# Cooperative Minibatching in Graph Neural Networks

**Muhammed Fatih Balın**[*] *balin@gatech.edu*
*School of Computational Science and Engineering*
*Georgia Institute of Technology, Atlanta, GA, USA*

**Dominique LaSalle** *dlasalle@nvidia.com*
*NVIDIA Corporation, Santa Clara, CA, USA*

**Ümit V. Çatalyürek**[†] *umit@gatech.edu*
*School of Computational Science and Engineering*
*Georgia Institute of Technology, Atlanta, GA, USA*

**Reviewed on OpenReview:** *https://openreview.net/forum?id=f6yMdmrD2g*

## Abstract

Training large scale Graph Neural Networks (GNNs) requires significant computational resources, and the process is highly data-intensive. One of the most effective ways to reduce resource requirements is minibatch training coupled with graph sampling. GNNs have the unique property that items in a minibatch have overlapping data. However, the commonly implemented Independent Minibatching approach assigns each Processing Element (PE, i.e., cores and/or GPUs) its own minibatch to process, leading to duplicated computations and input data access across PEs. This amplifies the Neighborhood Explosion Phenomenon (NEP), which is the main bottleneck limiting scaling. To reduce the effects of NEP in the multi-PE setting, we propose a new approach called Cooperative Minibatching. Our approach capitalizes on the fact that the size of the sampled subgraph is a concave function of the batch size, leading to significant reductions in the amount of work as batch sizes increase. Hence, it is favorable for processors equipped with a fast interconnect to work on a large minibatch together as a single larger processor, instead of working on separate smaller minibatches, even though global batch size is identical. We also show how to take advantage of the same phenomenon in serial execution by generating dependent consecutive minibatches. Our experimental evaluations show up to 4x bandwidth savings for fetching vertex embeddings, by simply increasing this dependency without harming model convergence. Combining our proposed approaches, we achieve up to 64% speedup over Independent Minibatching on single-node multi-GPU systems, using same resources.

## 1 Introduction

Graph Neural Networks (GNNs) have become de facto deep learning models for unstructured data, achieving state-of-the-art results on various application domains involving graph data such as recommendation systems (Wu et al., 2020; Ying et al., 2018), fraud detection (Liu et al., 2022; Patel et al., 2022), identity resolution (Xu et al., 2019), and traffic forecasting (Jiang & Luo, 2022). However, as the usage of technology continues to increase, the amount of data generated by these applications is growing exponentially, resulting in large and complex graphs that are infeasible or too time-consuming to train on a single processing element (Ying et al., 2018; Zhu et al., 2019). For example, some social media graphs are reaching billions of vertices and trillions of interactions (Ching et al., 2015). Efficient distributed training of GNNs is essential for extracting value from large-scale unstructured data that exceeds the cost of storing and processing such data.

[*]Part of this work was done during an internship at NVIDIA Corporation in the summer of 2022.
[†]Amazon Web Services. This publication describes work performed at the Georgia Institute of Technology and is not associated with AWS.

Due to the popularity of Deep Neural Networks (DNNs) and the need to support larger models and datasets, a great deal of research has focused on increasing the scale and efficiency of distributed DNN training. Techniques such as data parallelism (Ginsburg et al., 2017; Goyal et al., 2018), pipelining (Narayanan et al., 2019), and intra-layer parallelism (Dean et al., 2012) have been employed. Following the success of traditional distributed DNN training, the same techniques have also been adapted to GNN training, such as data parallelism (Gandhi & Iyer, 2021; Lin et al., 2020; Zheng et al., 2021; Zhu et al., 2019) and intra-layer parallelism (Tripathy et al., 2020).

The parallelization techniques mentioned earlier are used to scale both full-batch and minibatch training in a distributed setting. Minibatch training (Bertsekas, 1994) is the go-to method to train DNN models as it outperforms full-batch training in terms of convergence (Allen-Zhu & Hazan, 2016; Li et al., 2014; Keskar et al., 2016; Wilson & Martinez, 2003), and more recently has been shown to also offer the same benefit for GNNs (Zheng et al., 2021). In the distributed setting, minibatch training for DNNs using data parallelism is straightforward. The training samples are partitioned across the Processing Elements (PE) and they compute the forward/backward operations on their minibatches. The only communication required is an all-reduce operation for the gradients.

Unfortunately, minibatch training a GNN model is more challenging than a usual DNN model. GNNs turn a given graph encoding relationships into computational dependencies. Thus in an $L$-layer GNN model, each minibatch computation has a different structure as it is performed on the $L$-hop neighborhood of the minibatch vertices. Real-world graphs usually are power law graphs (Artico et al., 2020) with small diameters, thus, it is a challenge to train deep GNN models as the $L$-hop neighborhood grows exponentially w.r.t. $L$, reaching almost the whole graph within a few hops. This is especially the case for the node classification and link prediction scenarios where the graph can be large and a single connected component contains almost all the data.

Very large GNN datasets necessitate storing the graph and node embeddings on slower storage mediums. To enable GNN training efficiently in such a setting, several techniques have been proposed (Park et al., 2022; Waleffe et al., 2022). These studies assume that the graph and its features are stored on disks or SSDs and design their systems to reduce data transfers. The methods proposed in this paper directly apply to these settings by reducing bandwidth requirements, as seen in Section 4.

A single epoch of full-batch GNN training requires computation proportional to the number of layers $L$ and the size of the graph. However, minibatch training requires more operations to process a single epoch due to repeating calculations in the 2nd through $L$-th layers. As the batch size decreases, the number of repeated calculations increases. This is because the vertices and edges have to be processed each time they appear in the $L$-hop neighborhood. Thus, it is natural to conclude that using effectively larger batch sizes in GNNs reduces the number of computations and data accesses of an epoch in contrast to regular DNN models. Our contributions in this work utilizing this important observation can be listed as follows:

- Investigate work vs. batch size relationship and present theorems stating the cost of processing a minibatch is a concave function of the batch size (Theorems 3.1 and 3.2).
- Utilize this relationship by combining data and intra-layer parallelism to process a minibatch across multiple PEs for reduced work (Section 3.1), with identical global batch size. We call this new approach *Cooperative Minibatching*.
- Use the same idea to generate consecutive dependent minibatches to increase temporal vertex embedding access locality (Section 3.2). *Dependent Minibatching* can reduce the transfer amount of vertex embeddings up to 4×, without harming model convergence.
- Show that the two approaches are orthogonal. When combined, they result in up to 1.64× speedup over Independent Minibatching with identical global batch size.
- Further show the applicability of Cooperative Minibatching to the subgraph sampling methods and prove that sampled subgraph density is a nondecreasing function of the batch size (Theorem 3.3).

## 2 Background

A graph $\mathcal{G} = (V, E)$ consists of vertices $V$ and edges $E \subset V \times V$ along with optional edge weights $A_{ts} > 0, \forall (t \to s) \in E$. Given a vertex $s$, the 1-hop neighborhood $N(s)$ is defined as $N(s) = \{t | (t \to s) \in E\}$, and it can be naturally expanded to a set of vertices $S$ as $N(S) = \cup_{s \in S} N(s)$.

GNN models work by passing previous layer embeddings ($H$) from $N(s)$ to $s$, and then combining them using a nonlinear function $f^{(l)}$ at layer $l$, given initial vertex features $H^{(0)}$:

$$H_s^{(l+1)} = f^{(l)}(H_s^{(l)}, \{H_t^{(l)} \mid t \in N(s)\}) \tag{1}$$

If the GNN model has $L$ layers, then the loss is computed by taking the final layer $L$'s embeddings and averaging their losses over the set of training vertices $V_t \subset V$ for *full-batch* training. In $L$-layer full-batch training, the total number of vertices that needs to be processed is $L|V|$.

### 2.1 Minibatching in GNNs

In minibatch training, a random subset of training vertices, called *Seed Vertices*, is selected, and training is done over the (sampled) subgraph composed of $L$-hop neighborhood of the seed vertices. On each iteration, minibatch training computes the loss on seed vertices, which are random subsets of the training set $V_t$.

Given a set of vertices $S$, we define $l$-th layer expansion set, or the $l$-hop neighborhood $S^l$ as:

$$S^0 = S, \quad S^{(l+1)} = S^l \cup N(S^l) \tag{2}$$

For GNN computations, $S^l$ would also denote the set of the required vertices to compute (1) at each layer $l$. Using the same notation, $\{s\}^l$ denotes $l$-layer expansion set starting from single vertex $s \in V$.

For a single minibatch iteration, the total number of vertices that need to be processed is $\sum_{l=1}^{L} |S^l|$. There are $\frac{|V|}{|S^0|}$ minibatches assuming $V_t = V$. Since each $|S^l| \geq |S^0|$, and a single epoch of minibatch training needs to go over the whole dataset, the work $W(|S^0|)$ for a single epoch is:

$$W(|S^0|) = \frac{|V|}{|S^0|} \sum_{l=1}^{L} E[|S^l|] \geq \frac{|V|}{|S^0|} \sum_{l=1}^{L} |S^0| = L|V| \tag{3}$$

where $E[|S^l|]$ is the expected number of sampled vertices in layer $l$ and $|S^0|$ is the batch size. That is, the total amount of work to process a single epoch increases over full-batch training. The increase in work due to minibatch training is thus encoded in the ratios $\frac{E[|S^l|]}{|S^0|}, 1 \leq l \leq L$.

When sampling is used with minibatching, the minibatch subgraph may potentially become random. However, the same argument for the increasing total amount of work holds for them too, as seen in Figure 3.

### 2.2 Graph Sampling

We focus on samplers whose expected number of sampled vertices is a function of the batch size. Neighbor Sampling (NS) (Hamilton et al., 2017), LABOR (Balin & Çatalyürek, 2023) and RandomWalk (RW) Sampling (Ying et al., 2018) all have this property and they are all applied recursively for GNN models with multiple layers. Appendix A.1 describes the details of these sampling methods.

### 2.3 Independent Minibatching

Independent minibatching is commonly used in multi-GPU, and distributed GNN training frameworks (Cai et al., 2023; Gandhi & Iyer, 2021; Lin et al., 2020; Zheng et al., 2021; Zhu et al., 2019) to parallelize the training and allows scaling to larger problems. Each Processing Element (PE, e.g., GPUs, CPUs, or cores of multi-core CPU), starts with their own $S^0$ of size $b$ as the seed vertices, and compute $S^1, \ldots, S^L$ along

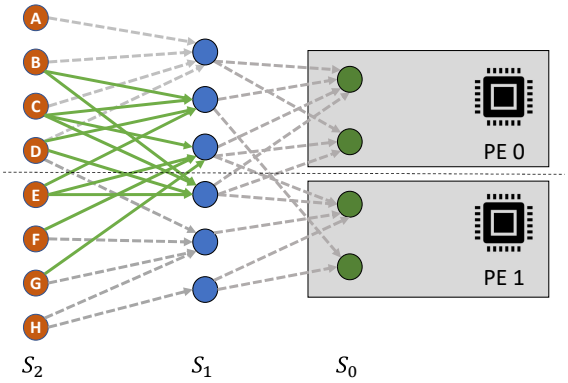



Figure 1: Minibatches of two processing elements may share edges in the second layer and vertices starting in the first layer. For independent minibatching, the solid green edges shared by both processing elements represent duplicate work, and input nodes B through G are duplicated along with the directed endpoints of green edges. However for cooperative minibatching, the vertices and edges are partitioned across the PEs with no duplication, and the edges crossing the line between the two PEs necessitate communication.

with the sampled edges to generate minibatches (see Figure 1). Computing $S^1, \ldots, S^L$ depends on the chosen sampling algorithm, such as the ones explained in Appendix A.1. It has the advantage that doing a forward/backward pass does not involve any communication with other PEs after the initial minibatch preparation stage at the expense of duplicate work.

## 3 Cooperative Minibatching

In this section, we present two theorems that show the work of an epoch will be monotonically nonincreasing with increasing batch sizes. After that, we propose two algorithms that can take advantage of this monotonicity. Due to lack of space, we provide their full proofs in Appendices A.2 and A.3.

**Theorem 3.1.** *The work per epoch* $\frac{E[|S^l|]}{|S^0|}$ *required to train a GNN model using minibatch training is monotonically nonincreasing as the batch size* $|S^0|$ *increases.*

*Proof.* (Idea) Given any random $S^0$ and $l \geq 1$, we have the corresponding set $S^l$. We can shrink $S^0$ by removing random elements $s \in S^0$ from it to get $S^0 - \{s\}$. If we compare $(S^0 - \{s\})^l$ to $S^l$, we will see that any difference will come from elements that are contained only in $\{s\}^l$. If we count such elements for each $s \in S^0$ and sum them, it will be bounded by $|S^l|$. More concretely if we let $S'^0 = S^0 - \{s\}$ be a random subset of $S^0$ with one element missing, we have:

$$\sum_{s \in S^0} |S^l| - |(S^0 - \{s\})^l| \leq |S^l| \iff |S^0||S^l| - |S^l| \leq \sum_{s \in S^0} |(S^0 - \{s\})^l| \iff |S^l||S'^0| \leq |S^0| E[|S'^l|] \quad (4)$$

□

Empirical evidence can be seen in Figures 3 and 6. In fact, the decrease is related to the cardinality of the following set:

$$T^l(S) = \{w \in S^l \mid w \in \{s\}^l, \exists! s \in S^0\} \quad (5)$$

When $T(S^0)$ is equal to $S^l$, the work is equal as well. In the next section, we further investigate the effect of $|T(S^0)|$ on $E[|S^l|]$.

In addition to the definition of $T(S)$ above, if we define the following set $T_2(S)$:

$$T_2^l(S) = \{w \in S^l \mid w \in \{s\}^l \cap \{s'\}^l, \exists! \{s, s'\} \subset S^0\} \quad (6)$$

**Theorem 3.2.** *The expected subgraph size* $E[|S^l|]$ *required to train a GNN model using minibatch training is a concave function of batch size,* $|S^0|$.

*Proof.* (Idea) Note that we have:

$$|T^l(S^0)| - 2|T_2^l(S^0)| = \sum_{\substack{S'^0 \subset S^0 \\ |S'^0|+1=|S^0|}} |T^l(S^0)| - |T^l(S'^0)| = |S^0||T^l(S^0)| - |S^0|E[|T^l(S'^0)|]$$

$$\implies (|S^0|-1)|T^l(S^0)| \leq |S^0|E[|T^l(S'^0)|] \implies \frac{|T^l(S^0)|}{|S^0|} \leq \frac{E[|T^l(S'^0)|]}{|S'^0|}$$

We have just shown that $\frac{E[|T^l(S_i^0)|]}{i}$ is monotonically nonincreasing as $i$ increases. If we let $S_i^0$ denote a random subset of $S^0$ of size $i$, we have:

$$E[|S^l|] = \sum_{i=1}^{|S^0|} \frac{E[|T^l(S_i^0)|]}{i} \tag{7}$$

As $E[|S^l|]$ is the sum of monotonically nonincreasing parts, we conclude that it is a concave function of the batch size, $|S^0|$. □

So, the slope of the expected number of sampled vertices flattens as batch size increases, see the last row in Figure 3 and the first row in Figure 6. Note that this implies work monotonicity as well.

### 3.1 Cooperative Minibatching

As explained in Section 2, Independent Minibatching can not take advantage of the reduction in work with increasing global batch sizes and number of PEs, because it uses separate small, *local*, batches of sizes $b$ on each PE for each step of training. On the other hand, one can also keep the *global* batch size constant, $bP = |S^0|$, and vary the number of processors $P$. As $P$ increases, Independent Minibatching will perform more and more duplicate work because the local batch size is a decreasing function, $b = \frac{|S^0|}{P}$, of $P$.

**Algorithm 1** Cooperative minibatching

**Input:** seed vertices $S_p^0$ for each PE $p \in P$, # layers $L$
**for all** $l \in \{0, \ldots, L-1\}$ **do** {Sampling}
  **for all** $p \in P$ **do in parallel**
    Sample next layer vertices $\tilde{S}_p^{l+1}$ and edges $E_p^l$ for $S_p^l$
    all-to-all to redistribute vertex ids for $\tilde{S}_p^{l+1}$ to get $S_p^{l+1}$
**for all** $p \in P$ **do in parallel** {Feature Loading}
  Load input features $H_p^L$ from Storage for vertices $S_p^L$
  all-to-all to redistribute $H_p^L$ to get $\tilde{H}_p^L$
**for all** $l \in \{L-1, \ldots, 0\}$ **do** {Forward Pass}
  **for all** $p \in P$ **do in parallel**
    **if** $l+1 < L$ **then**
      all-to-all to redistribute $H_p^{l+1}$ to get $\tilde{H}_p^{l+1}$
    Forward pass on bipartite graph $\tilde{S}_p^{l+1} \to S_p^l$ with edges $E_p^l$, input $\tilde{H}_p^{l+1}$ and output $H_p^l$
**for all** $p \in P$ **do in parallel**
  Compute the loss and initialize gradients $G_p^0$
**for all** $l \in \{0, \ldots, L-1\}$ **do** {Backward Pass}
  **for all** $p \in P$ **do in parallel**
    Backward pass on bipartite graph $S_p^l \to \tilde{S}_p^{(l+1)}$ with edges $E_p^l$, input $G_p^l$ and output $\tilde{G}_p^{l+1}$
    **if** $l+1 < L$ **then**
      all-to-all to redistribute $\tilde{G}_p^{l+1}$ to get $G_p^{l+1}$

Here, we propose the *Cooperative Minibatching* method that will take advantage of the work reduction with increasing batch sizes in multi-PE settings. In Cooperative Minibatching, a single global batch of size $bP$ will be processed by all the $P$ PEs in parallel, eliminating any redundant work across PEs.

We achieve this as follows: we first partition the graph in 1D fashion by logically assigning each vertex and its incoming edges to PEs as $V_p$ and $E_p$ for each PE $p$. Next, PE $p$ samples its batch seed vertices $S_p^l$ from the training vertices in $V_p$ for $l = 0$ of size $b$. Then using any sampling algorithm, PE $p$ samples the incoming edges $E_p^l$ from $E_p$ for its seed vertices. Each PE then computes the set of vertices sampled $\tilde{S}_p^{l+1} = \{t \mid (t \to s) \in E_p^l\}$. Note that, $\tilde{S}_p^{l+1}$ has elements residing on different PEs. The PEs exchange the vertex ids $\tilde{S}_p^{l+1}$ so that each PE receives the set $S_p^{l+1} \in V_p$. This process is repeated recursively for GNN models with multiple layers by using $S_p^{l+1}$ as the seed vertices for the next layer. The exchanged information is cached to be used during the forward/backward passes.

For the forward/backward passes, the same communication pattern used during cooperative sampling is used to send and receive input and intermediate layer embeddings before each GNN layer invocation. Algorithm 1 details cooperative sampling and cooperative forward/backward passes for a single GNN training iteration. Independent minibatching works the same except that it lacks the all-to-all operations and has $\tilde{A}_p^{l+1} = A_p^{l+1}$ for any given variable $A$ instead. The redistribution of vertices during sampling happens according to the initial graph partitioning and the rest of the redistribution operations follow the same communication pattern, always converting a variable $\tilde{A}_p^{l+1}$ into $A_p^{l+1}$ during the forward pass and $A_p^{l+1}$ into $\tilde{A}_p^{l+1}$ during sampling and the backward passes for any variable $A$. Cooperative Minibatching was first proposed in Balin et al. (2022; 2023) and was later re-discovered in Polisetty et al. (2023) as Cooperative Training, which was used during the feature loading and the forward/backward passes of their work while sampling of each minibatch was performed on a single CPU thread and the edges were split among the GPUs. We refer the reader to Appendix A.10 to see the relation between the approach proposed here and the work of Jia et al. (2020) on redundancy-free GCN aggregation.

Table 1: Algorithmic complexities of different stages of GNN training at layer $0 \leq l < L$ with $L$ total layers and batch size $B = |S^0|$ with $P$ PEs. Note that $|S_p^l(B)| = |S^l(B)|\frac{1}{P}$, $|E_p^l(B)| = |E^l(B)|\frac{1}{P}$ since the PEs process the partitioned subgraphs. Feature loading happens only at layer $L$.

| **Stage** | **Independent** | **Cooperative** |
|---|---:|---:|
| Sampling | $\mathcal{O}(\lvert S^l(\frac{B}{P})\rvert\frac{1}{\beta})$ | $\mathcal{O}(\lvert S_p^l(B)\rvert\frac{1}{\beta} + \lvert\tilde{S}_p^{l+1}(B)\rvert\frac{c}{\alpha})$ |
| Feature loading | $\mathcal{O}(\lvert S^L(\frac{B}{P})\rvert\frac{d\rho}{\beta})$ | $\mathcal{O}(\lvert S_p^L(B)\rvert\frac{d\rho}{\beta} + \lvert\tilde{S}_p^L(B)\rvert\frac{dc}{\alpha})$ |
| Forward/Backward | $\mathcal{O}(M(S^l(\frac{B}{P}), E^l(\frac{B}{P}), S^{l+1}(\frac{B}{P}))\frac{d}{\gamma})$ | $\mathcal{O}(M(S_p^l(B), E_p^l(B), \tilde{S}_p^{l+1}(B))\frac{d}{\gamma} + \lvert\tilde{S}_p^{l+1}(B)\rvert\frac{dc}{\alpha})$ |

**Complexity Analysis:** Let $M(V_1, E, V_2)$ denote the work to process a bipartite graph $V_2 \to V_1$ with edges $E$ for a given GNN model $M$. Assuming cross PE communication bandwidth $\alpha$, Storage (e.g., disk, network, or DRAM) to PE bandwidth as $\beta$ and PE memory bandwidth $\gamma$, and cache miss rate $\rho$, we have the time complexities given in Table 1 to perform different stages of GNN training per PE. We also use $d$ for embedding dimension and $c < 1$ for the cross edge ratio, note that $c \approx \frac{P-1}{P}$ for random partitioning, and smaller for smarter graph partitioning with $P$ standing for the number of PEs. Also the sizes of $\tilde{S}^l$ become smaller when graph partitioning is used due to increased overlap.

We see that $\gamma \approx 2$ TB/s , $\alpha \approx 300$ GB/s and $\beta \approx 30$ GB/s in today's modern multi-GPU systems (NVIDIA, 2020a). Due to $\alpha$ being relatively fast compared to $\frac{\gamma}{M}$ and $\beta$, our approach brings performance benefits. On newer systems, the all-to-all bandwidth continues to increase (NVIDIA, 2023), decreasing the cost of cooperation on a global mini-batch. However, on systems where the interconnect does not provide full bandwidth for all-to-all operations, our approach is limited in the speedup it can provide. Our approach is most applicable for systems with relatively fast all-to-all bandwidth $\frac{\alpha}{c}$ compared to $\frac{\gamma}{M}$ and $\beta$ and large $P$. In particular, starting from $P = 2$, cooperative starts to outperform independent even on F/B with the mag240M dataset and the R-GCN model in Section 4.3 and Tables 4 and 5. More discussion on this topic can be found in Appendix A.6.

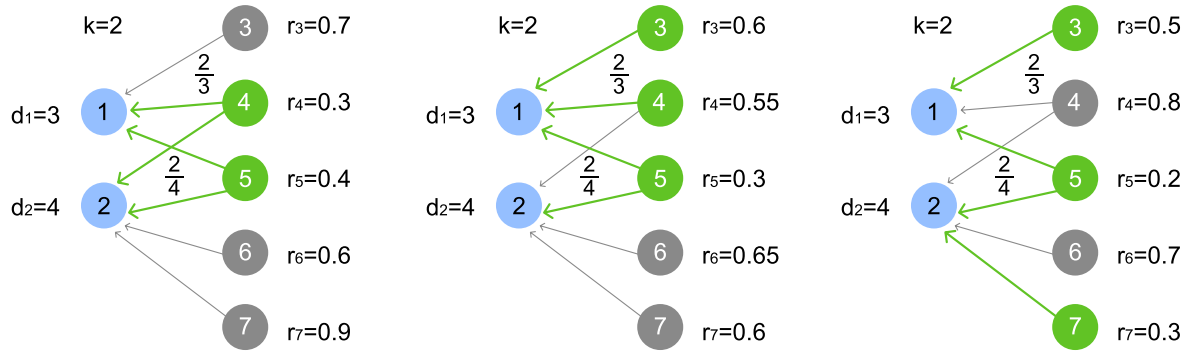



Figure 2: A smoothed dependent minibatching example for $\kappa = 2$. The middle minibatch is interpolated between the two independent minibatches on the left and the right by interpolating the random numbers used during sampling.

### 3.2 Cooperative Dependent Minibatching

Just as any parallel algorithm can be executed sequentially, we can reduce the number of distinct data accesses by having a single PE process $b$-sized parts of a single $\kappa b$-sized minibatch for $\kappa$ iterations. In light of Theorems 3.1 and 3.2, consider doing the following: choose $\kappa \in \mathbb{Z}^+$, then sample a batch $S^0$ of size $\kappa b$, i.e., $\kappa b = |S^0|$ to get $S^0, \ldots, S^L$. Then sample $\kappa$ minibatches $S_i^0$, of size $b = |S_i^0|$ from this batch of size $\kappa b$ to get $S_i^0, \ldots, S_i^L, \forall i \in \{0, \ldots, \kappa - 1\}$. In the end, all of the input features required for these minibatches will be a subset of the input features of the large batch, i.e. $S_i^j \subset S^j, \forall i, j$. This means that the collective input feature requirement of these $\kappa$ batches will be $|S^L|$, the same as our batch of size $\kappa b$. Hence, we can now take advantage of the concave growth of the work in Theorem 3.2 and Figure 3.

Note that, if one does not use any sampling algorithms and proceeds to use the full neighborhoods, this technique will not give any benefits, as by definition, the $l$-hop neighborhood of a batch of size $\kappa b$ will always be equal to the union of the $l$-hop neighborhoods of batches of sizes $b$. However for sampling algorithms, any overlapping vertex sampled by any two batches of sizes $b$ might end up with different random neighborhoods resulting in a larger number of sampled vertices. Thus, having a single large batch ensures that only a single random set of neighbors is used for any vertex processed over a period of $\kappa$ batches.

The approach described above has a nested iteration structure and the minibatches part of one $\kappa$ group will be significantly different than another group and this might slightly affect convergence. In Appendix A.7, we propose an alternative smoothing approach that does not require two-level nesting and still takes advantage of the same phenomenon for the NS and LABOR sampling algorithms.

The main idea of our smoothing approach is as follows: each time one samples the neighborhood of a vertex, normally it is done independently of any previous sampling attempts. If one were to do it fully dependently, then one would end up with an identical sampled neighborhood at each sampling attempt. What we propose is to do something inbetween, so that the sampled neighborhood of a vertex changes slowly over time. The speed of change in the sampled neighborhoods is $\frac{1}{\kappa}$, and after every $\kappa$ iterations, one gets fully independent new random neighborhoods for all vertices, see Figure 2. We will experimentally evaluate the locality benefits and the overall effect of this algorithm on convergence in Sections 4.2 and 4.3.1, and more details on our smoothing approach are discussed in Appendix A.7.

### 3.3 Relation to subgraph sampling methods

Existing subgraph sampling methods Chiang et al. (2019); Zeng et al. (2020); Hu et al. (2020b); Zeng et al. (2021); Fey et al. (2021); Shi et al. (2023) randomly sample a subset $S$ of the vertices and use the same subset for all layers of GNN training, $S = S^0 = \cdots = S^L$. The edges $S_E$ used for training are obtained by taking the vertex induced subgraph $S_E = \{t \to s \mid (t \to s) \in E \wedge t, s \in S\}$. We utilize the same observation as Theorems 3.1 and 3.2 and state:

**Theorem 3.3.** *The expected sampled subgraph density* $\frac{E[|S_E|]}{|S|}$ *is nondecreasing as the batch size* $|S|$ *increases.*

Thus, having multiple PEs cooperate and work on a larger minibatch together instead of processing independent smaller minibatches is theoretically expected to converge faster, as the higher density leads to better approximations. The proof of Theorem 3.3 is provided in Appendix A.4.

## 4 Experimental Evaluation

We first compare how the work to process an epoch changes w.r.t. to the batch size to empirically validate Theorems 3.1 and 3.2 for different graph sampling algorithms. Next, we show how dependent batches introduced in Section 3.2 benefits GNN training. We also show the runtime benefits of cooperative minibatching compared to independent minibatching in the multi-GPU setting. Finally, we show that these two techniques are orthogonal, can be combined to get multiplicative savings. Table 2 lists details of the datasets we used in experiments. Details on our experimental setup can be found in Appendix A.5.

Table 2: Traits of datasets used in experiments: numbers of vertices, edges, avg. degree, features, cached vertex embeddings, and training, validation and test vertex splits. Last column has # minibatches in an epoch during model training with 1024 batch size including validation.

| **Dataset** | $\lvert V\rvert$ | $\lvert E\rvert$ | $\frac{\lvert E\rvert}{\lvert V\rvert}$ | **# feats.** | **cache size** | **train - val - test (%)** | **# minibatches** |
|---|---|---|---|---|---|---|---|
| flickr | 89.2K | 900K | 10.09 | 500 | 70k | 50.00 - 25.00 - 25.00 | 65 |
| yelp | 717K | 14.0M | 19.52 | 300 | 200k | 75.00 - 10.00 - 15.00 | 595 |
| reddit | 233K | 115M | 493.56 | 602 | 60k | 66.00 - 10.00 - 24.00 | 172 |
| papers100M | 111M | 3.2B | 29.10 | 128 | 2M | 1.09 - 0.11 - 0.19 | 1300 |
| mag240M | 244M | 3.44B | 14.16 | 768 | 2M | 0.45 - 0.06 - 0.04 | 1215 |

### 4.1 Demonstrating monotonicity of work

We use three sampling approaches, NS, LABOR, and RW, to demonstrate that the work to process an epoch decreases as the batch size increases for the $L = 3$ layer case across these three different classes of sampling algorithms. We carried out our evaluation in two problem settings: node and edge prediction. For node prediction, a batch of training vertices is sampled with a given batch size. Then, the graph sampling algorithms described in Appendix A.1 are applied to sample the neighborhood of this batch. The top row of Figure 3 shows how many input vertices is required on average to process an epoch, specifically $\frac{E[|S^3|]}{|S^0|}$. For edge prediction, we add reverse edges to the graph making it undirected and sample a batch of edges. For each of these edges a random negative edge (an edge that is not part of $E$) with one endpoint coinciding with the positive edge is sampled. Then, all of the endpoints of these positive and negative edges are used as seed vertices to sample their neighborhoods. The bottom row of Figure 3 shows $E[|S^3|]$.

We can see that in all use cases, datasets and sampling algorithms, the work to process an epoch is monotonically decreasing as we proved in Theorem 3.1. We also see the plot of the expected number of vertices sampled, $E[|S^3|]$, is concave with respect to batch size as we already know from Theorem 3.2.

Another observation is that the concavity characteristic of $E[|S^3|]$ seems to differ for different sampling algorithms. In increasing order of concavity we have RW, NS, LABOR-0 and LABOR-*. The more concave a sampling algorithm's $E[|S^L|]$ curve is, the less it is affected from the NEP and more savings are available through the use of the proposed methods in Sections 3.1 and 3.2. Note that the differences would grow with a larger choice of layer count $L$.

### 4.2 Dependent Minibatches

We vary the batch dependency parameter $\kappa$ introduced in Section 3.2 for the LABOR-0 sampler with a batch size of 1024. Our expectation is that as consecutive batches become more dependent on each other,

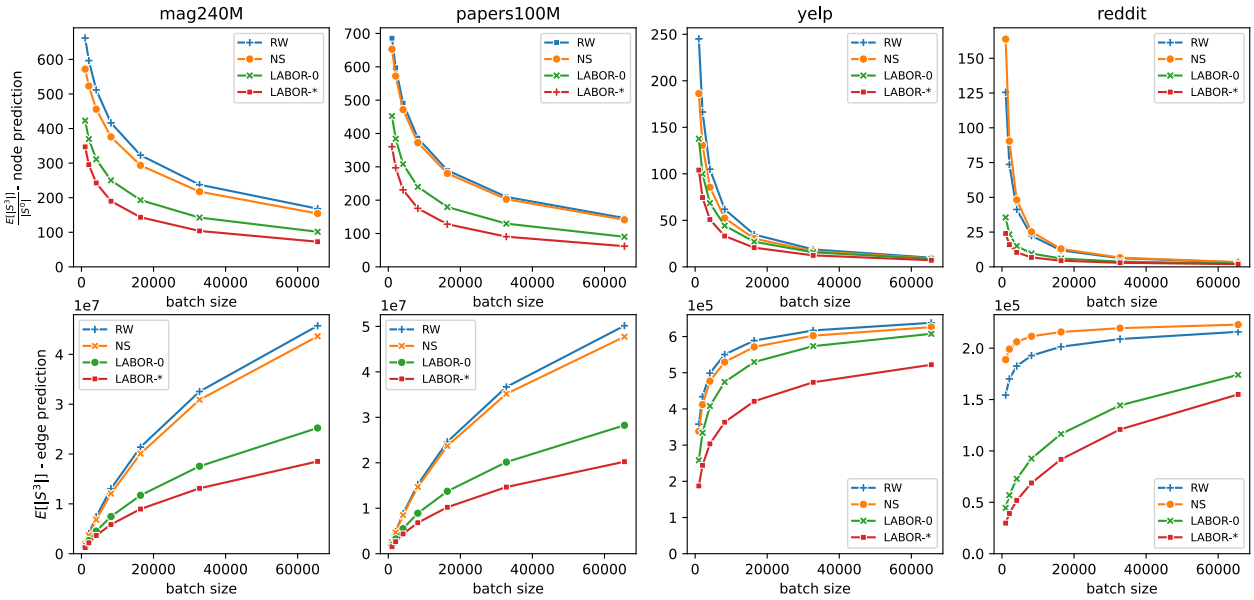



Figure 3: Monotonicity of the work. x-axis shows the batch size, y-axis shows $\frac{E[|S^3|]}{|S^0|}$ (see Theorem 3.1) for node prediction (top row) and $E[|S^3|]$ (see Theorem 3.2) for edge prediction (bottom row), where $E[|S^3|]$ denotes the expected number of sampled vertices in the 3rd layer and $|S^0|$ denotes the batch size. RW stands for Random Walks, NS for Neighbor Sampling, and LABOR-0/* for the two different variants of the LABOR sampling algorithm described in Section 2.2.

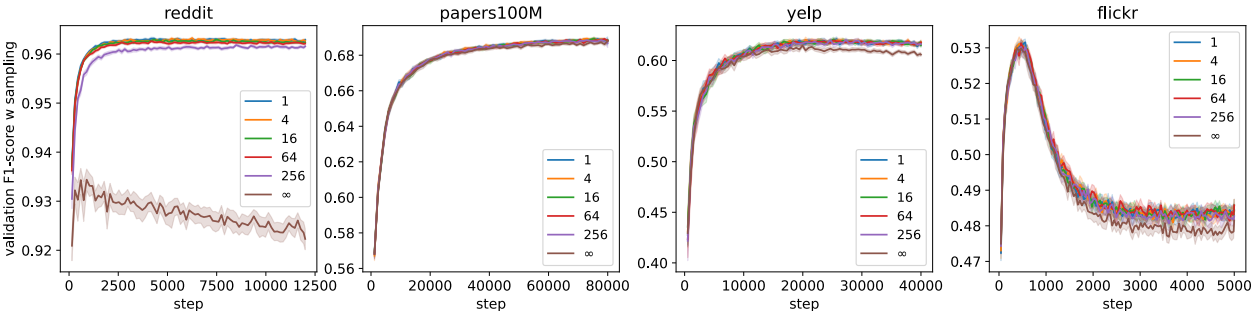
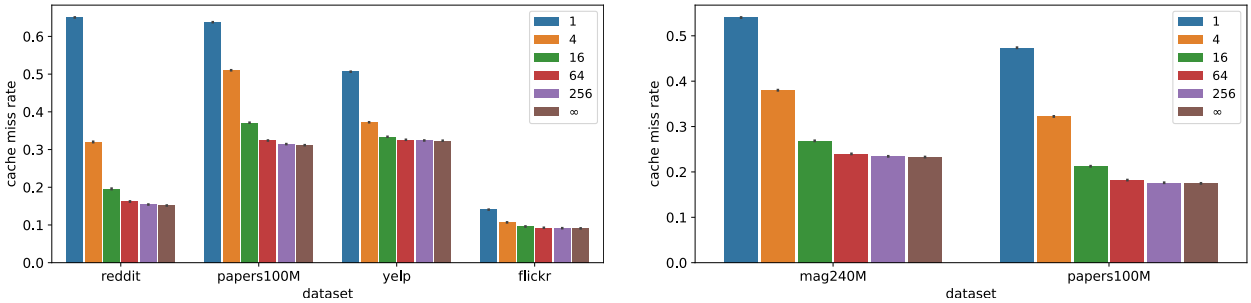



Figure 4: The validation F1-score with NS sampled neighborhoods trained with the LABOR-0 sampling algorithm with 1024 batch size and varying $\kappa$ dependent minibatches, $\kappa = \infty$ denotes infinite dependency, meaning the neighborhood sampled for a vertex stays static during training. See Figure 5a for cache miss rates. See Figure 8 for the training loss and F1-score with the dependent sampler.

the subgraphs used during consecutive steps of training would start overlapping with each other, in which case, the vertex embedding accesses would become more localized. We attempted to capture this increase in temporal locality in vertex embedding accesses by implementing an LRU cache to fetch them, the cache sizes used for different datasets is given in Table 2. Note that the cache miss rate is proportional to the amount of data that needs to be copied from the vertex embedding storage. The Figure 5a shows that as $\kappa$ increases, the cache miss rate across all datasets drops. On reddit, this is a drop from 64% to 16% on, a 4x improvement. We also observe that the improvement is monotonically increasing as a function of $\frac{|E|}{|V|}$ given in Table 2. Table 3 and Figure 4 show that training is not negatively affected across all datasets up to $\kappa = 256$ with less than 0.1% F1-score difference, after which point the validation F1-score with w/o sampling starts to diverge from the $\kappa = 1$ case. Runtime benefits of this approach can be observed by comparing the **Cache** and **Cache, $\boldsymbol{\kappa}$** columns in Table 4. Appendix A.8 has additional discussion about the effect of varying $\kappa$ and the last column of Table 2 shows the number of minibatches in an epoch during training.

Table 3: Test F1-scores at the highest validation F1-score corresponding to Figure 4, using early stopping. Averages of 40 runs are presented.

| $\kappa$ | reddit | papers100M | yelp | flickr |
|---|---|---|---|---|
| 1 | 96.72 ± 0.06 | 66.58 ± 0.17 | 63.40 ± 0.18 | 53.82 ± 0.21 |
| 4 | 96.71 ± 0.05 | 66.60 ± 0.13 | 63.39 ± 0.19 | 53.87 ± 0.18 |
| 16 | 96.70 ± 0.05 | 66.58 ± 0.20 | 63.42 ± 0.18 | 53.90 ± 0.17 |
| 64 | 96.67 ± 0.05 | 66.55 ± 0.20 | 63.35 ± 0.18 | 53.88 ± 0.20 |
| 256 | 96.65 ± 0.06 | 66.56 ± 0.14 | 63.35 ± 0.19 | 53.86 ± 0.24 |
| $\infty$ | 95.01 ± 0.30 | 66.46 ± 0.15 | 62.88 ± 0.18 | 53.84 ± 0.19 |




(a) 1 GPU, cache sizes are listed in Table 2.

(b) 4 cooperating GPUs, each having a cache of size 1M.

Figure 5: LRU-cache miss rates for LABOR-0 sampling algorithm with 1024 batch size per GPU and varying $\kappa$ dependent minibatches, $\kappa = \infty$ denotes infinite dependency.

### 4.3 Cooperative Minibatching

We use our largest datasets, mag240M and papers100M, as distributed training is motivated by large-scale datasets. We present our runtime results on systems equipped with NVIDIA GPUs, with 4 and 8 A100 80 GB (NVIDIA, 2021) and 16 V100 32GB (NVIDIA, 2020b), all with NVLink interconnect between the GPUs (600 GB/s for A100 and 300 GB/s for V100). The GPUs perform all stages of GNN training and the CPUs are only used to launch kernels for the GPUs. Feature copies are performed by GPUs as well, accessing pinned feature tensors over the PCI-e using zero-copy access. In cooperative minibatching, both data size and computational cost are shrinking with increasing numbers of GPUs, relative to independent minibatching. We use the GCN model for papers100M and the R-GCN model (Schlichtkrull et al., 2017) for mag240M. As seen in Table 4, cooperative minibatching reduces all the runtimes for different stages of GNN training, except for the F/B (forward/backward) times on papers100M where the computational cost is not high enough to hide the overhead of communication.

If we take the numbers in the **Total** columns from Table 4, divide independent runtimes by the corresponding cooperative ones, then we get Table 5. We can see that the theoretical decrease in work results in increasing speedup numbers with the increasing number of PEs, due to Theorem 3.1. We would like to point out that $\frac{E[|S^3|]}{|S^0|}$ curves in Figure 3 are responsible for these results. With $P$ PEs and $|S^0|$ global batch size, the work performed by independent minibatching vs cooperative minibatching can be compared by looking at $x = \frac{1}{P}|S^0|$ vs $x = |S^0|$ respectively.

We also ran experiments that show that graph partitioning using METIS (Karypis & Kumar, 1998) prior to the start of training can help the scenarios where communication overhead is significant. The forward-backward time decreases from 13.0ms to 12.0ms on papers100M with LABOR-0 on 4 NVIDIA A100 GPUs with partitioning due to reduced communication overhead using the same setup as Table 4.

Increasing the number of GPUs increases the advantage of cooperative minibatching compared to independent minibatching. The forward-backward time on mag240M with LABOR-0 is 200 (same as independent baseline), 194, 187 and 183 ms with 1, 2, 3 and 4 cooperating GPUs, respectively measured on the NVIDIA DGX

Table 4: Cooperative vs independent minibatching runtimes per minibatch (ms) on three different systems with 4 and 8 NVIDIA A100 80 GB GPUs, and 16 NVIDIA V100 32GB GPUs. I/C denotes whether independent or cooperative minibatching is used. Samp. is short for Graph Sampling, Feature Copy stands for vertex embedding copies over PCI-e and Cache denotes the runtime of copies performed with a cache that can hold $10^6$ vertex embeddings per A100 and $5 \times 10^5$ per V100. $\kappa$ denotes the use of batch dependency $\kappa = 64$. F/B means forward/backward. Total time is computed by the fastest available Feature Copy time, the sampling time, and the F/B time. $|S^0|$ is the global batch size and $b$ is the the batch size per GPU. $\alpha$ stands for cross GPU communication bandwidth (NVLink), $\beta$ for PCI-e bandwidth and $\gamma$ for GPU global memory bandwidth. Green was used to indicate the better result between independent and cooperative minibatching, while **Bold** was used to highlight the feature copy time included in the **Total** column.

| # GPUs, $\gamma$ $\alpha, \beta, \lvert S^0 \rvert$ | Dataset & Model | Sampler | I/C | Samp. | Feature Copy - | Feature Copy Cache | Feature Copy Cache, $\kappa$ | F/B | Total |
|---|---|---|---|---|---|---|---|---|---|
| 4 A100 $\gamma = 2$TB/s $\alpha = 600$GB/s $\beta = 64$GB/s $\lvert S^0 \rvert = 2^{12}$ $b = 1024$ | papers100M GCN | LABOR-0 | Indep | 21.7 | 18.4 | 16.8 | **11.2** | 8.9 | 41.8 |
| | | | Coop | 17.7 | 14.0 | 10.1 | **5.8** | 13.0 | **36.5** |
| | | NS | Indep | 16.1 | 26.5 | **22.1** | - | 10.1 | 48.3 |
| | | | Coop | 11.9 | 21.3 | **12.9** | - | 15.0 | **39.8** |
| | mag240M R-GCN | LABOR-0 | Indep | 26.0 | 57.9 | 56.0 | **41.0** | 199.9 | 266.9 |
| | | | Coop | 20.0 | 51.1 | 36.9 | **23.4** | 183.3 | **226.7** |
| | | NS | Indep | 14.4 | 78.0 | **71.2** | - | 223.0 | 308.6 |
| | | | Coop | 12.3 | 73.9 | **47.5** | - | 215.6 | **275.4** |
| 8 A100 $\gamma = 2$TB/s $\alpha = 600$GB/s $\beta = 64$GB/s $\lvert S^0 \rvert = 2^{13}$ $b = 1024$ | papers100M GCN | LABOR-0 | Indep | 21.3 | 21.1 | 18.7 | **12.0** | 9.3 | 42.6 |
| | | | Coop | 16.5 | 12.4 | 7.1 | **4.0** | 13.5 | **34.0** |
| | | NS | Indep | 15.8 | 31.0 | **24.5** | - | 10.3 | 50.6 |
| | | | Coop | 12.5 | 19.4 | **9.0** | - | 15.6 | **37.1** |
| | mag240M R-GCN | LABOR-0 | Indep | 30.6 | 70.1 | 66.2 | **46.8** | 202.1 | 279.5 |
| | | | Coop | 21.6 | 50.6 | 29.0 | **19.3** | 172.2 | **213.1** |
| | | NS | Indep | 15.0 | 94.9 | **80.9** | - | 224.8 | 320.7 |
| | | | Coop | 14.9 | 71.6 | **39.6** | - | 209.0 | **263.5** |
| 16 V100 $\gamma = 0.9$TB/s $\alpha = 300$GB/s $\beta = 32$GB/s $\lvert S^0 \rvert = 2^{13}$ $b = 512$ | papers100M GCN | LABOR-0 | Indep | 39.1 | 44.5 | 40.2 | **29.4** | 15.1 | 83.6 |
| | | | Coop | 26.9 | 22.7 | 10.4 | **4.9** | 19.1 | **50.9** |
| | | NS | Indep | 18.0 | 61.3 | **52.0** | - | 16.2 | 86.2 |
| | | | Coop | 19.2 | 34.9 | **13.0** | - | 21.3 | **53.5** |
| | mag240M R-GCN | LABOR-0 | Indep | 50.8 | 128.8 | 121.3 | **96.2** | 156.1 | 303.1 |
| | | | Coop | 29.2 | 78.1 | 42.8 | **23.5** | 133.3 | **186.0** |
| | | NS | Indep | 19.3 | 167.3 | **152.6** | - | 170.9 | 342.8 |
| | | | Coop | 19.3 | 116.1 | **53.1** | - | 160.4 | **232.8** |

Table 5: Runtime improvements of Cooperative Minibatching over Independent Minibatching compiled from the **Total** column of Table 4. This is a further improvement on top of the speedup of independent minibatching already achieves using the same number GPUs.

| Dataset & Model | Sampler | 4 GPUs | 8 GPUs | 16 GPUs |
|---|---|---|---|---|
| papers100M & GCN | LABOR-0 | 15% | 25% | 64% |
| | NS | 21% | 36% | 61% |
| mag240M & R-GCN | LABOR-0 | 18% | 31% | 63% |
| | NS | 12% | 22% | 47% |

Station A100 machine with R-GCN. When using a different GNN model such as GATs (Veličković et al., 2018), the forward-backward runtime is 190 ms on 1 GPU vs 172 ms on 4 GPUs on mag240M. The decrease in runtime with increasingly cooperating GPUs is due to the decrease in redundant work they have to perform. Even though the batch size per GPU is constant, the runtime goes down similar to the plots in the top row

of Figure 3, except that it follows, $\frac{kE[|S^2|]}{|S^0|}$, the average number of edges in the 3rd layer when a sampler with fanout $k$ is used.

Additionally, we demonstrate that there is no significant model convergence difference between independent vs cooperative minibatching in Appendix A.9.

#### 4.3.1 Cooperative-Dependent Minibatching

Table 6: Runtime improvements of Dependent Minibatching, for Independent and Cooperative Minibatching methods, compiled from the **Cache, $\boldsymbol{\kappa}$** and **Cache** columns of Table 4 with LABOR-0. Making consecutive minibatches dependent increases temporal locality, hence reducing cache misses.

| Dataset & Model | I/C | 4 GPUs | 8 GPUs | 16 GPUs |
|---|---|---|---|---|
| papers100M & GCN | Indep + Depend | 50% | 57% | 37% |
| | Coop + Depend | 74% | 78% | 112% |
| mag240M & R-GCN | Indep + Depend | 37% | 41% | 26% |
| | Coop + Depend | 58% | 50% | 82% |

We use the same experimental setup as Section 4.3 but vary the $\kappa$ parameter to show that cooperative minibatching can be used with dependent batches (Figure 5b). We use a cache size of 1M vertex embeddings per GPU. Cooperative feature loading effectively increases the global cache size since each PE (GPU) caches only the vertices assigned to them while independent feature loading can have duplicate entries across caches. For our largest dataset mag240M, on top of 1.4× reduced work due to cooperative minibatching alone, the cache miss rates were reduced by more than 2×, making the total improvement 3×. Runtime results for $\kappa \in \{1, 256\}$ are presented in Table 4, the Feature Copy **Cache** and **Cache, $\boldsymbol{\kappa}$** columns. Table 6 summarizes these results by dividing the runtimes in **Cache** by **Cache, $\boldsymbol{\kappa}$** and reporting percentage improvements.

## 5 Key insights of Cooperative Minibatching

In Section 2, the work $W(|S^0|)$ to process an epoch (full pass over the dataset) for a given minibatch size $|S^0|$ is characterized by the number of minibatches in an epoch ($\frac{|V|}{|S^0|}$) × the amount of work to process a single minibatch, which is approximated by the sum of the number of sampled vertices in each layer ($\sum_{l=1}^{L} |S^l|$). This can be seen in Equation (3).

Equation (3) only considers the number of processed vertices and it is good enough for our purposes. Since all the sampling algorithms we consider in Section 2.2 have fanout parameters $k$, the number of edges sampled for each seed vertex has an upper bound $k$. So, given vertices $S^l$ for the $l$th layer, the number of sampled edges in that same layer will be $\leq k|S^l|$. Clearly for almost any GNN model, the runtime complexity to process layer $l$ is linearly increasing w.r.t. the number of vertices ($|S^l|$) and edges ($\leq k|S^l|$) processed, so the runtime complexity is $\mathcal{O}(|S^l| + k|S^l|) = \mathcal{O}(|S^l|)$.

A more comprehensive analysis of the runtime complexities of Cooperative and Independent Minibatching approaches is provided in Appendix A.6, taking into account the exact numbers of sampled vertices ($|S^l|$), edges ($|E^l|$), and various communication bandwidths ($\alpha$, $\gamma$, $\beta$) and even graph partition quality $c$ and cache misses $\rho$.

As Cooperative Minibatching considers a single minibatch of size $B$ for all $P$ PEs, the growth of the number of sampled vertices and edges is characterized by $B$ as the minibatch size. In contrast, Independent Minibatching assigns different minibatches of sizes $\frac{B}{P}$ to each PE, so the growth of the sampled vertices and edges is characterized by $\frac{B}{P}$ as the minibatch size. Due to Theorems 3.1 and 3.2, the work $W(B)$ with respect to a minibatch size $B$ is a concave function, we have $W(B) \leq PW(\frac{B}{P})$, meaning that Cooperative Minibatching is theoretically faster than Independent Minibatching, if communication between PEs is infinitely fast. For empirical data, one can look at the first-row of Figure 3 at the $x$-axis $B$ for Cooperative and $\frac{B}{P}$ for Independent,

due to:

$$W(B) \leq PW(\frac{B}{P}) \iff \frac{W(B)}{B} \leq \frac{W(\frac{B}{P})}{\frac{B}{P}}$$

as proven by Theorem 3.1. As $\frac{E[|S^3|]}{|S^0|}$ curves in Figure 3 are decreasing, the work of Cooperative Minibatching is significantly less than Independent Minibatching. Finally, modern multi-GPU computer systems have very fast inter-GPU communication bandwidths, e.g. NVLink, which is why we are able to show favorable results compared to Independent Minibatching despite performing seemingly unnecessary communication.

## 6 Conclusions

In this paper, we experimentally and theoretically demonstrated that the cost of processing a minibatch is a concave function of batch size in GNNs, unlike DNNs where the cost scales linearly. We then proposed two approaches to take advantage of cost concavity. The first approach, which we call cooperative minibatching proposes to partition a minibatch between multiple PEs and process it cooperatively. This is in contrast to existing practice of having independent minibatches per PE, and avoids duplicate work that is a result of vertex and edge repetition across PEs. The second approach proposes the use of consecutive dependent minibatches, through which the temporal locality of vertex and edge accesses is manipulated. As batches get more dependent, the locality increases. We demonstrated this increase in locality by employing an LRU-cache for vertex embeddings on GPUs. Finally, we showed that these approaches can be combined without affecting convergence, and speed up multi-GPU GNN training by up to 64%.

## Acknowledgements

We would like to thank Kaan Sancak for their feedback on the manuscript, and also Murat Guney for their helpful discussions and for enabling this project during the author's time at NVIDIA Corporation. This work was partially supported by the NSF grant CCF-1919021.

## References

Zeyuan Allen-Zhu and Elad Hazan. Variance reduction for faster non-convex optimization. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 699–707. PMLR, 20–22 Jun 2016. URL `https://proceedings.mlr.press/v48/allen-zhua16.html`.

I. Artico, I. Smolyarenko, V. Vinciotti, and E. C. Wit. How rare are power-law networks really? In *Royal Society*, volume 476, 2020. URL `http://doi.org/10.1098/rspa.2019.0742`.

Muhammed Fatih Balin and Ümit Çatalyürek. Layer-neighbor sampling — defusing neighborhood explosion in gnns. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/51f9036d5e7ae822da8f6d4adda1fb39-Paper-Conference.pdf`. Also available as a Tech Rep on ArXiv.

Muhammed Fatih Balin, Dominique LaSalle, and Ümit V. Çatalyürek. Cooperative minibatching, August 2022. URL `https://github.com/dmlc/dgl/pull/4337`.

Muhammed Fatih Balin, Dominique LaSalle, and Ümit V. Çatalyürek. Cooperative minibatching in gnns, Jan 2023. URL `https://github.com/mfbalin/mfbalin.github.io/blob/main/public/papers/6621_Cooperative_Minibatching_.pdf`. ICML Submission #6621. (Rejected).

D.P. Bertsekas. Incremental least squares methods and the extended kalman filter. In *Proceedings of 1994 33rd IEEE Conference on Decision and Control*, volume 2, pp. 1211–1214 vol.2, 1994. doi: 10.1109/CDC.1994.411166.

Zhenkun Cai, Qihui Zhou, Xiao Yan, Da Zheng, Xiang Song, Chenguang Zheng, James Cheng, and George Karypis. Dsp: Efficient gnn training with multiple gpus. In *Proceedings of the 28th ACM SIGPLAN Annual Symposium on Principles and Practice of Parallel Programming*, PPoPP '23, pp. 392–404, 2023. doi: 10.1145/3572848.3577528.

Wei Lin Chiang, Yang Li, Xuanqing Liu, Samy Bengio, Si Si, and Cho Jui Hsieh. Cluster-GCN: An efficient algorithm for training deep and large graph convolutional networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 257–266. Association for Computing Machinery, jul 2019. doi: 10.1145/3292500.3330925.

Avery Ching, Sergey Edunov, Maja Kabiljo, Dionysios Logothetis, and Sambavi Muthukrishnan. One trillion edges: Graph processing at facebook-scale. *Proc. VLDB Endow.*, 8(12):1804–1815, aug 2015. doi: 10.14778/2824032.2824077.

Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc' aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, Quoc Le, and Andrew Ng. Large scale distributed deep networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL `https://proceedings.neurips.cc/paper/2012/file/6aca97005c68f1206823815f66102863-Paper.pdf`.

Matthias Fey, Jan E. Lenssen, Frank Weichert, and Jure Leskovec. Gnnautoscale: Scalable and expressive graph neural networks via historical embeddings. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3294–3304. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/fey21a.html`.

Swapnil Gandhi and Anand Padmanabha Iyer. P3: Distributed deep graph learning at scale. In *15th USENIX Symposium on Operating Systems Design and Implementation (OSDI 21)*, pp. 551–568, 2021.

Boris Ginsburg, Igor Gitman, and Yang You. Large batch training of convolutional networks with layer-wise adaptive rate scaling. Technical Report arXiv:1708.03888, ArXiv, September 2017. URL `http://arxiv.org/abs/1708.03888`.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. Technical Report arXiv:1706.02677, ArXiv, April 2018. URL `http://arxiv.org/abs/1706.02677`.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 1025–1035, 2017.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *Advances in Neural Information Processing Systems*, 2020-Decem(NeurIPS):1–34, 2020a.

Weihua Hu, Matthias Fey, Hongyu Ren, Maho Nakata, Yuxiao Dong, and Jure Leskovec. Ogb-lsc: A large-scale challenge for machine learning on graphs, 2021. URL `https://arxiv.org/abs/2103.09430`.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. Heterogeneous Graph Transformer. *The Web Conference 2020 - Proceedings of the World Wide Web Conference, WWW 2020*, pp. 2704–2710, 2020b. doi: 10.1145/3366423.3380027.

Zhihao Jia, Sina Lin, Rex Ying, Jiaxuan You, Jure Leskovec, and Alex Aiken. Redundancy-free computation for graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pp. 997–1005. Association for Computing Machinery, 2020. URL `https://doi.org/10.1145/3394486.3403142`.

Weiwei Jiang and Jiayun Luo. Graph neural network for traffic forecasting: A survey. *Expert Systems with Applications*, 207:117921, nov 2022. doi: 10.1016/j.eswa.2022.117921.

George Karypis and Vipin Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, 20(1):359–392, 1998. doi: 10.1137/S1064827595287997.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. URL `https://arxiv.org/abs/1412.6980`.

Mu Li, Tong Zhang, Yuqiang Chen, and Alexander J. Smola. Efficient mini-batch training for stochastic optimization. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pp. 661–670, 2014. doi: 10.1145/2623330.2623612.

Zhiqi Lin, Cheng Li, Youshan Miao, Yunxin Liu, and Yinlong Xu. Pagraph: Scaling gnn training on large graphs via computation-aware caching. In *Proceedings of the 11th ACM Symposium on Cloud Computing*, SoCC '20, pp. 401–415, 2020. doi: 10.1145/3419111.3421281.

Yajing Liu, Zhengya Sun, and Wensheng Zhang. Improving fraud detection via hierarchical attention-based graph neural network, 2022.

Deepak Narayanan, Aaron Harlap, Amar Phanishayee, Vivek Seshadri, Nikhil R. Devanur, Gregory R. Ganger, Phillip B. Gibbons, and Matei Zaharia. Pipedream: Generalized pipeline parallelism for dnn training. In *Proceedings of the 27th ACM Symposium on Operating Systems Principles*, SOSP '19, pp. 1–15, 2019. doi: 10.1145/3341301.3359646.

NVIDIA. NVIDIA DGX A100 system architecture: The universal system for AI infrastructure. Technical report, NVIDIA Corporation, July 2020a. URL `https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-dgx-a100-datasheet.pdf`.

NVIDIA. NVIDIA V100 | datasheet. Technical report, NVIDIA Corporation, January 2020b. URL `https://images.nvidia.com/content/technologies/volta/pdf/volta-v100-datasheet-update-us-1165301-r5.pdf`.

NVIDIA. NVIDIA A100 tensor core GPU | datasheet. Technical report, NVIDIA Corporation, June 2021. URL `https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-us-nvidia-1758950-r4-web.pdf`.

NVIDIA. NVIDIA DGX H100: The gold standard for AI infrastructure. Technical report, NVIDIA Corporation, March 2023. URL `https://resources.nvidia.com/en-us-dgx-systems/ai-enterprise-dgx`.

NVIDIA. NVIDIA GB200 NVL72: Powering the new era of computing. Technical report, NVIDIA Corporation, March 2024. URL `https://www.nvidia.com/en-us/data-center/gb200-nvl72/`.

Yeonhong Park, Sunhong Min, and Jae W. Lee. Ginex: Ssd-enabled billion-scale graph neural network training on a single machine via provably optimal in-memory caching. *Proc. VLDB Endow.*, 15(11):2626–2639, jul 2022. ISSN 2150-8097. doi: 10.14778/3551793.3551819.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

Vatsal Patel, Sutharshan Rajasegarar, Lei Pan, Jiajun Liu, and Liming Zhu. Evangcn: Evolving graph deep neural network based anomaly detection in blockchain. In Weitong Chen, Lina Yao, Taotao Cai, Shirui Pan, Tao Shen, and Xue Li (eds.), *Advanced Data Mining and Applications*, pp. 444–456, 2022.

Sandeep Polisetty, Juelin Liu, Kobi Falus, Yi Ren Fung, Seung-Hwan Lim, Hui Guan, and Marco Serafini. Gsplit: Scaling graph neural network training on large graphs via split-parallelism, 2023.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks, 2017.

Zhihao Shi, Xize Liang, and Jie Wang. LMC: Fast training of GNNs via subgraph sampling with provable convergence. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=5VBBA91N6n.

Alok Tripathy, Katherine Yelick, and Aydın Buluç. Reducing communication in graph neural network training. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2020.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

Roger Waleffe, Jason Mohoney, Theodoros Rekatsinas, and Shivaram Venkataraman. Mariusgnn: Resource-efficient out-of-core training of graph neural networks, 2022.

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. Deep graph library: A graph-centric, highly-performant package for graph neural networks, 2019. URL https://arxiv.org/abs/1909.01315.

D Randall Wilson and Tony R Martinez. The general inefficiency of batch training for gradient descent learning. *Neural networks*, 16(10):1429–1451, 2003.

Shiwen Wu, Fei Sun, Wentao Zhang, Xu Xie, and Bin Cui. Graph neural networks in recommender systems: A survey, 2020.

Nuo Xu, Pinghui Wang, Long Chen, Jing Tao, and Junzhou Zhao. MR-GNN: Multi-resolution and dual graph neural network for predicting structured entity interactions. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, aug 2019. doi: 10.24963/ijcai.2019/551.

Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L. Hamilton, and Jure Leskovec. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '18, pp. 974–983, 2018. doi: 10.1145/3219819.3219890.

Hanqing Zeng, Hongkuan Zhou, Ajitesh Srivastava, Rajgopal Kannan, and Viktor Prasanna. GraphSAINT: Graph sampling based inductive learning method. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJe8pkHFwS.

Hanqing Zeng, Muhan Zhang, Yinglong Xia, Ajitesh Srivastava, Andrey Malevich, Rajgopal Kannan, Viktor Prasanna, Long Jin, and Ren Chen. Decoupling the depth and scope of graph neural networks. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=dOMtHWYONZ.

Da Zheng, Xiang Song, Chengru Yang, Dominique LaSalle, Qidong Su, Minjie Wang, Chao Ma, and George Karypis. Distributed hybrid cpu and gpu training for graph neural networks on billion-scale graphs. *arXiv preprint arXiv:2112.15345*, 2021.

Rong Zhu, Kun Zhao, Hongxia Yang, Wei Lin, Chang Zhou, Baole Ai, Yong Li, and Jingren Zhou. Aligraph: A comprehensive graph neural network platform. *Proc. VLDB Endow.*, 12(12):2094–2105, aug 2019. doi: 10.14778/3352063.3352127.

# A Appendix

## A.1 Graph Sampling

Below, we review three different sampling algorithms for minibatch training of GNNs. Our focus in this work is samplers whose expected number of sampled vertices is a function of the batch size. All these methods are applied recursively for GNN models with multiple layers.

### A.1.1 Neighbor Sampling (NS)

Given a fanout parameter $k$ and a batch of seed vertices $S^0$, NS by (Hamilton et al., 2017) samples the neighborhoods of vertices randomly. Given a batch of vertices $S^0$, a vertex $s \in S^0$ with degree $d_s = |N(s)|$, if $d_s \leq k$, NS uses the full neighborhood $N(s)$, otherwise it samples $k$ random neighbors for the vertex $s$.

### A.1.2 LABOR Sampling

Given a fanout parameter $k$ and a batch of seed vertices $S^0$, LABOR-0 (Balin & Çatalyürek, 2023) samples the neighborhoods of vertices as follows. First, each vertex rolls a uniform random number $0 \leq r_t \leq 1$. Given batch of vertices $S^0$, a vertex $s \in S^0$ with degree $d_s = |N(s)|$, the edge $(t \to s)$ is sampled if $r_t \leq \frac{k}{d_s}$. Since different seed vertices $\in S^0$ end up using the same random variate $r_t$ for the same source vertex $t$, LABOR-0 samples fewer vertices than NS in expectation.

The LABOR-* algorithm is the importance sampling variant of LABOR-0 and samples an edge $(t \to s)$ if $r_t \leq c_s \pi_t$, where $\pi$ is importance sampling probabilities optimized to minimize the expected number of sampled vertices and $c_s$ is a normalization factor. LABOR-* samples fewer vertices than LABOR-0 in expectation.

Note that, choosing $k \geq \max_{s \in V} d_s$ corresponds to training with full neighborhoods for both NS and LABOR methods.

### A.1.3 RandomWalk (RW) Sampling

Given a walk length $o$, a restart probability $p$, number of random walks $a$, a fanout $k$, and a batch of vertices $S^0$, a vertex $s \in S^0$, a *Random Walk* (Ying et al., 2018) starts from $s$ and each step picks a random neighbor $s'$ from $N(s)$. For the remaining $o-1$ steps, the next neighbor is picked from $N(s')$ with probability $1-p$, otherwise it is picked from $N(s)$. This process is repeated $a$ times for each seed vertex and lastly, the top $k$ visited vertices become the *neighbors* of $s$ for the current layer.

Notice that random walks correspond to weighted neighbor sampling from a graph with adjacency matrix $\tilde{A} = \sum_{i=1}^{o} A^i$, where the weights of $\tilde{A}$ depend on the parameters $a$, $p$ and $k$. Random walks give us the ability to sample from $\tilde{A}$ without actually forming $\tilde{A}$.

## A.2 Work Monotonicity Theorem

**Theorem A.1.** *The work per epoch required to train a GNN model using minibatch training is monotonically nonincreasing as the batch size increases.*

*Proof.* Given any $n \geq 2$, let's say we uniform randomly sample without replacement $S^0 \subset V$, where $n = |S^0|$. Now note that for any $S'^0 \subset S^0$, using the definition in (2), we have $S'^l \subset S^l, \forall l$. We will take advantage of that and define $S'^0 = S^0 \setminus \{s\}$ in following expression.

$$
\begin{aligned}
\sum_{\substack{s\in S^0\\ S'^0=S^0\setminus\{s\}}} |S^l| - |S'^l| &= \sum_{\substack{s\in S^0\\ S'^0=S^0\setminus\{s\}}} \sum_{w\in S^l} \mathbb{1}[w \notin S'^l] \\
&= \sum_{\substack{s\in S^0\\ S'^0=S^0\setminus\{s\}}} \sum_{w\in \{s\}^l} \mathbb{1}[w \notin S'^l] \\
&= \sum_{\substack{s\in S^0\\ S'^0=S^0\setminus\{s\}}} \sum_{w\in \{s\}^l} \mathbb{1}[w \notin \{s'\}^l, \forall s' \in S'^0] \\
&= \sum_{w\in S^l} \sum_{\substack{s\in S^0\\ w\in\{s\}^l}} \mathbb{1}[w \notin \{s'\}^l, \forall s' \in S^0 \setminus \{s\}]
\end{aligned}
\tag{8}
$$

In the last expression, for a given $w \in S^l$, if there are two different elements $s, s' \in S^0$ such that $w \in \{s\}^l$ and $w \in \{s'\}^l$, then the indicator expression will be 0. It will be 1 only if $w \in \{s\}^l$ for a unique $s \in S^0$. So:

$$
\begin{aligned}
\sum_{w\in S^l} \sum_{\substack{s\in S^0\\ w\in\{s\}^l}} \mathbb{1}[w \notin \{s'\}^l, \forall s' \in S^0 \setminus \{s\}] &= \sum_{\substack{w\in S^l\\ \exists! s\in S^0, w\in\{s\}^l}} 1 \\
&= |\{w \in S^l \mid w \in \{s\}^l, \exists! s \in S^0\}| \leq |S^l|
\end{aligned}
\tag{9}
$$

Using this, we can get:

$$
\begin{aligned}
&\sum_{\substack{S'^0\subset S^0\\ |S'^0|+1=|S^0|}} |S^l| - |S'^l| \leq |S^l| \\
\iff\ & |S^0||S^l| - \sum_{\substack{S'^0\subset S^0\\ |S'^0|+1=|S^0|}} |S'^l| \leq |S^l| \\
\iff\ & |S^l|(|S^0|-1) \leq \sum_{\substack{S'^0\subset S^0\\ |S'^0|+1=|S^0|}} |S'^l| \\
\iff\ & |S^l|(|S^0|-1) \leq |S^0| E[|S'^l|] \\
\iff\ & \frac{|S^l|}{|S^0|} \leq \frac{E[S'^l]}{|S'^0|}
\end{aligned}
$$

Since $S^0$ was uniformly randomly sampled from $V$, its potential subsets $S'^0$ are also uniformly randomly picked from $V$ as a result. Then, taking an expectation for the random sampling of $S^0 \subset V$, we conclude that $\frac{E[|S^l|]}{|S^0|} \leq \frac{E[|S'^l|]}{|S'^0|}$, i.e., the expected work of batch size $n$ is not greater than the work of batch of size $n-1$. This implies that the work with respect to batch size is a monotonically nonincreasing function. □

### A.3 Overlap monotonicity

**Theorem A.2.** *The expected subgraph size $E[|S^l|]$ required to train a GNN model using minibatch training is a concave function of batch size, $|S^0|$.*

*Proof.* Given any $n \geq 2$, let's say we uniformly randomly sample without replacement $S^0 \subset V$ of size $n$. Note that we use the $T^l$ and $T_2^l$ as defined in Equations (5) and (6).

$$
\begin{aligned}
|T^l(S^0)| - 2|T_2^l(S^0)| &= \sum_{\substack{S'^0 \subset S^0 \\ |S'^0|+1=|S^0|}} |T^l(S^0)| - |T^l(S'^0)| \\
&= |S^0||T^l(S^0)| - |S^0|E[|T^l(S'^0)|] \\
\iff (|S^0|-1)|T^l(S^0)| &= |S^0|E[|T^l(S'^0)|] - 2|T_2^l(S^0)| \\
\iff \frac{|T^l(S^0)|}{|S^0|} &= \frac{E[|T^l(S'^0)|]}{|S'^0|} - \frac{2|T_2^l(S^0)|}{|S'^0||S^0|} \\
\implies \frac{|T^l(S^0)|}{|S^0|} &\leq \frac{E[|T^l(S'^0)|]}{|S'^0|}
\end{aligned}
\tag{10}
$$

where the first equality above is derived similar to Equations (8) and (9). Overall, this means that the overlap between vertices increases as the batch size increases. Utilizing our finding from Equations (8) and (9), we have:

$$
\begin{aligned}
\sum_{\substack{S'^0 \subset S^0 \\ |S'^0|+1=|S^0|}} |S^l| - |S'^l| &= |T^l(S^0)| \\
\implies |S^0||S^l| - |S^0|E[|S'^l|] &= |T^l(S^0)| \\
\implies |S^l| &= E[|S'^l|] + \frac{|T^l(S^0)|}{|S^0|} \\
\implies E[|S^l|] &= E[|S'^l|] + \frac{E[|T^l(S^0)|]}{|S^0|}
\end{aligned}
\tag{11}
$$

Note that the last step involved taking expectations for the random sampling of $S^0$. See the recursion embedded in the equation above, the expected size of the subgraph $S^l$ with batch size $|S^0|$ depends on the expected size of the subgraph $S'^l$ with batch size $|S^0|-1$. Expanding the recursion, we get:

$$
E[|S^l|] = \sum_{i=1}^{|S^0|} \frac{E[|T^l(S_i^0)|]}{i} \tag{12}
$$

where $S_i^0$ is a random subset of $S^0$ of size $i$. Since $\frac{E[|T^l(S_i^0)|]}{i}$ is monotonically nonincreasing as $i$ increases as we showed in (10), we conclude that $E[|S^l|]$ is a concave function of the batch size, $|S^0|$. □

### A.4 Subgraph density monotonicity

**Theorem A.3.** *The expected sampled subgraph density $\frac{E[|S_E|]}{|S|}$ is nondecreasing as the batch size $|S|$ increases.*

*Proof.* Given $n \geq 1$ and any probability distribution $p$ on $V$ such that $p_s$ denotes the probability of including $s$ in our minibatch $S$ of size $|S| = n$, the probability of including the edge $t \to s$ becomes $p_t p_s$. Since $p_s(n)$ is linear as a function of $n$, $\forall s \in V$, it can be decomposed as $p_s(n) = p'_s n$.

$$
\frac{E[|S_E|]}{|S|} = \frac{1}{|S|} \sum_{(t \to s) \in E} \mathbb{P}(t \in S \wedge s \in S) = \frac{1}{|S|} \sum_{(t \to s) \in E} p_t p_s = \frac{1}{n} \sum_{(t \to s) \in E} p'_t n p'_s n = n \sum_{(t \to s) \in E} p'_t p'_s \tag{13}
$$

We can see that the expected sampled subgraph density is a linear function of the batch size $|S| = n$ and this implies that it is nondecreasing as the batch size $|S|$ increases. □

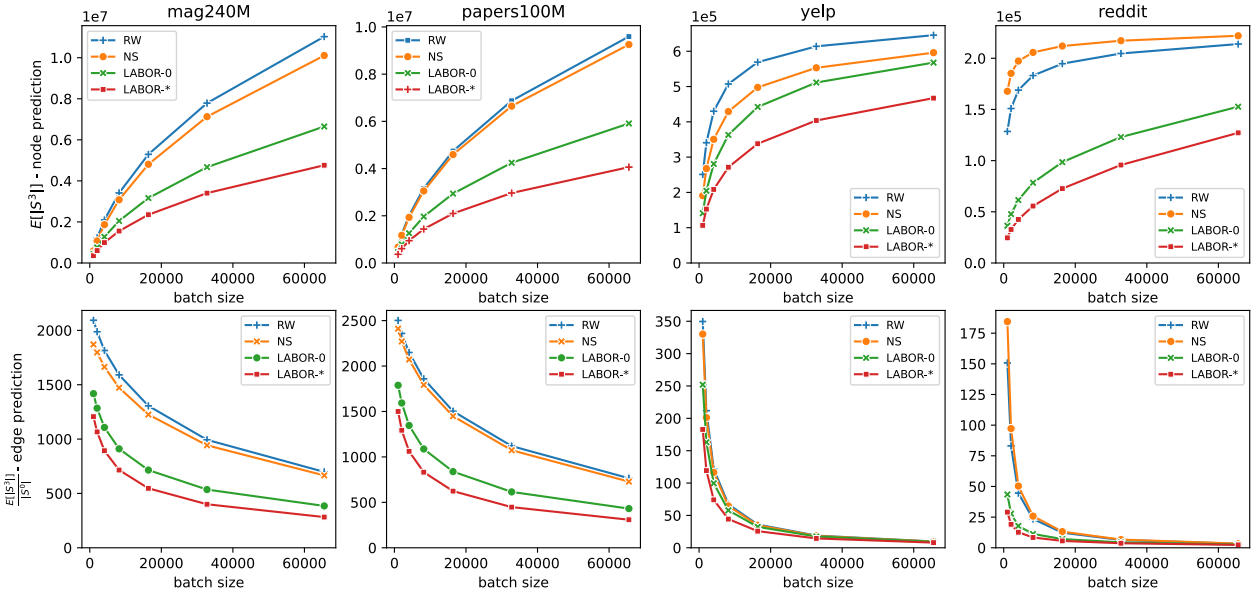



Figure 6: Monotonicity of the work. x axis shows the batch size, y axis shows $E[|S^3|]$ for node prediction (top row) and $\frac{E[|S^3|]}{|S^0|}$ for edge prediction (bottom row), where $E[|S^3|]$ denotes the expected number of vertices sampled in the 3rd layer and $|S^0|$ denotes the batch size. RW stands for Random Walks, NS stands for Neighbor Sampling, and LABOR-0/* stand for the two different variants of the LABOR sampling algorithm described in Section 2.2. Completes Figure 3.

### A.5 Experimental Setup

**Setup:** In our experiments, we use the following datasets: reddit (Hamilton et al., 2017), papers100M (Hu et al., 2020a), mag240M (Hu et al., 2021), yelp and flickr (Zeng et al., 2020), and their details are given in Table 2. We use Neighbor Sampling (NS) (Hamilton et al., 2017), LABOR Sampling (Balin & Çatalyürek, 2023) and Random Walks (RW) (Ying et al., 2018) to form minibatches. We used a fanout of $k = 10$ for the samplers. In addition, Random Walks used length of $o = 3$, restart probability $p = 0.5$ and number of random walks from each seed $a = 100$. All our experiments involve a GCN model with $L = 3$ layers (Hamilton et al., 2017), with 1024 hidden dimension for mag240M and papers100M and 256 for the rest. Additionally, papers100M and mag240M datasets were made undirected graphs for all experiments and this is reflected in the reported edge counts in Table 2. Input features of mag240M are stored with the 16-bit floating point type. We use the Adam optimizer (Kingma & Ba, 2014) with $10^{-3}$ learning rate in all the experiments.

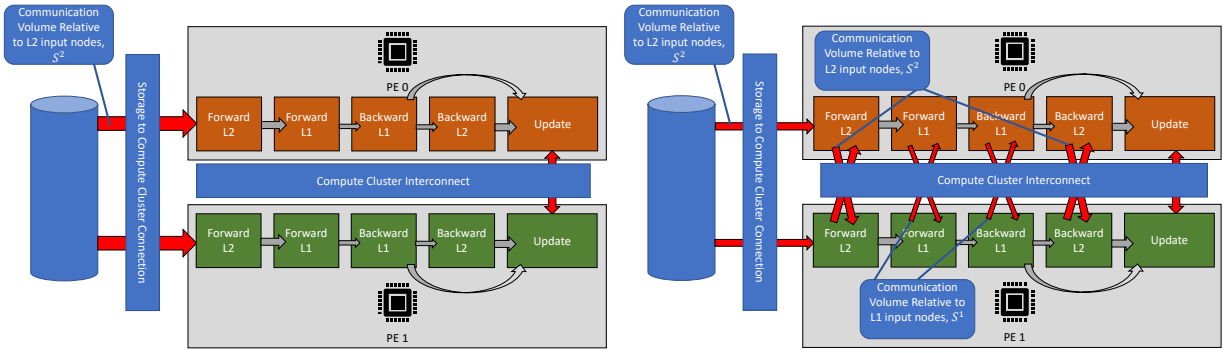
**Implementation:** We implemented [1][2] our experimental code using C++ and Python in the DGL framework (Wang et al., 2019) with the Pytorch backend (Paszke et al., 2019). All our experiments were repeated 50 times and averages are presented. Early stopping was used during model training runs. So as we go to the right along the x-axis, the variance of our convergence plots increases because the number of runs that were ongoing is decreasing.

### A.6 Complexity Analysis (cont.)

Our goal in this section is to empirically show the work reduction enjoyed by cooperative minibatching over independent minibatching by reporting the number of vertices and edges processed per PE. We also report the number of vertices that are communicated for cooperative minibatching during its all-to-all calls in Algorithm 1 and Figure 7b. The results are given in Table 7.

Looking at Tables 1 and 7, we make the following observations:

[1]Source code is available at `https://github.com/GT-TDAlab/dgl-coop/tree/dist_graph_squashed_wip_cache`
[2]Cooperative and dependent minibatching are also available in the GraphBolt GNN dataloading library.




(a) Independent Minibatching (b) Cooperative Minibatching

Figure 7: A comparison of Independent and Cooperative Minibatching approaches in the feature loading and forward-backward GNN stages. The thickness of the red arrows indicates the data volume. Due to redundant vertices across PEs, Independent Minibatching wastes (PCI-e) bandwidth for vertex embedding copies from Storage. Moreover, the PEs perform identical computations for redundant edges across PEs in Independent Minibatching.

Table 7: Average number of vertices and edges sampled in different layers with LABOR-0 per PE, reduced by taking the maximum over 4 PEs (All the numbers are in thousands, lower is better) with batch size $|S^0| = 1024$. $c|\tilde{S}^l|$ shows the number of vertices communicated at layer $l$. Papers and mag were used as short versions of papers100M/GCN and mag240M/R-GCN dataset model pairs respectively. Last column shows forward-backward (F/B) runtime in ms.

| Dataset | Part. | I/C | $\lvert S^3\rvert$ | $c\lvert\tilde{S}^3\rvert$ | $\lvert\tilde{S}^3\rvert$ | $\lvert E^2\rvert$ | $\lvert S^2\rvert$ | $c\lvert\tilde{S}^2\rvert$ | $\lvert\tilde{S}^2\rvert$ | $\lvert E^1\rvert$ | $\lvert S^1\rvert$ | F/B |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| papers | random | Indep | 463 | 0 | 463 | 730 | 74.8 | 0 | 74.8 | 93.6 | 9.63 | 8.9 |
| | random | Coop | 318 | 311 | 463 | 608 | 62.4 | 56.8 | 82.8 | 89.9 | 9.28 | 13.0 |
| | metis | Coop | 328 | 179 | 402 | 615 | 63.1 | 34.0 | 73.8 | 90.8 | 9.35 | 12.0 |
| mag | random | Indep | 443 | 0 | 443 | 647 | 67.9 | 0 | 67.9 | 82.0 | 8.78 | 199.9 |
| | random | Coop | 324 | 310 | 459 | 566 | 59.8 | 53.1 | 77.3 | 80.4 | 8.62 | 183.3 |
| | metis | Coop | 334 | 178 | 419 | 576 | 60.6 | 31.0 | 71.3 | 81.8 | 8.80 | 185.1 |

1. All runtime complexities for cooperative minibatching scales with $|S^l_p(B)| = |S^l(B)|\frac{1}{P}$ and $|E^l_p(B)| = |E^l(B)|\frac{1}{P} \leq |S^l(B)|\frac{k}{P}$ and for independent minibatching with $|S^l(\frac{B}{P})|$ and $|E^l(\frac{B}{P})| \leq |S^l(\frac{B}{P})|k$, for a sampler with fanout $k$. Since $E[|S^l(B)|]$ is a concave function, $E[|S^l(B)|]\frac{1}{P} \leq E[|S^l(\frac{B}{P})|]$, and this corresponds to looking at Figure 3 first row with $x = B$ for coop and $x = \frac{B}{P}$ for independent if one wanted to guess how their runtime would change with changing $B$ and $P$. For an example, all the runtime numbers we have provided in the Table 4 are for 4 GPUs. Going from 4 to 8 would increase the edge of cooperative over independent even more, see Table 5.

2. Sampling and Feature loading over PCI-e requires $\alpha \gg \beta$ for cooperative to get a speedup over independent.

3. In order for cooperative F/B to improve against independent, we need that $\frac{\alpha}{c} > \frac{\gamma}{M}$.

4. Cross edge ratio $c$ reduces all communication between PEs. In particular, graph partitioning will lower both $c$ and $|\tilde{S}^l_p(B)|$, lowering the communication overhead, see $\boldsymbol{c|\tilde{S}^l|}$ columns in Table 7.

5. The model complexity $M$ is small for the GCN model (papers100M) but large for the R-GCN model (mag240M), as shown by the F/B runtime numbers in Table 2. Also, the communication overhead between the two scenarios is similar, meaning communication can take from upto 30% to less than a few percent depending on $M$. For the papers100M dataset, communication makes up more of the

runtime, so graph partitioning helps bring the F/B runtime down. However, the load imbalance caused by graph partitioning slows down the F/B runtime despite lowered communication costs for the mag240M dataset.

### A.7 Smoothed Dependent Minibatching

As described in Section 2, NS algorithm works by using the random variate $r_{ts}$ for each edge $(t \to s)$. Being part of the same minibatch means that a single random variate $r_{ts}$ will be used for each edge. To generate these random variates, we initialize a Pseudo Random Number Generator (PRNG) with a random seed $z$ along with $t$ and $s$ to ensure that the first rolled random number $r_{ts}$ from the PRNG stays fixed when the random seed $z$ is fixed. Given random seeds $z_1$ and $z_2$, let's say we wanted to use $z_1$ for the first $\kappa$ iterations and would later switch to the seed $z_2$. This switch can be made smoothly by interpolating between the random seeds $z_1$ and $z_2$ while ensuring that the resulting sampled random numbers are uniformly distributed. If we are processing the batch number $i < \kappa$ in the group of $\kappa$ batches, then we want the contribution of $z_2$ to be $c = \frac{i}{\kappa}$, while the contribution of $z_1$ is $1 - c$. We can sample $n^1_{ts} \sim \mathcal{N}(0, 1)$ with seed $z_1$ and $n^2_{ts} \sim \mathcal{N}(0, 1)$ with seed $z_2$. Then we combine them as

$$n_{ts}(c) = \cos(\frac{c\pi}{2})n^1_{ts} + \sin(\frac{c\pi}{2})n^2_{ts}$$

note that $n_{ts}(0) = n^1_{ts}$, $n_{ts}(1) = n^2_{ts}$ and $n_{ts}(c) \sim \mathcal{N}(0, 1), \forall c \in \mathbb{R}$ also, then we can set $r_{ts} = \Phi(n_{ts}(c))$, where $\Phi(x) = \mathbb{P}(Z \leq x)$ for $Z \sim \mathcal{N}(0, 1)$, to get $r_{ts} \sim U(0, 1)$ that the NS algorithm can use. Dropping $s$ from all the notation above gives the version for LABOR. In this way, the random variates change slowly as we are switching from one group of $\kappa$ batches to another. When $i = \kappa$, we let $z_1 \leftarrow z_2$ and initialize $z_2$ with another random seed. To use this approach, only the random variate generation logic needs modification, making its implementation for any sampling algorithm straightforward compared to the nested approach initially described. Figure 2 shows an example of this approach for $\kappa = 2$.

### A.8 Dependent batches (cont.)

Looking at the training loss and validation F1-score with sampling curves in Figure 8, we notice that the performance gets better as $\kappa$ is increased. This is due to the fact that a vertex's neighborhood changes slower and slower as $\kappa$ is increased, the limiting case being $\kappa = \infty$, in which case the sampled neighborhoods are unchanging. This makes training easier so $\kappa = \infty$ case leads the pack in the training loss and validation F1-score with sampling curves.

### A.9 Comparing a single batch vs $P$ independent batches convergence

We investigate whether training with a single large batch in $P$-GPU training shows any convergence differences to the current approach of using $P$ separate batches for each of the GPUs. We use a global batch size of 4096 and divide a batch into $P \leq 8$ independent batches, with each batch having a size of $\frac{4096}{P}$. We use NS and LABOR-0 samplers with fanouts of $k = 10$ for each of the 3 layers. Figure 9 shows that there are no significant differences between the two approaches, we present the results averaged over the samplers to save space.

### A.10 Redundancy Free GCN aggregation

Jia et al. (2020) proposes a method of reducing processing during neighborhood aggregation by finding common sub-neighborhoods among aggregated vertices, whether using full-batch or minibatch training. That is, if two vertices have the neighborhoods $\{A, B, C\}$ and $\{B, C, D\}$, and a summation operator is used for aggregation, then instead of computing four additions: $A + B + C$ and $B + C + D$ concurrently, the three additions $BC = B + C$, $A + BC$, and $BC + D$ can be computed. This approach is orthogonal to the approaches proposed in Section 3 in that it reduces redundant aggregation steps, whereas our approach reduces redundant input nodes and edges in parallel computations. As such, the two approaches could be employed together.

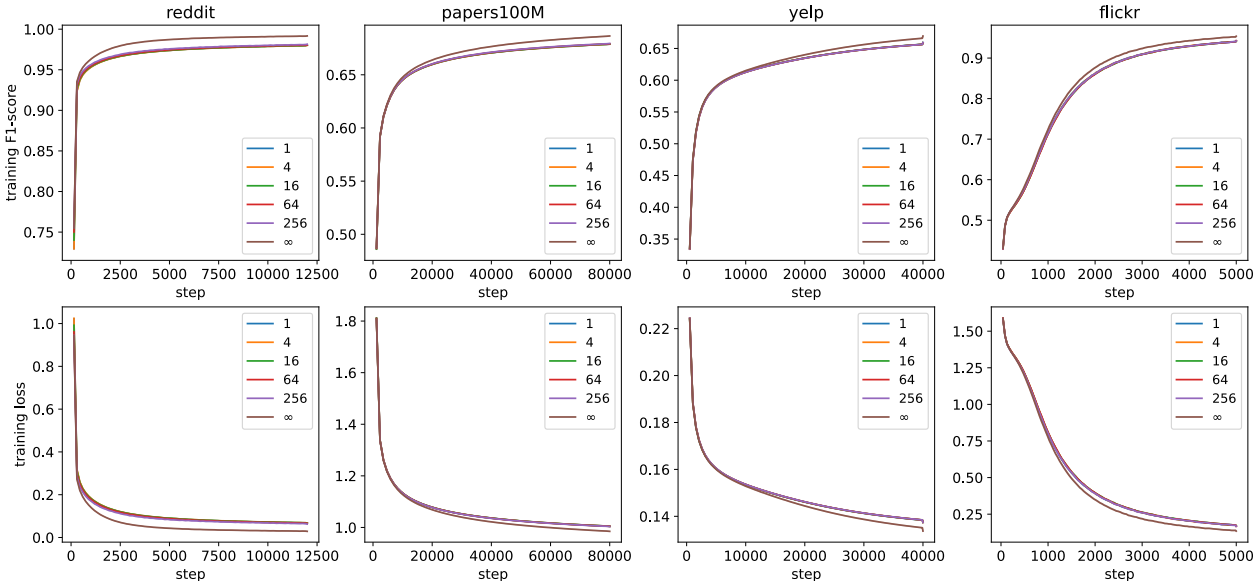



Figure 8: LABOR-0 sampling algorithm with 1024 batch size and varying $\kappa$ dependent minibatches, $\kappa = \infty$ denotes infinite dependency, meaning the neighborhood sampled for a vertex stays static during training. The first row shows the training F1-score with the dependent sampler. The second row shows the training loss curve. Completes Figure 4.

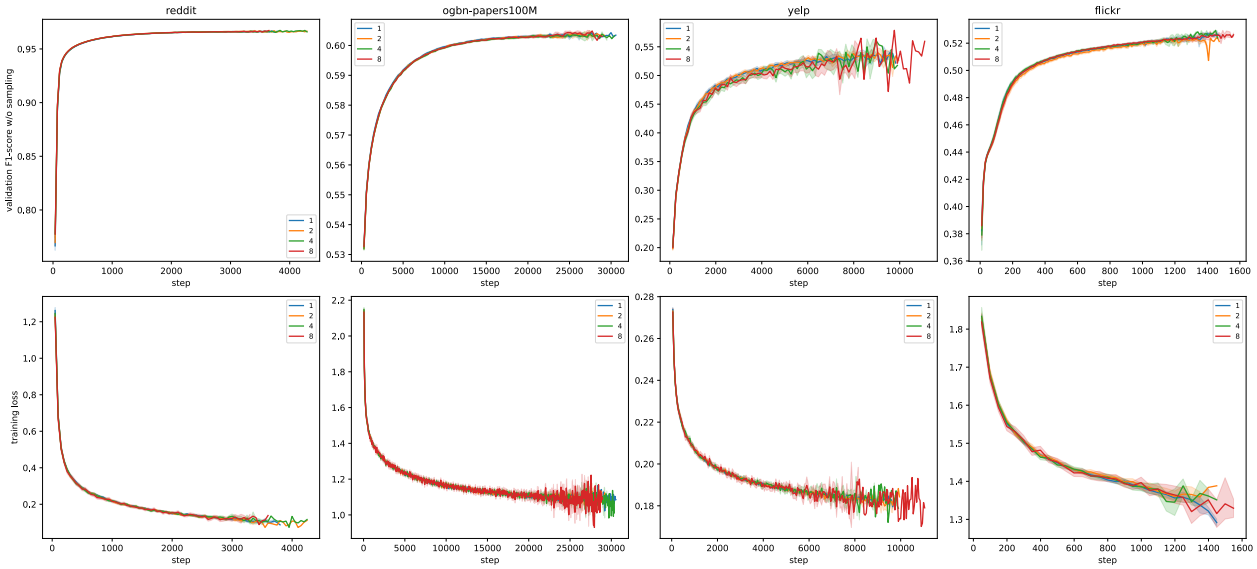



Figure 9: Convergence difference between cooperative vs independent minibatching with a global batch size of 4096 averaged over Neighbor and LABOR-0 samplers.

### A.11 Limitations

Our proposed Cooperative Minibatching approach requires a relatively fast interconnect between the Processing Elements. Modern multi-GPU systems usually have such interconnect between the GPUs. Nowadays, such fast interconnects are being extended to the multi-node setting, for up to 72 GPUs connected via NVLink (NVIDIA, 2024).

On distributed training, each computing node equipped with multiple cores and/or GPUs, the interconnect between cores and/or GPUs of the same node, are also relatively much faster than the interconnect among nodes.

The proposed dependent minibatching approach does not have any such limitation.