

Adversarial Signed Graph Learning with Differential Privacy

Haobin Ke

The Hong Kong Polytechnic
University

Hong Kong, China

haobin.ke@connect.polyu.hk

Sen Zhang

The Hong Kong Polytechnic
University

Hong Kong, China

senzhang@polyu.edu.hk

Qingqing Ye

The Hong Kong Polytechnic
University

Hong Kong, China

qqing.ye@polyu.edu.hk

Xun Ran

The Hong Kong Polytechnic
University

Hong Kong, China

qi-xun.ran@connect.polyu.hk

Haibo Hu*

The Hong Kong Polytechnic
University

Hong Kong, China

haibo.hu@polyu.edu.hk

Abstract

Signed graphs with positive and negative edges can model complex relationships in social networks. Leveraging on balance theory that deduces edge signs from multi-hop node pairs, signed graph learning can generate node embeddings that preserve both structural and sign information. However, training on sensitive signed graphs raises significant privacy concerns, as model parameters may leak private link information. Existing protection methods with differential privacy (DP) typically rely on edge or gradient perturbation for unsigned graph protection. Yet, they are not well-suited for signed graphs, mainly because edge perturbation tends to cascading errors in edge sign inference under balance theory, while gradient perturbation increases sensitivity due to node interdependence and gradient polarity change caused by sign flips, resulting in larger noise injection. In this paper, motivated by the robustness of adversarial learning to noisy interactions, we present ASGL, a privacy-preserving adversarial signed graph learning method that preserves high utility while achieving node-level DP. We first decompose signed graphs into positive and negative subgraphs based on edge signs, and then design a gradient-perturbed adversarial module to approximate the true signed connectivity distribution. In particular, the gradient perturbation helps mitigate cascading errors, while the subgraph separation facilitates sensitivity reduction. Further, we devise a constrained breadth-first search tree strategy that fuses with balance theory to identify the edge signs between generated node pairs. This strategy also enables gradient decoupling, thereby effectively lowering gradient sensitivity. Extensive experiments on real-world datasets show that ASGL achieves favorable privacy-utility trade-offs across multiple downstream tasks. Our code and data are available in <https://github.com/KHBDL/ASGL-KDD26>.

*Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD '26, Jeju, ON, Korea

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YYY/MM

<https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

Keywords

Differential privacy, Adversarial signed graph learning, Constrained breadth first search-trees, Balanced theory.

ACM Reference Format:

Haobin Ke, Sen Zhang, Qingqing Ye, Xun Ran, and Haibo Hu. 2026. Adversarial Signed Graph Learning with Differential Privacy. In *Proceedings of the 32st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1 (KDD '26), August 9–13, 2026, Jeju, ON, Korea*. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

1 Introduction

The signed graph is a common and widely adopted graph structure that can represent both positive and negative relationships using signed edges [4, 6, 30]. For example, in online social networks shown in Fig. 1, while user interactions reflect positive relationships (e.g., like, trust, friendship), negative relationships (e.g., dislike, distrust, complaint) also exist. Signed graphs provide more expressive power than unsigned graphs to capture such complex user interactions.

Recently, some studies [16, 22, 26] have explored signed graph learning methods, aiming to obtain low-dimensional vector representations of nodes that preserve key signed graph properties: neighbor proximity and structural balance. These embeddings are subsequently applied to downstream tasks such as edge sign prediction, node clustering, and node classification. Among existing signed graph learning methods, *balance theory* [3] has proven effective in identifying the edge signs between the source node and multi-hop neighbor nodes. It is leveraged in graph neural network (GNN)-based models to guide message passing across signed edges, ensuring that information aggregation is aligned with the node proximity [7, 17, 18]. Moreover, to enhance the robustness and generalization capability of deep learning models, the adversarial graph embedding model [21, 33] learns the underlying connectivity distribution of signed graphs by generating high-quality node embeddings that preserve signed node proximity.

Despite their ability to effectively capture signed relationships between nodes, graph learning models remain vulnerable to link stealing attacks [13, 35, 42], which aim to infer the existence of links between arbitrary node pairs in the training graph. For instance, in online social graphs, such attacks may reveal whether two users share a friendly or adversarial relationship, compromising user privacy and damaging personal or professional reputations.

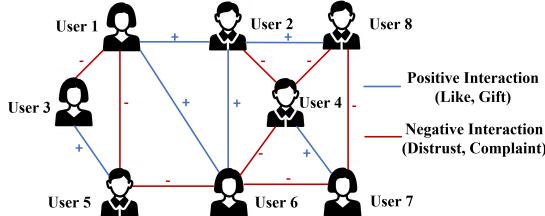


Figure 1: A signed social graph with blue edges for positive links and red edges for negative links.

Differential privacy (DP) [9] is a rigorous privacy framework that guarantees statistically indistinguishable outputs regardless of any individual data presence. Such guarantee is achieved through sufficient perturbation while maintaining provable privacy bounds and computational feasibility. Existing privacy-preserving graph learning methods with DP can be categorized into two types based on the perturbation mechanism: one applies edge perturbation [23] to protect the link information by modifying the graph structure, and the other adopts gradient perturbation [36, 37] to obscure the relationships between nodes during model training. However, these methods are not well-suited for signed graph learning due to the following two challenges:

- **Cascading error:** As illustrated in Fig. 2, balance theory facilitates the inference of the edge sign between two unconnected nodes by computing the product of edge signs along a path. However, existing methods that use edge perturbation to protect link information may alter the sign of any edge along the path, thereby leading to incorrect inference of edge signs under balance theory. Such a local error can further propagate along the path, resulting in cascading errors in edge sign inference.
- **High sensitivity:** While gradient perturbation methods without directly perturbing edges may mitigate cascading errors, they are still ill-suited for signed graph learning because the node interdependence in signed graphs leads to high gradient sensitivity.¹ Furthermore, edge change may induce sign flips that reverse gradient polarity within the loss function (see Eq. (10) for details), resulting in higher sensitivity compared to unsigned graphs. This increased sensitivity requires larger noise for privacy protection, thereby reducing the data utility.

To address these challenges, we turn to an adversarial learning-based approach for private signed graph learning. The core motivation is that this adversarial method generates node embeddings by approximating the true connectivity distribution, making it naturally robust to noisy interactions during optimization. As a result, we propose ASGL, a differentially private adversarial signed graph learning method that achieves high utility while maintaining node-level differential privacy. Within ASGL, the signed graph is first decomposed into positive and negative subgraphs based on edge signs. These subgraphs are then processed through an adversarial learning module within shared model parameters, enabling both positive and negative node pairs to be mapped into a unified embedding space while effectively preserving signed proximity. Based on this, we develop the adversarial learning module with differentially private stochastic gradient descent (DPSGD), which generates

¹The presence or absence of a node affects gradient updates of itself and its neighbors.

private node embeddings that closely approximate the true signed connectivity distribution. In particular, *the gradient perturbation helps mitigate cascading errors, while the subgraph separation avoids gradient polarity reversals induced by edge sign flips within the loss function, thereby reducing the sensitivity to changes in edge signs*. Considering that node interdependence further increases gradient sensitivity, we design a constrained breadth-first search (BFS) tree strategy within adversarial learning. *This strategy integrates balance theory to identify the edge signs between generated node pairs, while also constraining the receptive fields of nodes to enable gradient decoupling, thereby effectively lowering gradient sensitivity and reducing noise injection*. Our main contributions are listed as follows:

- We present a privacy-preserving adversarial learning method for signed graphs, called ASGL. To our best knowledge, it is the **first** work that can ensure the node-level differential privacy of signed graph learning while preserving high data utility.
- To mitigate cascading errors, we develop the adversarial learning module with DPSGD, which generates private node embeddings that closely approximate the true signed connectivity distribution. This approach avoids direct perturbation of the edge structure, which helps mitigate cascading errors and prevents gradient polarity reversals in the loss function.
- To further reduce the sensitivity caused by complex node relationships, we design a constrained breadth-first search tree strategy that integrates balance theory to identify edge signs between generated node pairs. This strategy also constrains the receptive fields of nodes, enabling gradient decoupling and effectively lowering gradient sensitivity.
- Extensive experiments demonstrate that our method achieves favorable privacy-accuracy trade-offs and significantly outperforms state-of-the-art methods in edge sign prediction and node clustering tasks. Additionally, we conduct link stealing attacks, demonstrating that ASGL exhibits stronger resistance to such attacks across all datasets.

The remainder of our work is organized as follows. Section 2 describes the preliminaries of our solution. The problem statement is introduced in Section 3. Our proposed solution and its privacy analysis are presented in Section 4. The experimental results are reported in Section 5. We discuss related works in Section 6, followed by conclusion in Section 7.

2 Preliminaries

In this section, we provide an overview of signed graphs, differential privacy, and DPSGD. Additionally, the vanilla adversarial graph learning is introduced in **App. A**, and the frequently used notations are summarized in Table 5 (See **App. B**).

2.1 Signed Graph with Balance Theory

A signed graph is denoted as $\mathcal{G} = (V, E^+, E^-)$, where V is the set of nodes, and E^+/E^- represent positive and negative edge sets, respectively. An edge $e_{ij} = (v_i, v_j) \in E^+/E^-$ represents the positive/negative link between node pair $(v_i, v_j) \in V$, respectively. Notably, $E^+ \cap E^- = \emptyset$ ensures that any node pair cannot maintain both positive and negative relationships simultaneously. The objective of signed graph embedding is to learn a mapping function $f: V \rightarrow \mathbb{R}^k$ that projects each node $v \in V$ into a low k -dimensional

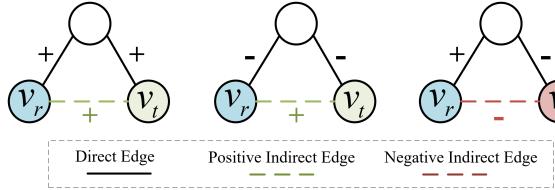


Figure 2: The signs of multi-hop connection based on balanced theory.

vector while preserving both the structural properties of the original signed graph. In other words, node pairs connected by positive edges should be embedded closely, while those connected by negative edges should be placed farther apart in the embedding space.

Balance theory [3] is a well-established standard to describe the signed relationships of unconnected node pairs. It is commonly summarized by four intuitive rules: “A friend of my friend is my friend,” “A friend of my enemy is my enemy,” “An enemy of my friend is my enemy,” and “An enemy of my enemy is my friend.” Based on these rules, the balance theory can deduce signs of the multi-hop connection. As shown in Fig. 2, given a path $P_{rt} : v_r \rightarrow v_t$ from rooted node v_r to target node v_t , the sign of the indirect relationships between v_r and v_t can be inferred by iteratively applying balance theory. Specifically, the sign of the multi-hop connection corresponds to the product of the signs of the edges along the path.

2.2 Differential Privacy

Differential Privacy (DP) [8] provides a rigorous mathematical framework for quantifying the privacy guarantees of algorithms operating on sensitive data. Informally, it bounds how much the output distribution of a mechanism can change in response to small changes in its input. When applying DP to signed graph data, the definition of adjacent databases typically considers two signed graphs, \mathcal{G} and \mathcal{G}' , which are regarded as adjacent graphs if they differ by at most one edge or one node with its associated edges.

Definition 1 (Edge (Node)-level DP [12]). *Given $\epsilon > 0$ and $\delta > 0$, a graph analysis mechanism \mathcal{M} satisfies edge- or node-level (ϵ, δ) -DP, if for any two adjacent graph datasets \mathcal{G} and \mathcal{G}' that only differ by an edge or a node with its associated edges, and for any possible algorithm output $S \subseteq \text{Range}(\mathcal{M})$, it holds that*

$$\Pr[\mathcal{M}(\mathcal{G}) \in S] \leq e^\epsilon \Pr[\mathcal{M}(\mathcal{G}') \in S] + \delta. \quad (1)$$

Here, ϵ is the privacy budget (i.e., privacy cost), where smaller values indicate stronger privacy protection but greater utility reduction. The parameter δ denotes the probability that the privacy guarantee may not hold, and is typically set to be negligible. In other words, δ allows for a negligible probability of privacy leakage, while ensuring the privacy guarantee holds with high probability.

Remark 1. Note that satisfying node-level DP is much more challenging than satisfying edge-level DP, as removing a single node may, in the worst case, remove $|V| - 1$ edges, where $|V|$ denotes the total number of nodes. Consequently, node-level DP requires injecting substantially more noise.

Two fundamental properties of DP are useful for the privacy analysis of complex algorithms: (1) **Post-Processing Property** [9]: If a mechanism $\mathcal{M}(\mathcal{G})$ satisfies (ϵ, δ) -DP, then for any function f that

indirectly queries the private dataset \mathcal{G} , the composition $f(\mathcal{M}(\mathcal{G}))$ also satisfies (ϵ, δ) -DP; (2) **Composition Property** [9]: If $\mathcal{M}(\mathcal{G})$ and $f(\mathcal{G})$ satisfy (ϵ_1, δ_1) -DP and (ϵ_2, δ_2) -DP, respectively, then the combined mechanism $\mathcal{F}(\mathcal{G}) = (\mathcal{M}(\mathcal{G}), f(\mathcal{G}))$ which outputs both results, satisfies $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$ -DP.

2.3 DPSGD

A common approach to differentially private training combines noisy stochastic gradient descent with the Moments Accountant (MA) [1]. This approach, known as DPSGD, has been widely adopted for releasing private low-dimensional representations, as MA effectively mitigates excessive privacy loss during iterative optimization. Formally, for each sample x_i in a batch of size B , we compute its gradient $\nabla \mathcal{L}_i(\theta)$, denoted as $\nabla(x_i)$ for simplicity. **Gradient sensitivity** refers to the maximum change in the output of the gradient function resulting from a change in a single sample. To control the sensitivity of $\nabla(x_i)$, the ℓ_2 norm of each gradient is clipped by a threshold C . These clipped gradients are then aggregated and perturbed with Gaussian noise $\mathcal{N}(0, \sigma^2 C^2 \mathbf{I})$ to satisfy the DP guarantee. Finally, the average noisy gradient $\tilde{\nabla}_B$ is used to update the model parameters θ . This process is given by:

$$\tilde{\nabla}_B \leftarrow \frac{1}{B} \left(\sum_{i=1}^B \text{Clip}_C(\nabla(x_i)) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I}) \right). \quad (2)$$

Here, $\text{Clip}_C(\nabla(x_i)) = \nabla(x_i) / \max(1, \frac{\|\nabla(x_i)\|_2}{C})$.

3 Problem Definition and Existing Solutions

3.1 Problem Definition

Instead of publishing a sanitized version of original node embeddings, we aim to release a privacy-preserving ASGL model trained on raw signed graph data with node-level DP guarantees, enabling data analysts to generate task-specific node embeddings.

Threat Model. We consider a black-box attack [13], where the attacker can query the trained model and observe its outputs with no access to its internal architecture or parameters. The attacker attempts to infer the presence of specific nodes or edges in the training graph solely from model outputs. This setting reflects a more practical attack surface compared to the white-box scenario [14].

Privacy Model. Signed graph data encodes both positive and negative relationships between nodes, which differs from tabular or image data. Therefore, it is necessary to adapt the standard definition of node-level DP (See Definition 1) to ensure black-box adversaries cannot determine whether a specific node and its associated signed edges are present in the training data. To this end, we define the differentially private adversarial signed graph learning model as follows.

Definition 2 (Adversarial signed graph learning model under node-level DP). *The vanilla process of graph adversarial learning is illustrated in App. A, let θ_D denote the discriminator parameters, and its r -th row element corresponds to the k -dimensional vector \mathbf{d}_{v_r} of node v_r , that is $\mathbf{d}_{v_r} \in \theta_D$. The discriminator module L_D satisfies node-level (ϵ, δ) -DP if two adjacent signed graphs \mathcal{G} and \mathcal{G}' only differ in one node with its associated signed edges, and for all possible*

$\theta_s \subseteq Range(L_D)$, we have

$$\Pr[L_D(\mathcal{G}) \in \theta_s] \leq e^\epsilon \Pr[L_D(\mathcal{G}') \in \theta_s] + \delta, \quad (3)$$

where θ_s denotes the set comprising all possible values of θ_D .

In particular, the generator G is trained based on the feedback from the differentially private discriminator D . According to the post-processing property of DP [19, 40], the generator module L_G also satisfies node-level (ϵ, δ) -DP. Leveraging the robustness to post-processing property, the privacy guarantee is preserved in the generated signed node embeddings and their downstream usage.

3.2 Existing Solutions

To our best knowledge, existing differentially private graph learning methods follow two main tracks: **gradient perturbation** and **edge perturbation**. In the first category, Yang et al. [37] introduce a privacy-preserving generative model that incorporates generative adversarial networks (GAN) or variational autoencoders (VAE) with DPSGD to protect edge privacy, while Xiang et al. [36] design a node sampling mechanism that adds Laplace noise to per-subgraph gradients, achieving node-level DP. For the edge perturbation-based methods, Lin et al. [23] use randomized response to perturb the adjacency matrix for edge-level privacy, and EDGERAND [13] perturbs the graph structure while preserving sparsity by clipping the adjacency matrix according to a privacy-calibrated graph density.

Limitation. The aforementioned solutions are not directly applicable to signed graphs. This is primarily because edge perturbation can lead to cascading errors when inferring edge signs under balance theory. Moreover, gradient perturbation often suffers from high sensitivity caused by complex node dependencies and gradient polarity reversal from edge sign flips, leading to excessive noise and degraded model utility.

4 Our Proposal: ASGL

To tackle the above limitations, we present ASGL, a DP-based adversarial signed graph learning model that integrates a constrained BFS-tree strategy to achieve favorable utility-privacy tradeoffs.

4.1 Overview

The ASGL framework, illustrated in Fig. 3, comprises three steps:

- **Private Adversarial Signed Graph Learning.** The signed graph \mathcal{G} is first split into positive and negative subgraphs, \mathcal{G}^+ and \mathcal{G}^- , based on edge signs. Subsequently, two discriminators, D^+ and D^- , sharing parameters θ_D , are trained to distinguish real from fake positive and negative edges. Guided by D^+ and D^- , two generators G^+ and G^- with shared parameters θ_G generate node embeddings that approximate the true connectivity distribution. To ensure node-level DP, we apply gradient perturbation during discriminator training instead of directly perturbing edges. This strategy mitigates cascading errors and prevents gradient polarity reversals caused by edge sign flips, thereby reducing gradient sensitivity. By the post-processing property, the generators also preserve node-level DP.
- **Optimization via Constrained BFS-tree.** To further reduce gradient sensitivity and the required noise scale, ASGL employs a constrained BFS-tree strategy. By empirically limiting

the number and length of paths, each node's receptive field is restricted, which reduces node dependency and enables gradient decoupling. This significantly lowers gradient sensitivity and enhances model utility under differential privacy constraints.

- **Privacy Accounting and Complexity Analysis.** The complete training process for ASGL is outlined in Algorithm 2 (see App. F.3). Based on this, we present a comprehensive privacy accounting and computational complexity analysis for ASGL.

4.2 Private Adversarial Signed Graph Learning

Motivated by [21, 33], a signed graph \mathcal{G} is first divided into a positive subgraph \mathcal{G}^+ and a negative subgraph \mathcal{G}^- according to edge signs. Let $N(v_r)$ be the set of neighbor nodes directly connected to node v_r . We denote the true positive and negative connectivity distributions of v_r over its neighborhood $N(v_r)$ as the conditional probabilities $p_{\text{true}}^+(\cdot|v_r)$ and $p_{\text{true}}^-(\cdot|v_r)$, which capture the preference of v_r to connect with other nodes in V . The adversarial learning for the signed graph \mathcal{G} is conducted by two adversarial learning modules:

Generators G^+ and G^- : Through optimizing the shared parameters θ_G , generators G^+ and G^- aim to approximate the underlying true connectivity distribution and generate the most likely but unconnected nodes $v_t \notin N(v_r)$ that are relevant to a given node v_r . To this end, we estimate the relevance probabilities of these *fake*² node pairs. Specifically, for the implementation of G^+ , given the fake positive node pairs $(v_r, v_t)^+$, we use the graph softmax function [33] to calculate the fake positive connectivity probability:

$$p_{\text{fake}}^+(v_t|v_r) = G^+(v_t|v_r; \theta_G) = \sigma(\mathbf{g}_{v_t}^\top \mathbf{g}_{v_r}) = \frac{1}{1 + \exp(-\mathbf{g}_{v_t}^\top \mathbf{g}_{v_r})}, \quad (4)$$

where $\mathbf{g}_{v_t}, \mathbf{g}_{v_r} \in \mathbb{R}^k$ are the k -dimensional vectors of nodes v_t and v_r , respectively, and θ_G is the union of all \mathbf{g}_v 's. The output $G^+(v_t|v_r; \theta_G)$ increases with the decrease of the distance between v_r and v_t in the embedding space of the generator G^+ . Similarly, for the generator G^- , given the fake negative node pairs $(v_r, v_t)^-$, we estimate their fake negative connectivity probability:

$$p_{\text{fake}}^-(v_t|v_r) = G^-(v_t|v_r; \theta_G) = 1 - \sigma(\mathbf{g}_{v_t}^\top \mathbf{g}_{v_r}) = \frac{\exp(-\mathbf{g}_{v_t}^\top \mathbf{g}_{v_r})}{1 + \exp(-\mathbf{g}_{v_t}^\top \mathbf{g}_{v_r})}. \quad (5)$$

Here, Eq. (5) ensures that node pairs with higher negative connectivity probabilities are mapped farther apart in the embedding space of G^- . Since generators G^+ and G^- share the parameters θ_G , they jointly learn the proximity and separation of positive and negative node pairs in a unified embedding space, respectively.

Notably, the aforementioned fake node pairs $(v_r, v_t)^+$ and $(v_r, v_t)^-$ are sampled by a breadth-first search (BFS)-tree strategy [3]. Compared to depth-first search (DFS) [32], BFS ensures more uniform exploration of neighboring nodes and can be integrated with random walk techniques [25] to optimize computational efficiency. Specifically, we perform BFS on the positive subgraph \mathcal{G}^+ to construct a BFS-tree $T_{v_r}^+$ rooted from node v_r . Then, we calculate the positive relevance probability of node v_r with its neighbors $v_k \in N(v_r)$:

$$p_{T_{v_r}^+}^+(v_k|v_r) = \frac{\exp(\mathbf{g}_{v_k}^\top \mathbf{g}_{v_r})}{\sum_{v_k \in N(v_r)} \exp(\mathbf{g}_{v_k}^\top \mathbf{g}_{v_r})}, \quad (6)$$

²The term “Fake” indicates that although a node v selected by the generator is relevant to v_r , there is no actual edge between them.

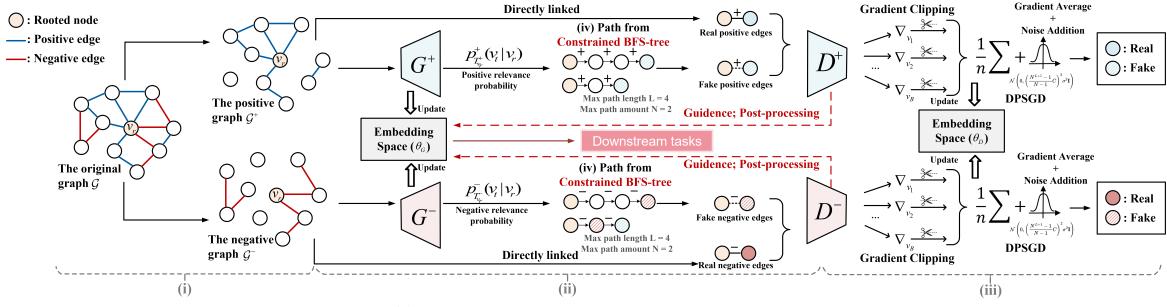


Figure 3: Overview of the ASGL framework: (i) The process decomposes a signed graph into positive and negative subgraphs, (ii) then maps node pairs into a unified embedding space while preserving signed proximity. To ensure privacy, (iii) adversarial learning module with DPSGD generates private node embeddings that approximate true connectivity without cascading errors. (iv) A constrained BFS-tree strategy manages node receptive field, reduces gradient noise, and improves model utility.

which is actually a softmax function over $\mathcal{N}(v_r)$. To further sample node pairs unconnected in $T_{v_r}^+$ as fake positive edges, we perform a random walk at $T_{v_r}^+$: Starting from the root node v_r , a path P_{rt} : $v_r \rightarrow v_t$ is built by iteratively selecting the next node based on the transition probabilities defined in Eq. (6). The resulting unconnected node pair $(v_r, v_t)^+$ is treated as a fake positive edge, and App. E provides an example of this process. Given the node pair $(v_r, v_t)^+$, the generator G^+ estimates $p_{\text{fake}}^+(v_t | v_r)$ according to Eq. (4).

Similarly, we also establish a BFS-tree $T_{v_r}^-$ rooted at node v_r in the negative subgraph G^- . To obtain the negative node pair $(v_r, v_t)^-$, we perform a random walk on $T_{v_r}^-$ according to the following transition probability (i.e., negative relevance probability):

$$p_{T_{v_r}^-}^-(v_k | v_r) = \frac{1 - \exp(\mathbf{g}_{v_k}^\top \mathbf{g}_{v_r})}{\sum_{v_k \in \mathcal{N}(v_r)} (1 - \exp(\mathbf{g}_{v_k}^\top \mathbf{g}_{v_r}))}. \quad (7)$$

In particular, the edge sign of the negative node pair $(v_r, v_t)^-$ depends on the length of the path P_{rt} : $v_r \rightarrow v_t$. According to the balance theory introduced in Section 2.1, the edge signs of multi-hop node pairs correspond to the product of the edge signs along the path. Accordingly, the rules for generating fake negative edges within P_{rt} are defined as follows: (1) If the path length of P_{rt} is odd, a node pair $(v_r, v_t)^-$ for the rooted node v_r and the last node v_t is selected as a fake negative pair; (2) If the path length of P_{rt} is even, a node pair $(v_r, v_t)^-$ for the rooted node v_r and the second last node v_t is selected as a fake negative pair. The resulting node pair $(v_r, v_t)^-$ is then used to compute $p_{\text{fake}}^-(v_t | v_r)$ according to Eq. (5).

Discriminators D^+ and D^- : This module tries to distinguish between real node pairs and fake node pairs synthesized by the generators G^+ and G^- . Accordingly, the discriminators D^+ and D^- estimate the likelihood that positive and negative edges exists between v_r and $v \in V$, respectively, denoted as:

$$D^+(v_r, v | \theta_D) = \sigma(\mathbf{d}_v^\top \mathbf{d}_{v_r}) = \frac{1}{1 + \exp(-\mathbf{d}_v^\top \mathbf{d}_{v_r})}, \quad (8)$$

$$D^-(v, v_r | \theta_D) = 1 - \sigma(\mathbf{d}_v^\top \mathbf{d}_{v_r}) = \frac{\exp(-\mathbf{d}_v^\top \mathbf{d}_{v_r})}{1 + \exp(-\mathbf{d}_v^\top \mathbf{d}_{v_r})}, \quad (9)$$

where $\mathbf{d}_v, \mathbf{d}_{v_r} \in \mathbb{R}^k$ are vectors corresponding to the v -th and v_r -th rows of shared parameters θ_D , respectively. $\sigma(\cdot)$ represents the sigmoid function of the inner product of these two vectors.

In summary, given real positive and real negative edges sampled from $p_{\text{true}}^+(\cdot | v_r)$ and $p_{\text{true}}^-(\cdot | v_r)$, along with fake positive and fake negative edges generated from generators G^+/G^- , the adversarial

learning pairs (D^+, G^+) and (D^-, G^-) , operating on the positive subgraph G^+ and the negative subgraph G^- , respectively, engage in a four-player mini-max game with the joint loss function:

$$\begin{aligned} & \min_{\theta_G} \max_{\theta_D} L(G^+, G^-, D^+, D^-) \\ &= \sum_{v_r \in V^+} \left(\left(\mathbb{E}_{v \sim p_{\text{true}}^+(\cdot | v_r)} \right) [\log D^+(v, v_r | \theta_D)] \right. \\ & \quad \left. + \left(\mathbb{E}_{v \sim G^+(\cdot | v_r; \theta_G)} \right) [\log (1 - D^+(v, v_r | \theta_D))] \right) \\ &+ \sum_{v_r \in V^-} \left(\left(\mathbb{E}_{v \sim p_{\text{true}}^-(\cdot | v_r)} \right) [\log D^-(v, v_r | \theta_D)] \right. \\ & \quad \left. + \left(\mathbb{E}_{v \sim G^-(\cdot | v_r; \theta_G)} \right) [\log (1 - D^-(v, v_r | \theta_D))] \right). \end{aligned} \quad (10)$$

Based on Eq. (10), the parameters θ_D and θ_G are updated alternately by maximizing and minimizing the joint loss function. Competition between G and D results in mutual improvement until the fake node pairs generated by G are indistinguishable from the real ones, thus approximating the true connectivity distribution. Lastly, the learned node embeddings $\mathbf{d}_v \in \theta_G$ are used in downstream tasks.

How to Achieve DP? Given real and fake positive/negative edges of the node v_i , the corresponding node embedding $\mathbf{d}_{v_i} \in \theta_D$ is updated by ascending gradients of the joint loss function in Eq. (10):

$$\frac{\partial L_D}{\partial \mathbf{d}_{v_i}} = \begin{cases} \frac{\partial \log D^+(v_i, v_j | \theta_D) / \partial \mathbf{d}_{v_i}}{\partial \mathbf{d}_{v_i}} = [1 - \sigma(\mathbf{d}_{v_j}^\top \mathbf{d}_{v_i})] \mathbf{d}_{v_j}, & \text{if } (v_i, v_j) \text{ is a real positive edge from } G^+; \\ \frac{\partial \log (1 - D^+(v_i, v_j | \theta_D)) / \partial \mathbf{d}_{v_i}}{\partial \mathbf{d}_{v_i}} = -\sigma(\mathbf{d}_{v_j}^\top \mathbf{d}_{v_i}) \mathbf{d}_{v_j}, & \text{if } (v_i, v_j) \text{ is a fake positive edge from } G^+; \\ \frac{\partial \log D^-(v_i, v_j | \theta_D) / \partial \mathbf{d}_{v_i}}{\partial \mathbf{d}_{v_i}} = -\sigma(\mathbf{d}_{v_j}^\top \mathbf{d}_{v_i}) \mathbf{d}_{v_j}, & \text{if } (v_i, v_j) \text{ is a real negative edge from } G^-; \\ \frac{\partial \log (1 - D^-(v_i, v_j | \theta_D)) / \partial \mathbf{d}_{v_i}}{\partial \mathbf{d}_{v_i}} = [1 - \sigma(\mathbf{d}_{v_j}^\top \mathbf{d}_{v_i})] \mathbf{d}_{v_j}, & \text{if } (v_i, v_j) \text{ is a fake negative edge from } G^-. \end{cases} \quad (11)$$

According to Definition 2, to achieve node-level differential privacy in adversarial signed graph learning, it is necessary to add the Gaussian noise to the sum of clipped gradients over a batch of nodes. The resulting noisy gradient $\tilde{\nabla}L_D$ is formulated as:

$$\tilde{\nabla}L_D = \frac{1}{B} \left(\sum_{v_i \in V_B} \text{Clip}_C \left(\frac{\partial L_D}{\partial \mathbf{d}_{v_i}} \right) + \mathcal{N}(0, B^2 C^2 \sigma^2 \mathbf{I}) \right), \quad (12)$$

where V_B denotes the batch set of nodes, with batch size $B = |V_B|$. C is the clipping threshold to control gradient sensitivity. The fact that the gradient sensitivity reaches BC is explained in Section 4.3.

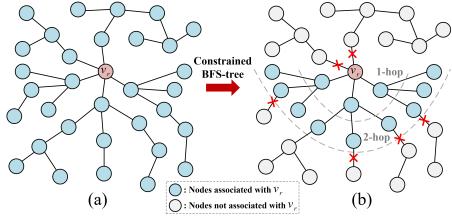


Figure 4: The receptive field of node v_r within a batch is illustrated in two cases: (a) An unconstrained BFS tree, and the receptive field size of v_r is $B = |V_B| = 34$; (b) A constrained BFS tree with path length $L = 2$, path amount $N = 3$ of each node, and the receptive field size of v_r is $\sum_{l=0}^L N^l = 13$.

Remark 2. To achieve node-level DP, we perturb discriminator gradients instead of signed edges, avoiding cascading errors and gradient polarity reversals from edge sign flips (see Eq. (10)), which reduces gradient sensitivity. Furthermore, generators also preserve DP under discriminator guidance via the post-processing property of DP.

4.3 Optimization via Constrained BFS-Tree

According to Eq. (11), in graph adversarial learning, the interdependence among samples implies that modifying a single node v_i may affect the gradients of multiple other nodes v_j within the same batch. This interdependence also exists among the fake node pairs generated along the BFS-tree paths. Consequently, in the worst-case illustrated in Fig. 4(a), all node samples within a batch may become interrelated due to the BFS-tree, resulting in the gradient sensitivity of discriminators D as high as BC . Such high sensitivity necessitates injecting substantial noise to satisfy node-level DP, hindering effective optimization and reducing model utility.

To address the aforementioned challenge, we introduce the **constrained BFS-tree strategy**: As illustrated in Algorithm 1 (see App. F.2), when performing a random walk on the BFS-tree $T_{v_r}^+$ or $T_{v_r}^-$ rooted at $v_r \in V_{tr}$ to generate multiple unique paths, we also limit both the number of sampled paths and their lengths by N and L . Following this, the training set of subgraphs S_{tr} composed of constrained paths is obtained. The rationale behind these settings is discussed below.

Theorem 1. By constraining both the number and length of paths generated via random walks on the BFS-trees to N and L , respectively, the gradient sensitivity Δ_g of the discriminator can be reduced from BC to $\frac{N^{L+1}-1}{N-1}C$.³ Thus, the noisy gradient $\tilde{\nabla}L_D$ of discriminator within a mini-batch \mathcal{B}_t is denoted as:

$$\tilde{\nabla}L_D = \frac{1}{|\mathcal{B}_t|} \left(\sum_{v \in \mathcal{B}_t} \text{Clip}_C \left(\frac{\partial L_D}{\partial d_v} \right) + \mathcal{N} \left(0, \Delta_g^2 \sigma^2 I \right) \right), \quad (13)$$

where the gradient sensitivity $\Delta_g = \frac{N^{L+1}-1}{N-1}C$.

Proof of Theorem 1. Let the sum of clipped gradients of batch subgraphs be $g_t(\mathcal{G}) = \sum_{v \in \mathcal{B}_t} \text{Clip}_C \left(\frac{\partial L_D}{\partial d_v} \right)$, where \mathcal{B}_t represents any choice of batch subgraphs from S_{tr} . Consider a node-level adjacent graph \mathcal{G}' formed by removing a node v^* with its associated edges

³Empirical results in Section 5 demonstrate that our ASGL achieves satisfactory performance even with a relatively small receptive field. Specifically, when setting $N = 3$ and $L = 4$, that is, $\frac{N^{L+1}-1}{N-1} = 121 < B = 256$, the ASGL method still performs good model utility.

from \mathcal{G} , we obtain their training sets of subgraphs S_{tr} and S'_{tr} via the SAMPLE-SUBGRAPHS method in Algorithm 1, denoted as:

$$\begin{aligned} S_{tr} &= \text{SAMPLE-SUBGRAPHS}(\mathcal{G}, V_{tr}, N, L), \\ S'_{tr} &= \text{SAMPLE-SUBGRAPHS}(\mathcal{G}', V_{tr}, N, L). \end{aligned} \quad (14)$$

The only subgraphs that differ between S_{tr} and S'_{tr} are those that involve the node v^* . Let $S(v^*)$ denote the set of such subgraphs, i.e., $S(v^*) = S_{tr} \setminus S'_{tr}$. According to Lemma 2 in App. G, the number of such subgraphs $S(v^*)$ is at most $R_{N,L}$. Thus, in any mini-batch training, the only gradient terms $\frac{\partial L_D}{\partial d_v}$ affected by the removal of node v^* are those associated with the subgraphs in $(S(v^*) \cap \mathcal{B}_t)$:

$$\begin{aligned} g_t(\mathcal{G}) - g_t(\mathcal{G}') &= \sum_{v \in \mathcal{B}_t} \text{Clip}_C \left(\frac{\partial L_D}{\partial d_v} \right) - \sum_{v' \in \mathcal{B}'_t} \text{Clip}_C \left(\frac{\partial L_D}{\partial d_{v'}} \right) \quad (15) \\ &= \sum_{v, v' \in (S(v^*) \cap \mathcal{B}_t)} [\text{Clip}_C \left(\frac{\partial L_D}{\partial d_v} \right) - \text{Clip}_C \left(\frac{\partial L_D}{\partial d_{v'}} \right)], \end{aligned}$$

where $\mathcal{B}'_t = \mathcal{B}_t \setminus (S(v^*) \cap \mathcal{B}_t)$. Since each gradient term is clipped to have an ℓ_2 -norm of at most C , it holds that:

$$\|\text{Clip}_C \left(\frac{\partial L_D}{\partial d_v} \right) - \text{Clip}_C \left(\frac{\partial L_D}{\partial d_{v'}} \right)\|_F \leq C. \quad (16)$$

In the worst case, all subgraphs in $S(v^*)$ appear in \mathcal{B}_t , so we bound the ℓ_2 -norm of the following quantity based on Lemma 3 in App. G:

$$\|g_t(\mathcal{G}) - g_t(\mathcal{G}')\|_F \leq C \cdot R_{N,L} = C \cdot \frac{N^{L+1} - 1}{N - 1}. \quad (17)$$

The same reasoning applies when \mathcal{G}' is obtained by adding a new node v^* to \mathcal{G} . Since \mathcal{G} and \mathcal{G}' are arbitrary node-level adjacent graphs, the proof is complete.

4.4 Privacy and Complexity Analysis

The complete training process for ASGL is outlined in Algorithm 2 (see App. F.3). In this section, we present a comprehensive privacy analysis and computational complexity analysis for ASGL.

Privacy Accounting. In this section, we adopt the functional perspective of Rényi Differential Privacy (RDP; see App. C) to analyze privacy budgets of ASGL, as summarized below:

Theorem 2. Given the number of training set N_{tr} , number of epochs n^{epoch} , number of discriminators' iterations n^{iter} , batch size B_d , maximum path length L , and maximum path number N , over $T = n^{\text{epoch}} n^{\text{iter}}$ iterations, Algorithm 2 satisfies node-level $(\alpha, 2T\gamma)$ -RDP, where $\gamma = \frac{1}{\alpha-1} \ln \left(\sum_{i=0}^{R_{N,L}} \beta_i \left(\exp \frac{\alpha(\alpha-1)^2}{2\sigma^2 R_{N,L}^2} \right) \right)$, $R_{N,L} = \frac{N^{L+1}-1}{N-1}$ and $\beta_i = \binom{R_{N,L}}{i} \left(\frac{N_{tr} - R_{N,L}}{B_d - i} \right) / \binom{N_{tr}}{B_d}$. Please refer to App. I for the proof.

Complexity Analysis. To analyze the time complexity of training ASGL (App. F.3), we break down the major computations. The outer loop runs for n^{epoch} epochs, and in each epoch, the discriminators D^+ and D^- are trained for n^{iter} iterations. Each iteration samples a batch of B_d real and fake edges to update θ_D , with DP cost updates incurring complexity $\mathcal{O}(B_d k \xi)$, where ξ is the sampling probability and k is the embedding dimension [2, 40]. Thus, each epoch of D^+ or D^- costs $\mathcal{O}(n^{\text{iter}} B_d k (1 + \xi))$. For the generators G^+ and G^- , each iteration samples B_g fake edges to update θ_G , resulting in per-epoch complexity $\mathcal{O}(n^{\text{iter}} B_g k)$. In total, ASGL's overall time complexity over n^{epoch} epochs is: $\mathcal{O}(2n^{\text{epoch}} n^{\text{iter}} (B_d + B_g)(1 + \xi)k)$.

Table 1: Overview of the datasets

Datasets	Nodes	Edges	Positive Edges	Negative Edges
Bitcoin-Alpha	3,783	14,081	12,769 (90.7%)	1,312 (9.3%)
Bitcoin-OTC	5,881	21,434	18,281 (85.3%)	3,153 (14.7%)
WikiRfA	11,258	185,627	144,451 (77.8%)	41,176 (22.2%)
Slashdot	13,182	36,338	30,914 (85.1%)	5,424 (14.9%)
Epinions	131,828	841,372	717,690 (85.3%)	123,682 (14.7%)

This complexity is linear in the number of iterations and batch size, demonstrating the scalability of ASGL for large-scale graphs.

5 Experiments

In this section, some experiments are designed to answer the following questions: (1) How do key parameters affect the performance of ASGL (See Section 5.2)? (2) How much does the privacy budget affect the performance of ASGL and other private signed graph learning models in edge sign prediction (See Section 5.3)? (3) How much does the privacy budget affect the performance of ASGL and other baselines in node clustering (See Section 5.4)? (4) How resilient is ASGL to defense link stealing attacks (See Section 5.5)?

5.1 Experimental Settings

Datasets. To comprehensive evaluate our ASGL method, we conduct extensive experiments on five real-world datasets, namely Bitcoin-Alpha⁴, Bitcoin-OTC⁴, WikiRfA⁴, Slashdot⁵ and Epinions⁴. These datasets are regarded as undirected signed graphs, with their detailed statistics summarized in Table 1 and App. J.1.

Competitive Methods. To the best of our knowledge, this work is the first to address the problem of differentially private signed graph learning while aiming to preserve model utility. Due to the absence of prior studies in this area, we construct baselines by integrating four state-of-the-art signed graph learning methods—SGCN [7], SiGAT [17], LSNE [31], and SDGNN [18]—with the DPSGD mechanism. Since these models primarily leverage structural information, we further include the private graph learning method GAP [28], using Truncated SVD-generated spectral features [7] as input to ensure a fair comparison involving node features.

Evaluation Metrics. For edge sign prediction tasks, we follow the evaluation procedures in [17, 18, 21]. Specifically, we first generate embedding vectors for all nodes in the training set using each comparative method. Then, we train a logistic regression classifier using the concatenated embeddings of node pairs as input features. Finally, we use the trained classifier to predict edge signs in the test set for each method. Considering the class imbalance between positive and negative edges (see Table 1), we adopt the *area under curve* (AUC) as the evaluation metric to ensure a fair comparison.

For node clustering, to fairly evaluate the clustering effect of node embeddings, we compute the average cosine distance for both positive and negative node pairs: $CD^+ = \sum_{(v_i, v_j) \in E^+} \text{Cos}(\mathbf{Z}_i, \mathbf{Z}_j) / |E^+|$ and $CD^- = \sum_{(v_n, v_m) \in E^-} \text{Cos}(\mathbf{Z}_n, \mathbf{Z}_m) / |E^-|$, where \mathbf{Z}_i is the node embedding generated by each comparative method, and $\text{Cos}(\cdot)$ represents the cosine distance between node embeddings. Then we

⁴Collected in <https://snap.stanford.edu/data>.

⁵Collected in <https://www.aminer.cn>.

propose the *symmetric separation index* (SSI) to measure the clustering degree between the embeddings of positive and negative node pairs in the test set, denoted as $SSI = 1/(|CD^+ - 1| + |CD^- + 1|)$. A higher SSI indicates better structural proximity, with positive node pairs more tightly clustered and negative pairs more clearly separated in the unified embedding space.

Parameter Settings. For both edge sign prediction and node clustering tasks, we set the dimensionality of all node embeddings, d_v and g_v , to 128, following standard practice in prior work [20, 21]. ASGL adopts DPSGD-based optimization, where the total number of training epochs is determined by the moments accountant (MA) [8], which offers tighter privacy tracking across multiple iterations. We set the iteration number n^{iter} to 10 for Bitcoin-Alpha and Bitcoin-OTC, 15 for WikiRfA and Slashdot, and 20 for Epinions. Since all comparative methods are trained using DPSGD, their number of training epochs depends on the privacy budget. As discussed in Section 5.2, the maximum path number N and path length L are varied to analyze their impact on ASGL’s utility. For privacy parameters, we follow [1, 29, 40] by fixing $\delta = 10^{-5}$ and $C = 1$, and vary the privacy budget $\epsilon \in \{1, 2, \dots, 6\}$ to evaluate utility under different privacy levels. To ensure fair comparison, we modify the official GitHub implementations of all baselines and adopt the best hyperparameter settings reported in their original papers. To minimize random errors, each experiment is repeated five times.

5.2 Impact of Key Parameters

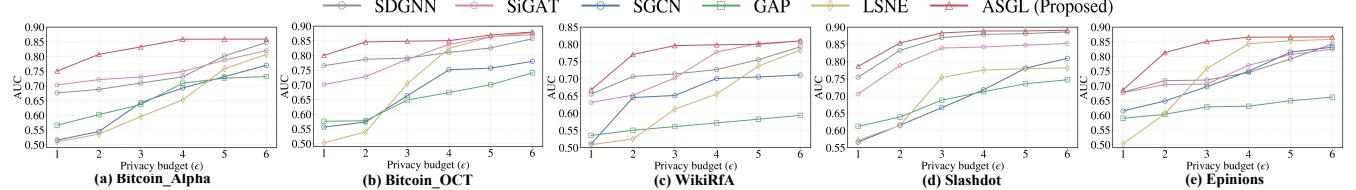
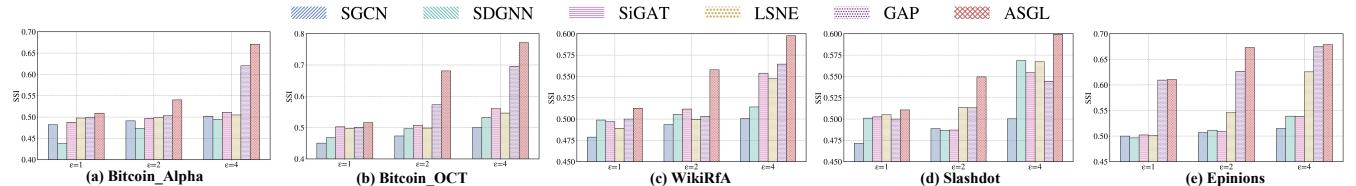
In this section, we perform experiments on two datasets by varying the maximum number N and the maximum length L of paths in the BFS-trees, providing a rationale for parameter selection.

5.2.1 The effect of the parameter N . As discussed in Section 4.3, the greater the number of neighbors a rooted node has, the more paths can be obtained through random walks. Therefore, the maximum number of paths N also depends on the node degrees. As shown in Fig. 8 (see App. J.2), for the Bitcoin-Alpha and Slashdot datasets, most nodes in signed graphs have degrees below 3. In addition, we investigate the impact of N by varying its value within $\{2, 3, 4, 5, 6\}$. As shown by the average AUC results in Table 2, the proposed ASGL method achieves optimal edge prediction performance at $N = 3$ for Bitcoin-Alpha and $N = 4$ for Slashdot. Considering both gradient sensitivity and computational efficiency, we adopt $N = 3$ for subsequent experiments.

Table 2: Summary of average AUC with different maximum path counts N under $\epsilon = 3$ and $L = 3$. (BOLD: Best)

Dataset	$N = 2$	$N = 3$	$N = 4$	$N = 5$	$N = 6$
Bitcoin-Alpha	0.8025	0.8562	0.8557	0.8498	0.8553
Slashdot	0.7723	0.8823	0.8888	0.8871	0.8881

5.2.2 The effect of the parameter L . In this experiment, we evaluate the impact of the path length L on the utility of ASGL by varying its value. As shown in Table 3, ASGL achieves the best performance on both datasets when $L = 4$. This result is closely aligned with the structural characteristics of the signed graphs: As summarized in Fig. 9 (see App. J.2), most node pairs in these datasets exhibit maximum path lengths of 3 or 4. Therefore, in subsequent experiments, we set $L = 4$, as it adequately covers the receptive field of most nodes.

Figure 5: AUC vs. Privacy cost (ϵ) of private signed graph learning methods in edge sign prediction.Figure 6: Symmetric separation index (SSI) vs. Privacy cost (ϵ) of private signed graph learning methods in node clustering.Table 3: Summary of average AUC with different path lengths L under $\epsilon = 3$ and $N = 3$. (BOLD: Best)

Dataset	$L = 1$	$L = 2$	$L = 4$	$L = 6$	$L = 8$
Bitcoin-Alpha	0.7409	0.8443	0.8587	0.8545	0.8516
Slashdot	0.7629	0.8290	0.8833	0.8809	0.8807

5.3 Impact of Privacy Budget on Edge Sign Prediction

To evaluate the effectiveness of different private graph learning methods on edge sign prediction, we compare their AUC scores under privacy budgets ϵ ranging from 1 to 6, as shown in Fig. 5 and Table 6 (see App. J.3). The proposed ASGL consistently outperforms all baselines across all privacy levels and datasets, owing to its ability to generate node embeddings that preserve connectivity distributions while satisfying DP guarantees. Although SDGNN achieves sub-optimal performance, it exhibits a noticeable gap from ASGL under limited privacy budgets ($\epsilon < 4$). SiGAT, SGCN, and LSNE employ the moments accountant (MA) to mitigate excessive privacy budget consumption, yet still suffer from poor convergence and degraded utility under limited privacy budgets. GAP adopts aggregation perturbation to ensure node-level DP, but its performance is limited due to noisy neighborhood information, hindering its ability to capture structural information for edge prediction tasks.

5.4 Impact of Privacy Budget on Node Cluster

To further examine the capability of ASGL in preserving signed node proximity, we conduct a fair comparison across multiple private graph learning methods using the SSI metric. As shown in Fig. 6 and Table 7 (see App. J.4), ASGL consistently outperforms all baselines across different datasets and privacy budgets, demonstrating that ASGL is capable of generating node embeddings that effectively preserve signed node proximity. Notably, GAP achieves the second-best clustering performance on most datasets (excluding Slashdot), benefiting from its ability to leverage node features for clustering nodes. Nevertheless, to guarantee node-level DP, GAP needs to repeatedly query sensitive graph information in every training iteration, resulting in significantly higher privacy costs.

Table 4: The average AUC of LSA on different comparisons and datasets. (BOLD: Best resilience against LSA)

ϵ	Dataset	GAP	SGCN	LSNE	SiGAT	SDGNN	ASGL
1	Bitcoin-Alpha	0.5072	0.7091	0.5079	0.5145	0.5404	0.5053
	Bitcoin-OTC	0.5081	0.7118	0.5119	0.5409	0.5660	0.5466
	Slashdot	0.5538	0.8232	0.5551	0.5609	0.5460	0.5325
	WikiRFA	0.5148	0.5424	0.5427	0.5293	0.5470	0.5302
	Epinions	0.7877	0.6329	0.5114	0.5129	0.5188	0.5092
3	Bitcoin-Alpha	0.5547	0.7514	0.5533	0.5542	0.5598	0.5430
	Bitcoin-OTC	0.5655	0.7273	0.5684	0.5734	0.5765	0.5612
	Slashdot	0.5742	0.8394	0.6267	0.5730	0.6464	0.5634
	WikiRFA	0.5276	0.5466	0.5542	0.5696	0.5772	0.5624
	Epinions	0.7981	0.6456	0.5588	0.5629	0.5665	0.5542

5.5 Resilience Against Link Stealing Attack

To assess the effectiveness of ASGL in preserving the privacy of edge information, we perform link stealing attacks (LSA) across all datasets and compare the resilience of all methods to such attacks in edge sign prediction tasks. The LSA setup is detailed in App. J.5. Attack performance is measured by the AUC score, averaged over five independent runs. Table 4 summarizes the effectiveness of LSA on various trained target models and datasets. It can be observed that as the privacy budget ϵ increases, the average AUC of LSA consistently improves, indicating the reduced privacy protection of target models and an increased success rate of the attack. Overall, the average AUC of the attack is close to 0.50 in most cases, indicating the unsuccessful edge inference and the robustness of DP against such an attack. When $\epsilon = 3$, ASGL demonstrates stronger resistance to LSA across most datasets, with AUC values consistently below 0.57. This suggests that ASGL offers defense performance comparable to other differentially private graph learning methods.

6 Related Work

Signed graph learning. In recent years, deep learning approaches have been increasingly adopted for signed graph learning. For example, SiNE [34] extracts signed structural information based on balance theory and designs an objective function to learn signed node proximity. Furthermore, the GNN model [7] and its variants [17, 18]

are used to learn signed relationships between nodes in multi-hop neighborhoods. However, these GNNs-based methods depend on the message-passing mechanism, which is sensitive to noisy interactions between nodes [41]. To address this issue, Lee *et al.* [21] extends the adversarial framework to signed graphs by generating both positive and negative node embeddings. Still, these signed graph learning models are vulnerable to user-linkage attacks.

Private graph learning. Recent works have increasingly focused on developing DP methods to address privacy leakage in GNNs. For instance, Daigavane *et al.* [10] propose a DP-GNN method based on gradient perturbation. However, this method fails to balance utility and privacy due to excessive noise. Furthermore, GAP [28] and DPRA [38] are proposed to ensure the privacy of sensitive node embeddings by perturbing node aggregations. Despite their success in node classification, the private node information is repeatedly queried in the training process of GAP, which consumes more privacy budgets to implement DPSGD. DPRA is not well-suited for signed graph embedding learning, as its edge perturbation strategy introduces cascading errors under balance theory.

7 Conclusion

In this paper, we propose ASGL that achieves strong model utility while providing node-level DP guarantees. To address the cascading error and gradient polarity reversals from edge sign flips, ASGL separately processes positive and negative subgraphs within a shared embedding space using a DPSGD-based adversarial mechanism to learn high-quality node embeddings. To further reduce gradient sensitivity, we introduce a constrained BFS-tree strategy that limits node receptive fields and enables gradient decoupling. This effectively reduces the required noise scale and enhances model performance. Extensive experiments demonstrate that ASGL achieves a favorable privacy-utility trade-off. Our future work is to extend the ASGL framework by considering edge directions and weights.

References

- [1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 308–318.
- [2] Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. 2023. Differentially private optimization on large model at small cost. In *International Conference on Machine Learning*. 3192–3218.
- [3] Dorwin Cartwright and Frank Harary. 1956. Structural balance: a generalization of Heider's theory. *Psychological review* 63, 5 (1956), 277.
- [4] Yiqi Chen, Tieyun Qian, Huan Liu, and Ke Sun. 2018. Bridge Enhanced Signed Directed Network Embedding. In *Proceedings of the 27th ACM international conference on information and knowledge management*. 773–782.
- [5] Ameya Daigavane, Gagan Madan, Aditya Sinha, Abhradeep Guha Thakurta, Gaurav Aggarwal, and Prateek Jain. 2022. Node-Level Differentially Private Graph Neural Networks. In *ICLR 2022 Workshop on PAIR'2Struct*.
- [6] Tyler Derr. 2020. Network analysis with negative links. In *Proceedings of the 13th international conference on web search and data mining*. 917–918.
- [7] Tyler Derr, Yao Ma, and Jiliang Tang. 2018. Signed graph convolutional networks. In *2018 IEEE International Conference on Data Mining (ICDM)*. 929–934.
- [8] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*. 265–284.
- [9] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 3–4 (2014), 211–407.
- [10] Vitaly Feldman, Ilya Mironov, Kunal Talwar, and Abhradeep Thakurta. 2018. Privacy amplification by iteration. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 521–532.
- [11] Catherine Forbes, Merran Evans, Nicholas Hastings, and Brian Peacock. 2011. *Statistical distributions*. John Wiley & Sons.
- [12] Michael Hay, Chao Li, Jerome Miklau, and David Jensen. 2009. Accurate estimation of the degree distribution of private networks. In *2009 Ninth IEEE International Conference on Data Mining*. 169–178.
- [13] Xinlei He, Jinyuan Jia, Michael Backes, Neil Zhenqiang Gong, and Yang Zhang. 2021. Stealing links from graph neural networks. In *30th USENIX security symposium (USENIX security 21)*. 2669–2686.
- [14] Zecheng He, Tianwei Zhang, and Ruby B Lee. 2019. Model inversion attacks against collaborative inference. In *Proceedings of the 35th Annual Computer Security Applications Conference*. 148–162.
- [15] Renxuan Hou, Qingqing Ye, Xun Ran, Sen Zhang, and Haibo Hu. 2025. PrivIM: Differentially Private Graph Neural Networks for Influence Maximization. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*. 3467–3479.
- [16] Junjie Huang, Huawei Shen, Qi Cao, Shuchang Tao, and Xueqi Cheng. 2021. Signed bipartite graph neural networks. In *Proceedings of the 30th ACM international conference on information & knowledge management*. 740–749.
- [17] Junjie Huang, Huawei Shen, Liang Hou, and Xueqi Cheng. 2019. Signed graph attention networks. In *International Conference on Artificial Neural Networks*. 566–577.
- [18] Junjie Huang, Huawei Shen, Liang Hou, and Xueqi Cheng. 2021. SDGNN: Learning node representation for signed directed networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 196–203.
- [19] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. 2018. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*.
- [20] Junghwan Kim, Haekyu Park, Ji-Eun Lee, and U Kang. 2018. Side: representation learning in signed directed networks. In *Proceedings of the 2018 world wide web conference*. 509–518.
- [21] Yeon-Chang Lee, Nayoun Seo, Kyungsik Han, and Sang-Wook Kim. 2020. Asine: Adversarial signed network embedding. In *Proceedings of the 43rd international ACM sigir conference on research and development in information retrieval*. 609–618.
- [22] Rong Li, Long Xu, Songbai Liu, Junkai Ji, Lingjie Li, Qiuzhen Lin, and Lijia Ma. 2025. Structure Balance and Gradient Matching-Based Signed Graph Condensation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 12121–12129.
- [23] Wanyu Lin, Baochun Li, and Cong Wang. 2022. Towards private learning on decentralized graphs with local differential privacy. *IEEE Transactions on Information Forensics and Security* 17 (2022), 2936–2946.
- [24] Ilya Mironov. 2017. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium*. 263–275.
- [25] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 701–710.
- [26] John Pougué-Biyong, Akshay Gupta, Aria Haghighi, and Ahmed El-Kishky. 2023. Learning stance embeddings from signed social graphs. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 177–185.
- [27] Xun Ran, Qingqing Ye, Haibo Hu, Xin Huang, Jianliang Xu, and Jie Fu. 2024. Differentially private graph neural networks for link prediction. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. 1632–1644.
- [28] Sina Sajadmanesh, Ali Shahin Shamsabadi, Aurélien Bellet, and Daniel Gatica-Perez. 2023. GAP: Differentially Private Graph Neural Networks with Aggregation Perturbation. In *32nd USENIX Security Symposium (USENIX Security 23)*. 3223–3240.
- [29] Reza Shokri and Vitaly Shmatikov. 2015. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*. 1310–1321.
- [30] Lin Shu, Erxin Du, Yaomin Chang, Chuan Chen, Zibin Zheng, Xingxing Xing, and Shaofeng Shen. 2021. Sgcl: Contrastive representation learning for signed graphs. In *Proceedings of the 30th ACM international conference on information & knowledge management*. 1671–1680.
- [31] Wenzhuo Song, Shengsheng Wang, Bo Yang, You Lu, Xuehua Zhao, and Xueyan Liu. 2018. Learning node and edge embeddings for signed networks. *Neurocomputing* 319 (2018), 42–54.
- [32] Robert Tarjan. 1972. Depth-first search and linear graph algorithms. *SIAM journal on computing* 1, 2 (1972), 146–160.
- [33] Hongwei Wang, Jia Wang, Jialin Wang, Miao Zhao, Weinan Zhang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018. Graphgan: Graph representation learning with generative adversarial nets. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.
- [34] Suhang Wang, Jiliang Tang, Charu Aggarwal, Yi Chang, and Huan Liu. 2017. Signed network embedding in social media. In *Proceedings of the 2017 SIAM international conference on data mining*. SIAM, 327–335.
- [35] Fan Wu, Yunhui Long, Ce Zhang, and Bo Li. 2022. Linkteller: Recovering private edges from graph neural networks via influence analysis. In *2022 ieee symposium on security and privacy (sp)*. 2005–2024.
- [36] Zihang Xiang, Tianhao Wang, and Di Wang. 2024. Preserving node-level privacy in graph neural networks. In *2024 IEEE Symposium on Security and Privacy (SP)*.

- IEEE, 4714–4732.
- [37] Carl Yang, Haonan Wang, Ke Zhang, Liang Chen, and Lichao Sun. 2021. Secure Deep Graph Generation with Link Differential Privacy. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conferences on Artificial Intelligence Organization, 3271–3278.
 - [38] Qiuchen Zhang, Hong kyu Lee, Jing Ma, Jian Lou, Carl Yang, and Li Xiong. 2024. DPGR: Decoupled graph neural networks with node-level differential privacy. In *Proceedings of the ACM Web Conference 2024*, 1170–1181.
 - [39] Sen Zhang, Haibo Hu, Qingqing Ye, and Jianliang Xu. 2025. PrivDPR: Synthetic Graph Publishing with Deep PageRank under Differential Privacy. *arXiv preprint arXiv:2501.02354* (2025).
 - [40] Sen Zhang, Qingqing Ye, Haibo Hu, and Jianliang Xu. 2025. AdvSGM: Differentially Private Graph Learning via Adversarial Skip-gram Model. In *2025 IEEE 41th International Conference on Data Engineering (ICDE)*, 3494–3507.
 - [41] Zeyu Zhang, Jiamou Liu, Kaiqi Zhao, Song Yang, Xianda Zheng, and Yifei Wang. 2023. Contrastive learning for signed bipartite graphs. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1629–1638.
 - [42] Zaixi Zhang, Qi Liu, Zhenya Huang, Hao Wang, Chengqiang Lu, Chuanren Liu, and Enhong Chen. 2021. GraphMI: Extracting Private Graph Data from Graph Neural Networks. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 3749–3755.

A Adversarial Learning on Graph

The adversarial learning model for graph embedding [33] is illustrated as follows. Let $\mathcal{N}(v_r)$ be the node set directly connected to v_r . We denote the underlying true connectivity distribution of node v_r as the conditional probability $p(v|v_r)$, which captures the preference of v_r to connect with other nodes $v \in V$. In other words, the neighbor set $\mathcal{N}(v_r)$ can be interpreted as a set of observed nodes drawn from $p(v|v_r)$. The adversarial learning for the graph G is conducted by the following two modules:

Generator G : Through optimizing the generator parameters θ_G , this module aims to approximate the underlying true connectivity distribution and generate (or select) the most likely nodes $v \in V$ that are relevant to v_r . Specifically, the *fake*⁶ (i.e., estimated) connectivity distribution of node v_r is calculated as:

$$p'(v|v_r) = G(v|v_r; \theta_G) = \frac{\exp(\mathbf{g}_v^\top \mathbf{g}_{v_r})}{\sum_{v \neq v_r} \exp(\mathbf{g}_v^\top \mathbf{g}_{v_r})}, \quad (18)$$

where $\mathbf{g}_v, \mathbf{g}_{v_r} \in \mathbb{R}^k$ are the k -dimensional vectors of nodes v and v_r , respectively, and θ_G is the union of all \mathbf{g}_v 's. To update θ_G in each iteration, a set of node pairs (v, v_r) , not necessarily directly connected, is sampled according to $p'(v|v_r)$. The key purpose of generator G is to deceive the discriminator D , and thus its loss function L_G is determined as follows:

$$L_G = \min \sum_{r=1}^{|V|} \mathbb{E}_{v \sim G(\cdot|v_r; \theta_G)} [\log(1 - D(v_r, v | \theta_D))], \quad (19)$$

where the discriminant function $D(\cdot)$ estimates the probability that a given node pairs (v, v_r) are considered *real*, i.e., directly connected.

Discriminator D : This module tries to distinguish between real node pairs and fake node pairs synthesized by the generator G . Accordingly, the discriminator estimates the probability that an edge exists between v_r and v , denoted as:

$$D(v_r, v | \theta_D) = \sigma(\mathbf{d}_v^\top \mathbf{d}_{v_r}) = \frac{1}{1 + \exp(-\mathbf{d}_v^\top \mathbf{d}_{v_r})}, \quad (20)$$

⁶The term “Fake” indicates that although a node v selected by the generator is relevant to v_r , there is no actual edge between them.

where $\mathbf{d}_v, \mathbf{d}_{v_r} \in \mathbb{R}^k$ are the k -dimensional vectors corresponding to the v -th and v_r -th rows of discriminator parameters θ_D , respectively. $\sigma(\cdot)$ represents the sigmoid function of the inner product of these two vectors. Given the sets of real and fake node pairs, the loss function of D can be derived as:

$$\begin{aligned} L_D = \max \sum_{r=1}^{|V|} & (\mathbb{E}_{v \sim p(\cdot|v_r)} [\log D(v, v_r | \theta_D)]) \\ & + \mathbb{E}_{v \sim G(\cdot|v_r; \theta_G)} [\log(1 - D(v_r, v | \theta_D))]. \end{aligned} \quad (21)$$

In summary, the generator G and discriminator D operate as two adversarial components: the generator G aims to fit the true connectivity distribution $p(v|v_r)$, generating candidate nodes v that resemble the real neighbors of v_r to deceive the discriminator D . In contrast, the discriminator D seeks to distinguish whether a given node is a true neighbor of v_r or one generated by G . Formally, D and G are engaged in a two-player minimax game with the following loss function:

$$\begin{aligned} \min_{\theta_G} \max_{\theta_D} L(G, D) = \sum_{r=1}^{|V|} & (\mathbb{E}_{v \sim p(\cdot|v_r)} [\log D(v, v_r | \theta_D)]) \\ & + \mathbb{E}_{v \sim G(\cdot|v_r; \theta_G)} [\log(1 - D(v_r, v | \theta_D))]. \end{aligned} \quad (22)$$

Based on Eq. (22), the parameters θ_D and θ_G are updated by alternately maximizing and minimizing the loss function $L(G, D)$. Competition between G and D results in mutual improvement until G becomes indistinguishable from the true connectivity distribution.

B Notation Introduction

The frequently used notations are summarized in Table 5.

Table 5: Notation Summary

Symbol	Description
$\mathcal{G}, \mathcal{G}^+, \mathcal{G}^-$	Signed graph, positive subgraph, negative subgraph
V, E^+, E^-	Node set, negative and positive edge sets
$\mathcal{N}(v_r)$	Neighbor node set of node v_r
θ_D	Shared parameters of discriminators D^+ and D^-
θ_G	Shared parameters of generators G^+ and G^-
\mathbf{d}_{v_r}	Node embedding for node v_r of Discriminators
\mathbf{g}_{v_r}	Node embedding for node v_r of Generators
N, L	Maximum number and length of generated path
ϵ, δ	Privacy parameters
$N(0, \sigma^2)$	Gaussian distribution with standard deviation σ^2
P_{rt}	A path from rooted node v_r to target node v_t
$T_{v_r}^+, T_{v_r}^-$	Positive and negative BFS-trees rooted from v_r
$p_{\text{true}}^+(\cdot v_r)$	Positive connectivity distributions of $(v_r, v) \in E^+$
$p_{\text{true}}^-(\cdot v_r)$	Negative connectivity distributions of $(v_r, v) \in E^-$
$p_{T_{v_r}^+}^+(v v_r)$	Positive relevance probability between v_r and v
$p_{T_{v_r}^-}^-(v v_r)$	Negative relevance probability between v_r and v

C Rényi Differential Privacy

Since standard DP can be overly strict for deep learning, we follow prior work [27, 39] and adopt an alternative definition—Rényi Differential Privacy (RDP) [24]. RDP offers tighter and more efficient composition bounds, enabling more accurate estimation of cumulative privacy cost over multiple queries on graphs.

Algorithm 1: SAMPLE-SUBGRAPHS by Constrained BFS-trees

Input: Graph $\mathcal{G} = \{\mathcal{G}^+, \mathcal{G}^-\}$; The training set of nodes V_{tr} ; The maximum path length L ; The maximum path number N .
Output: The training set of subgraphs S_{tr} .

```

1 for  $v_r \in V_{tr}$  do
2   Construct BFS-trees  $T_{v_r}^+$  (or  $T_{v_r}^-$ ) rooted from the node  $v_r$  on  $\mathcal{G}^+$  (or  $\mathcal{G}^-$ );
3   for  $n = 0; n < N$  do
4     Based on the positive and negative relevance probability in Eqs. (6) and (7), conduct the random walk at  $T_{v_r}^+$  (or  $T_{v_r}^-$ ) to form a path  $P_{rt}^{(n)+}$  (or  $P_{rt}^{(n)-}$ ) of length  $L$ ;
5     Add all nodes  $v$  (excluding those in  $\mathcal{N}(v_r)$ ) along the path  $P_{rt}^{(n)+}$  (or  $P_{rt}^{(n)-}$ ) as a fake edge  $(v_r, v)$  to the corresponding subgraph set  $S_{tr}^+$  (or  $S_{tr}^-$ );
6     Drop  $P_{rt}^{(n)+}$  (or  $P_{rt}^{(n)-}$ ) from  $T_{v_r}^+$  (or  $T_{v_r}^-$ ).
7   end
8 end
9 Return  $S_{tr} = \{S_{tr}^+, S_{tr}^-\}$ ;

```

Definition 3 (Rényi Differential Privacy [24]). *The Rényi divergence quantifies the similarity between output distributions of a mechanism and is defined as:*

$$D_\alpha(P\|Q) = \frac{1}{\alpha-1} \log \left(\sum_x P(x)^\alpha Q(x)^{1-\alpha} \right), \quad (23)$$

where $P(x)$ and $Q(x)$ are probability distributions over the output space. $\alpha > 1$ denotes the order of the divergence, and its choice allows for different levels of sensitivity to the output distribution. Accordingly, an algorithm \mathcal{M} satisfies (α, ϵ) -RDP if, for any two adjacent graphs \mathcal{G} and \mathcal{G}' , the following condition holds $D_\alpha(\mathcal{M}(\mathcal{G})\|\mathcal{M}(\mathcal{G}')) \leq \epsilon$.

Since RDP is an extension of DP, it can be converted into (ϵ, δ) -DP based on Proposition 3 in [24], as outlined below.

Lemma 1 (Conversion from RDP to DP [24]). *If a mechanism \mathcal{M} satisfies (α, ϵ) -RDP, it also satisfies $(\epsilon + \log(1/\delta)/(\alpha-1), \delta)$ -DP for any $\delta \in (0, 1)$.*

D Gaussian Mechanism

Let f be a function that maps a graph \mathcal{G} to k -dimensional node vectors $Z \in \mathbb{R}^{|V| \times k}$. To ensure the RDP guarantees of f , it is common to inject Gaussian noise into its output [24]. The noise scale depends on the sensitivity of f , defined as $\Delta_f = \max_{\mathcal{G}, \mathcal{G}'} \|f(\mathcal{G}) - f(\mathcal{G}')\|_2$. Specifically, the privatized mechanism is defined as $\mathcal{M}(\mathcal{G}) = f(\mathcal{G}) + \mathcal{N}(0, \sigma^2 I)$, where $\mathcal{N}(0, \sigma^2 I)$ is the Gaussian distribution with zero mean and standard deviation σ^2 . This results in an (α, ϵ) -RDP mechanism \mathcal{M} for all $\alpha > 1$ with $\epsilon = \alpha \Delta_f^2 / 2\sigma^2$.

E BFS-tree Strategy

Fig. 7 provides an illustrative example of the BFS-tree strategy: Let v_{r_0} be the rooted node. We first compute the transition probabilities between v_{r_0} and its neighbors $\mathcal{N}(v_{r_0})$. The next node v_{r_1} is then sampled as the first step of the walk, in proportion to these transition probabilities. Similarly, the next node v_{r_2} is selected based on the transition probabilities between v_{r_1} and its neighbors $\mathcal{N}(v_{r_1})$. The random walk continues until it reaches the terminal node v_{r_n} , and

unconnected node pairs $(v_{r_0}, v_{r_k})^+$ for $k = 2, 3, \dots, n$ are regarded as fake positive edges.

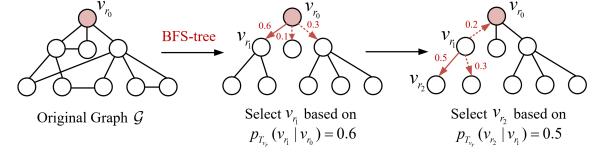


Figure 7: Random-walk-based edge generation for generator G^+ or G^- . Red digits denote the transition probabilities (Eqs. (6) and (7)), and red arrows indicate the walk directions.

F Details of Algorithm

F.1 The Parameter Update of Generators

Given fake positive/negative edges (v_r, v_t) from G^+ / G^- , the gradient of joint loss function (Eq. (10)) with respect to θ_G is derived via the policy gradient [33]:

$$\nabla L_G = \begin{cases} \sum_{r=1}^{|V^+|} [\nabla_{\theta_G} \log G^+(v_t | v_r; \theta_G) \log (1 - D^+(v_t, v_r))], \\ \text{if } (v_r, v_t) \text{ is a fake positive edge;} \\ \sum_{r=1}^{|V^-|} \nabla_{\theta_G} \log G^-(v_t | v_r; \theta_G) \log (1 - D^-(v_t, v_r)), \\ \text{if } (v_r, v_t) \text{ is a fake negative edge.} \end{cases} \quad (24)$$

F.2 SAMPLE-SUBGRAPHS by Constrained BFS-trees

As shown in Algorithm 1, during the random walk on the BFS tree $T_{v_r}^+$ or $T_{v_r}^-$ rooted at $v_r \in V_{tr}$, we generate multiple unique paths while constraining their number and length by parameters N and L , respectively. This process yields a training subgraph set S_{tr} composed of constrained paths.

F.3 The training of ASGL

The training process of ASGL is outlined in Algorithm 2 and consists of the following main steps:

(1) **Signed graph decomposition and subgraph sampling:** Given an input signed graph \mathcal{G} , we first divide it into a positive subgraph \mathcal{G}^+ and a negative subgraph \mathcal{G}^- based on edge signs. Then, for each node $v_r \in V_{tr}$, constrained BFS trees are constructed from \mathcal{G}^+ and \mathcal{G}^- , respectively, to generate a set of training subgraphs $S_{tr} = \{S_{tr}^+, S_{tr}^-\}$ by limiting the maximum number of paths N and the maximum path length L . These subgraphs are used to sample fake edges for adversarial training.

(2) **Edge sampling for adversarial learning:** For each node v_r , we sample real edges from \mathcal{G}^+ and \mathcal{G}^- , and fake edges from S_{tr}^+ and S_{tr}^- . These edges are organized into four sets:

- E_D^+ : real and fake positive edges for training D^+ .
- E_G^+ : fake positive edges for training G^+ .
- E_D^- : real and fake negative edges for training D^- .
- E_G^- : fake negative edges for training G^- .

(3) **Adversarial training with DPSGD:** The training is performed over n^{epoch} epochs. In each epoch:

- **Discriminator training:** For each discriminator D^+ and D^- , we perform n^{iter} iterations. In each iteration, a batch of B_d real and fake edges is sampled. The discriminator parameters θ_D are updated using gradient descent with noise addition according

Algorithm 2: ASGL Algorithm

Input: Graph \mathcal{G} ; Training set of nodes V_{tr} ; Maximum path length L ; Maximum path number N ; Batch-size B_d and B_g of sampled edges in discriminators and generators; Number of epochs n^{epoch} ; Number of iterations for generators and discriminators per epoch n^{iter} ; Privacy parameters δ, ϵ, σ .

Output: Privacy-preserving node embedding $\mathbf{g}_v \in \theta_G$ for downstream tasks.

```

1 According to edge signs, divide  $\mathcal{G}$  into  $\mathcal{G}^+$  and  $\mathcal{G}^-$ ;
2 Generate the training subgraph set  $S_{tr} = \{S_{tr}^+, S_{tr}^-\}$  based on SAMPLE-SUBGRAPHS( $\mathcal{G}, V_{tr}, N, L$ ) in Algorithm 1;
3 for  $v_r \in V_{tr}$  do
4   Sample all real positive edges  $(v_r, v_t)^+$  from  $\mathcal{G}^+$ ;
5   Sample all fake positive edges  $(v_r, v'_t)^+$  from  $S_{tr}^+$ ;
6   Sample all real negative edges  $(v_r, v_t)^-$  from  $\mathcal{G}^-$ ;
7   Sample all fake negative edges  $(v_r, v'_t)^-$  from  $S_{tr}^-$ ;
8    $E_D^+.add((v_r, v_t)^+, (v_r, v'_t)^+)$ ,  $E_G^+.add((v_r, v'_t)^+)$ ,
9    $E_D^-.add((v_r, v_t)^-, (v_r, v'_t)^-)$ ,  $E_G^-.add((v_r, v'_t)^-)$ ;
10 end
11 for epoch = 0; epoch <  $n^{epoch}$  do
12   Train the discriminator  $D^+$ :
13   for iter = 0; iter <  $n^{iter}$  do
14     Sample  $B_d$  real and fake positive edges from  $E_D^+$ ;
15     Update  $\theta_D$  via Eqs. (8) and (11), and achieve gradient perturbation via Eq. (13);
16     Calculate privacy spent  $\hat{\delta}$  given the target  $\epsilon$ ;
17     Stop optimization if  $\hat{\delta} \geq \delta$ .
18   end
19   Train the generator  $G^+$ :
20   for iter = 0; iter <  $n^{iter}$  do
21     Subsample  $B_g$  fake positive edges from  $E_G^+$ ;
22     Update  $\theta_G$  via Eqs. (4) and (24).
23   end
24   Train the discriminator  $D^-$ :
25   for iter = 0; iter <  $n^{iter}$  do
26     Subsample  $B_d$  real and fake negative edges from  $E_D^-$ ;
27     Update  $\theta_D$  via Eqs. (9) and (11), and achieve gradient perturbation via Eq. (13);
28     Calculate privacy spent  $\hat{\delta}$  given the target  $\epsilon$ ;
29     Stop optimization if  $\hat{\delta} \geq \delta$ .
30   end
31   Train the generator  $G^-$ :
32   for iter = 0; iter <  $n^{iter}$  do
33     Subsample  $B_g$  fake negative edges from  $E_G^-$ ;
34     Update  $\theta_G$  via Eqs. (5) and (24).
35   end
36 end
37 Return privacy-preserving node embedding  $\mathbf{g}_v \in \theta_G$ ;

```

to the DPSGD mechanism (Eq. (13)), ensuring node-level DP. The privacy budget $\hat{\delta}$ is tracked, and training stops early if $\hat{\delta} > \delta$.

- Generator training: Each generator G^+ and G^- is trained for n^{iter} iterations. In each iteration, a batch of B_g fake edges is sampled, and the generator parameters θ_G are updated by maximizing the generator objective (Eq. (24)).

(4) Embedding output for downstream tasks: After all epochs, the generator parameters θ_G encode the privacy-preserving node embeddings $\mathbf{g}_v \in \theta_G$, which are used for downstream tasks such as edge sign prediction and node clustering.

G Details of Lemma

The following lemmas are used for proving Theorem 1:

Lemma 2 (Receptive field of a node). *As shown in Fig. 4(b), we define the **receptive field** of a node as the region (i.e., the set of nodes) over which it can exert influence. Accordingly, for a **subgraph** constructed from paths sampled on constrained BFS-trees (Fig. 4(b)), the maximum receptive field size of v_r is given by $R_{N,L} = \sum_{l=0}^L N^l = \frac{N^{L+1}-1}{N-1} \leq B$.*

Lemma 3. *Let S_{tr} denote the training set of subgraphs constructed from constrained BFS-tree paths, and $S(v) \subset S_{tr}$ denote the subgraph subset that contains the node v . Since $R_{N,L}$ represents the upper bound on the number of occurrences of any node in S_{tr} , it follows that $|S(v)| \leq R_{N,L}$. The proof of Lemma 3 is illustrated in App. H.*

H Proof of Lemma 3

Proof. We proceed by induction [5] on the path length L of the BFS-tree.

Base case: When $L = 0$, each sampled subgraph $S(v)$ contains exactly the training node $v \in V_{tr}$ itself. Thus, every node appears in one subgraph, trivially satisfying the bound $|S(v)| = R_{N,0} = 1$.

Inductive hypothesis: Assume that for some fixed $L \geq 0$, any $v \in V_{tr}$ appears in at most $R_{N,L}$ subgraphs constructed from constrained BFS-tree paths. Let $S^L(v)$ denote a subgraph set with L path length. Thus, the hypothesis is $|S^L(v)| \leq R_{N,L}$ for any v .

Inductive step: We further show that the above hypothesis also holds for $L + 1$ path length: Let $T_{u'}$ represent the L -length BFS-tree rooted at u' . If $T_{u'} \in S^{L+1}(v)$, there must exist node u such that $u \in T_{u'}$ and $T_u \in S^L(v)$. According to the setting of Algorithm 1, the number of such nodes u is at most N . By the hypothesis, there are at most $R_{N,L} - 1$ such $u' \neq v$ such that $T_{u'} \in S^{L+1}(v)$. Based on these upper bounds, we can derive the upper bound matching the inductive hypothesis for $L + 1$:

$$|S^{L+1}(v)| \leq N \cdot (R_{N,L} - 1) + 1 = \frac{N^{L+2} - 1}{N - 1} = R_{N,L+1}. \quad (25)$$

By induction, the Lemma 3 holds for all $L \geq 0$.

I Proof of Theorem 2

The following lemmas are used for proving Theorem 2:

Lemma 4 (Adaptation of Lemma 5 from [15]). *Let $\mathcal{N}(\mu, \sigma^2)$ represent the Gaussian distribution with mean μ and standard deviation σ^2 , it holds that:*

$$\mathcal{D}_\alpha (\mathcal{N}(\mu, \sigma^2) \parallel \mathcal{N}(0, \sigma^2)) = \frac{\alpha \mu^2}{2\sigma^2} \quad (26)$$

Lemma 5 (Adaptation of Lemma 25 from [10]). *Assume μ_0, \dots, μ_n and η_0, \dots, η_n are probability distributions over some domain Z such that their Rényi divergences satisfy: $\mathcal{D}_\alpha(\mu_0 \parallel \eta_0) \leq \epsilon_0, \dots, \mathcal{D}_\alpha(\mu_n \parallel \eta_n) \leq \epsilon_n$ for some given $\epsilon_0, \dots, \epsilon_n$. Let ρ be a probability distribution over $\{0, \dots, n\}$. Denoted by μ_ρ (η_ρ , respectively) the probability distribution*

on Z obtained by sampling i from ρ and then randomly sampling from μ_i and η_i , we have:

$$\mathcal{D}_\alpha(\mu_\rho \| \eta_\rho) \leq \ln \mathbb{E}_{i \sim \rho} [e^{\varepsilon_i(\alpha-1)}] = \frac{1}{\alpha-1} \ln \sum_{i=0}^n \rho_i e^{\varepsilon_i(\alpha-1)} \quad (27)$$

Proof of Theorem 2. Consider any minibatch \mathcal{B}_t randomly sampled from the training subgraph set S_{tr} of Algorithm 2 at iteration t . For a subset $S(v^*) \subset S_{tr}$ containing node v^* , its size is bounded by $R_{N,L}$ (Lemma 3). Define the random variable β as $|S(v^*) \cap \mathcal{B}_t|$, and its distribution follows the hypergeometric distribution $\text{Hypergeometric}(|S_{tr}|, R_{N,L}, |\mathcal{B}_t|)$ [11]:

$$\beta_i = P[\beta = i] = \frac{|\mathcal{B}_t|=B_d}{|S_{tr}|=N_{tr}} \frac{\binom{R_{N,L}}{i} \binom{N_{tr}-R_{N,L}}{B_d-i}}{\binom{N_{tr}}{B_d}}. \quad (28)$$

Next, consider the training of the discriminators (Lines 12–18 and 24–30 in Algorithm 2). Let \mathcal{G} and \mathcal{G}' be two adjacent graphs differing only in the presence of node v^* and its associated signed edges. Based on the gradient perturbation applied in Lines 15 and 27 of Algorithm 2, we have:

$$\begin{aligned} \tilde{g}_t &= g_t + \mathcal{N}\left(0, \sigma^2 \Delta_g^2 \mathbf{I}\right) = \sum_{v \in \mathcal{B}_t} \text{Clip}_C\left(\frac{\partial L_D}{\partial \mathbf{d}_v}\right) + \mathcal{N}\left(0, \sigma^2 \Delta_g^2 \mathbf{I}\right) \\ \tilde{g}'_t &= g'_t + \mathcal{N}\left(0, \sigma^2 \Delta_g^2 \mathbf{I}\right) = \sum_{v' \in \mathcal{B}'_{tr}} \text{Clip}_C\left(\frac{\partial L_D}{\partial \mathbf{d}_{v'}}\right) + \mathcal{N}\left(0, \sigma^2 \Delta_g^2 \mathbf{I}\right), \end{aligned} \quad (29)$$

where $\Delta_g = R_{N,L} C = \frac{N^{L+1}-1}{N-1} C$ (Theorem 1). \tilde{g}_t and \tilde{g}'_t denote the noisy gradients of \mathcal{G} and \mathcal{G}' , respectively. When $\beta = i$, their Rényi divergences can be upper bounded as:

$$\begin{aligned} &\mathcal{D}_\alpha(\tilde{g}_{t,i} \| \tilde{g}'_{t,i}) \\ &= \mathcal{D}_\alpha(g_{t,i} + \mathcal{N}(0, \sigma^2 \Delta_g^2 \mathbf{I}) \| g'_{t,i} + \mathcal{N}(0, \sigma^2 \Delta_g^2 \mathbf{I})) \\ &= \mathcal{D}_\alpha(\mathcal{N}(g_{t,i}, \sigma^2 \Delta_g^2 \mathbf{I}) \| \mathcal{N}(g'_{t,i}, \sigma^2 \Delta_g^2 \mathbf{I})) \\ &\stackrel{(a)}{=} \mathcal{D}_\alpha(\mathcal{N}((g_{t,i} - g'_{t,i}), \sigma^2 \Delta_g^2 \mathbf{I}) \| \mathcal{N}(0, \sigma^2 \Delta_g^2 \mathbf{I})) \\ &\stackrel{(b)}{\leq} \sup_{\|\Delta_i\|_2 \leq iC} \mathcal{D}(\mathcal{N}(\Delta_i, \sigma^2 \Delta_g^2 \mathbf{I}) \| \mathcal{N}(0, \sigma^2 \Delta_g^2 \mathbf{I})) \\ &\stackrel{(c)}{=} \sup_{\|\Delta_i\|_2 \leq iC} \frac{\alpha \|\Delta_i\|_2^2}{2\Delta_g^2 \sigma^2} = \frac{\alpha i^2}{2R_{N,L}^2 \sigma^2}, \end{aligned} \quad (30)$$

where $\Delta_i = g_{t,i} - g'_{t,i}$. (a) leverages the property that Rényi divergence remains unchanged under invertible transformations [15], while (b) and (c) are derived from Theorem 1 and Lemma 4, respectively. Based on Lemma 5, we derive that:

$$\begin{aligned} \mathcal{D}_\alpha(\tilde{g}_t \| \tilde{g}'_t) &\leq \ln \mathbb{E}_{i \sim \beta} \left[\exp\left(\frac{\alpha i^2 (\alpha-1)}{2R_{N,L}^2 \sigma^2}\right) \right] \\ &= \frac{1}{\alpha-1} \ln \left(\sum_{i=0}^{R_{N,L}} \beta_i \exp\left(\frac{\alpha i^2 (\alpha-1)}{2R_{N,L}^2 \sigma^2}\right) \right) = \gamma. \end{aligned} \quad (31)$$

Here, β_i is illustrated in Eq. (28). Based on the composition property of DP, after $T = n^{\text{epoch}} \cdot n^{\text{iter}}$ iterations, the discriminators satisfy node-level $(\alpha, 2T\gamma)$ -RDP. Moreover, owing to the post-processing property of DP, the generators G^+ and G^- inherit the same privacy guarantee as the discriminators. Therefore, Algorithm 2 obeys node-level $(\alpha, 2T\gamma)$ -RDP, and the proof of Theorem 2 is completed.

J Additional Details of Experiments

J.1 Dataset Introduction

The detailed introduction of all datasets is as follows.

- Bitcoin-Alpha and Bitcoin-OTC are trust networks among Bitcoin users, aimed at preventing transactions with fraudulent or high-risk users. In these networks, user relationships are represented by positive (trust) and negative (distrust) edges.
- Slashdot is a social network derived from user interactions on a technology news site, where relationships are annotated as positive (friend) or negative (enemy) edges.
- WikiRfA is a voting network for electing managers in Wikipedia, where edges denote positive (supporting vote) or negative (opposing vote) relationships between users.
- Epinions is a product review site where users can establish both trust and distrust relationships with others.

J.2 The Distribution of Node Degrees and Path Lengths

The findings for the distribution of node degrees and path lengths in the Bitcoin-Alpha and Slashdot datasets are shown in Figs. 8 and 9.

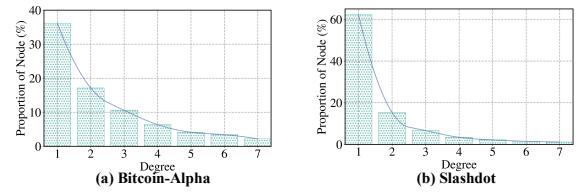


Figure 8: Distribution of node degrees.

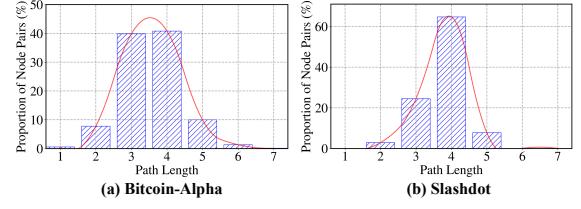


Figure 9: Distribution of path lengths.

J.3 The detailed results of Edge Sign Prediction

The average AUC results under different values of ϵ and datasets for edge prediction tasks are detailed in Table 6.

J.4 The detailed results of node clustering

The average SSI results under different values of ϵ and datasets for node clustering tasks are detailed in Table 7.

J.5 The Setup of Link Stealing Attack

Motivated by [13], we assume that the adversary has black-box access to the node embeddings produced by the target signed graph learning model, but not to its internal parameters or gradients. The adversary also possesses an auxiliary graph dataset comprising

Table 6: Summary of average AUC with different ϵ and datasets for edge sign prediction tasks. (BOLD: Best)

Dataset	Method	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 3$	$\epsilon = 4$	$\epsilon = 6$
Bitcoin-OTC	SDGNN	0.7655	0.7872	0.7913	0.8105	0.8571
	SiGAT	0.7011	0.7282	0.7869	0.8379	0.8706
	SGCN	0.5565	0.5740	0.6634	0.7516	0.7801
	GAP	0.5763	0.5782	0.6486	0.6741	0.7411
	LSNE	0.5030	0.5405	0.7041	0.8239	0.8776
	ASGL	0.8004	0.8462	0.8488	0.8505	0.8801
Bitcoin-Alpha	SDGNN	0.6761	0.6883	0.7098	0.7308	0.8476
	SiGAT	0.7033	0.7215	0.7303	0.7488	0.8207
	SGCN	0.5157	0.5450	0.6433	0.6930	0.7702
	GAP	0.5664	0.6025	0.6367	0.7091	0.7320
	LSNE	0.5112	0.5361	0.5959	0.6524	0.8069
	ASGL	0.7505	0.8075	0.8589	0.8591	0.8592
WikiRfA	SDGNN	0.6558	0.7066	0.7142	0.7267	0.7930
	SiGAT	0.6313	0.6525	0.7023	0.7777	0.8099
	SGCN	0.5107	0.6456	0.6515	0.7008	0.7110
	GAP	0.5356	0.5506	0.5612	0.5717	0.5937
	LSNE	0.5086	0.5253	0.6119	0.6553	0.7832
	ASGL	0.6680	0.7706	0.7963	0.7986	0.8100
Slashdot	SDGNN	0.7547	0.8325	0.8697	0.8788	0.8862
	SiGAT	0.7061	0.7886	0.8392	0.8424	0.8527
	SGCN	0.5662	0.6151	0.6662	0.7181	0.8093
	GAP	0.6121	0.6389	0.6879	0.7126	0.7471
	LSNE	0.5717	0.6144	0.7541	0.7753	0.7816
	ASGL	0.7861	0.8539	0.8887	0.8890	0.8910
Epinions	SDGNN	0.6788	0.7180	0.7201	0.7455	0.8428
	SiGAT	0.6772	0.7046	0.7063	0.7702	0.8253
	SGCN	0.6152	0.6487	0.6974	0.7502	0.8318
	GAP	0.5899	0.6034	0.6288	0.6310	0.6618
	LSNE	0.5033	0.6055	0.7590	0.8434	0.8585
	ASGL	0.6869	0.8134	0.8513	0.8658	0.8666

Table 7: Summary of average SSI with different ϵ and datasets for node clustering tasks. (BOLD: Best)

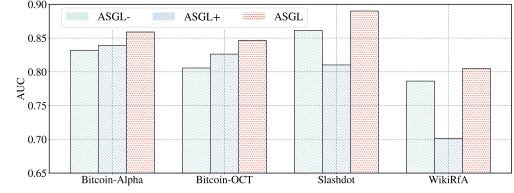
ϵ	Dataset	SGCN	SDGNN	SiGAT	LSNE	GAP	ASGL
1	Bitcoin-Alpha	0.4819	0.4378	0.4877	0.4977	0.4988	0.5091
	Bitcoin-OTC	0.4505	0.4677	0.5025	0.4970	0.5008	0.5160
	Slashdot	0.4715	0.5011	0.5025	0.5052	0.5005	0.5107
	WikiRfA	0.4788	0.4988	0.4968	0.4890	0.5003	0.5126
	Epinions	0.5001	0.4965	0.5022	0.5013	0.6095	0.6106
2	Bitcoin-Alpha	0.4910	0.4733	0.4969	0.4985	0.5032	0.5402
	Bitcoin-OTC	0.4733	0.4968	0.5075	0.4986	0.5729	0.6810
	Slashdot	0.4888	0.4864	0.4871	0.5134	0.5132	0.5494
	WikiRfA	0.4934	0.5054	0.5117	0.4996	0.5032	0.5577
	Epinions	0.5068	0.5116	0.5086	0.5463	0.6263	0.6732
4	Bitcoin-Alpha	0.5019	0.4948	0.5112	0.5049	0.6204	0.6707
	Bitcoin-OTC	0.5005	0.5325	0.5612	0.5465	0.6953	0.7713
	Slashdot	0.5003	0.5685	0.5545	0.5671	0.5444	0.5994
	WikiRfA	0.5005	0.5142	0.5538	0.5476	0.5644	0.5977
	Epinions	0.5148	0.5389	0.5386	0.6255	0.6747	0.6787

node pairs that partially overlap in distribution with the target

graph. Some of these node pairs belong to the training graph (members), while others are from the test graph (non-members). For each node pair, a feature vector is constructed by concatenating their embeddings. Finally, these feature vectors, along with their corresponding member or non-member labels, are then used to train a logistic regression classifier to infer whether an edge exists between any two nodes of the target graph. To simulate this link stealing attack, each dataset is partitioned into target training, auxiliary training, target test, and auxiliary test sets with a 5:2:2:1 ratio.

J.6 Effectiveness of Adversarial Learning with Edge Signs.

To verify the effectiveness of adversarial learning with signed edges, we also compare our ASGL with its variants, denoted as ASGL⁺ and ASGL⁻. Specifically, ASGL⁺ and ASGL⁻ only operate on the positive graph \mathcal{G}^+ and the negative graph \mathcal{G}^- , respectively. Fig. 10 presents the average AUC scores of ASGL, ASGL⁺, and ASGL⁻ across all datasets. It can be observed that ASGL significantly outperforms both ASGL⁺ and ASGL⁻ in all cases. These results demonstrate that our privacy-preserving adversarial learning framework with edge signs is more effective in representing signed graphs compared to its variants that neglect edge sign information.

**Figure 10: Comparison between ASGL, ASGL⁺, and ASGL⁻.**