 sigma_ista 값은 표 4와 같다.

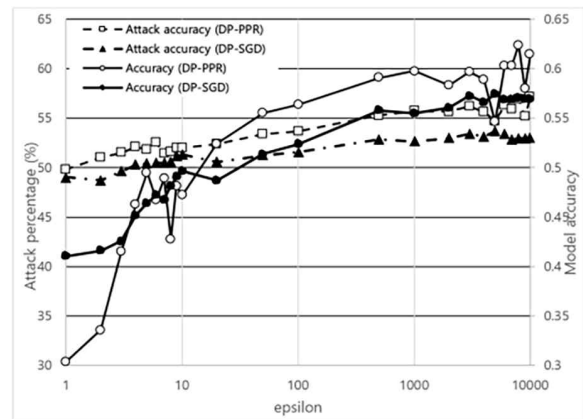


그림 4. 프라이버시 예산 ϵ 에 따른 정확도와 공격 정밀도(DP-PPR vs DP-SGD)

Fig. 4. Performance vs. Privacy (ϵ): DP-PPR and DP-SGD

표 4. DP-PPR(가우시안 메커니즘)에서 사용한 ϵ - σ 대응표

Table 4. ϵ - σ mapping in DP-PPR(Gaussian mechanism)

ϵ 값	sigma_ista
1	0.190
2	0.107
3	0.072
4	0.057
5	0.046
6	0.040
7	0.0345
8	0.0305

3) DP-SGD (학습 단계) 성능

학습 단계에서는 DPAR의 GNN 모델 학습 단계에 차등 프라이버시를 적용하는 DP-SGD(Differentially Private Stochastic Gradient Descent)의 성능을 분석한다. DP-SGD는 학습 과정에서 그래디언트에 클리핑을 적용한 후 가우시안 노이즈를 주입하는 방식이다. 실험은 프라이버시 예산(ϵ)을 1부터 8까지 동일한 범위에서 변화시키며 수행했으며(seed=42), 각 ϵ 값은 σ_{sgd} 파라미터를 통해 조정했다. clip_bound_sgd 와 delta_sgd 는 고정값으로 사용했으며, 그 외 모든 조건은 베이 스텐 모델과 동일하다. 그림 4의 DP-SGD 곡선을 DP-PPR과 비교 분석한 결과, 두 가지 주요한 특징이 관찰되었다. 첫째, 동일한 ϵ 변화 구간에 대해, DP-SGD는 DP-PPR보다 정확도 하락폭이 비교적 완만하게 나타났다. 둘째, 흥미롭게도 프라이버시 제약이 약한 높은 ϵ 구간에서도 DP-SGD의 전반적인 성능이 DP-PPR의 성능을 상회하지는 못했다. 각 ϵ 값에 해당하는 σ_{sgd} 값은 표 5와 같다.

표 5. DP-SGD(가우시안 메커니즘)에서 사용한 ϵ - σ 대응표
Table 5. ϵ - σ mapping in DP-SGD(Gaussian mechanism)

ϵ	σ_{sgd}
1	12.48
2	7.60
3	5.90
4	5.00
5	4.3
6	3.90
7	3.60
8	3.39

4) 실험 파라미터 설정

본 연구의 모든 실험에서 공통적으로 고정된 파라미터는 다음과 같다: 데이터 분할 방식(8:1:1 층화추출, seed=42), GNN 모델 구조(hidden_size, n_layers, dropout), 최적화 관련 하이퍼파라미터(learning_rate, weight_decay, batch_size), 그리고 추론 관련 파라미터(alpha, top_k, n_prop_inference)이다. 차등 프라이버시의 강도를 조절하는 실험은 다음과 같은 규칙에 따라 수행했다.

- DP-PPR (전처리): delta와 clip_bound_lista 값을 고정 한 상태에서, 노이즈의 크기를 결정하는 sigma_ista 값을 조정하여 목표 ϵ 값을 달성했다.
- DP-SGD (학습): delta_sgd와 clip_bound_sgd 값을 고정한 상태에서, sigma_sgd 값을 조정하여 목표 ϵ 값을 제어했다.

본 연구에서 사용된 차등 프라이버시의 핵심 파라미터 ϵ 는 가우시안 메커니즘(Gaussian Mechanism)의 노이즈 크기(σ), 데이터 샘플링 비율, 그리고 총 학습 단계(step) 수를 고려하여 산출되었다. (ϵ , δ)-DP의 정확한 계산을 위해 Rényi

Differential Privacy(RDP)를 기반으로 한 모멘트 회계 방식을 사용했으며, δ 값은 $1e-5$ 로 고정하였다.

DP-PPR의 경우, 전처리 단계에서 한 번의 노이즈 주입이 발생하므로 ϵ 는 σ_{ista} 값에 의해 직접적으로 결정된다. DP-SGD의 ϵ 산출에 사용된 주요 파라미터는 다음과 같다.

- 노이즈 승수 (Noise Multiplier, σ_{sgd}): 그래디언트에 추가되는 가우시안 노이즈의 표준편차. 표 5에 명시된 값을 사용했다.
- 샘플링 비율(Sampling Ratio, q): 각 학습 단계에서 사용되는 데이터의 비율로, 배치 크기/총 학습 데이터 수로 계산된다. 본 실험에서는 $q = \text{batch_size} / n_{\text{train_samples}}$ 이다.
- 총 학습 단계(Total Steps, T): 에포크(epoch) 수 \times (총 학습 데이터 수 / 배치 크기)로 계산된다.
- 델타 (δ): 프라이버시 보장이 실패할 미세한 확률로, $1e-5$ 로 설정했다.

위 파라미터들을 모멘트 회계 함수에 입력하여 각 σ_{sgd} 값에 해당하는 ϵ 값을 도출하였다. 이러한 파라미터와 회계 절차를 통해 본 실험의 프라이버시 수준은 재현 가능하다.

5) 독립적 평가의 당위성

본 연구가 DP-PPR과 DP-SGD의 효과를 독립적으로 분리하여 평가한 이유는 다음과 같은 두 가지 실험적 고려사항에 기반한다.

첫째, 효과의 독립적 분석이다. DPAR의 전처리 단계(구조 정보 보호)와 학습 단계(특정 정보 보호)에 적용되는 노이즈는 모델 성능에 미치는 영향이 상이할 것으로 가정된다. 따라서 각 메커니즘을 단독으로 적용했을 때의 프라이버시 예산(ϵ)과 성능 간의 관계를 명확히 규명하기 위해 독립적인 실험을 설계했다.

둘째, 파라미터 탐색 공간의 제어이다. 두 메커니즘의 ϵ 값을 동시에 변화시킬 경우, 탐색해야 할 파라미터 조합의 수가 기하급수적으로 증가하여 효율적인 분석이 어렵다. 이에 본 연구에서는 각 메커니즘의 단독 효과를 우선적으로 분석하고, 두 메커니즘의 최적 조합을 찾는 연구는 향후 과제로 남겨두었다.

4-2 공격 결과

본 절에서는 DPAR 프레임워크의 프라이버시 보호 성능을 정량적으로 평가하기 위해 수행한 멤버십 추론 공격 및 속성 추론 공격의 결과를 기술한다. 멤버십 추론 공격은 실제 서비스 환경을 모사하기 위해 블랙박스(black-box) 설정을 가정하여 수행했다. 구체적으로, 공격 대상과 동일한 아키텍처를 갖는 10개의 새도우 모델을 학습시킨 후, 이 모델들의 출력값을 기반으로 로지스틱 회귀(Logistic Regression) 기반의 공격 모델을 구축했다.

실험 결과, 놀랍게도 프라이버시 보호 기술(DP-SGD,

DP-PPR)의 적용 여부와 관계없이 모든 실험 조건에서 공격 정확도가 무작위 추측 수준인 약 0.5로 측정되었다. 이는 DPAR 프레임워크가 차등 프라이버시 메커니즘을 추가하지 않더라도, MIA에 대해 내재적으로 높은 견고성(robustness)을 가짐을 시사한다. 이러한 결과의 원인은 DPAR의 내장된 정규화 기법에서 찾을 수 있다. 멤버십 추론 공격은 주로 모델의 과적합(overfitting) 현상을 악용하여 학습 데이터의 특징을 추론한다[18]. DPAR 프레임워크는 과적합을 억제하기 위해 L1 정규화와 드롭아웃(dropout)을 기본적으로 사용하는데, 이러한 장치들이 결과적으로 멤버십 정보의 유출을 효과적으로 방지하여 공격을 무력화한 것으로 판단된다.

앞선 멤버십 추론 공격(MIA) 실험에서는 DPAR 프레임워크의 내재적 정규화 기능으로 인해, 차등 프라이버시 기술이 제공하는 추가적인 보호 효과를 분리하여 측정하는 데 한계가 있었다. 이에 본 연구는 GNN 아키텍처의 고유한 특성을 직접적으로 공략하는 속성 추론 공격(Attribute Inference Attack)을 수행하여 프라이버시 강도를 보다 심층적으로 평가했다.

속성 추론 공격은 대상 노드의 이웃 노드 정보(그래프 구조 및 공개 속성)를 바탕으로 알려지지 않은 민감 속성을 추론하는 것을 목표로 한다. GNN은 메시지 패싱(message passing)을 통해 이웃 노드의 속성 정보를 집계하여 중심 노드의 표현을 학습하는 모델이므로, 이웃의 정보가 중심 노드에 유출될 본질적인 위험을 내포한다. 따라서 속성 추론 공격은 GNN의 핵심 동작 원리가 야기하는 프라이버시 취약점을 직접적으로 평가할 수 있는 합리적인 벤치마크라 할 수 있다.

본 속성 추론 공격은 공격자가 DPAR 모델 학습에 사용된 그래프 구조 및 공개 속성과 함께, 모델이 출력하는 최종 예측 직전의 로짓(logit) 값을 획득할 수 있는 회색 상자(gray-box) 시나리오를 가정한다. 공격자는 이 로짓 값을 신뢰도 점수로 활용하여, 모델의 주 예측 대상인 레이블을 제외한 35개의 다른 속성을 추론했다. 평가는 35개 속성에 대한 평균 예측 정확도를 지표로 사용했으며, DPAR의 두 가지 프라이버시 보호 기법(DP-PPR, DP-SGD)의 효과를 개별적으로 분석했다.

실험 결과, 프라이버시 보호 기술이 없는 베이스라인 모델에 대한 공격 정확도는 57.12%로 측정되었다. 이는 무작위 추측을 상회하지만, 공격의 성공률이 모델 자체의 예측 성능에 의해 제한되는 상한선을 가짐을 시사한다. 본 연구는 이 베이스라인 정확도를 기준으로 각 프라이버시 기법의 방어 성능을 평가했다. DP-PPR을 적용한 경우, 프라이버시 예산(ϵ)과 공격 정확도 간의 명확한 상충 관계가 나타났다. $\epsilon=1$ 의 가장 강력한 보호 설정에서 공격 정확도는 베이스라인 대비 5.95%p 감소한 51.17%를 기록했다. 이후 ϵ 값이 증가함에 따라 방어 성능이 점차 약화되어, $\epsilon=10,000$ 에서는 베이스라인과 거의 동일한 57.10%의 공격 정확도를 보였다.

반면, DP-SGD는 ϵ 값의 변화에 크게 의존하지 않는 안정적이고 강력한 방어 성능을 보였다. $\epsilon=1$ 설정에서는 공격 정

표 6. MIA 결과
Table 6. Result of MIA

$\epsilon = 1$	NO-DP
Accuracy	0.5324

표 7. 속성 추론 공격의 공격성능결과
Table 7. Result of attribute inference attack

	DP-PPR
Baseline	57.12
$\epsilon = 1$	51.17
$\epsilon = 2$	51.05
$\epsilon = 3$	51.49
$\epsilon = 4$	52.09
$\epsilon = 5$	51.83
$\epsilon = 6$	52.51
$\epsilon = 7$	51.45
$\epsilon = 8$	51.58
$\epsilon = 9$	51.96
$\epsilon = 10$	51.98
$\epsilon = 20$	52.39
$\epsilon = 50$	53.36
$\epsilon = 100$	53.66
$\epsilon = 500$	55.28
$\epsilon = 1000$	55.69
$\epsilon = 2000$	55.68
$\epsilon = 3000$	56.19
$\epsilon = 4000$	55.62
$\epsilon = 5000$	54.75
$\epsilon = 6000$	56.27
$\epsilon = 7000$	55.90
$\epsilon = 8000$	56.99
$\epsilon = 9000$	55.15
$\epsilon = 10000$	57.10

확도가 베이스라인 대비 9.95%p 하락한 47.17%까지 감소했다. ϵ 값이 10,000으로 증가했을 때도 공격 정확도는 53.04% 수준에 머물러, DP-SGD가 속성 추론 공격에 대해 일관된 내성을 가짐을 확인했다.

4-3 유용성과 프라이버시의 상충 관계 분석

본 연구에서는 프라이버시 수준(ϵ)을 1부터 10,000까지 조정하며 모델의 유용성과 프라이버시 견고성 간의 상충 관계를 분석했다. 모델 유용성의 경우, ϵ 값이 감소함에 따라(프라이버시 강화) 정확도가 하락하는 전형적인 상충 관계가 관찰되었다. 속성 추론 공격에 대한 방어 성능 역시 ϵ 값에 비례하는 경향을 보였으나, 두 메커니즘(DP-PPR, DP-SGD) 모두 전 구간에서 공격 정확도를 50%대로 안정적으로 억제하는 높은 견고성을 나타냈다.

모델의 유용성과 프라이버시 보호의 최적 균형점(sweet spot)을 탐색하기 위해, 모델 성능(정확도) 곡선과 공격 성공률 곡선 간의 차이를 분석했다. 그 결과, ϵ 값이 7,000에서 8,000 사이 구간에서 두 지표 간의 균형이 가장 이상적인 것으로 나타났다. 해당 구간에서 모델 성능은 약 60%에 도달하며 높은 유용성을 확보하는 반면, 공격 성공률은 약 55% 수준으로 억제되었다. 이는 다른 구간과 비교했을 때 모델의 유용성이 공격 성공률을 가장 큰 폭(각각 4.46%p, 5.43%p)으로 상회하는 지점이다.

이러한 결과는 실용적 관점에서 중요한 시사점을 제공한다. DP-SGD는 ϵ 값과 무관하게 강력한 방어 성능을 보였지만 전반적인 모델 유용성이 낮았다. 반면, DP-PPR은 최적 구간에서 높은 모델 유용성을 달성하면서도 충분한 수준의 프라이버시 보호가 가능함을 보였다. 따라서 제한된 성능 저하 내에서 효과적인 프라이버시를 확보해야 하는 실제 환경에서는 DP-PPR을 중심으로 프라이버시 예산을 조정하는 것이 합리적인 전략이 될 수 있다.

V. 결론 및 향후 연구방향

본 연구는 실제 비트코인 거래 네트워크 데이터에 차등 프라이버시 GNN 모델인 DPAR을 적용하여, 프라이버시 예산(ϵ)의 변화에 따른 모델 유용성과 프라이버시 견고성 간의 상충 관계를 실증적으로 분석했다. 실험을 통해 도출된 주요 결과는 다음과 같다.

첫째, DPAR 프레임워크는 차등 프라이버시 기술의 적용 여부와 관계없이 멤버십 추론 공격(MIA)에 대해 내재적인 견고성을 보였다. 모든 실험 조건에서 공격 성공률이 무작위 추측 수준인 약 50%로 측정되었으며, 이는 모델에 기본적으로 적용된 L1 정규화 및 드롭아웃이 과적합을 효과적으로 억제하여 정보 유출을 방지한 결과로 해석된다.

둘째, GNN의 구조적 특성을 직접 공략하는 속성 추론 공격에서는 DP-PPR과 DP-SGD가 상이한 방어 특성을 나타냈다. DP-PPR은 ϵ 값과 비례하여 방어 성능이 결정되는 명확한 상충 관계를 보인 반면, DP-SGD는 ϵ 값에 크게 의존하지 않는 안정적인 강력한 방어 성능을 입증했다.

셋째, 모델 유용성과 프라이버시 보호 간의 최적 균형점 분석을 통해, DP-PPR 적용 시 ϵ 값이 7,000에서 8,000 사이 일 때 모델 정확도는 약 60% 수준을 유지하면서도 속성 추론 공격 성공률을 55% 수준으로 억제했다. 이는 다른 구간 대비 모델의 유용성이 공격 성공률을 가장 큰 폭(각각 4.46%p, 5.43%p)으로 상회하는 지점이다. 다만, 향후 연구에서 ϵ 값을 1, 2, 5, 10, 15, 20과 같이 낮은 구간에서 더 세분화하여 실험을 진행할 필요가 있다.

본 연구의 결과는 차등 프라이버시 기술을 실제 금융 시스템에 적용하는 데 있어 중요한 학술적, 실용적 시사점을 제공한다. 가장 중요한 의의는 프라이버시 예산(ϵ)과 실질적인 보

안 강도의 관계가 비선형적일 수 있음을 보였다는 점이다. 이론적으로 매우 큰 ϵ 값(약한 프라이버시)으로 여겨지는 7,000~8,000 구간에서도 정교한 속성 추론 공격을 효과적으로 방어할 수 있음을 입증함으로써, '강력한 프라이버시는 항상 낮은 유용성을 동반한다'는 고정관념에 도전한다. 이는 프라이버시 수준을 단순히 이론적인 ϵ 값으로만 판단할 것이 아니라, 구체적인 공격 시나리오와 모델 아키텍처를 고려한 실증적 평가가 중요함을 시사한다.

또한, 본 연구는 AI 모델의 아키텍처 자체가 프라이버시의 첫 번째 방어선이 될 수 있음을 보여준다. MIA에 대한 DPAR의 내재적 견고성은, 차등 프라이버시와 같은 명시적인 보호 기술을 적용하기에 앞서 과적합을 방지하는 견고한 모델 설계를 갖추는 것이 중요함을 강조한다. 이는 향후 프라이버시 보호 모델을 설계할 때 고려해야 할 중요한 원칙을 제시한다.

본 연구는 의미 있는 결과를 도출했음에도 불구하고 다음과 같은 한계점을 가지며, 이는 향후 연구의 중요한 방향이 될 수 있다.

첫째, 모델의 표현력 한계이다. 본 연구에서는 프라이버시 예산 소모를 고려하여 비교적 단순한 신경망 구조를 사용했다. 하지만 높은 ϵ 값에서도 견고성이 유지됨을 확인한 만큼, 향후에는 그래프 트랜스포머(Graph Transformer)와 같이 더 복잡하고 표현력이 높은 모델 아키텍처를 적용하여 유용성을 극대화하는 연구가 필요하다.

둘째, 정적 그래프(static graph) 분석의 한계이다. 실제 금융 거래는 시간에 따라 동적으로 변화하지만, 본 연구는 특정 시점의 스냅샷인 정적 그래프만을 다루었다. 향후 연구는 거래의 시간적 순서와 패턴을 학습할 수 있는 Temporal GNN(TGN)이나 Dynamic GNN에 차등 프라이버시를 결합하여, 시간에 따라 진화하는 이상 거래 패턴을 탐지하는 방향으로 확장되어야 한다. 또한 모델이 학습한 정보를 바탕으로 원본 그래프의 연결 관계를 복원하는 그래프 구조 재구성 공격(Graph Reconstruction Attack)이나 특정 노드들 간의 연결 여부를 추론하는 링크 추론 공격(Link Inference Attack)에 대한 방어 성능을 추가적으로 검증하는 연구가 필요하다.

셋째, 거래 익명화 기술 미반영이다. 본 연구는 다수의 거래를 의도적으로 혼합하여 송수신 관계를 모호하게 만드는 코인조인(CoinJoin)과 같은 특수 거래를 분석에서 제외했다. 이러한 익명화 기술은 GNN 기반 분석에 큰 도전 과제이므로, 향후에는 이러한 의도적 노이즈가 포함된 환경에서도 강건하게 작동하는 프라이버시 보호 모델을 개발하는 연구가 요구된다.

넷째, 계산 효율성 분석의 한계이다. 본 연구는 차등 프라이버시 적용에 따른 유용성과 프라이버시 간의 상충 관계를 실증적으로 분석하는 데 초점을 맞추었다. 따라서 실제 수억 개 이상의 노드를 가진 대규모 그래프 환경에서의 학습 시간, 메모리 사용량 등 계산 자원 소모에 대한 심층적인 분석은 다루지 않았다. DPAR 모델을 실제 금융 시스템에 적용하기 위해서는 확장성 검증이 필수적이므로, 향후 연구에서는 모델의

계산 효율성을 측정하고 최적화하는 방안에 대한 탐구가 필요하다.

결론적으로 본 연구는 실제 금융 데이터에 대한 프라이버시 보호 AI 기술의 적용 가능성을 탐색하고 실질적인 지침을 제공하는 foundational work로서 의의를 갖는다. 향후 연구는 동적 그래프의 확장과 고도화된 모델 아키텍처 적용을 통해, 더욱 복잡하고 현실적인 금융 환경에서 신뢰할 수 있는 AI 시스템을 구축하는 방향으로 나아가야 할 것이다.

감사의 글

본 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-지역지능화혁신인재양성사업의 지원을 받아 수행된 연구(IITP-2025-RS-2024-00436765)로서, 관계부처에 감사드립니다.

참고문헌

- [1] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System," *Whitepaper*, 2008. <https://bitcoin.org/bitcoin.pdf>
- [2] S. Meiklejohn, M. Pomarole, G. Jordan, K. Levchenko, D. McCoy, G. M. Voelker, and S. Savage, "A Fistful of Bitcoins: Characterizing Payments Among Men with No Names," in *Proceedings of the 2013 Conference on Internet Measurement Conference (IMC)*, Barcelona: Spain, pp. 127-140, October 2013. <https://doi.org/10.1145/2504730.2504747>
- [3] S. Foley, J. R. Karlsen, and T. J. Putniņš, "Sex, Drugs, and Bitcoin: How Much Illegal Activity Is Financed through Cryptocurrencies?," *Review of Financial Studies*, Vol. 32, No. 5, pp. 1798-1853, 2019. <https://doi.org/10.1093/rfs/hhz015>
- [4] S. X. Rao, S. Zhang, Z. Han, Z. Zhang, W. Min, Z. Chen, ... and C. Zhang, "xFraud: Explainable Fraud Transaction Detection," *Proceedings of the VLDB Endowment*, Vol. 15, No. 3, pp. 427-436, 2022. <https://doi.org/10.14778/3494124.3494128>
- [5] Y. Qi, X. Lin, and J. Wu, "Differentially Private Graph Neural Network with Importance-Grained Noise Adaption," arXiv:2308.04943, August 2023. <https://doi.org/10.48550/arXiv.2308.04943>
- [6] Q. Zhang, H. K. Lee, J. Ma, J. Lou, C. Yang, and L. Xiong, "DPA: Decoupled Graph Neural Networks with Node-Level Differential Privacy," in *Proceedings of the ACM Web Conference 2024*, Singapore, pp. 1170-1181, May 2024. <https://doi.org/10.1145/3589334.3645607>
- [7] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," arXiv:1609.02907, 2017. <https://arxiv.org/abs/1609.02907>
- [8] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," arXiv:1710.10903, 2018. <https://arxiv.org/abs/1710.10903>
- [9] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep Learning with Differential Privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, Vienna: Austria, pp. 308-318, October 2016. <https://doi.org/10.1145/2976749.2978318>
- [10] H. Schnoering and M. Vazirgiannis, "Bitcoin Research with a Transaction Graph Dataset," *Scientific Data*, Vol. 12, No. 1, 404, 2025. <https://doi.org/10.1038/s41597-025-04684-8>
- [11] C. Dwork and A. Roth, "The Algorithmic Foundations of Differential Privacy," *Foundations and Trends in Theoretical Computer Science*, Vol. 9, No. 3-4, pp. 211-407, 2014. <https://doi.org/10.1561/04000000042>
- [12] F. McSherry and K. Talwar, "Mechanism Design via Differential Privacy," in *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, Providence: RI, pp. 94-103, 2007. <https://doi.org/10.1109/FOCS.2007.66>
- [13] I. Mironov, "Rényi Differential Privacy," in *Proceedings of the 2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, Santa Barbara: CA, pp. 263-275, 2017. <https://doi.org/10.1109/CSF.2017.11>
- [14] Y.-X. Wang, B. Balle, and S. P. Kasiviswanathan, "Subsampled Rényi Differential Privacy and Analytical Moments Accountant," in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1226-1235, 2019. <http://proceedings.mlr.press/v89/wang19b.html>
- [15] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," *Computer Networks and ISDN Systems*, Vol. 30, No. 1-7, pp. 107-117, 1998. [https://doi.org/10.1016/S0169-7552\(98\)00110-X](https://doi.org/10.1016/S0169-7552(98)00110-X)
- [16] R. Andersen, F. Chung, and K. Lang, "Local Graph Partitioning Using PageRank Vectors," in *Proceedings of the 2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, Berkeley: CA, pp. 475-486, 2006. <https://doi.org/10.1109/FOCS.2006.44>
- [17] D. Cheng, Y. Zou, S. Xiang, and C. Jiang, "Graph Neural Networks for Financial Fraud Detection: A Review," *Frontiers of Computer Science*, Vol. 19, 199609, 2025. <https://doi.org/10.1007/s11704-024-40474-y>
- [18] R. Shokri, M. Stronati, C. Song, and V. Shmatikov,

“Membership Inference Attacks Against Machine Learning Models,” in *Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP)*, San Jose: CA, pp. 3-18, May 2017. <https://doi.org/10.1109/SP.2017.41>

[19] B. Jayaraman and D. Evans, “Are Attribute Inference Attacks Just Imputation?,” in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, Los Angeles: CA, pp. 1569-1582, November 2022. <https://doi.org/10.1145/3548606.3560663>

[20] K. Fountoulakis, F. Roosta-Khorasani, J. Shun, X. Cheng, and M. W. Mahoney, “Variational Perspective on Local Graph Clustering,” *Mathematical Programming*, Vol. 174, No. 1, pp. 553-573, March 2019. <https://doi.org/10.1007/s10107-017-1214-8>



우에다 마사토(Masato Ueda)

2021년 : 선문대학교 AI소프트웨어학과 (학사)

2025년~현 재: 선문대학교 AI소프트웨어학과 석사과정
 ※ 관심분야 : 정보보호(Privacy), AIOps, GNN 등

콘도 코이치(Koichi Kondo)



2021년 : 선문대학교 AI소프트웨어학과 (학사)

2025년~현 재: 선문대학교 AI소프트웨어학과 석사과정
 ※ 관심분야 : LLM, NLP, 정보보호(Privacy) 등

황석형(Suk-Hyung Hwang)



1994년 : 일본Osaka University 공학 석사(소프트웨어공학 전공)
 1997년 : 일본Osaka University 공학 박사(소프트웨어공학 전공)

1997년 3월~현 재: 선문대학교 AI소프트웨어학과 교수
 ※ 관심분야 : Software Engineering, Formal Concept Analysis, QA4AI, Software2.0&3.0



정영애(Young-Ae Jung)

2000년 : 호서대학교 대학원 이학석사 (인공지능 전공)
 2007년 : 단국대학교 대학원 이학박사 (AI-소프트웨어공학 전공)

2009년 3월~2024년 2월: 선문대학교 IT교육학부 교수
 2024년 3월~현 재: 선문대학교 AI소프트웨어학과 교수
 ※ 관심분야 : AI-Software Engineering, Computer Vision, AIOps, Text Mining, Privacy



황세웅(Se-Woong Hwang)

2015년 : 연세대학교 대학원 정보시스템 석사
 2021년 : 연세대학교 대학원 정보시스템 박사

2022년 9월~현 재: 선문대학교 AI소프트웨어학과 교수
 ※ 관심분야 : collaborative AI, Sensor-based Application Technology, System Intelligence