

Global Neighbor Sampling for Mixed CPU-GPU Training on Giant Graphs

Jialin Dong^{*†}
jialind@g.ucla.edu

University of California, Los Angeles
USA

Lin F. Yang[‡]
linyaf@ee.ucla.edu

University of California, Los Angeles
USA

Da Zheng[†]

dzzhen@amazon.com
Amazon
USA

George Karypis[‡]
gkarypis@amazon.com

Amazon
USA

ABSTRACT

Graph neural networks (GNNs) are powerful tools for learning from graph data and are widely used in various applications such as social network recommendation, fraud detection, and graph search. The graphs in these applications are typically large, usually containing hundreds of millions of nodes. Training GNN models on such large graphs efficiently remains a big challenge. Despite a number of sampling-based methods have been proposed to enable mini-batch training on large graphs, these methods have not been proved to work on truly industry-scale graphs, which require GPUs or *mixed-CPU-GPU training*. The state-of-the-art sampling-based methods are usually not optimized for these real-world hardware setups, in which data movement between CPUs and GPUs is a bottleneck. To address this issue, we propose Global Neighborhood Sampling that aims at training GNNs on giant graphs specifically for mixed-CPU-GPU training. The algorithm samples a global cache of nodes periodically for all mini-batches and stores them in GPUs. This global cache allows in-GPU importance sampling of mini-batches, which drastically reduces the number of nodes in a mini-batch, especially in the input layer, to reduce data copy between CPU and GPU and mini-batch computation without compromising the training convergence rate or model accuracy. We provide a highly efficient implementation of this method and show that our implementation outperforms an efficient node-wise neighbor sampling baseline by a factor of $2 \times -4 \times$ on giant graphs. It outperforms an efficient implementation of LADIES with small layers by a factor of $2 \times -14 \times$ while achieving much higher accuracy than LADIES. We also theoretically analyze the proposed algorithm and show that with cached node data of a proper size, it enjoys a comparable convergence rate as the underlying node-wise sampling method.

^{*}The work was performed during an internship at AWS Shanghai AI Lab.

[†]Both authors contributed equally to this research.

[‡]Corresponding author.

Permission to make digital or hard copies of part or all 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 third-party components of this work must be honored. For all other uses, contact the owner/author(s).

KDD '21, August 14–18, 2021, Virtual Event, Singapore

© 2021 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-8332-5/21/08.

<https://doi.org/10.1145/3447548.3467437>

CCS CONCEPTS

• **Computing methodologies** → **Batch learning**; • **Theory of computation** → **Machine learning theory**.

KEYWORDS

Graph Neural Networks, Neighbor Sampling, Mixed CPU-GPU Training

ACM Reference Format:

Jialin Dong, Da Zheng, Lin F. Yang, and George Karypis. 2021. Global Neighbor Sampling for Mixed CPU-GPU Training on Giant Graphs. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), August 14–18, 2021, Virtual Event, Singapore*. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3447548.3467437>

1 INTRODUCTION

Many real world data come naturally in the form of graphs; e.g., social networks, gene expression networks and knowledge graphs. In recent years, Graph Neural Networks (GNNs) [4, 7, 13] have been proposed to learn from such graph-structured data and have achieved outstanding performance. Yet, in many applications, graphs are usually large, containing hundreds of millions to billions of nodes and tens to hundreds of billions of edges. Learning on such giant graphs is challenging due to the limited amount of memory available on a single GPU or a single machine. As such, mini-batch training is developed to train GNN models on such giant graphs. However, due to the connectivities between nodes, computing node embeddings with multi-layer GNNs usually involves in many nodes in a mini-batch. This leads to substantial computation and data movement between CPUs and GPUs and makes training inefficient.

To remedy this issue, various GNN sampling methods have been developed to reduce the number of nodes in a mini-batch [1, 4, 9, 15, 17]. Node-wise neighbor sampling used by GraphSage [4] samples a fixed number of neighbors for each node independently. Even though it reduces the number of neighbors in a mini-batch, the number of nodes in a layer still grows exponentially. FastGCN [1] and LADIES [17] sample a fixed number of nodes in each layer, which results in isolated nodes when used on large graphs. In addition, LADIES requires significantly more computation on sampling and potentially slows down the overall training speed. The work by Liu et al. [9] alleviates the neighborhood explosion and reduces

the sampling variance by applying a bandit sampler. However, this method leads to very large sampling overhead and does not scale to large graphs. These sampling methods are usually evaluated on small to medium-size graphs. When applying them to industry-scale graphs, they have suboptimal performance or substantial computation overhead as we discovered in our experiments.

To address some of these problems and reduce training time, LazyGCN [11] periodically samples a *mega-batch* of nodes and edges and reuses it to sample further mini-batches. By loading each mega-batch in GPU memory once, LazyGCN can mitigate data movement/preparation overheads. However, LazyGCN requires very large mega-batches (cf. Figure 4) to match the accuracy of standard GNN training, which makes it impractical for graphs with hundreds of millions of nodes.

We design an efficient and scalable sampling method that takes into account the characteristics of training hardware into consideration. GPUs are the most efficient hardware for training GNN models. Due to the small GPU memory size, state-of-the-art GNN frameworks, such as DGL [14] and Pytorch-Geometric [3], keep the graph data in CPU memory and perform mini-batch computations on GPUs when training GNN models on large graphs. We refer to this training strategy as *mixed CPU-GPU training*. The main bottleneck of mixed CPU-GPU training is data copy between CPUs and GPUs (cf. Figure 1). Because mini-batch sampling occurs in CPU, we need a low-overhead sampling algorithm to enable efficient training. Motivated by the hardware characteristics, we developed the *Global Neighborhood Sampling* (GNS) approach that samples a global set of nodes periodically for all mini-batches. The sampled set is small so that we can store all of their node features in GPU memory and we refer this set of nodes as *cache*. The cache is used for neighbor sampling in a mini-batch. Instead of sampling any neighbors of a node, GNS gives the priorities of sampling neighbors that exist in the cache. This is a fast way of biasing node-wise neighbor sampling to reduce the number of distinct nodes of each mini-batch and increase the overlap between mini-batches. When coupled with GPU cache, this method drastically reduces the amount of data copy between GPU and CPU to speed up training. In addition, we deploy importance sampling that reduces the sampling variance and also allows us to use a small cache size to train models.

We develop a highly optimized implementation of GNS and compare it with efficient implementations of other sampling methods provided by DGL, including node-wise neighbor sampling and LADIES. We show that GNS achieves state-of-the-art model accuracy while speeding up training by a factor of $2 \times -4 \times$ compared with node-wise sampling and by a factor of $2 \times -14 \times$ compared with LADIES.

The main contributions of the work are described below:

- (1) We analyze the existing sampling methods and demonstrate their main drawbacks on large graphs.
- (2) We develop an efficient and scalable sampling algorithm that addresses the main overhead in mixed CPU-GPU mini-batch training and show a substantial training speedup compared with efficient implementations of other training methods.
- (3) We demonstrate that this sampling algorithm can train GNN models on graphs with over 111 millions of nodes and 1.6 billions of edges.

2 BACKGROUND

In this section, we review GNNs and several state-of-the-art sampling-based training algorithms, including node-wise neighbor sampling methods and layer-wise importance sampling methods. The fundamental concepts of mixed-CPU-GPU training architecture is introduced. We discuss the limitations of state-of-the-art sampling methods in mixed-CPU-GPU training.

2.1 Existing GNN Training Algorithms

Full-batch GNN Given a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, the input feature of node $v \in \mathcal{V}$ is denoted as $\mathbf{h}_v^{(0)}$, and the feature of the edge between node v and u is represented as \mathbf{w}_{uv} . The representation of node $v \in \mathcal{V}$ at layer ℓ can be derived from a GNN model given by:

$$\mathbf{h}_v^\ell = g(\mathbf{h}_v^{\ell-1}, \bigcup_{u \in \mathcal{N}(v)} f(\mathbf{h}_u^{\ell-1}, \mathbf{h}_v^{\ell-1}, \mathbf{w}_{uv})), \quad (1)$$

where f , \bigcup and g are pre-defined or parameterized functions for computing feature data, aggregating data information, and updating node representations, respectively. For instance, in GraphSage training [4], the candidate aggregator functions include mean aggregator [7], LSTM aggregator [5], and max pooling aggregator [10]. The function g is set as nonlinear activation function.

Given training dataset $\{(\mathbf{x}_i, y_i)\}_{v_i \in \mathcal{V}_s}$, the parameterized functions will be learned by minimizing the loss function:

$$\mathcal{L} = \frac{1}{|\mathcal{V}_s|} \sum_{v_i \in \mathcal{V}_s} \ell(y_i, \mathbf{z}_i^L), \quad (2)$$

where $\ell(\cdot, \cdot)$ is a loss function, \mathbf{z}_i^L is the output of GNN with respect to the node $v_i \in \mathcal{V}_s$ where \mathcal{V}_s represents the set of training nodes. For full-batch optimization, the loss function is optimized by gradient descent algorithm where the gradient with respect to each node $v_i \in \mathcal{V}_s$ is computed as $\frac{1}{|\mathcal{V}_s|} \sum_{v_i \in \mathcal{V}_s} \nabla \ell(y_i, \mathbf{z}_i^L)$. During the training process, full-batch GNN requires to store and aggregate representations of all nodes across all layers. The expensive computation time and memory costs prohibit full-batch GNN from handling large graphs. Additionally, the convergence rate of full-batch GNN is slow because model parameters are updated only once at each epoch.

Mini-batch GNN To address this issue, a mini-batch training scheme has been developed which optimizes via mini-batch stochastic gradient descent $\frac{1}{|\mathcal{V}_B|} \sum_{v_i \in \mathcal{V}_B} \nabla \ell(y_i, \mathbf{z}_i^L)$ where $\mathcal{V}_B \in \mathcal{V}_s$. These methods first uniformly sample a set of nodes from the training set, known as *target nodes*, and sample neighbors of these target nodes to form a mini-batch. The focus of the mini-batch training methods is to reduce the number of neighbor nodes for aggregation via various sampling strategies to reduce the memory and computational cost. The state-of-the-art sampling algorithm is discussed in the sequel.

Node-wise Neighbor Sampling Algorithms. Hamilton et al. [4] proposed an unbiased sampling method to reduce the number of neighbors for aggregation via neighbor sampling. It randomly selects at most s_{node} (defined as *fan-out* parameter) neighborhood nodes for every target node; followed by computing the representations of target nodes via aggregating feature data from the sampled neighborhood nodes. Based on the notations in (1), the

representation of node $v \in \mathcal{V}$ at layer ℓ can be described as follows:

$$\mathbf{h}_v^\ell = g\left(\mathbf{h}_v^{\ell-1}, \bigcup_{u \in \mathcal{N}_\ell(v)} f\left(\frac{1}{s_{\text{node}}} \mathbf{h}_u^{\ell-1}, \mathbf{h}_v^{\ell-1}\right)\right), \quad (3)$$

where $\mathcal{N}_\ell(v)$ is the sampled neighborhood nodes set at ℓ -th layer such that $|\mathcal{N}_\ell(v)| = s_{\text{node}}$. The neighbor sampling procedure is repeated recursively on target nodes and their sampled neighbors when dealing with multiple-layer GNN. Even though node-wise neighbor sampling scheme addresses the memory issue of GNN, there exists excessive computation under this scheme because the scheme still results in exponential growth of neighbor nodes with the number of layers. This yields a large volume of data movement between CPU and GPU for mixed CPU-GPU training.

Layer-wise Importance Sampling Algorithms. To address the scalability issue, Chen et al. [1] proposed an advanced layer-wise method called FastGCN. Compared with node-wise sampling method, it yields extra variance when sampling a fixed number of nodes for each layer. To address the variance issue, it performs degree-based importance sampling on each layer. The representation of node $v \in \mathcal{V}$ at layer ℓ of FastGCN model is described as follows:

$$\mathbf{h}_v^\ell = g\left(\mathbf{h}_v^{\ell-1}, \bigcup_{u \in q(v)} f\left(\frac{1}{s_{\text{layer}}} \mathbf{h}_u^{\ell-1}/q_u, \mathbf{h}_v^{\ell-1}\right)\right), \quad (4)$$

where the sample size denotes as s_{layer} , $q(v)$ is the distribution over $v \in \mathcal{V}$ and q_u is the probability assigned to node u . A major limitation is that FastGCN performs sampling on every layer independently, which yields approximate embeddings with large variance. Moreover, the subgraph sampled by FastGCN is not representative to the original graph. This leads to poor performance and the number of sampled nodes required to guarantee convergence during training process is large.

The work by Zhou et al. [17] proposed a sampling algorithm known as LAYER-DEPENDENT IMPORTANCE SAMPLING (LADIES) to address the limitation of FastGCN and exploit the connection between different layers. Specifically, at ℓ -th layer, LADIES samples nodes reachable from the nodes in the previous layer. However, this method [17] comes with cost. To ensure the node connectivity between layers, the method needs to extract and merge the entire neighborhood of all nodes in the previous layer and compute the sampling probability for all candidate nodes in the next layer. Thus, this sampling method has significantly higher computation overhead. Furthermore, when applying this method on a large graph, it still constructs a mini-batch with many isolated nodes, especially for nodes in the first layer (Table 5).

LazyGCN. Even though layer-wise sampling methods effectively address the neighborhood explosion issue, they failed to investigate computational overheads in preprocessing data and loading fresh samples during training. Ramezan et al. [11] proposed a framework called LazyGCN which decouples the frequency of sampling from the sampling strategy. It periodically samples *mega-batches* and effectively reuses *mega-batches* to generate mini-batches and alleviate the preprocessing overhead. There are some limitations in the LazyGCN setting. First, this method requires large mini-batches to guarantee model accuracy. For example, their experiments on Yelp and Amazon dataset use the batch size of 65,536. This batch size is close to the entire training set, yielding overwhelming overhead in

a single mini-batch computation. Its performance deteriorates for smaller batch sizes even with sufficient epochs (Figure 4). Second, their evaluation is based on inefficient implementations with very large sampling overhead. In practice, the sampling computation overhead is relatively low in the entire mini-batch computation (Figure 1) when using proper development tools.

Even though both GNS and LazyGCN cache data in GPU to accelerate computation in mixed CPU-GPU training, they use very different strategies for caching. LazyGCN caches the entire graph structure of multiple mini-batches sampled by node-wise neighbor sampling or layer-wise sampling and suffers from the problems in these two algorithms. Due to the exponential growth of the neighborhood size in node-wise neighbor sampling, LazyGCN cannot store a very large mega-batch in GPU and can generate few mini-batches from the mega-batch. Our experiments show that LazyGCN runs out of GPU memory even with a small mega-batch size and mini-batch size on large graphs (OAG-paper and OGBN-papers100M in Table 2). Layer-wise sampling may result in many isolated nodes in a mini-batch. In addition, LazyGCN uses the same sampled graph structure when generating mini-batches from mega-batches, this potentially leads to overfit. In contrast, GNS cache nodes and use the cache to reduce the number of nodes in a mini-batch; GNS always sample a different graph structure for each mini-batch and thus it is less likely to overfit.

2.2 Mixed CPU-GPU training

Due to limited GPU memory, state-of-the-art GNN framework (e.g., DGL [14] and Pytorch Geometric [3]) train GNN models on large graph data by storing the whole graph data in CPU memory and performing mini-batch computation on GPUs. This allows users to take advantage of large CPU memory and use GPUs to accelerate GNN training. In addition, mixed CPU-GPU training makes it easy to scale GNN training to multiple GPUs or multiple machines [16].

Mixed CPU-GPU training strategy usually involves six steps: 1) sample a mini-batch from the full graph, 2) slice the node and edge data involved in the mini-batch from the full graph, 3) copy the above sliced data to GPU, 4) perform forward computation on the mini-batch, 5) perform backward propagation, and 6) run the optimizer and update model parameters. Steps 1–2 are done by the CPU, whereas steps 4–6 are done by the GPU.

We benchmark the mini-batch training of GraphSage [4] models with node-wise neighbor sampling provided by DGL, which provides very efficient neighbor sampling implementation and graph kernel computation for GraphSage. Figure 1 shows the breakdown of the time required to train GraphSage on the OGBN-products graph and the OAG-paper graph (see Table 2 for dataset information). Even though sampling happens in CPU, its computation accounts for 10% or less with sufficient optimization and parallelization. However, the speed of copying node data in CPU (step 2) is limited by the CPU memory bandwidth and moving data to GPU (step 3) is limited by the PCIe bandwidth. Data copying accounts for most of the time required by mini-batch training. For example, the training spends 60% and 80% of the per mini-batch time in copying data from CPU to GPU on OGBN-products and OAG-paper, respectively. The training on OAG-paper takes significantly more time on data copy because OAG-paper has 768-dimensional BERT

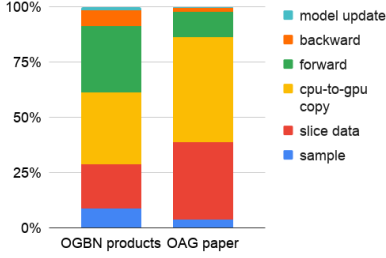


Figure 1: Runtime breakdown (%) of each component in mini-batch training for an efficient GraphSage implementation in DGL.

embeddings [2] as node features, whereas OGBN-products uses 100-dimensional node features.

These results show that when the different components of mini-batch training are highly-optimized, the main bottleneck of mixed CPU-GPU training is data copy (both data copy in CPU and between CPU and GPUs). To speed up training, it is essential to reduce the overhead of data copy, without significantly increasing the overhead of other steps.

3 GLOBAL NEIGHBOR SAMPLING (GNS)

To overcome the drawbacks of the existing sampling algorithms and tackle the unique problems in mixed CPU-GPU training, we developed a new sampling approach, called *Global Neighborhood Sampling* (GNS), that has low computational overhead and reduces the number of nodes in a mini-batch without compromising the model accuracy and convergence rate. Like node-wise and layer-wise sampling, GNS uses mini-batch training to approximate the full-batch GNN training on giant graphs.

3.1 Overview of GNS

Instead of sampling neighbors independently like node-wise neighbor sampling, GNS periodically samples a global set of nodes following a probability distribution \mathcal{P} to assist in neighbor sampling. \mathcal{P}_i defines the probability of node i in the graph being sampled and placed in the set. Because GNS only samples a small number of nodes to form the set, we can copy all of the node features in the set to GPUs. Thus, we refer to the set of nodes as *node cache* C . When sampling neighbors of a node, GNS prioritizes the sampled neighbors from the cache and samples additional neighbors outside the cache only if the cache does not provide sufficient neighbors.

Because the nodes in the cache are sampled, we can compute the node sampling probability from the probability of a node appearing in the cache, i.e., \mathcal{P} . We rescale neighbor embeddings by importance sampling coefficients $p_u^{(\ell)}$ from \mathcal{P} in the mini-batch forward propagation so that the expectation of the aggregation of sampled neighbors is the same as the aggregation of the full neighborhood.

$$\mathbb{E} \left(\sum_{u \in \mathcal{N}_\ell(v)} p_u^{(\ell)} * h_u^\ell \right) = \sum_{u \in \mathcal{N}(v)} h_u^\ell \quad (5)$$

Algorithm 1 illustrates the entire training process.

Algorithm 1: Minibatch Training with Global Neighbor sampling for GNN on Node Classification

Input : Graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$;

list of target nodes of minibatches $\{\mathcal{B}_1, \dots, \mathcal{B}_M\}$;

input features $\{\mathbf{x}_v, \forall v \in \mathcal{V}\}$;

number of epochs T ;

depth L ; weight matrices $\mathbf{W}^\ell, \forall \ell \in \{1, \dots, L\}$;

cache sampling probability \mathcal{P} ;

nonlinear activation function g ;

differentiable aggregator functions

$f_\ell, \forall \ell \in \{1, \dots, L\}$.

Output: Vector representations \mathbf{z}_v for all $v \in \mathcal{B}$

1: **for** $t = 0$ to T **do**

2: $C \leftarrow \text{sample_cache}(\mathcal{V}, \mathcal{P}, \{\mathcal{B}_1, \dots, \mathcal{B}_M\})$

3: **for** $\mathcal{B} \in \{\mathcal{B}_1, \dots, \mathcal{B}_M\}$ **do**

4: $\mathcal{B}^L \leftarrow \mathcal{B}$

5: **for** $\ell = L \dots 1$ **do**

6: $\mathcal{B}^{\ell-1} \leftarrow \{\}$

7: **for** $u \in \mathcal{B}^\ell$ **do**

8: $\mathcal{N}_\ell(u), \mathcal{P}_\ell(u) \leftarrow \text{sample}(\mathcal{N}(u), C)$

9: $\mathcal{B}^{\ell-1} \leftarrow \mathcal{B}^{\ell-1} \cup \mathcal{N}_\ell(u)$;

10: $\mathcal{P}^{\ell-1} \leftarrow \mathcal{P}^{\ell-1} \cup \mathcal{P}_\ell(u)$

11: **end for**

12: **end for**

13: $\mathbf{h}_u^0 \leftarrow \mathbf{x}_v, \forall v \in \mathcal{B}^0$

14: **for** $\ell = 1 \dots L$ **do**

15: **for** $u \in \mathcal{B}^\ell$ **do**

16: Compute importance sampling coefficients $p_{u'}^{(\ell-1)}$ for $\forall u' \in \mathcal{N}_\ell(u)$

17: $\mathbf{h}_{\mathcal{N}(u)}^\ell \leftarrow f_\ell(\{p_{u'}^{(\ell-1)} \mathbf{h}_{u'}^{\ell-1}, \forall u' \in \mathcal{N}_\ell(u)\})$

18: $\mathbf{h}_u^\ell \leftarrow g(\mathbf{W}^\ell \cdot (\mathbf{h}_u^{\ell-1}, \mathbf{h}_{\mathcal{N}(u)}^\ell))$

19: **end for**

20: **end for**

21: **end for**

22: **end for**

In the remaining sections, we first discuss the cache sampling in Section 3.2. We explain the sampling procedure in Section 3.3. To reduce the variance in GNS, an importance sampling scheme is further developed in Section 3.4. We then establish the convergence rate of GNS which is inspired by the paper [11]. It shows that under mild assumption, GNS enjoys comparable convergence rate as underlying node-wise sampling in training, which is demonstrated in Section 3.5. The notations and definition used in the following is summarized in Table 1.

3.2 Sample Cache

GNS periodically constructs a cache of nodes C to facilitate neighbor sampling in mini-batch construction. GNS uses a biased sampling approach to select a set of nodes C that, with high probability, can be reached from nodes in the training set. The features of the nodes in the cache are loaded into GPUs beforehand.

Ideally, the cache needs to meet two requirements: 1) in order to keep the entire cache in the GPU memory, the cache has to be

Table 1: Summary of Notations and Definitions

$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	\mathcal{G} denotes the graph consist of set of $ \mathcal{V} $ nodes and $ \mathcal{E} $ edges.
L, K	L is the total number of layers in GCN, and K is the dimension of embedding vectors (for simplicity, assume it is the same across all layers).
b, s_{node}, s_{layer}	For batch-wise sampling, b denotes the batch size, s_{node} is the number of sampled neighbors per node for node-wise sampling, and s_{layer} is number of sampled nodes per layer for layer-wise sampling.
$\mathcal{N}(v)$	Denotes the set of neighbors of node $v \in \mathcal{V}$.
$\mathcal{N}_\ell(v)$	Denotes the set of sampled neighbors of node $v \in \mathcal{V}$ at ℓ -th layer.
$\mathcal{N}_\ell^s(v)$	Denotes the set of sampled neighbors of node $v \in \mathcal{V}$ at ℓ -th layer.
$\mathcal{N}_C(v)$	Denotes the set of neighbors of node $v \in \mathcal{V}$ in the cache.
C, p_v^{cache}	Denotes the set of cached nodes which are sampled from \mathcal{V} corresponding to the probability of p_v^{cache} for $v \in \mathcal{V}$.
$p_v^{(\ell)}$	Denotes importance sampling coefficients with respect to the node $v \in \mathcal{V}$ at ℓ -th layer in Algorithm 1.
$\mathcal{V}_S, \mathcal{V}_S $	Denotes the training set and the size of the training set.
target node	The node in the mini-batch where the mini-batch is sampled at random from the training node set.
$\mathcal{V}_B, \mathcal{V}_B $	Denotes the set of target nodes and the number of target nodes in a mini-batch.

sufficiently small; 2) in order to have sampled neighbors to come from the cache, the nodes in the cache have to be reachable from the nodes in the training set with a high probability.

Potentially, we can uniformly sample nodes to form the cache, which may require a large number of nodes to meet requirement 2. Therefore, we deploy two approaches to define the sampling probability for the cache. If majority of the nodes in a graph are in the training set, we define the sampling probability based on node degree. For node i , the probability of being sampled in the cache is given by

$$p_i = \deg(i) / \sum_{k \in \mathcal{V}} \deg(k). \quad (6)$$

For a power-law graph, we only need to maintain a small cache of nodes to cover majority of the nodes in the graph.

If the training set only accounts for a small portion of the nodes in the graph, we use short random walks to compute the sampling probability. Define $\mathcal{N}_\ell(v)$ as the number of sampled neighbor nodes corresponding to node $v \in \mathcal{V}$ in each layer,

$$\mathbf{d} = [\mathcal{N}_\ell(v_1)/\deg(v_1), \dots, \mathcal{N}_\ell(v_{|\mathcal{V}|})/\deg(v_{|\mathcal{V}|})]^\top, \quad (7)$$

The node sampling probability $P^\ell \in \mathbb{R}^{|\mathcal{V}|}$ for the ℓ -th layer is represented as

$$P^\ell = (\mathbf{D}\mathbf{A} + \mathbf{I})P^{\ell-1}, \quad (8)$$

where \mathbf{A} is the adjacency matrix and $\mathbf{D} = \text{diag}(\mathbf{d})$. P^0 is

$$p_i^0 = \begin{cases} \frac{1}{|\mathcal{V}_S|}, & \text{if } i \in \mathcal{V}_S \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

The sampling probability for the cache is set as P^L , where L is the number of layers in the multi-layer GNN model.

As the experiments will later show (cf. Section 4), the size of the cache C can be as small as 1% of the number of nodes ($|\mathcal{V}|$) without compromising the model accuracy and convergence rate.

3.3 Sample Neighbors with Cache

When sampling k neighbors for a node, GNS first restricts sampled neighbor nodes from the cache C . If the number of neighbors sampled from the cache is less than k , it samples remaining neighbors uniformly at random from its own neighborhood.

A simple way of sampling neighbors from the cache is to compute the overlap of the neighbor list of a node with the nodes in the cache. Assuming one lookup in the cache has $O(1)$ complexity, this algorithm will result in $O(|\mathcal{E}|)$ complexity, where $|\mathcal{E}|$ is the number of edges in the graph. However, this complexity is significantly larger than the original node-wise neighbor sampling $O(\sum_{i \in \mathcal{V}_B} \min(k, |\mathcal{N}(i)|))$ in a power-law graph, where \mathcal{V}_B is the set of target nodes in a mini-batch and $|\mathcal{N}(i)|$ is the number of neighbors of node i . Instead, we construct an induced subgraph \mathcal{S} that contains the nodes in the cache and their neighbor nodes. This is done once, right after we sample nodes in the cache. For an undirected graph, this subgraph contains the neighbors of all nodes that reside in the cache. During neighbor sampling, we can get the cached neighbors of node i by reading the neighborhood $\mathcal{N}_\mathcal{S}(i)$ of node i in the subgraph. Constructing the subgraph \mathcal{S} is much more lightweight, usually $\ll O(|\mathcal{E}|)$.

We parallelize the sampling computations with multiprocessing. That is, we create a set of processes to sample mini-batches independently and send them back to the trainer process for mini-batch computation. The construction of subgraphs \mathcal{S} for multiple caches C can be parallelized.

3.4 Importance Sampling Coefficient

The nodes in the cache are sampled non-uniformly at random. So are the neighbors of a node. To approximate the expectation of the uniform sampling method, we assign importance weights to the nodes, thereby rescaling the neighbor features for aggregation.

$$\mathbf{h}_{\mathcal{N}(u)}^\ell \leftarrow f_\ell(\{1/p_{u'}^{(\ell-1)} \cdot \mathbf{h}_{u'}^{\ell-1}, \forall u' \in \mathcal{N}_\ell(u)\}). \quad (10)$$

To establish the importance sampling coefficient, we begin with computing the probability of the sampled node $u' \in \mathcal{N}_\ell(u)$ being contained in the cache, given by

$$p_{u'}^C = 1 - (1 - p_{u'})^{|\mathcal{C}|}, \quad (11)$$

where the sampling probability $p_{u'}$ refers to (6) and $|\mathcal{C}|$ denotes the size of cache set. The importance sampling coefficient $p_{u'}^{(\ell-1)}$ can be represented as

$$p_{u'}^{(\ell-1)} = p_{u'}^C \frac{k}{\max\{k, \mathcal{N}_C(i)\}}. \quad (12)$$

3.5 Theoretical Analysis

In this section, we establish the convergence rate of GNS which is inspired by the work of Ramezani et al. [11]. It shows that under mild assumption, GNS enjoys comparable convergence rate as underlying node-wise sampling in training. Here, the convergence rate of GNS mainly depends on the graph degree and the size of cached set. We focus on a two-layer GCN for simplicity and denote the loss functions of full-batch, mini-batch and proposed GNS as

$$J(\theta) = \frac{1}{N} \sum_{i \in \mathcal{V}} f_i \left(\frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \frac{1}{|\mathcal{N}(j)|} \sum_{k \in \mathcal{N}(j)} g_{jk}(\theta) \right) \quad (13)$$

$$J_B(\theta) = \frac{1}{B} \sum_{i \in \mathcal{V}_B} f_i \left(\frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \frac{1}{|\mathcal{N}(j)|} \sum_{k \in \mathcal{N}(j)} g_{jk}(\theta) \right) \quad (14)$$

$$\begin{aligned} \tilde{J}_B(\theta) = & \frac{1}{B} \sum_{i \in \mathcal{V}_B} f_i \left(\frac{1}{|\mathcal{N}_2^u(i) \cap \mathcal{C}|} \sum_{j \in \mathcal{N}_2^u(i) \cap \mathcal{C}} \frac{1}{|\mathcal{N}_1^u(j) \cap \mathcal{C}|} \right. \\ & \left. \sum_{k \in \mathcal{N}_1^u(j) \cap \mathcal{C}} p_k^{(1)} g_{jk}(\theta) \right), \end{aligned} \quad (15)$$

respectively, where the outer and inner layer function are defined as $f(\cdot) \in \mathbb{R}$ and $g(\cdot) \in \mathbb{R}^n$, and their gradients as $\nabla f(\cdot) \in \mathbb{R}^n$ and $\nabla g(\cdot) \in \mathbb{R}^{n \times n}$, respectively. Specifically, the function $g_{jk}(\cdot)$ depends on the nodes contained in two layers. For simplicity, we denote $\mathcal{N}^C(j) := \mathcal{N}_1^u(j) \cap \mathcal{C}$. We denote $|\mathcal{N}(i)| = N^i$, $|\mathcal{N}_\ell^u(i)| = N_\ell^i$, $|\mathcal{N}^C(j)| = N_C^j$ in the following.

The following assumption gives the Lipschitz continuous constant of the gradient of composite function $J(\theta)$, which plays a vital role in theoretical analysis.

ASSUMPTION 1. Suppose $f(\cdot)$ is L_f -Lipschitz continuous, $g(\cdot)$ is L_g -Lipschitz continuous, $\nabla f(\cdot)$ is L'_f -Lipschitz continuous, $\nabla g(\cdot)$ is L'_g -Lipschitz continuous.

THEOREM 1. Denote N_ℓ^i as the number of the neighborhood nodes with respect to $i \in \mathcal{V}$ sampling uniformly at random at ℓ -th layer. The cached nodes in the set \mathcal{C} with the size of $|\mathcal{C}|$ are sampled without replacement according to p_v^{cache} . The dimension of node feature is denoted as n and the size of mini-batch is denoted as B . Define $\tilde{C} = |\mathcal{C}|/|\mathcal{V}|$ and $C_d = \sum_{v_i \in \mathcal{V}} \deg(v_i)/|\mathcal{V}|$ with the constant $c > 0$. Under Assumption 1, with probability exceeding $1 - \delta$, GNS optimized

by stochastic gradient descent can achieve

$$\mathbb{E} \left[\|\nabla J(\hat{\theta})\|^2 \right] \leq \mathcal{O} \left(\sqrt{\frac{\text{MSE}}{t}} \right), \quad (16)$$

where $\hat{\theta} = \min_t \mathbb{E} [\|\nabla J(\theta_t)\|]$ with $\theta_t = \{\mathbf{W}_t^\ell\}_{\ell=1}^L$ and

$$\begin{aligned} \text{MSE} \leq & \mathcal{O} \left(L_f'^2 \frac{\log(4n/\delta) + 1/2}{B} \right) + \mathcal{O} \left(L_f'^2 L_g^4 \frac{\log(4n/\delta) + 1/2}{c\tilde{C}C_d N_1^j N_2^i} \right) \\ & + \mathcal{O} \left(L_g'^2 L_f^2 \frac{\log(4n/\delta)}{c\tilde{C}C_d N_1^j N_2^i} \right). \end{aligned} \quad (17)$$

PROOF. The details on the proof of Theorem 1 is provided in Appendix A. \square

Variance of GNS We aim to derive the average variance of the embedding for the output nodes at each layer. Before moving forward, we provide several useful definitions. Let B denote the size of the nodes in one layer. Consider the underlying embedding: $\mathbf{Z} = \mathbf{L}\mathbf{H}\Theta$, where \mathbf{L} is the Laplacian matrix, \mathbf{H} denotes the feature matrix and Θ is the weight matrix, Let $\tilde{\mathbf{Z}} \in \mathbb{R}^{B \times d}$ with the dimension of feature d denote the estimated embedding derived from the sample-based method. Denote $\mathbf{P} \in \mathbb{R}^{B \times |\mathcal{V}|}$ as the row selection matrix which samples the embedding from the whole embedding matrix. The variance can be represented as $\mathbb{E}[\|\tilde{\mathbf{Z}} - \mathbf{P}\mathbf{Z}\|_F^2]$. Denote $\mathbf{L}_{i,*}$ as the i -th row of matrix \mathbf{L} , $\mathbf{L}_{*,j}$ is the j -th column of matrix \mathbf{L} , and $L_{i,j}$ is the element at the position (i, j) of matrix \mathbf{L} .

For each node at each layer, its embedding is estimated based on its neighborhood nodes established from the cached set \mathcal{C} . Based on the Assumption 1 in [17], we have

$$\begin{aligned} & \mathbb{E} [\|\tilde{\mathbf{Z}} - \mathbf{P}\mathbf{Z}\|_F^2] \\ &= \sum_{i=1}^{|\mathcal{V}|} q_i \cdot \mathbb{E} [\|\tilde{\mathbf{Z}}_{i,*} - \mathbf{Z}_{i,*}\|_2^2] \\ &= \sum_{i=1}^{|\mathcal{V}|} q_i \|\mathbf{L}_{i,*}\|_0 \left(\sum_{j=1}^{|\mathcal{V}|} p_{ij} \cdot s_j \|\mathbf{L}_{i,j} \mathbf{H}_{j,*} \Theta\|_2^2 - \|\mathbf{L}_{i,*} \mathbf{H}\Theta\|_F^2 \right) \\ &= \sum_{i=1}^{|\mathcal{V}|} q_i \|\mathbf{L}_{i,*}\|_0 \left(\sum_{j=1}^{|\mathcal{V}|} p_{ij} \cdot s_j \|\mathbf{L}_{i,j} \mathbf{H}_{j,*} \Theta\|_2^2 - \right. \\ & \quad \left. q_i \cdot p_{ij} \cdot s_j \|\mathbf{L}\mathbf{H}\Theta\|_F^2 \right) \end{aligned} \quad (18)$$

where q_i is the probability of node i being contained in the first layer via neighborhood sampling and p_{ij} is the importance sampling coefficient related to node i and j . Moreover, s_j is the probability of node i being in the cache set \mathcal{C} .

Under Assumption 1 and 2 in [17] such that $\|\mathbf{H}_{i,*} \Theta\|_2 \leq \gamma$ for all $i \in [|\mathcal{V}|]$ and $\|\mathbf{L}_{i,*}\|_0 \leq \frac{C}{|\mathcal{V}|} \sum_{i=1}^{|\mathcal{V}|} \|\mathbf{L}_{i,*}\|_0$ and the definition of the importance sampling coefficient, we arrive

$$\begin{aligned} & \mathbb{E} [\|\tilde{\mathbf{Z}} - \mathbf{P}\mathbf{Z}\|_F^2] \\ & \leq \sum_{i=1}^{|\mathcal{V}|} q_i \|\mathbf{L}_{i,*}\|_0 \sum_{j=1}^{|\mathcal{V}|} p_{ij} \cdot s_j \|\mathbf{L}_{i,j} \mathbf{H}_{j,*} \Theta\|_2^2 \\ & \leq C \sum_{i=1}^{|\mathcal{V}|} \sum_{j=1}^{|\mathcal{V}|} q_i \cdot p_{ij} \cdot s_j \|\mathbf{L}_{i,j} \mathbf{H}_{j,*} \Theta\|_2^2 \end{aligned}$$

$$\leq \frac{CB_{\text{out}}C_dY\|L\|_F^2}{|\mathcal{V}|}, \quad (19)$$

where C_d denotes the average degree and B_{out} is the size of nodes at the output layer.

3.6 Summary and Discussion

GNS shares many advantages of various sampling methods and is able to avoid their drawbacks. Like node-wise neighbor sampling, it samples neighbors on each node independently and, thus, can be implemented and parallelized efficiently. Due to the cache, GNS tends to avoid the neighborhood explosion in multi-layer GNN. GNS maintains a global and static distribution to sample the cache, which requires only one-time computation and can be easily amortized during the training. In contrast, LADIES computes the sampling distribution for every layer in every mini-batch, which makes the sampling procedure expensive. Even though GNS constructs a mini-batch with more nodes than LADIES, forward and backward computation on a mini-batch is not the major bottleneck in many GNN models for mixed CPU-GPU training. Even though both GNS and LazyGCN deploy caching to accelerate computation in mixed CPU-GPU training, they use cache very differently. GNS uses cache to reduce the number of nodes in a mini-batch to reduce computation and data movement between CPU and GPUs. It captures majority of connectivities of nodes in a graph. LazyGCN caches and reuses the sampled graph structure and node data. This requires a large mega-batch size to achieve good accuracy, which makes it difficult to scale to giant graphs. Because LazyGCN uses node-wise sampling or layer-wise sampling to sample mini-batches, it suffers from the problems inherent to these two sampling algorithms. For example, as shown in the experiment section, LazyGCN cannot construct a mega-batch with node-wise neighbor sampling on large graphs.

4 EXPERIMENTS

4.1 Datasets and Setup

We evaluate the effectiveness of GNS under inductive supervise setting on the following real-world large-scale datasets: Yelp [15], and Amazon [15], OAG-paper¹, OGBN-products [6], OGBN-papers100M [6]. OAG-paper is the paper citation graph in the medical domain extracted from the OAG graph[12]. On each of the datasets, the task is to predict labels of the nodes in the graphs. Table 2 provides various statistics for these datasets.

For each trial, we run the algorithm with ten epochs on Yelp, Amazon OGBN-products, OGBN-Papers100M dataset and each epoch proceeds for $\frac{\# \text{train set}}{\text{batch size}}$ iterations. For the OAG-paper dataset, we run the algorithm with three epochs. We compare GNS with node-wise neighbor sampling (used by GraphSage), LADIES and LazyGCN for training 3-layer GraphSage. The detailed settings with respect to these four methods are summarized as follows:

- **GNS:** GNS is implemented by DGL [14] and we apply GNS on all layers to sample neighbors. The sampling fan-outs of each layer are 15, 10 for the third and second layer. We sample nodes in the first layer (input layer) only from the cache. The size of cached set is $1\% \cdot |\mathcal{V}|$.

- **Node-wise neighbor sampling (NS)**² [4]: NS is implemented by DGL [14]. The sampling fan-outs of each layer are 15, 10 and 5.
- **LADIES**³ [17]: LADIES is implemented by DGL [14]. We sample 512 and 5000 nodes for LADIES per layer, respectively.
- **LazyGCN**⁴ [11]: we use the implementation provided by the authors. We set the recycle period size as $R = 2$ and the recycling growth rate as $\rho = 1.1$. The sampler of LazyGCN is set as nodewise sampling with 15 neighborhood nodes in each layer.

GNS, NS and LADIES are parallelized with multiprocessing. For all methods, we use the batch size of 1000. We use two metrics to evaluate the effectiveness of sampling methods: micro F1-score to measure the accuracy and the average running time per epoch to measure the training speed.

We run all experiments on an AWS EC2 g4dn.16xlarge instance with 32 CPU cores, 256GB RAM and one NVIDIA T4 GPU.

4.2 Experiment Results

We evaluate test F1-score and average running time per epoch via using different methods in the case of large-scale dataset. As is shown in Table 3, GNS can obtain the comparable accuracy score compared to NS with $2 \times -4 \times$ speed in training time, using a small cache. The acceleration attributes to the smaller number of nodes in a mini-batch, especially a smaller number of nodes in the input layer (Table 4). This significantly reduces the time of data copy between CPU and GPUs as well as reducing the computation overhead in a mini-batch (Figure 2). In addition, a large number of input nodes have been cached in GPU, which further reduces the time used in data copy between CPU to GPU. GNS scales well to giant graphs with 100 million nodes as long as CPU memory of the machine can accommodate the graph. In contrast, LADIES cannot achieve the state-of-the-art model accuracy and its training speed is actually slower than NS on many graphs. Our experiments also show that LazyGCN cannot achieve good model accuracy with a small mini-batch size, which is not friendly to giant graphs. In addition, The LazyGCN implementation provided by the authors fails to scale to giant graphs (e.g., OAG-paper and OGBN-products) due to the out-of-memory error even with a small mega-batch. Our method is robust to a small mini-batch size and can easily scale to giant graphs.

We plot the convergence rate of all of the training methods on OGBN-products based on the test F1-scores (Figure 3). In this study, LADIES samples 512 nodes per layer and GNS caches 1% of nodes in the graph. The result indicates that GNS achieves similar convergence and accuracy as NS even with a small cache, thereby confirming the theoretical analysis in Section 3.5, while LADIES and LazyGCN fail to converge to good model accuracy.

One of reasons why LADIES suffers poor performance is that it tends to construct a mini-batch with many isolated nodes, especially for nodes in the first layer (Table 5). When training three-layer GCN on OGBN-products with LADIES, the percentage of isolated nodes

¹https://s3.us-west-2.amazonaws.com/dgl-data/dataset/OAG/oag_max_paper.dgl

²<https://github.com/dmlc/dgl/tree/master/examples/pytorch/graphsage>

³<https://github.com/BarclayII/dgl/tree/ladies/examples/pytorch/ladies>

⁴<https://github.com/MortezaRamezani/lazygcn>

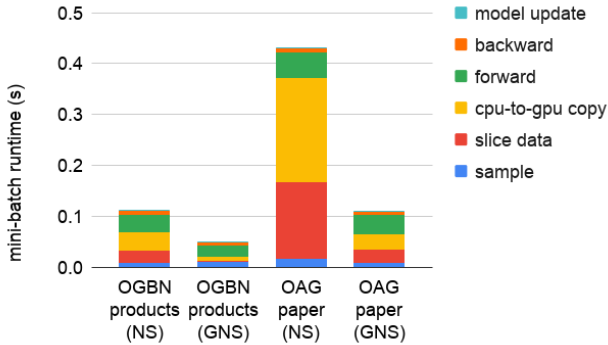
Table 2: Dataset statistics.

Dataset	Nodes	Edges	Avg. Deg	Feature	Classes	Multiclass	Train / Val / Test
Yelp	716,847	6,977,410	10	300	100	Yes	0.75 / 0.10 / 0.15
Amazon	1,598,960	132,169,734	83	200	107	Yes	0.85 / 0.05 / 0.10
OAG-paper	15,257,994	220,126,508	14	768	146	Yes	0.43 / 0.05 / 0.05
OGBN-products	2,449,029	123,718,280	51	100	47	No	0.10 / 0.02 / 0.88
OGBN-Papers100M	111,059,956	3,231,371,744	30	128	172	No	0.01 / 0.001 / 0.002

Table 3: Performance of different sampling approaches.

Dataset(hidden layer dimension)	Metric	NS	LADIES (512)	LADIES (5000)	LazyGCN	GNS
Yelp(512)	F1-Score(%)	62.54	59.32	61.04	35.58	63.20
	Time per epoch (s)	58.5	62.1	237.9	1248.7	23.1
Amazon(512)	F1-Score(%)	76.69	76.46	77.05	31.08	76.13
	Time per epoch (s)	89.5	613.4	3234.2	3280.2	42.8
OAG-paper(256)	F1-Score(%)	50.23	43.51	46.72	N/A	49.23
	Time per epoch (s)	3203.2	2108.0	7956.0		819.4
OGBN-products(256)	F1-Score(%)	78.44	70.32	75.36	69.78	78.01
	Time per epoch (s)	25.6	45.4	223.5	264.2	11.9
OGBN-Papers100M(256)	F1-Score(%)	63.61	57.94	59.23	N/A	63.31
	Time per epoch (s)	462.2	152.7	313.2		98.5

The results were obtained by training a 3-layer GraphSage with hidden state dimension as 512 on Yelp and Amazon dataset, and 256 on the rest, using four methods. We update the model with a mini-batch size of 1000 and ADAM optimizer with a learning rate of 0.003 for all training methods. We use the efficient implementation of node-wise neighbor sampling, LADIES and GNS in DGL, parallelized with multiprocessing. The number of sampling workers is 4. The column labeled “LADIES(512)” means sampling 512 nodes at each layer and “LADIES(5000)” means sampling 5000 nodes in each layer. In GNS, the size of cached was 1% of $|\mathcal{V}|$. LazyGCN runs out of GPU memory on OAG-paper and OGBN-papers100M.

**Figure 2: Runtime breakdown (s) of each component in mini-batch training of NS and GNS on OGBN-products and OAG-paper graphs.**

in the first layer under different numbers of layerwise sampled nodes is illustrated in Tables 5.

LazyGCN requires a large batch size to train GCN on large graphs, which usually leads to out of memory. Under the same setting of the paper [11], we investigate the performance of nodewise lazyGCN on Yelp dataset with different mini-batch sizes. As shown in Figure 4, LazyGCN performs poorly at the small mini-batch size. This may

Table 4: The average number of input nodes in a mini-batch of NS and GNS as well as the average number of input nodes from the cache of GNS.

	#input nodes (NS)	#input nodes (GNS)	#cached nodes (GNS)
Yelp	151341	24150	5796
Amazon	132288	19063	13986
OGBN-products	433928	88137	21552
OAG-paper	408854	102984	56422
OGBN-Papers100M	507405	155128	111923

Table 5: Percentage of isolated training nodes in LADIES.

# of sampled nodes/layer	256	512	1000	5000	10000
% of isolated target nodes	52.7	45.2	24.0	3.9	0

Percentage of isolated nodes in the first layer when training three-layer GCN on OGBN-products with LADIES.

be caused by training with less representative graph data in mega-batch when recycling a small batch.

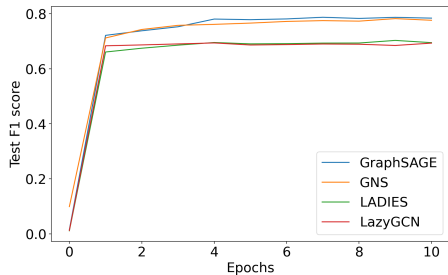


Figure 3: Comparison of the accuracy (F1 score) v.s. epochs.

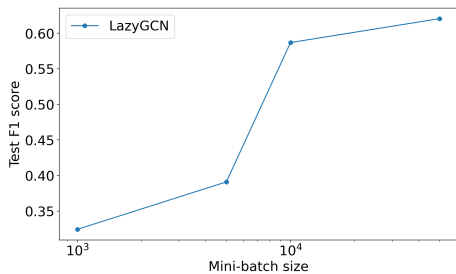


Figure 4: The effect of mini-batch size on the performance of LazyGCN on the Yelp dataset.

Table 6: GNS sensitivity to update period and cache size.

Size of cache	cache update period size P			
	$P = 1$	$P = 2$	$P = 5$	$P = 10$
$ \mathcal{V} \times 1\%$	78.34	78.40	78.17	77.54
$ \mathcal{V} \times .1\%$	78.04	77.31	76.16	74.71
$ \mathcal{V} \times .01\%$	76.29	72.83	71.60	71.21

Performance in terms of test-set F1-score for different cache sizes and update periods.

4.3 Hyperparameter study

In this section, we explore the effect of various parameters in GNS on OGBN-products dataset. Table 6 summarizes the test F1-score with respect to different values of cache update period sizes P and cache sizes. A cache size as small as 0.01% can still achieve a fairly good accuracy. As long as the cache size is sufficiently large (e.g., $1\% \cdot |\mathcal{V}|$), properly reducing the frequency of updating the cache (e.g., $P = 1, 2, 5$) does not affect performance. Note that it is better to get a .1% sample every epoch, than a single 1% sample every 10 epochs, and .01% sample every epoch that the 0.1% sample every 10 epochs.

5 CONCLUSIONS

In this paper, we propose a new effective sampling framework to accelerate GNN mini-batch training on giant graphs by removing the main bottleneck in mixed CPU-GPU training. GNS creates a global

cache to facilitate neighbor sampling and periodically updates the cache. Therefore, it reduces data movement between CPU and GPU. We empirically demonstrate the advantages of the proposed algorithm in convergence rate, computational time, and scalability to giant graphs. Our proposed method has a significant speedup in training on large-scale dataset. We also theoretically analyze GNS and show that even with a small cache size, it enjoys a comparable convergence rate as the node-wise sampling method.

REFERENCES

- [1] Jie Chen, Tengfei Ma, and Cao Xiao. 2018. FastGCN: Fast Learning with Graph Convolutional Networks via Importance Sampling. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. <https://openreview.net/forum?id=rytstxWAW>
- [2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *abs/1810.04805* (2019).
- [3] Matthias Fey and Jan Eric Lenssen. 2019. Fast Graph Representation Learning with PyTorch Geometric. *CoRR* abs/1903.02428 (2019).
- [4] William L. Hamilton, Zitao Ying, and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*. 1025–1035. <http://papers.nips.cc/paper/6703-inductive-representation-learning-on-large-graphs>
- [5] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [6] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2020. Open graph benchmark: Datasets for machine learning on graphs. *arXiv preprint arXiv:2005.00687* (2020).
- [7] Thomas N. Kipf and Max Welling. [n.d.]. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations (ICLR-17)*.
- [8] Jonas Moritz Kohler and Aurelien Lucchi. 2017. Sub-sampled Cubic Regularization for Non-convex Optimization. In *International Conference on Machine Learning*. 1895–1904.
- [9] Ziqi Liu, Zhengwei Wu, Zhiqiang Zhang, Jun Zhou, Shuang Yang, Le Song, and Yuan Qi. 2020. Bandit Samplers for Training Graph Neural Networks. *abs/2006.05806* (2020).
- [10] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 652–660.
- [11] Morteza Ramezani, Weilin Cong, Mehrdad Mahdavi, Anand Sivasubramanian, and Mahmut Kandemir. 2020. GCN meets GPU: Decoupling “When to Sample” from “How to Sample”. *Advances in Neural Information Processing Systems* 33 (2020).
- [12] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnet-Miner: Extraction and Mining of Academic Social Networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Las Vegas, Nevada, USA). New York, NY, USA.
- [13] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. [n.d.]. Graph Attention Networks. In *6th International Conference on Learning Representations (ICLR-18)*.
- [14] Minjie Wang, Da Zheng, Zihao Ye, Quan Gan, Mufei Li, Xiang Song, Jinjing Zhou, Chao Ma, Lingfan Yu, Yu Gai, Tianjun Xiao, Tong He, George Karypis, Jinyang Li, and Zheng Zhang. 2019. Deep Graph Library: A Graph-Centric, Highly-Performant Package for Graph Neural Networks. *arXiv preprint arXiv:1909.01315* (2019).
- [15] Hanqing Zeng, Hongkuan Zhou, Ajitesh Srivastava, Rajgopal Kannan, and Viktor Prasanna. 2019. Graphsaint: Graph sampling based inductive learning method. *arXiv preprint arXiv:1907.04931* (2019).
- [16] Da Zheng, Chao Ma, Minjie Wang, Jinjing Zhou, Qidong Su, Xiang Song, Quan Gan, Zheng Zhang, and George Karypis. 2020. DistDGL: Distributed Graph Neural Network Training for Billion-Scale Graphs. *arXiv preprint arXiv:2010.05337* (2020).
- [17] Difan Zou, Ziniu Hu, Yewen Wang, Song Jiang, Yizhou Sun, and Quanquan Gu. 2019. Layer-dependent importance sampling for training deep and large graph convolutional networks. In *Advances in Neural Information Processing Systems*. 11249–11259.

A PROOF OF THEOREM 1

By extending Lemma 4 in [11], we conclude that $\mathbb{E}[\|\nabla J(\hat{\boldsymbol{\theta}})\|^2]$ depends on $\frac{1}{T} \sum_{t=1}^T \mathbb{E}[\|\nabla \tilde{J}_{\mathcal{B}}(\boldsymbol{\theta}_t) - \nabla J(\boldsymbol{\theta}_t)\|^2]$. To complete the proof of theorem 1, we first present few vital lemmas. To be specific, Lemma 1 characterizes the bounds on $\frac{1}{T} \sum_{t=1}^T \mathbb{E}[\|\nabla \tilde{J}_{\mathcal{B}}(\boldsymbol{\theta}_t) - \nabla J(\boldsymbol{\theta}_t)\|^2]$.

LEMMA 1. Denote N_ℓ^i as the number of the neighborhood nodes with respect to $i \in \mathcal{V}$ sampling uniformly at random at ℓ -th layer. The cached nodes in the set \mathcal{C} with the size of $|\mathcal{C}|$ are sampled without replacement according to p_v^{cache} . The dimension of node feature is denoted as n and the size of min-batch is denoted as B . Define $\tilde{C} = |\mathcal{C}|/|\mathcal{V}|$ and $C_d = \sum_{v_i \in \mathcal{V}} \deg(v_i)/|\mathcal{V}|$ with the constant $c > 0$. Under Assumption 1, the expected mean-square error of stochastic gradient $\nabla \tilde{J}_{\mathcal{B}}(\boldsymbol{\theta})$ derived from Algorithm 1 to the full gradient is bounded by

$$\begin{aligned} \text{MSE} &:= \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[\left\| \nabla \tilde{J}_{\mathcal{B}}(\boldsymbol{\theta}_t) - \nabla J(\boldsymbol{\theta}_t) \right\|^2 \right] \\ &\leq \mathcal{O} \left(L_f'^2 \frac{\log(4n/\delta) + 1/2}{B} \right) + \mathcal{O} \left(L_f'^2 L_g^4 \frac{\log(4n/\delta) + 1/2}{c\tilde{C}C_d N_1^j N_2^i} \right) \\ &\quad + \mathcal{O} \left(L_g'^2 L_f^2 \frac{\log(4n/\delta)}{c\tilde{C}C_d N_1^j N_2^i} \right). \end{aligned} \quad (\text{A.1})$$

PROOF. According to the inequality, $\|\mathbf{a} + \mathbf{b}\|^2 \leq 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2$, MSE (A.1) can be decomposed into two parts:

$$\begin{aligned} \text{MSE} &:= \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[\left\| \nabla \tilde{J}_{\mathcal{B}}(\boldsymbol{\theta}_t) - \nabla J(\boldsymbol{\theta}_t) \right\|^2 \right] \\ &\leq \frac{2}{T} \sum_{t=1}^T \mathbb{E} \left[\left\| \nabla \tilde{J}_{\mathcal{B}}(\boldsymbol{\theta}_t) - \nabla J_{\mathcal{B}}(\boldsymbol{\theta}_t) \right\|^2 \right] \\ &\quad + \frac{2}{T} \sum_{t=1}^T \mathbb{E} \left[\left\| \nabla J_{\mathcal{B}}(\boldsymbol{\theta}_t) - \nabla J(\boldsymbol{\theta}_t) \right\|^2 \right] \end{aligned} \quad (\text{A.2})$$

Two terms in (A.2) are bounded by Lemma 2 and 3, respectively. \square

LEMMA 2. Based on the notations in Lemma 1, with probability exceeding $1 - \delta$ we have

$$\begin{aligned} &\mathbb{E}[\|\nabla \tilde{J}_{\mathcal{B}}(\boldsymbol{\theta}) - \nabla J_{\mathcal{B}}(\boldsymbol{\theta})\|^2] \\ &\leq 128L_f'^2 L_g^4 \frac{\log(4n/\delta) + 1/2}{c\tilde{C}C_d N_1^j N_2^i} + 64L_g'^2 L_f^2 \frac{\log(4n/\delta)}{c\tilde{C}C_d N_1^j N_2^i}. \end{aligned} \quad (\text{A.3})$$

PROOF. The proof of Lemma 2 is provided in Appendix B. \square

LEMMA 3. Based on the notations in Lemma 1, with probability exceeding $1 - \delta$ we have

$$\mathbb{E}[\|\nabla J_{\mathcal{B}}(\boldsymbol{\theta}) - \nabla J(\boldsymbol{\theta})\|^2] \leq 128L_f'^2 \frac{\log(4n/\delta) + 1/2}{B}. \quad (\text{A.4})$$

PROOF. Besides the definition of $\nabla J_{\mathcal{B}}(\boldsymbol{\theta})$ (B.1), $\nabla J(\boldsymbol{\theta})$ is given as

$$\nabla J(\boldsymbol{\theta}) = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} A_1^i A_2^i, \quad (\text{A.5})$$

where A_1^i, A_2^i are represented by (B.2) and (B.3), respectively.

For simplicity, we denote

$$\mathbb{E}_{\mathcal{V}_{\mathcal{B}} \sim \mathcal{V}} \left[\mathbb{E}_{j \sim \mathcal{N}(i), \forall i \in \mathcal{V}_{\mathcal{B}}} \left[\mathbb{E}_{k \sim \mathcal{N}(j), \forall j \in \mathcal{N}(i)} [\cdot] \right] \right]$$

as $\mathbb{E}[\cdot]$.

$$\mathbb{E}[\|\nabla J_{\mathcal{B}}(\boldsymbol{\theta}) - \nabla J(\boldsymbol{\theta})\|^2] \quad (\text{A.6})$$

$$\leq \mathbb{E} \left[\left\| \frac{1}{B} \sum_{i \in \mathcal{V}_{\mathcal{B}}} A_1^i A_2^i - \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} A_1^i A_2^i \right\|^2 \right], \quad (\text{A.7})$$

$$\leq 128L_f'^2 \frac{\log(4n/\delta) + 1/2}{c\tilde{C}C_d N_1^j N_2^i}, \quad (\text{A.8})$$

where the last inequality comes from Lemma 4. \square

B PROOF OF LEMMA 2

To bound $\mathbb{E}[\|\nabla \tilde{J}_{\mathcal{B}}(\boldsymbol{\theta}) - \nabla J_{\mathcal{B}}(\boldsymbol{\theta})\|^2]$, we begin with the definition of $\nabla \tilde{J}_{\mathcal{B}}(\boldsymbol{\theta})$ and $\nabla J_{\mathcal{B}}(\boldsymbol{\theta})$, which is given by

$$\nabla J_{\mathcal{B}}(\boldsymbol{\theta}) = \frac{1}{B} \sum_{i \in \mathcal{V}_{\mathcal{B}}} A_1^i A_2^i, \quad (\text{B.1})$$

where

$$A_1^i = \nabla f_i \left(\frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \frac{1}{|\mathcal{N}(j)|} \sum_{k \in \mathcal{N}(j)} g_{jk}(\boldsymbol{\theta}) \right), \quad (\text{B.2})$$

$$A_2^i = \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \frac{1}{|\mathcal{N}(j)|} \sum_{k \in \mathcal{N}(j)} \nabla g_{jk}(\boldsymbol{\theta}). \quad (\text{B.3})$$

$$\nabla \tilde{J}_{\mathcal{B}}(\boldsymbol{\theta}) = \frac{1}{B} \sum_{i \in \mathcal{V}_{\mathcal{B}}} B_1^i B_2^i \quad (\text{B.4})$$

where

$$B_1^i = \nabla f_i \left(\frac{1}{|\mathcal{N}_2^C(i)|} \sum_{j \in \mathcal{N}_2^C(i)} \frac{1}{|\mathcal{N}_1^C(j)|} \sum_{k \in \mathcal{N}_1^C(j)} p_k^{(1)} g_{jk}(\boldsymbol{\theta}) \right), \quad (\text{B.5})$$

$$B_2^i = \frac{1}{|\mathcal{N}_2^C(i)|} \sum_{j \in \mathcal{N}_2^C(i)} \frac{1}{|\mathcal{N}_1^C(j)|} \sum_{k \in \mathcal{N}_1^C(j)} p_k^{(1)} \nabla g_{jk}(\boldsymbol{\theta}). \quad (\text{B.6})$$

For simplicity, we denote

$$\mathbb{E}_{\mathcal{V}_{\mathcal{B}} \sim \mathcal{V}} \left[\mathbb{E}_{j \sim \mathcal{N}(i), \forall i \in \mathcal{V}_{\mathcal{B}}} \left[\mathbb{E}_{k \sim \mathcal{N}(j), \forall j \in \mathcal{N}(i)} [\cdot] \right] \right]$$

as $\mathbb{E}[\cdot]$.

Based on the inequalities $\left\| \frac{1}{n} \sum_{i=1}^n \mathbf{a}_i \right\| \leq \frac{1}{n} \sum_{i=1}^n \|\mathbf{a}_i\|$, $\|\mathbf{a} + \mathbf{b}\|^2 \leq 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2$, and $\|\mathbf{a}\mathbf{b}\| \leq \|\mathbf{a}\|\|\mathbf{b}\|$, we arrive

$$\begin{aligned} &\mathbb{E}[\|\nabla \tilde{J}_{\mathcal{B}}(\boldsymbol{\theta}) - \nabla J_{\mathcal{B}}(\boldsymbol{\theta})\|^2] \\ &= 2\mathbb{E} \left[\left\| B_1^i \right\|^2 \right] \mathbb{E} \left[\left\| B_2^i - A_2^i \right\|^2 \right] \\ &\quad + 2\mathbb{E} \left[\left\| B_1^i - A_1^i \right\|^2 \right] \mathbb{E} \left[\left\| A_2^i \right\|^2 \right]. \end{aligned} \quad (\text{B.7})$$

We shall bound two terms in (B.7).

(1) The first term: for $\mathbb{E} \left[\left\| B_1^i \right\|^2 \right]$, we have

$$\mathbb{E} \left[\left\| B_1^i \right\|^2 \right] \leq L_f^2. \quad (\text{B.8})$$

In terms of $\mathbb{E} \left[\|B_2^i - A_2^i\|^2 \right]$, we have

$$\begin{aligned} & \mathbb{E} \left[\|B_2^i - A_2^i\|^2 \right] \\ &= L_f'^2 \mathbb{E} \left[\left\| \sum_{j \in \mathcal{N}_2^C(i)} \sum_{k \in \mathcal{N}_1^C(j)} \frac{p_k^{(1)}}{N_{C_2}^i N_{C_1}^j} \nabla g_{jk}(\boldsymbol{\theta}) \right. \right. \\ & \quad \left. \left. - \sum_{j \in \mathcal{N}(i)} \sum_{k \in \mathcal{N}(j)} \frac{1}{(N^i)^2} \nabla g_{jk}(\boldsymbol{\theta}) \right\|^2 \right] \\ &\leq 64L_g'^2 \frac{\log(4n/\delta)}{c\tilde{C}C_d N_1^j N_2^i}, \end{aligned} \quad (\text{B.9})$$

where the last inequality comes from Lemma 4.

(2) The second term: for $\mathbb{E} \left[\|A_2^i\|^2 \right]$, we have

$$\mathbb{E} \left[\|A_2^i\|^2 \right] \leq L_g^2. \quad (\text{B.10})$$

In terms of $\mathbb{E} \left[\|B_1^i - A_1^i\|^2 \right]$, we have

$$\begin{aligned} & \mathbb{E} \left[\|B_1^i - A_1^i\|^2 \right] \\ &= L_f'^2 \mathbb{E} \left[\left\| \sum_{j \in \mathcal{N}_2^C(i)} \sum_{k \in \mathcal{N}_1^C(j)} \frac{p_k^{(1)}}{N_{C_2}^i N_{C_1}^j} g_{jk}(\boldsymbol{\theta}) \right. \right. \\ & \quad \left. \left. - \sum_{j \in \mathcal{N}(i)} \sum_{k \in \mathcal{N}(j)} \frac{1}{(N^i)^2} g_{jk}(\boldsymbol{\theta}) \right\|^2 \right] \\ &\leq 128L_g^2 L_f'^2 \frac{\log(4n/\delta) + 1/2}{c\tilde{C}C_d N_1^j N_2^i}, \end{aligned} \quad (\text{B.11})$$

where the last inequality comes from Lemma 4.

By integrating inequalities (B.8), (B.11), (B.10), (B.9), it yields

$$\begin{aligned} & \mathbb{E} [\|\nabla \tilde{J}_{\mathcal{B}}(\boldsymbol{\theta}) - \nabla J_{\mathcal{B}}(\boldsymbol{\theta})\|^2] \\ &\leq 128L_f'^2 L_g^4 \frac{\log(4n/\delta) + 1/2}{c\tilde{C}C_d N_1^j N_2^i} + 64L_g'^2 L_f^2 \frac{\log(4n/\delta)}{c\tilde{C}C_d N_1^j N_2^i}. \end{aligned} \quad (\text{B.12})$$

LEMMA 4. *Based on the notations in Lemma 1, with probability $1 - \delta$ we have*

$$\begin{aligned} & \mathbb{E} \left[\left\| \sum_{j \in \mathcal{N}_2^C(i)} \sum_{k \in \mathcal{N}_1^C(j)} \frac{p_k^{(1)}}{N_{C_2}^i N_{C_1}^j} g_{jk}(\boldsymbol{\theta}) \right. \right. \\ & \quad \left. \left. - \mathbb{E}_{j \sim \mathcal{N}(i), k \sim \mathcal{N}(j)} [g_{jk}(\boldsymbol{\theta})] \right\|^2 \right] \\ &\leq 8\sqrt{2}L_g \sqrt{\frac{\log(4n/\delta) + 1/2}{c\tilde{C}C_d N_1^j N_2^i}} \quad \forall i \in \mathcal{V} \quad (\text{B.13}) \\ & \mathbb{E} \left[\left\| \sum_{j \in \mathcal{N}_2^C(i)} \sum_{k \in \mathcal{N}_1^C(j)} \frac{p_k^{(1)}}{N_{C_2}^i N_{C_1}^j} \nabla g_{jk}(\boldsymbol{\theta}) \right. \right. \\ & \quad \left. \left. - \mathbb{E}_{j \sim \mathcal{N}(i), k \sim \mathcal{N}(j)} [\nabla g_{jk}(\boldsymbol{\theta})] \right\|^2 \right] \end{aligned}$$

$$\leq 8L_g' \sqrt{\frac{\log(4n/\delta)}{c\tilde{C}C_d N_1^j N_2^i}} \quad \forall i \in \mathcal{V}, \quad (\text{B.14})$$

where $C_d = \sum_{v_i \in \mathcal{V}} \deg(v_i) / |\mathcal{V}|$ with the constant $c > 0$.

PROOF. The proof is derived from the proof of Lemma 6 in [8] and Lemma 10 and 11 in [11]. Based on the definition of $p_k^{(1)}$, for $i \in \mathcal{V}$, we have

$$\begin{aligned} & \mathbb{E} \left[\sum_{j \in \mathcal{N}_2^C(i)} \sum_{k \in \mathcal{N}_1^C(j)} \frac{p_k^{(1)}}{N_{C_2}^i N_{C_1}^j} g_{jk}(\boldsymbol{\theta}) \right] \\ &= \mathbb{E}_{j \sim \mathcal{N}(i), k \sim \mathcal{N}(j)} [g_{jk}(\boldsymbol{\theta})]. \end{aligned} \quad (\text{B.15})$$

Similarly, there is

$$\begin{aligned} & \mathbb{E} \left[\sum_{j \in \mathcal{N}_2^C(i)} \sum_{k \in \mathcal{N}_1^C(j)} \frac{p_k^{(1)}}{N_{C_2}^i N_{C_1}^j} \nabla g_{jk}(\boldsymbol{\theta}) \right] \\ &= \mathbb{E}_{j \sim \mathcal{N}(i), k \sim \mathcal{N}(j)} [\nabla g_{jk}(\boldsymbol{\theta})]. \end{aligned} \quad (\text{B.16})$$

The value of $N_{C_1}^j$ and $N_{C_2}^j$ depend on the predefined number of neighborhood nodes sampled uniformly at random in each layer, i.e., N_2^i, N_1^j , the ratio of cached nodes to the whole graph nodes, i.e., $\tilde{C} = |C| / |\mathcal{V}|$ and the sampling probability \mathcal{P} . Thus, $N_{C_1}^j$ can be approximated by $N_{C_1}^j = c\tilde{C}C_d N_1^j$ where $C_d = \sum_{v_i \in \mathcal{V}} \deg(v_i) / |\mathcal{V}|$ with the constant $c > 0$. Based on the above, the proof can be completed by extending the proof of Lemma 10 and 11 in [11].

Given the sampled set $\mathcal{S}_1, \mathcal{S}_2$ and

$$V_{\mathcal{S}}(\boldsymbol{\theta}) = \frac{1}{|\mathcal{S}_1| \cdot |\mathcal{S}_2|} \sum_{i \in \mathcal{S}_1} \sum_{j \in \mathcal{S}_2} V_{ij}(\boldsymbol{\theta}),$$

where $V_{ij}(\boldsymbol{\theta}) \in \mathbb{R}^n$ is L_v -Lipschitz continuous for all i, j , the proof can be completed via Bernstein's bound with a sub-Gaussian tail, given by we have

$$\begin{aligned} & \mathbb{P} (\|V_{\mathcal{S}}(\boldsymbol{\theta}) - \mathbb{E}[V_{\mathcal{S}}(\boldsymbol{\theta})]\| \geq \epsilon) \\ &\leq 4n \cdot \exp \left(-\frac{\epsilon^2 |\mathcal{S}_1| \cdot |\mathcal{S}_2|}{64L_v^2} + \frac{1}{2} \right), \end{aligned} \quad (\text{B.17})$$

where $\epsilon \leq 2L_v$.

Finally, let δ as the upper bound of Bernstein inequality

$$\delta = 4n \cdot \exp \left(-\frac{\epsilon^2 |\mathcal{S}_1| \cdot |\mathcal{S}_2|}{64L_v^2} + \frac{1}{2} \right). \quad (\text{B.18})$$

Therefore, we have

$$\epsilon = 8\sqrt{2}L_v \sqrt{\frac{\log(4n/\delta) + 1/2}{|\mathcal{S}_1| \cdot |\mathcal{S}_2|}}. \quad (\text{B.19})$$

The inequality (B.14) can be clarified in similar way. \square