
COSTA: Covariance-Preserving Feature Augmentation for Graph Contrastive Learning

Yifei Zhang¹, Hao Zhu^{2,3}, Zixing Song¹, Piotr Koniusz^{3,2}, Iriwn King¹

The Chinese University of Hong Kong¹, Hong Kong SAR, China
Australian National University² and Data61/CSIRO³, Canberra, Australia
{yfzhang,zxsong,king}@cse.cuhk.edu.hk
allenhaozhu@gmail.com; piotr.koniusz@data61.csiro.au

Abstract

Graph contrastive learning (GCL) improves graph representation learning, leading to SOTA on various downstream tasks. The graph augmentation step is a vital but scarcely studied step of GCL. In this paper, we show that the node embedding obtained via the graph augmentations is highly biased, somewhat limiting contrastive models from learning discriminative features for downstream tasks. Thus, instead of investigating graph augmentation in the input space, we alternatively propose to perform augmentations on the hidden features (feature augmentation). Inspired by so-called matrix sketching, we propose **COSTA**, a novel **CO**variance-**preS**erving **feaTure** space **A**ugmentation framework for GCL, which generates augmented features by maintaining a “good sketch” of original features. To highlight the superiority of feature augmentation with COSTA, we investigate a single-view setting (in addition to multi-view one) which conserves memory and computations. We show that the feature augmentation with COSTA achieves comparable/better results than graph augmentation based models.

1 Introduction

Many Graph Neural Networks (GNNs) [17, 19, 43, 30, 31, 39] focus on (semi-)supervised learning, which requires access to abundant labels. Recent trends in Self-Supervised Learning (SSL) have resulted in several methods that do not require labels [18, 13]. Among SSL methods, Contrastive Learning (CL) already achieved comparable performance with its supervised counterparts on many tasks [3, 9]. Recently, CL has been applied to the graph domain. A typical Graph Contrastive Learning (GCL) method constructs multiple graph views by stochastic augmentation of the input to learn representations by contrasting positive samples with negative samples [47, 26, 48]. However, the irregular structure of graphs complicates the adaptation of augmentation techniques used on images and prevents extending of theoretical analysis for vision-based contrastive learning to graphs. Thus, many works focus on the empirical design of hand-crafted graph augmentations (GA) for graph contrastive learning (*i.e.*, random edge/node/attribute dropping) [41, 47, 48]. Notably, some latest works point out that random data augmentations are problematic as their noise may not be relevant to downstream tasks [33, 34]. In certain scenarios (*i.e.*, recommendation systems [40, 5, 4]), GCL achieves the desired performance gain under extremely sparse GAs (with an edge dropout rate 0.9) [38] but method [42] achieves similar results without GAs. Such observations naturally raise the question: are there better augmentation strategies for GCL other than GA?

To this end, we show that the embeddings obtained with GA are highly biased compared to the embeddings obtained with feature augmentation (FA), that is, the embeddings obtained with FA (*e.g.*, an injection of random noise into the embedding) exhibit the so-called weak law of large numbers (WLLN). Specifically, for any error $\varepsilon \geq 0$, $\lim_{k \rightarrow \infty} \mathcal{P}(\|\mathbb{E}(\tilde{\mathbf{x}}^{(k)}) - \mathbf{x}\|_1 > \varepsilon) = 0$, where

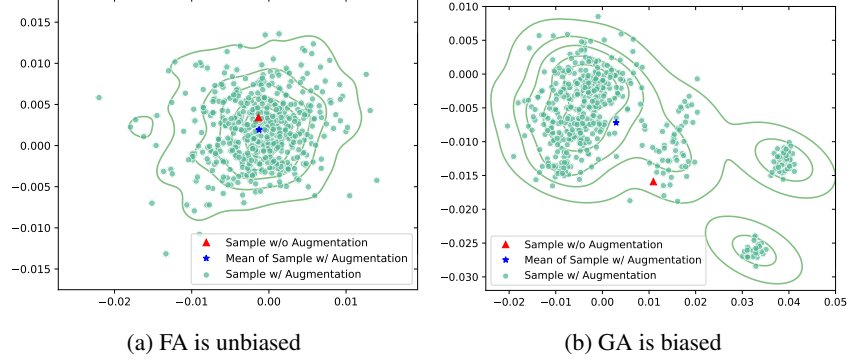


Figure 1: The distribution of node embeddings on Cora is generated by $500\times$ graph augmentations. **1a** corresponds to the feature augmentation (Gaussian noise injection). **1b** corresponds to the graph augmentation (edge permutation & attribute masking). We use 2D embeddings for visualization.

$\tilde{x}^{(k)}$ denotes the embedding obtained by augmenting x , and x is the original embedding without augmentation. For the i.i.d. random variables, as the sample size k increases, the expectation of embedding after augmentation, $\mathbb{E}(\tilde{x}^{(k)})$, tends toward the real mean x (embedding without augmentation). In contrast to FA, GA violates the weak law of large numbers. As shown in Figure **1a**, the population mean for the embedding obtained by the FA is in the densest region and in the proximity of the embedding of the original sample (without augmentation). In contrast, we cannot see such a trend in the case of the GA in Figure **1b**. In other words, GA introduces some bias, whereas FA produces unbiased embeddings.

We assert that a successful contrastive objective should promote similarity/dissimilarity between features of encoded attributes by implicitly grouping/separating related/unrelated nodes according to their attribute space, respectively. However, as the GA strategy results in the bias (Figure **1b**), attraction/separation of embeddings in the feature space does not necessarily result in an optimal attraction/separation of desired nodes in the attribute space, which may result in suboptimal pre-training for downstream tasks. Figure 2 and Section 4.1 further illustrate and motivate the above two scenarios. Furthermore, the adoption of GA in GCL often increases the complexity as GCL compares the node features obtained from multiple views (e.g., multiple network streams) to obtain correlated views of the same graph. However, this strategy is prohibitive on large graphs as, in the worst-case scenario, multi-view GCL requires a time and space complexity quadratic w.r.t. the number of views and nodes. Thus, apart from the multi-view setting, we also investigate a single-view GCL setting.

Our Contributions. Instead of the GA, we propose to perform augmentation on the hidden feature vectors (feature augmentation). Inspired by matrix sketching, we propose **COSTA**, a novel **CO**variance-**pre**Serving **fea**Ture space **A**ugmentation framework for GCL, which produces augmented features by generating a “good sketch” of original features. To highlight the superiority of feature augmentation, apart from the multi-view setting, we show many results in the single-view setting, which conserves the memory usage and computations. We empirically show that COSTA (even the single-view variant, *i.e.*, COSTA_{SV}) achieves comparable or better results than other GA strategies.

Our contributions are threefold:

- i. We point out the issue of bias introduced by the topology graph augmentation in the GCL framework, and we advocate feature augmentation strategies to prevent the aforementioned bias.
- ii. Inspired by matrix sketching, we propose COSTA, a simple and effective covariance-preserving feature augmentation framework for GCL, which generates augmented features by generating a “good sketch” (variance is bounded) of original features.
- iii. As an alternative to the multi-view GCL setting, we propose the single-view GCL setting, which produces equivalent or better results than the multi-view GCL while requiring less memory and incurring shorter computations.

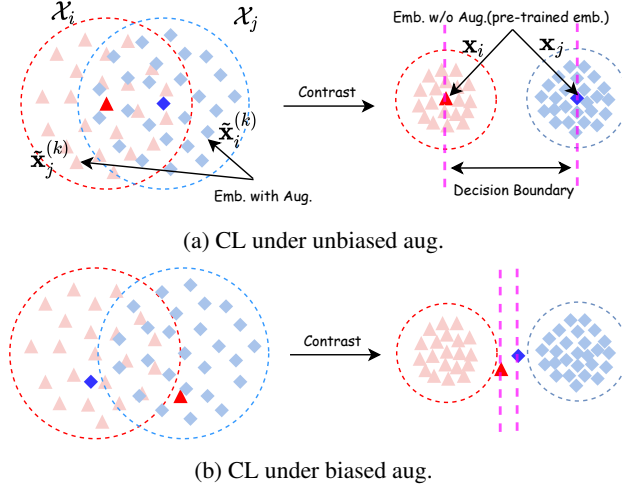


Figure 2: (Fig. 2a) Our strategy results in an unbiased augmentation strategy. (Fig. 2b) A counter-example illustrates the problem of using a biased augmentation strategy in CL.

To our best knowledge, this is the first work which considers feature augmentation (in the single-view setting) in GCL with the matrix sketching step performing feature augmentation.

2 Related Works

2.1 Data Augmentation

Input Space Augmentation, studied across many domains, usually refers to the augmentation performed in the input space. In computer vision, image transformations such as rotation, flipping, color jitters, translation, and noise injection [28] as well as more recent cut-off and random erasure [6] are very popular. In neural language processing, input space augmentations include token-level random augmentations such as synonym replacement, word swapping, word insertion, and deletion [37]. In the graph domain, input space augmentation is referred to as graph augmentation. Attribute masking, edge permutation, and node dropout are common graph augmentation strategies [41]. Adaptive graph augmentations based on node centrality and PageRank centrality were studied by Zhu *et al.* [48] and Page *et al.* [24] respectively with the goal of masking different edges with varying probability. We discuss the negative effect of graph augmentation (*e.g.*, edge and node removal) later in the text.

Feature Augmentation strategies generate augmented samples in the feature space instead of the input space [8]. Wang *et al.* [36] augment features in the hidden feature space, resulting in a feature representation that corresponds to another sample with the same class label but different semantics. So-called instance augmentations add perturbations to original instances [36]. Many few-shot learning approaches [14] estimate the “analogy transformations” between examples of known classes to apply them to examples of novel classes. Finally, feature augmentations are popular in many research domains, *i.e.*, semi-supervised learning, one-shot learning, and few-shot learning. However, no prior work has combined FA with contrastive learning in the graph domain the way COSTA performs FA.

2.2 Graph Contrastive Learning

Inspired by contrastive methods in vision and NLP [16, 3, 9], CL has also been adapted to the graph domain. By adapting DeepInfoMax [1] to graph representation learning, DGI [35] learns embedding by maximizing the mutual information to discriminate between nodes of original and corrupted graphs. REFINE [44] uses a simple negative sampling term inspired by skip-gram models. Fisher-Bures Adversarial GCN [32] uses adversarial perturbations of graph Laplacian. Inspired by SimCLR [3], GRACE [47] correlates graph views by pushing closer representations of the same node in different views and pushing apart representations of different nodes. Another example of a SimCLR strategy is the recent GraphCL method [12]. In contrast to GRACE, which learns node

embedding, GraphCL learns embeddings for graph-level tasks. The above multi-view methods suffer from the large memory and computational footprint, respectively. Although COLES [46] proposes a robust single-view GCL approach, it works the best with linear GNNs such as S²GC [45]. Thus, apart from multi-view COSTA, we also study the single-view GCL setting with FA.

3 Preliminaries

3.1 Notations

In this paper, a graph with node features is denoted as $G = (\mathcal{V}, \mathcal{E}, \mathbf{X})$, where \mathcal{V} is the vertex set, \mathcal{E} is the edge set, and $\mathbf{X} \in \mathbb{R}^{n \times d}$ is the feature matrix (*i.e.*, the i -th row of \mathbf{X} is the feature vector \mathbf{x}_i of node v_i). Let $n = |\mathcal{V}|$ and $m = |\mathcal{E}|$ be the numbers of vertices and edges respectively. We use $\mathbf{A} \in \{0, 1\}^{n \times n}$ to denote the adjacency matrix of G , *i.e.*, the (i, j) -th entry in \mathbf{A} is 1 if and only if there is an edge between v_i and v_j . The degree of a node v_i , denoted as d_i , is the number of edges incident with v_i . The degree matrix \mathbf{D} is a diagonal matrix, and its i -th diagonal entry is d_i . For a d -dimensional vector, $\mathbf{x} \in \mathbb{R}^d$, $\|\mathbf{x}\|_2$ is the Euclidean norm of \mathbf{x} . We use x_i to denote the i -th entry of \mathbf{x} , and $\text{diag}(\mathbf{x}) \in \mathbb{R}^{d \times d}$ is a diagonal matrix such that the i -th diagonal entry is x_i . We use $\mathbf{A}_{i:}$ and $\mathbf{A}_{:i}$ to denote the i -th row and column of \mathbf{A} respectively, and A_{ij} for the (i, j) -th entry of \mathbf{A} . The trace of a square matrix \mathbf{A} is denoted by $\text{Tr}(\mathbf{A})$, which is the sum along the diagonal of \mathbf{A} . The singular value decomposition of \mathbf{A} is denoted as $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ where $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_n]$, $\mathbf{\Sigma} = \text{diag}(\sigma_1, \dots, \sigma_d)$, and $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_d]$. We use $\|\mathbf{A}\|_2$ to denote the spectral norm of \mathbf{A} , which is the largest singular value σ_{\max} . We use $\|\mathbf{A}\|_F$ for the Frobenius norm, which is $\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} |a_{ij}|^2} = \sqrt{\text{Tr}(\mathbf{A}^\top \mathbf{A})} = \sqrt{\sum_{i=1}^d \sigma_i^2(\mathbf{A})}$.

3.2 Multi-view Graph Contrastive Learning (MV-GCL)

Following the conventions presented in [47, 48], MV-GCL learns node representations by maximizing the mutual information (MI) between views of the same graph. Below, we introduce components of MV-GCL: (i) graph augmentation, (ii) GNN-based encoders, (iii) projection head, and (iv) a contrastive loss.

Graph Augmentation. \mathcal{T}_{GA} generates augmented $(\tilde{\mathbf{A}}, \tilde{\mathbf{X}})$ by directly adding random perturbations to the original graph (\mathbf{A}, \mathbf{X}) . Different augmented graphs are constructed given one input (\mathbf{A}, \mathbf{X}) , yielding correlated views $(\tilde{\mathbf{A}}_i, \tilde{\mathbf{X}}_i)$ that represent the augmented adjacent matrix and node features in the i -th view. In the common GCL setting [47, 48], the graph structure is augmented via edge permutation. Node features are augmented via attribute masking.

Graph Neural Network Encoders. The GNN encoder $f_i : \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times d} \rightarrow \mathbb{R}^{n \times d}$ extracts hidden node features $\mathbf{H}_i \in \mathbb{R}^{n \times d}$ from the i -th augmented graph $(\tilde{\mathbf{A}}_i, \tilde{\mathbf{X}}_i)$. Usually, multiple encoders are applied to obtain the hidden node features \mathbf{H}_i of different views as:

$$\mathbf{H}_1 = f_1(\tilde{\mathbf{A}}_1, \tilde{\mathbf{X}}_1), \dots, \mathbf{H}_k = f_k(\tilde{\mathbf{A}}_k, \tilde{\mathbf{X}}_k). \quad (1)$$

The GNN encoders are implemented as a two-layer Graph Convolution Network (GCN):

$$\begin{aligned} \text{GCN}_l(\mathbf{X}, \mathbf{A}) &= \sigma \left(\hat{\mathbf{D}}^{-\frac{1}{2}} \hat{\mathbf{A}} \hat{\mathbf{D}}^{-\frac{1}{2}} \mathbf{X} \mathbf{W}_l \right), \\ f(\mathbf{X}, \mathbf{A}) &= \text{GCN}_2(\text{GCN}_1(\mathbf{X}, \mathbf{A}), \mathbf{A}), \end{aligned} \quad (2)$$

where $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ is the adjacency matrix with self-loops, \mathbf{D} is the degree matrix, $\sigma(\cdot)$ is an activation function, *e.g.*, $\text{ReLU}(\cdot) = \max(0, \cdot)$, and \mathbf{W}_l is a trainable weight matrix for the l -th layer.

Feature Augmentation \mathcal{T}_{FA} . We apply feature augmentation on \mathbf{H} . We elaborate on the proposed FA and detail its properties in Section 4.2. FA results in the augmented feature maps fed into the projection head describe below.

Projection Head. The projection head $\theta(\cdot)$ is a small network that maps representations to the space where contrastive loss is applied. It is implemented as a multi-layer perceptron (MLP) with one hidden layer to obtain $\mathbf{Z}_i = \theta(\mathbf{H}_i) = \mathbf{W}^{(2)} \sigma(\mathbf{W}^{(1)} \mathbf{H}_i)$, where σ is the ReLU non-linearity. As described in [3], it is beneficial to define the contrastive loss on \mathbf{Z}_i rather than \mathbf{H}_i .

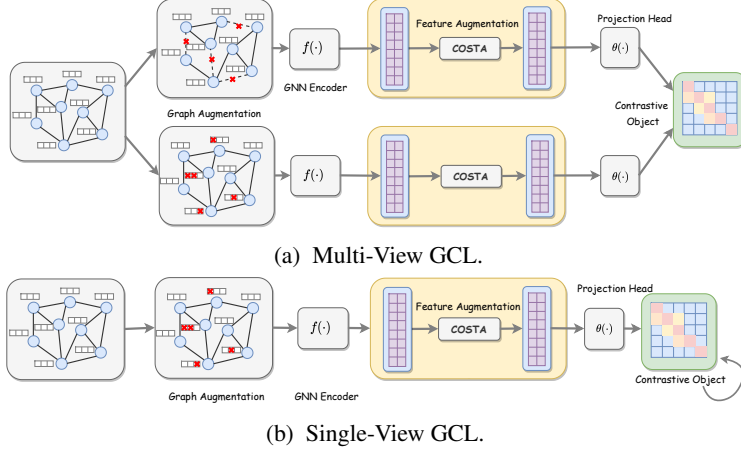


Figure 3: The illustrations of MV-GCL (standard) in Fig. 3a and SV-GCL (simplified) in Fig 3b. For simplicity, the architecture of MV-GCL is shown using two views only and SV-GCL is the trivial case where two views are the same. MV-GCL contrast two views while SV-GCL perform self-contrast.

Contrastive Loss. Let two feature matrices $U \in \mathbb{R}^{n \times d}$ and $V \in \mathbb{R}^{n \times d}$, where $U = Z_1$ and $V = Z_2$ are node features obtained from two different views. Then for any node i , its embedding generated in one view, $U_{i:}$, is treated as the anchor, its embedding generated in another view, $V_{i:}$, forms the positive sample. Remaining node embeddings $U_{j:}$ and $V_{j:}$ such that $j \neq i$ (from two views) are naturally regarded as negative samples. The contrastive loss function \mathcal{L} for all positive pairs is defined as:

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^n \left[(U_{i:}^\top, V_{i:}) / \tau - \log \left(\sum_{j=1}^n e^{(U_{i:}^\top, V_{j:}) / \tau} + \sum_{j=1}^n e^{(U_{i:}^\top, U_{j:}) / \tau} \right) \right]. \quad (3)$$

Note that computing Eq. (3) is both memory costly and time consuming as it requires the computation of three large similarity matrices, UV^\top , UU^\top , $VV^\top \in \mathbb{R}^{n \times n}$ for two views. Therefore, the memory consumption and runtime depend on the number of views multiplied by the number of nodes, making such a multi-view setting challenging to run on large-scale graphs.

Single-view Graph Contrastive Learning (SV-GCL). To validate the effectiveness of graph augmentation and feature augmentation, apart from MV-GCL, we use a special case of multi-view GCL that shares the same augmented graph for two views and is thus equivalent to single-view Graph Contrastive Learning (SV-GCL). We note that SV-GCL has a computational advantage, *i.e.*, only the features of a single view are calculated, and distances between features within the view. Distances between views are not needed. SV-GCL also provides a fairer way to compare the effectiveness of graph augmentations and feature augmentations as otherwise the multi-view setting would be the reason for the performance gain rather than the graph augmentation strategy.

4 Methodology

Section 4.1 presents our motivation. Section 4.2 presents **COSTA**, **CO**variance **pRe**serving **fEa**ture space **A**ugmentation framework. Section 4.3 relates COSTA to the problem of matrix sketching, which generates desired augmented samples with theoretical guarantee.

4.1 Motivation

We motivate COSTA with a simple experimental example, which shows that the node embedding obtained under graph augmentation is highly biased compared to feature augmentation. Inspired by WLLN (the weak law of large numbers explained in the introduction), below we quantify the bias introduced by the data augmentation $\mathcal{T}(\cdot)$ as follows. Let x_i be the original embedding of the i -th node and $\tilde{\mathcal{X}}_i$ be its augmentation set where each embedding $\tilde{x}_i^{(k)} \in \tilde{\mathcal{X}}_i$ is obtained by stochastic

transformation, *i.e.*, $\mathcal{T}(f(\cdot))$ or $f(\mathcal{T}(\cdot))$. Let the transformation distribution of \mathbf{x}_i be $\tilde{\mathcal{T}}(\mathbf{x}_i)$, then:

$$\text{Bias}(\mathcal{T}(\mathbf{x}_i)) = \left\| \mathbb{E}_{\tilde{\mathbf{x}}_i \sim \tilde{\mathcal{T}}(\mathbf{x}_i)}(\tilde{\mathbf{x}}_i) - \mathbf{x}_i \right\|_2 \approx \left\| \frac{1}{|\tilde{\mathcal{X}}_i|} \sum_{k=1}^{|\tilde{\mathcal{X}}_i|} \tilde{\mathbf{x}}_i^{(k)} - \mathbf{x}_i \right\|_2. \quad (4)$$

Following Eq. (4), given a randomly chosen node (Cora dataset), we randomly generate $|\tilde{\mathcal{X}}| = 500$ augmented samples by the graph augmentation (GA) by edge permutation, and the feature augmentation (FA) by directly adding random noise to \mathbf{x} . In both cases, we use the same encoder to obtain the node embeddings. The output layer of the encoder has two dimensions to facilitate visualization. Figure 1 depicts the distribution of node embeddings for both types of augmentations. Clearly, the expectation of node embeddings $\mathbb{E}(\mathcal{T}_{\text{FA}}(\mathbf{x}))$ (the blue star in Figure 1a) obtained by the FA converges to its original embedding \mathbf{x} (the red triangle in Figure 1a), whereas expectation of node embeddings $\mathbb{E}(\mathcal{T}_{\text{GA}}(\mathbf{x}))$ obtained via the GA deviates far from \mathbf{x} , indicating the following:

$$\text{Bias}(\mathcal{T}_{\text{GA}}(\mathbf{x})) \gg \text{Bias}(\mathcal{T}_{\text{FA}}(\mathbf{x})). \quad (5)$$

As shown, GA-based contrastive learning suffers from optimizing the biased setting. A standard contrastive loss aims to maximize the similarity of embeddings within the same augmentation set (*e.g.*, $\tilde{\mathbf{x}}_i^{(1)}, \tilde{\mathbf{x}}_i^{(2)} \in \tilde{\mathcal{X}}_i$) and minimize the similarity of embeddings between different sets (*e.g.*, $\tilde{\mathbf{x}}_i^{(1)} \in \tilde{\mathcal{X}}_i$ and $\tilde{\mathbf{x}}_j^{(1)} \in \tilde{\mathcal{X}}_j$). To perform well, such an objective requires the augmented embeddings to adhere to the unbiased case described above because as the bias tends to zero, the expectation of augmented embeddings converges to \mathbf{x}_i , *i.e.*, $\mathbb{E}(\tilde{\mathbf{x}}_i) \rightarrow \mathbf{x}_i$. Thus, pushing away $\tilde{\mathbf{x}}_i^{(k)}$ from $\tilde{\mathbf{x}}_j^{(k)}$ if $i \neq j$ separates embeddings of different instances, whereas the augmented embeddings $\tilde{\mathbf{x}}_i^{(k)}$ concentrate around \mathbf{x}_i . In contrast, when the bias is large, *i.e.*, $\|\mathbb{E}(\tilde{\mathbf{x}}_i) - \mathbf{x}_i\|_2 \gg 0$, separating augmented embeddings of different instance (*i.e.*, $\tilde{\mathbf{x}}_i^{(k)}, \tilde{\mathbf{x}}_j^{(k)}, i \neq j$) may not increase the discrimination of learned embeddings (*i.e.*, $\mathbf{x}_i, \mathbf{x}_j$) for downstream tasks. Figure 2 (bottom) illustrates such a case.

This motivates us to explore new ways of performing augmentations for GCL. Instead of explicitly eliminating the bias in the current GCL, we apply feature augmentations as an alternative.

4.2 Covariance-Preserving Feature Augmentation

Previous section indicates the graph augmentation produces the bias. We adopt the feature augmentation for GCL loss because FA lets us control the variance and reduce the bias.

Thus, we propose **the feature augmentation framework** in which the augmented feature matrix $\tilde{\mathbf{X}} \in \mathbb{R}^{k \times d}$, given the original feature matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, is obtained via:

$$\begin{aligned} \tilde{\mathbf{X}} &= \mathbf{P}\mathbf{X} + \mathbf{E}, \\ \text{such that } \|\mathbf{X}^\top \mathbf{X} - \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}\|_2 &\leq \varepsilon \text{Tr}(\mathbf{X}^\top \mathbf{X}). \end{aligned} \quad (6)$$

$\mathbf{P} \in \mathbb{R}^{k \times n}$ in Eq. (6) denotes an affine transformation, \mathbf{E} is the random noise matrix and ε is the error which controls the quality of approximation $\|\mathbf{X}^\top \mathbf{X} - \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}\|_2$. Note that the affine transformation can be either deterministic or stochastic.

Connection to the Gaussian Noise Injection. One special case of COSTA is the Gaussian noise injection [36, 42] which produces the augmented feature matrix $\tilde{\mathbf{X}}$ by adding random noise sampled as $\tilde{\mathbf{X}}_{i:} \sim \mathcal{N}(\mathbf{X}_{i:}, \varepsilon \mathbf{I})$, where $\varepsilon \geq 0$ controls the strength of noise. This is equivalent to setting $\mathbf{P} = \mathbf{I}$ and $\mathbf{E}_{i:} \sim \mathcal{N}(0, \varepsilon \mathbf{I})$ in Eq. (6).

4.3 Feature Augmentation via Matrix Sketching

Definition 4.1 (Matrix Sketching [21]). Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be the given feature matrix, $\mathbf{P} \in \mathbb{R}^{k \times n}$ be a sketching matrix, *e.g.*, random projection or row selection matrix. The sketch of \mathbf{X} is defined as $\tilde{\mathbf{X}} = \mathbf{P}\mathbf{X} \in \mathbb{R}^{k \times d}$. Usually, $\tilde{\mathbf{X}}$ contains fewer rows than \mathbf{P} , where $k \ll n$ but $\tilde{\mathbf{X}}$ still preserves many properties of \mathbf{P} .

Eq. (6) performs matrix sketching. Obtaining the augmented feature matrix $\tilde{\mathbf{X}}$ requires a good sketch of \mathbf{X} such that second-order statistics of the original and sketched matrices are similar. In what follows, we use SVD, random row selection, or random projection to form a sketch of \mathbf{X} . We prove that $\tilde{\mathbf{X}}$ obtained by sketching satisfies $\|\mathbf{X}^\top \mathbf{X} - \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}\|_2 \leq \varepsilon \text{Tr}(\mathbf{X}^\top \mathbf{X})$ for small $\varepsilon \geq 0$.

Matrix Sketching via SVD. One solution for Eq. (6) can be obtained through the singular value decomposition (SVD) where:

$$\mathbf{P} = \mathbf{U}^\top, \mathbf{X} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^\top. \quad (7)$$

Lemma 4.2. Let $\tilde{\mathbf{X}} = \mathbf{P} \mathbf{X}$ and $\mathbf{X} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^\top$ where $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_n]$, $\mathbf{\Sigma} = \text{diag}(\sigma_1, \dots, \sigma_d)$, $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_d]$. Then $\|\mathbf{X}^\top \mathbf{X} - \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}\|_2$ is bounded as:

$$\|\mathbf{X}^\top \mathbf{X} - \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}\|_2 \leq \frac{\sigma_{k+1}}{\sigma_{\max}} \text{Tr}(\mathbf{X}^\top \mathbf{X}). \quad (8)$$

Proof. See Appendix A. □

Remark 4.3. The upper bound of $\|\mathbf{X}^\top \mathbf{X} - \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}\|_2$ is controlled by the $(k+1)$ -th largest eigenvalue σ_{k+1} . Usually, $\frac{\sigma_{k+1}}{\sigma_{\max}}$ is small as $\sigma_{k+1} \ll \sigma_{\max}$ even when k is small. However, SVD is computationally intensive and unsuitable for decomposition of large feature matrices.

Random Row Selection (RS). Randomized algorithms trade accuracy for efficiency and strive for high accuracy and low runtime. A sketch of a matrix $\tilde{\mathbf{X}}$ can be constructed via randomly stacking the rows of the original matrix \mathbf{X} . Random row selection employs a small subset of rows based on a pre-defined probability distribution $\mathcal{P}(i)$ to form a sketch. The random assignment matrix $\mathbf{P} \in \mathbb{R}^{k \times n}$ stacks one-hot vectors, i.e., $\mathbf{P} = \{\mathbf{e}_i \in \mathbb{R}^n | \mathcal{P}(i) = \frac{\|\mathbf{X}_{i,:}\|_2}{\|\mathbf{X}\|_F}\} \in \mathbb{R}^{k \times n}$, where \mathbf{e}_i indicates that the i -th row is selected.

Lemma 4.4. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$. Let $\tilde{\mathbf{X}} \in \mathbb{R}^{m \times d}$ be a matrix whose rows are randomly selected from rows of $\mathbf{X} \in \mathbb{R}^{n \times d}$. It holds that:

$$\mathcal{P}\left(\|\mathbf{X}^\top \mathbf{X} - \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}\|_2 \leq \varepsilon \text{Tr}(\mathbf{X}^\top \mathbf{X})\right) \geq 1 - e^{-\left(\frac{(\varepsilon\sqrt{k}-1)^2}{8}\right)}. \quad (9)$$

Proof. See Appendix A. □

Remark 4.5. The failure probability $\delta_{RS} = e^{-\left(\frac{(\varepsilon\sqrt{k}-1)^2}{8}\right)}$ is exponentially decreasing with the error ε meaning that we can bound $\|\mathbf{X}^\top \mathbf{X} - \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}\|_2$ given small $\varepsilon \geq 0$ with a high probability $1 - \delta_{RS}$.

Random Projection (RP). A sketch of matrix can be RP. The projection matrix is defined as $\mathbf{P} \in \mathbb{R}^{n \times k}$ whose entry p_{ij} is sampled from $\mathcal{N}(0, 1)$. Ideally, we expect \mathbf{P} to provide a stable sketch that approximately preserves the distance between all pairs of columns in the original matrix. As the computation of dense matrix \mathbf{P} is time-consuming, a sparse version from Appendix B) can be used.

Lemma 4.6. let $\tilde{\mathbf{X}} = \frac{1}{\sqrt{k}} \mathbf{P} \mathbf{X}$ where p_{ij} is the (i, j) -th element of \mathbf{P} , and $p_{ij} \sim \mathcal{N}(0, 1)$. For $\varepsilon \in (0, 1)$, it holds that:

$$\mathcal{P}\left(\|\mathbf{X}^\top \mathbf{X} - \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}\|_2 \leq \varepsilon \text{Tr}(\mathbf{X}^\top \mathbf{X})\right) \geq 1 - e^{-\frac{\varepsilon^2 k}{8}}. \quad (10)$$

Proof. See Appendix A. □

Remark 4.7. Note that the failure probability of the random projection δ_{PR} is less than δ_{RS} when $k > \varepsilon^{-2}$.

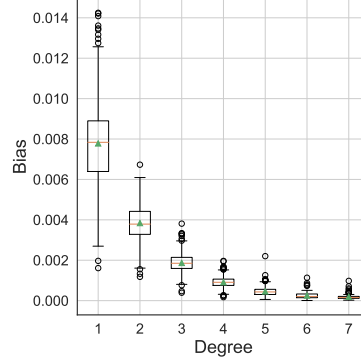


Figure 4: The bias vs. the node degree.

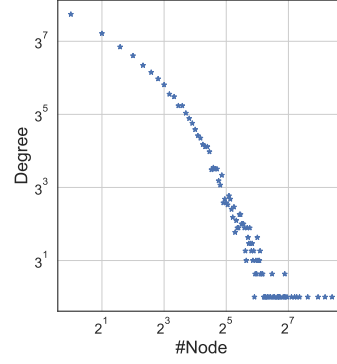


Figure 5: Node degrees obey the power law distribution.

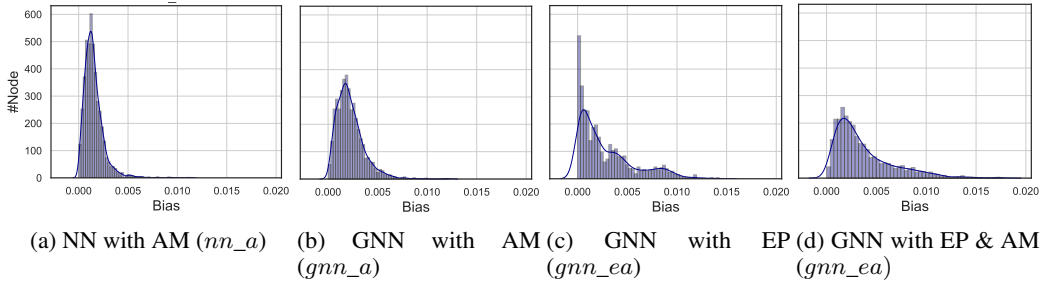


Figure 6: The bias distribution of all nodes on the Cora dataset w.r.t. different augmentation strategies and encoder .

5 Experiments

Below, we provide details of experimental settings, and we discuss our results. We answer the following research questions (RQ):

- **RQ1:** What is the bias problem in graph augmentation?
- **RQ2:** Does the proposed feature augmentation work for problems of practical interest? How does its accuracy/speed compare with MV-GCL models that adopt the graph augmentation strategy, and with other models? Does SV-GCL perform well in comparison to MV-GCL models?
- **RQ3:** What is the performance of the feature augmentation given different matrix sketching schemes? What are the major factors that contribute to the success of the proposed feature augmentation method?
- **RQ4:** How is the effectiveness affected by the number of augmented samples?

5.1 The Bias Problem of Graph Augmentation with GNNs (RQ1)

In Section 4.1, we intuitively point out the pitfall of the topology GA by providing an illustrative example. Based on that observation, we hypothesize that the topology GA introduces a substantial bias into the node embeddings used by the contrastive loss, which deteriorates the quality of features from the pre-training step, thus affecting downstream tasks. In this section, we conduct a quantitative analysis of this problem.

Experimental Protocol. We adopt the edge perturbation and attribute masking, the most commonly used strategies for the graph augmentation [47, 48, 35]. To show the difference between the attribute GA and the topology GA, we use MLP and GNN to encode the original features. For a fair comparison, MLP and GNN share the fixed weights. We only change the type of augmentations used to obtain the node embedding. Specifically, We denote GNN with the edge perturbation (EP) as *gnn_e*, GNN with

Table 1: Results (Cora, CiteSeer, and WikiCS) given a common testbed for different feature augmentation strategies realized by Eq. (6). The symbol e_i denotes the one-hot vector and \mathcal{O} is the zero matrix.

Feature Augmentation	Type	$P \in \mathbb{R}^{k \times n}$	$E \in \mathbb{R}^{k \times d}$	Cora	CiteSeer	WikiCS
Gaussian Noise Injection	Stochastic	$P = I$	$E \sim \mathcal{N}(0, 1)$	0.8271	0.7134	0.7823
SVD	Deterministic	$P = U^\top, X = U\Sigma V^\top$	$E = \mathcal{O}$	0.8269	0.7142	0.7814
Random Selection	Stochastic	$P = \{e_i \mathcal{P}(i) = \frac{\ X_{i,:}\ _2}{\ X\ _F}\}$	$E = \mathcal{O}$	0.8245	0.7121	0.7811
Random Projection	Stochastic	$P \sim \mathcal{N}(0, 1)$	$E = \mathcal{O}$	0.8425	0.7247	0.7911

Table 2: Node classification in terms of accuracy (%) with standard deviation. The highest performance is highlighted in boldface. COSTA_{MV} and COSTA_{SV} denote the variants of multi-view and single-view setting respectively, OOM indicates Out-Of-Memory.

Method	Training Data	Wiki-CS	Amazon-Computers	Amazon-Photo	Coauthor-CS	Coauthor-Physics
Raw features	X	71.98 \pm 0.00	73.81 \pm 0.00	78.53 \pm 0.00	90.37 \pm 0.00	93.58 \pm 0.00
Node2vec	A	71.79 \pm 0.05	84.39 \pm 0.08	89.67 \pm 0.12	85.08 \pm 0.03	91.19 \pm 0.04
DeepWalk	A	74.35 \pm 0.06	85.68 \pm 0.06	89.44 \pm 0.11	84.61 \pm 0.22	91.77 \pm 0.15
DeepWalk	X, A	77.21 \pm 0.03	86.28 \pm 0.07	90.05 \pm 0.08	87.70 \pm 0.04	94.90 \pm 0.09
GAE	X, A	70.15 \pm 0.01	85.27 \pm 0.19	91.62 \pm 0.13	90.01 \pm 0.71	94.92 \pm 0.07
VGAE	X, A	75.63 \pm 0.19	86.37 \pm 0.21	92.20 \pm 0.11	92.11 \pm 0.09	94.52 \pm 0.00
DGI	X, A	75.35 \pm 0.14	83.95 \pm 0.47	91.61 \pm 0.22	92.15 \pm 0.63	94.51 \pm 0.52
GMI	X, A	74.85 \pm 0.08	82.21 \pm 0.31	90.68 \pm 0.17	OOM	OOM
MVGRL	X, A	77.52 \pm 0.08	87.52 \pm 0.11	91.74 \pm 0.07	92.11 \pm 0.12	95.33 \pm 0.03
GRACE	X, A	78.31 \pm 0.05	87.80 \pm 0.23	92.53 \pm 0.16	92.95 \pm 0.03	95.72 \pm 0.03
GCA	X, A	78.23 \pm 0.04	87.54 \pm 0.49	92.24 \pm 0.21	92.95 \pm 0.13	95.73 \pm 0.03
G-BT	X, A	76.83 \pm 0.73	87.93 \pm 0.36	92.46 \pm 0.35	92.91 \pm 0.25	95.25 \pm 0.13
COSTA_{SV}	X, A	79.03 \pm 0.05	88.26 \pm 0.03	92.30 \pm 0.25	92.95 \pm 0.12	95.74 \pm 0.02
COSTA_{MV}	X, A	79.12 \pm 0.02	88.32 \pm 0.03	92.56 \pm 0.45	92.94 \pm 0.10	95.60 \pm 0.02

attribute masking (AM) as gnn_a , GNN with AM and EP as gnn_ea , MLP with AM as nn_a . For each node from Cora, we generate 500 augmented samples in four variants to compute their bias by Eq. (4).

Bias Patterns of Graph Augmentations. Figure 6 shows distributions containing bias. It is obvious that gnn_ea contains more node embeddings with a larger bias compared to nn_ea , judging by the distribution shift towards right. This confirms our hypothesis that the graph augmentation introduces the bias. Moreover, the long-tailed trend is mainly caused by the edge perturbation. Figure 4 shows that nodes with a small degree exhibit more bias, which confirms our insight that for node embeddings of low-degree nodes, obtained via GNN, removing any edges perturbs these embedding significantly. Real-world graphs (citation network, social network, *etc.*) follow the power-law distribution (shown in Figure 5) meaning a few of nodes connect with the majority of the edges, whereas the majority of nodes have only a few of edges (low degree). For example, 54%, 68%, 71%, 53% nodes of Cora, CitSeer, PubMed, and DBLP datasets have less than 3 edges. Thus, applying the graph augmentation in contrastive learning results in a large bias.

5.2 Comparison with the State-of-the-Art Methods (RQ2)

In this section, we compare COSTA to other baseline models to answer **RQ2**. We use the same experimental setup as the representative MV-GCL method (*i.e.*, GCA [48], and GRACE [47]) to perform a fair comparison to these methods. Unless stated otherwise, the random projection is employed as the default setting of COSTA as it balances well between accuracy and efficiency. A detailed comparison between different feature augmentations given different matrix sketching schemes is shown in Section 5.3.

Datasets. To evaluate our method, we adopt nine commonly used benchmark datasets in the previous works [47, 48, 35], including citation networks (Cora, CiteSeer, Pubmed, DBLP, Coauthor-CS, and Coauthor-Physics) and social networks (Wiki-CS, Amazon-Computers, Amazon-Photo) [17, 29, 22, 23]. Detailed descriptions and statistics are given in Appendix C. Apart from Wiki-CS adopting the public split, other datasets are randomly divided into 10%, 10%, 80% for training, validation, and testing.

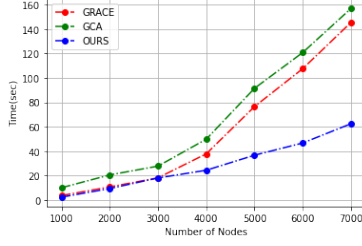


Figure 7: The runtime comparison of single-view COSTA (COSTA_{SV}) vs. MV-GCL models such as GCA and GRACE.

Method	Training Data	Cora	Citeseer	Pubmed	DBLP
Raw features	X	64.8	64.6	84.8	71.6
node2vec	A	74.8	52.3	80.3	78.8
DeepWalk	A	75.7	50.5	80.5	75.9
DeepWalk	X, A	73.1	47.6	83.7	78.1
GAE	X, A	76.9	60.6	82.9	81.2
VGAE	X, A	78.9	61.2	83.0	81.7
DGI	X, A	82.6 ± 0.4	68.8 ± 0.7	86.0 ± 0.1	83.2 ± 0.1
GRACE	X, A	83.3 ± 0.4	72.1 ± 0.5	86.7 ± 0.1	84.2 ± 0.1
GCA	X, A	82.8 ± 0.3	71.5 ± 0.3	86.0 ± 0.2	83.1 ± 0.2
COSTA_{SV}	X, A	84.3 ± 0.3	72.8 ± 0.3	86.2 ± 0.1	84.3 ± 0.1
COSTA_{MV}	X, A	84.3 ± 0.2	72.9 ± 0.3	86.0 ± 0.2	84.5 ± 0.1

Table 3: Results on Cora, CiteSeer, PubMed, and DBLP. We use the same setting as the setting used for experiments in Table 2.

Evaluation Protocol. For each experiment, we adopt the same evaluation scheme as in [35, 47, 48], where each model is firstly trained in an unsupervised manner on the whole graph with node features. Then, we transform the raw features into the resulting embeddings with the use of the trained encoder. Next, we train an ℓ_2 -regularized logistic regression classifier from the Scikit-Learn library [25] with the use of embeddings obtained in the previous step. We also perform a grid search over the regularization parameter with the following values $\{2^{-10}, 2^{-9}, \dots, 2^{-1}\}$. We compute the classification accuracy and report the mean and standard deviations for 20 model initializations and splits.

Baselines. To compare COSTA with previous works, we choose the representative baselines from traditional graph self-supervised learning, autoencoder-based model, and contrastive-based graph self-supervised learning. Methods include (i) Random walk based models: Deepwalk [27] and node2vec [11], (ii) Autoencoder Based models: GAE and VGAE [18], (iii) the contrastive-based models including Deep Graph Infomax (DGI) [35], Graphical Mutual Information Maximization (GMI) [26], Graph Barlow Twins (G-BT) [2] and Multi-View Graph Representation Learning (MVGRL) [15], GRACE [47] and GCA [48]. Note that all the contrastive models use the topology graph augmentation by default. For all baselines, we report their performance based on the official implementations and we use default hyper-parameters from original papers.

Main Results. Tables 2 and 3 show that COSTA achieves competitive performance compared to the baseline methods, and even surpasses them on most datasets. These results demonstrate that COSTA is an effective framework leveraging the advantage of feature augmentations. Specifically, the superiority of COSTA is confirmed by the fact that both single- and multi-view COSTA variants, COSTA_{SV} and COSTA_{MV} , outperform several MV-GCL models that use the topology graph augmentation (*i.e.*, GCA, GRACE, MVGRL) on several datasets (Cora, CiteSeer, DBLP, Wiki-CS, Amazon-Computers, AM-Photo and Coauthor-Physics) and achieve comparable results on PubMed, and Coauthor-CS datasets. We note that datasets on which COSTA_{SV} does not achieve SOTA have a small number of nodes with low node degrees (*i.e.*, only around 10% of nodes in Amazon-photo and Coauthor-Physics have the degree less than 3, meaning less bias is introduced by the topology GA. However, COSTA_{SV} requires less runtime and memory to achieve the comparable performance. This is attributed to the single-view design (SV-GCL) of COSTA_{SV} (Section 3). We also note that COSTA_{MV} typically outperforms COSTA_{SV} by a small margin which suggests that single-view augmentation strategies are a good choice for GA. In addition, we note that most of the contrastive learning models outperform models based on the reconstruction error (*i.e.*, GAE, VGAE, DeepWalk, Node2Vec), which reflects the superiority of contrastive learning.

Running time. We measure the runtime to further validate the practicality of the single-view COSTA (COSTA_{SV}) in terms of time complexity. We mainly compare it to GCA and GRACE (Amazon-Computers dataset) because both GCA and GRACE are the representative models for the multi-view contrastive learning framework utilizing graph augmentations. Note that GCA uses adaptive graph augmentations. To form Figure 7, we sampled a subgraph with a fixed number of nodes from 1,000 to 8,000. Figure 7 shows the training time for 1,000 epochs given different numbers of nodes. The figure shows that our method is faster than the other two models. What stands out is that the gap between them becomes more apparent as the number of nodes increases. COSTA_{SV} becomes 2 times faster than the other models at $\geq 5,000$ nodes. We attribute this to the single-view setting with the

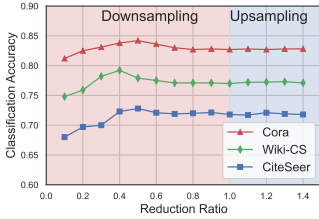


Figure 8: Accuracy w.r.t. sketching ratios.

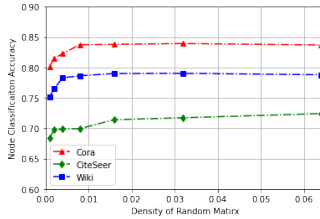


Figure 9: Accuracy w.r.t. matrix densities.

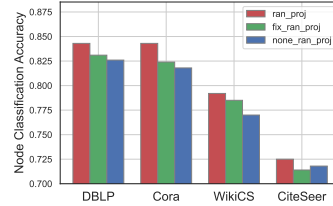


Figure 10: Ablation study on different variants of random projection.

Table 4: Ablation study: the topology graph augmentation vs. the feature augmentation (single-view COSTA).

Graph Aug.	Feature Aug.	WikiCS	Cora	CiteSeer
×	×	74.31	80.04	71.35
✓	×	77.67	82.13	71.93
×	✓	79.03	84.30	72.81
✓	✓	79.07	84.33	72.90

feature augmentation. Note that COSTA_{SV} computes the node feature matrix once before feeding it into the projection head, and the feature augmentation effectively can be understood as reducing the number of nodes, as our feature augmentation acts on columns of hidden feature matrix. Thus, we form one relatively small similarity matrix for the contrastive loss. In contrast, the MV-GCL framework incurs a higher complexity. Our experiments confirm that COSTA_{SV} is efficient in practice. Moreover, COSTA_{SV} can be accelerated by employing sketching by a sparse matrix without sacrificing its performance, as shown in Figure 9 and Appendix B.

5.3 Ablations/Performance Analysis (RQ3 & 4)

Below we use the single-view COSTA (COSTA_{SV}), as multi-view COSTA has a similar performance. We firstly show the superiority of random projection by comparing it with other matrix sketching variants. Subsequently, we ablate and discuss the factors that lead to the success of the random projection. Finally, we show the effect of the number of augmented samples that influence the error bound.

Comparison with Other Feature Augmentations. We compare the random projection with other matrix sketching strategies. Table 1 uses the common tested based on COSTA_{SV} : only the definition of \mathbf{P} and \mathbf{E} are varied. We note that the random projection works consistently better than all the other strategies as the random projection introduces a lesser error compared to the random selection and random noise strategies. It is somewhat surprising that although the random projection is not the optimal solution to Eq. (6), it still outperforms the SVD-based sketching. Such a good performance comes from the following facts: (i) the error bound of random projection is sufficiently small to maintain a good sketch; and (ii) compared to SVD, which is deterministic, random projection adds stochastic perturbations to the model (variance is a source of feature augmentations), which serves as a regularization.

Ablations w.r.t. the Augmentation Type in GCL. Below, we investigate different augmentation types, *i.e.*, the topology graph augmentation and the feature augmentation. To minimize other factors other than the augmentation strategy that might affect the results (*i.e.*, multi-view setting), we opt for the single-view COSTA. We only replace the augmentation type. Table 4 shows that without any augmentations, the model performs badly, showing the necessity of data augmentations. Furthermore, we observe that applying either the topology graph augmentation or the feature augmentation can improve results. However, the improvement of feature augmentation is larger compared to the topology graph augmentation, highlighting the effectiveness of COSTA.

Ablations on the Random Projection. To see where the performance improvement comes from, we conduct an ablation study on the COSTA_{SV} with the random projection. Firstly, we fix the random projection so that it remains the same in each epoch, denoting this variant as *ran_proj_fix*, and then we eliminate the random projection completely (*none_ran_proj*). We experiment on Cora, CiteSeer, PubMed, and DBLP. Figure 10 shows that fixing the random projection throughout the experiment causes the performance drop in all four datasets. In contrast, forming a new random projection matrix per epoch generates a variety of feature augmentations obeying the variance bound of random projection. As projecting is performed along columns of the hidden feature matrix, this is an equivalent of drawing different new nodes (obeying the mean and variance) with each change of random projection. Notably, removing the random projection completely from training decreases the accuracy on three datasets, as expected.

Downsampling vs. Upsampling the Node Dimension. According to the error bound derived in Eq. (10), the bound is related to the number of augmented samples k . The use of COSTA lets us control the number of augmented features by adjusting the number of rows of $\mathbf{P} \in \mathbb{R}^{k \times n}$ in Eq. (6). Downsampling and upsampling are applied by setting $k < |\mathcal{V}|$ and $k > |\mathcal{V}|$ respectively. We denote the $\frac{k}{|\mathcal{V}|}$ as the reduction ratio. Figure 8 shows the relation between the performance and the reduction ratio. Specifically, we obtain the best performance for downsampling (the sketched rows are fewer than the number of nodes), which also accelerates computations of the contrastive loss (smaller similarity/dissimilarity matrices).

6 Conclusions

We have quantitatively and qualitatively analyzed the problems stemming from the topology graph augmentation of current GCL methods, and we have shown that such a strategy suffers from the bias problem. To overcome this bias, we have proposed the feature augmentation framework COSTA. We theoretically proved that the quality of augmented features obtained via COSTA are guaranteed and COSTA accelerates the speed of GCL by working well in the single-view mode. Our results are equivalent or better than results of the standard contrastive multi-view graph augmentation models that rely on topology-based augmentations.

7 Acknowledgements

This work was supported by the National Key Research and Development Program of China (No. 2018AAA0100204) and the Research Grants Council of the Hong Kong Special Administrative Region, China (CUHK 2410021, Research Impact Fund, No. R5034-18).

References

- [1] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910*, 2019.
- [2] Piotr Bielak, Tomasz Kajdanowicz, and Nitesh V Chawla. Graph barlow twins: A self-supervised representation learning framework for graphs. *arXiv preprint arXiv:2106.02466*, 2021.
- [3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
- [4] Yankai Chen, Menglin Yang, Yingxue Zhang, Mengchen Zhao, Ziqiao Meng, Jianye Hao, and Irwin King. Modeling scale-free graphs with hyperbolic geometry for knowledge-aware recommendation. In *WSDM ’22: The Fifteenth ACM International Conference on Web Search and Data Mining*. ACM, 2022.
- [5] Yankai Chen, Yaming Yang, Yujing Wang, Jing Bai, Xiangchen Song, and Irwin King. Attentive knowledge-aware graph convolutional networks with collaborative guidance for personalized recommendation. In *The 38th IEEE International Conference on Data Engineering*, 2022.

- [6] Terrance DeVries and Graham W Taylor. Dataset augmentation in feature space. *arXiv preprint arXiv:1702.05538*, 2017.
- [7] Petros Drineas, Ravi Kannan, and Michael W Mahoney. Fast monte carlo algorithms for matrices i: Approximating matrix multiplication. *SIAM Journal on Computing*, 36(1):132–157, 2006.
- [8] Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*, 2021.
- [9] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.
- [10] Gene H Golub, Alan Hoffman, and Gilbert W Stewart. A generalization of the eckart-young-mirsky matrix approximation theorem. *Linear Algebra and its applications*, 88:317–327, 1987.
- [11] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.
- [12] Hakim Hafidi, Mounir Ghogho, Philippe Ciblat, and Ananthram Swami. Graphcl: Contrastive self-supervised learning of graph representations. *arXiv preprint arXiv:2007.08025*, 2020.
- [13] 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*, pages 1025–1035, 2017.
- [14] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3018–3027, 2017.
- [15] Kaveh Hassani and Amir Hosein Khasahmadi. Contrastive multi-view representation learning on graphs. In *International Conference on Machine Learning*, pages 4116–4126. PMLR, 2020.
- [16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.
- [17] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.
- [18] Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.
- [19] Piotr Koniusz and Hongguang Zhang. Power normalizations in fine-grained image, few-shot image and graph classification. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- [20] Ping Li, Trevor J Hastie, and Kenneth W Church. Very sparse random projections. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 287–296, 2006.
- [21] Edo Liberty. Simple and deterministic matrix sketching. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 581–588, 2013.
- [22] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52, 2015.
- [23] Péter Mernyei and Cătălina Cangea. Wiki-cs: A wikipedia-based benchmark for graph neural networks. *arXiv preprint arXiv:2007.02901*, 2020.

- [24] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, 1999.
- [25] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.
- [26] Zhen Peng, Wenbing Huang, Minnan Luo, Qinghua Zheng, Yu Rong, Tingyang Xu, and Junzhou Huang. Graph representation learning via graphical mutual information maximization. In *Proceedings of The Web Conference 2020*, 2020.
- [27] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710, 2014.
- [28] Connor Shorten and Taghi M Khoshgohfar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019.
- [29] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246, 2015.
- [30] Zixing Song, Ziqiao Meng, Yifei Zhang, and Irwin King. Semi-supervised multi-label learning for graph-structured data. In *CIKM*, pages 1723–1733. ACM, 2021.
- [31] Zixing Song, Xiangli Yang, Zenglin Xu, and Irwin King. Graph-based semi-supervised learning: A comprehensive review. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21, 2022.
- [32] Ke Sun, Piotr Koniusz, and Zhen Wang. Fisher-bures adversary graph convolutional networks. *Conference on Uncertainty in Artificial Intelligence*, 115:465–475, 2019.
- [33] Susheel Suresh, Pan Li, Cong Hao, and Jennifer Neville. Adversarial graph augmentation to improve graph contrastive learning. *CoRR*, abs/2106.05819, 2021.
- [34] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020.
- [35] Petar Velickovic, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. *ICLR (Poster)*, 2019.
- [36] Yulin Wang, Xuran Pan, Shiji Song, Hong Zhang, Gao Huang, and Cheng Wu. Implicit semantic data augmentation for deep networks. *Advances in Neural Information Processing Systems*, 32:12635–12644, 2019.
- [37] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019.
- [38] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. Self-supervised graph learning for recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 726–735, 2021.
- [39] Menglin Yang, Ziqiao Meng, and Irwin King. Featurenorm: L2 feature normalization for dynamic graph embedding. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 731–740. IEEE, 2020.
- [40] Menglin Yang, Min Zhou, Jiahong Liu, Defu Lian, and Irwin King. Hrcf: Enhancing collaborative filtering via hyperbolic geometric regularization. In *Proceedings of the ACM Web Conference 2022*, pages 2462–2471, 2022.

- [41] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in Neural Information Processing Systems*, 33:5812–5823, 2020.
- [42] Junliang Yu, Hongzhi Yin, Xin Xia, Tong Chen, Lizhen Cui, and Nguyen Quoc Viet Hung. Are graph augmentations necessary? simple graph contrastive learning for recommendation. *arXiv preprint arXiv:2112.08679*, 2022.
- [43] Yifei Zhang, Hao Zhu, Ziqiao Meng, Piotr Koniusz, and Irwin King. Graph-adaptive rectified linear unit for graph neural networks. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 1331–1339. ACM, 2022.
- [44] Hao Zhu and Piotr Koniusz. Refine: Random range finder for network embedding. In *ACM Conference on Information and Knowledge Management*, 2021.
- [45] Hao Zhu and Piotr Koniusz. Simple spectral graph convolution. In *International Conference on Learning Representations*, 2021.
- [46] Hao Zhu, Ke Sun, and Peter Koniusz. Contrastive laplacian eigenmaps. *Advances in Neural Information Processing Systems*, 34, 2021.
- [47] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Deep graph contrastive representation learning. *arXiv preprint arXiv:2006.04131*, 2020.
- [48] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Graph contrastive learning with adaptive augmentation. In *Proceedings of the Web Conference 2021*, pages 2069–2080, 2021.

— Appendices —

A Error Bounds of Different Feature Augmentations Preserving Second-order Statistics

Proof of Lemma 4.2 (Matrix Sketching via SVD).

Proof. According to the Eckart-Young-Mirsky theorem [10], $\mathbf{X}_k = \mathbf{P}^\top \mathbf{P} \mathbf{X}$ is the best k -rank approximation of \mathbf{X} and $\|\mathbf{X} - \mathbf{X}_k\|_2 = \sigma_{k+1}$. Thus, we have:

$$\begin{aligned}
 \|\mathbf{X}^\top \mathbf{X} - \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}\|_2 &= \|\mathbf{X}^\top \mathbf{X} - \mathbf{X}^\top \mathbf{P}^\top \mathbf{P} \mathbf{X}\|_2, \\
 &= \|\mathbf{X}^\top (\mathbf{X} - \mathbf{P}^\top \mathbf{P} \mathbf{X})\|_2 = \|\mathbf{X}^\top (\mathbf{X} - \mathbf{X}_k)\|_2, \\
 &\leq \|\mathbf{X}^\top\|_2 \|\mathbf{X} - \mathbf{X}_k\|_2 = \|\mathbf{X}\|_2^2 \frac{\|\mathbf{X} - \mathbf{X}_k\|_2}{\|\mathbf{X}\|_2}, \\
 &\leq \frac{\sigma_{k+1}}{\sigma_1} \|\mathbf{X}\|_F^2 = \frac{\sigma_{k+1}}{\sigma_1} \text{Tr}(\mathbf{X}^\top \mathbf{X}).
 \end{aligned}$$

□

Proof of Lemma 4.4 (Random Row Selection). To prove this Lemma, we use the following theorem from [7].

Theorem A.1. Let $\mathbf{A} \in \mathbb{R}^{d \times n}$, $\mathbf{B} \in \mathbb{R}^{r \times d}$ and $k \in \mathbb{Z}^+$ such that $1 \leq k \leq n$ and $\{p_i\}_{i=1}^n$ be probability distribution over rows of \mathbf{A} and columns of \mathbf{B} such that $p_i \geq \frac{\beta \|\mathbf{A}_{i,:}\|_2 \|\mathbf{B}_{:,i}\|_2}{\sum_{j=1}^n \|\mathbf{A}_{j,:}\|_2 \|\mathbf{B}_{:,j}\|_2}$ for some positive constant $\beta \leq 1$. If matrix $\mathbf{C} \in \mathbb{R}^{d \times k}$ is constructed by sampling columns of \mathbf{A} according to $\{p_i\}_{i=1}^n$ and matrix $\mathbf{D} \in \mathbb{R}^{k \times r}$ is constructed by picking same rows of \mathbf{B} , then with probability at least $1 - \delta$:

$$\|\mathbf{AB} - \mathbf{CD}\|_F^2 \leq \frac{\mu^2}{\beta c} \|\mathbf{A}\|_F^2 \|\mathbf{B}\|_F^2,$$

where $\delta \in (0, 1)$, $\mu = 1 + \sqrt{(8/\beta) \log(1/\delta)}$.

We set $\mathbf{A} = \mathbf{X}^\top \in \mathbb{R}^{d \times n}$, $\mathbf{B} = \mathbf{X} \in \mathbb{R}^{n \times d}$, $\mathbf{C} = \tilde{\mathbf{X}}^\top \in \mathbb{R}^{d \times k}$ and $\mathbf{D} = \tilde{\mathbf{X}} \in \mathbb{R}^{k \times d}$. Note that the distribution $p_i = \frac{\beta \|\mathbf{A}_{i:}\|_2 \|\mathbf{B}_{i:}\|_2}{\sum_{j=1}^n \|\mathbf{A}_{j:}\|_2 \|\mathbf{B}_{j:}\|_2} = \frac{\|\tilde{\mathbf{X}}_{i:}\|}{\|\mathbf{X}\|_F}$ holds with $\beta = 1$. Using theorem A.1, we obtain bound:

$$\begin{aligned} \|\mathbf{X}^\top \mathbf{X} - \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}\|_F^2 &\leq \frac{\mu^2}{k} \|\mathbf{A}\|_F^4, \\ \Rightarrow \|\mathbf{X}^\top \mathbf{X} - \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}\|_F &\leq \frac{\mu}{\sqrt{k}} \|\mathbf{A}\|_F^2 \quad (\text{As } \|\mathbf{X}\|_2 \leq \|\mathbf{X}\|_F), \\ \Rightarrow \|\mathbf{X}^\top \mathbf{X} - \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}\|_2 &\leq \frac{\mu}{\sqrt{k}} \text{Tr}(\mathbf{X}^\top \mathbf{X}), \end{aligned} \quad (11)$$

where $\delta \in (0, 1)$, $\mu = 1 + \sqrt{8 \log(1/\delta)}$. By setting $\frac{\mu}{\sqrt{k}} = \varepsilon$, we have:

$$\mathcal{P}(\|\mathbf{X}^\top \mathbf{X} - \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}\|_F \leq \varepsilon \text{Tr}(\mathbf{X}^\top \mathbf{X})) \geq 1 - e^{-\left(\frac{(\varepsilon \sqrt{k} - 1)^2}{8}\right)}, \quad (12)$$

which completes the proof.

Proof of Lemma 4.6 (Random Projection). We prove the lemma by showing the following inequality holds for any $\mathbf{x} \in \mathbb{R}^n$:

$$\mathcal{P}\left((1 - \varepsilon)\|\mathbf{x}\|_2^2 \leq \left\|\frac{1}{\sqrt{k}} \mathbf{P} \mathbf{x}\right\|_2^2 \leq (1 + \varepsilon)\|\mathbf{x}\|_2^2\right) \geq 1 - e^{-\left(\frac{k\varepsilon^2}{8}\right)}. \quad (13)$$

We firstly show that $\mathbb{E}\left[\left\|\frac{1}{\sqrt{k}} \mathbf{A} \mathbf{x}\right\|_2^2\right] = \|\mathbf{x}\|_2^2$:

$$\begin{aligned} \mathbb{E}\left[\frac{1}{\sqrt{k}} \mathbf{P} \mathbf{x}\right] &= \mathbb{E}\left[\sum_i \frac{1}{k} \sum_j (P_{ij} x_j)^2\right] = \mathbb{E}\left[\sum_i \frac{1}{k} \sum_{j,j'} (P_{ij} P_{ij'} x_j x_{j'})\right] \\ &= \sum_i \frac{1}{k} \mathbb{E}\left[\sum_j P_{ij}^2 x_j^2\right] = \sum_i \frac{1}{k} \sum_j x_j^2 \\ &= \|\mathbf{x}\|_2^2. \end{aligned} \quad (14)$$

This implies $\mathbf{X}_j \sim \mathcal{N}(0, 1)$ where $X_j = \frac{\mathbf{A}_{j:} \mathbf{x}}{\|\mathbf{x}\|_2}$. Then we obtain:

$$\begin{aligned} \mathcal{P}\left(\left\|\frac{1}{\sqrt{k}} \mathbf{P} \mathbf{x}\right\|_2^2 > (1 + \varepsilon)\|\mathbf{x}\|_2^2\right) &= \mathcal{P}\left(\sum_{j=1}^k X_j^2 > (1 + \varepsilon)k\right), \\ &= \mathcal{P}\left(e^{t \sum_{j=1}^k X_j^2} > e^{t(1+\varepsilon)k}\right) \\ &\leq \frac{\mathbb{E}\left[e^{t \sum_{j=1}^k X_j^2}\right]}{e^{t(1+\varepsilon)k}} \quad (\text{apply Markov's inequality}) \\ &= \frac{\prod_{j=1}^k \mathbb{E}\left[e^{t X_j^2}\right]}{e^{t(1+\varepsilon)k}} \quad (\text{as } X_j \text{ is i.i.d}) \\ &= \frac{(\mathbb{E}\left[e^{t X_j^2}\right])^k}{e^{t(1+\varepsilon)k}} \quad (\text{for } 1 < t < 1/2, \text{ it holds that } \mathbb{E}\left[e^{t X_i^2}\right] \leq \left(\frac{1}{\sqrt{1-2t}}\right)) \\ &\leq \left(\frac{1}{\sqrt{1-2t}}\right)^k \left(\frac{1}{e^{t(1+\varepsilon)}}\right)^k \\ &= \left(e^{-\varepsilon + \ln(1+\varepsilon)}\right)^{k/2} \quad (\text{using } \ln(1 + \varepsilon) \leq \varepsilon - \frac{\varepsilon^2}{4}) \\ &\leq e^{-\varepsilon^2 k/8}. \end{aligned} \quad (15)$$

Thus, we have:

$$\mathcal{P}\left(\left\|\frac{1}{\sqrt{k}} \mathbf{P} \mathbf{x}\right\|_2^2 \leq (1 + \varepsilon)\|\mathbf{x}\|_2^2\right) \leq 1 - e^{-\varepsilon^2 k/8}. \quad (16)$$

In the similar way, it is easy to prove that:

$$\mathcal{P}(\|\frac{1}{\sqrt{k}}\|\mathbf{P}\mathbf{x}\|_2^2 \geq (1-\varepsilon)\|\mathbf{x}\|_2^2) \leq 1 - e^{-\varepsilon^2 k/8}. \quad (17)$$

Thus, it holds that:

$$\mathcal{P}((1-\varepsilon)\|\mathbf{x}\|_2^2 \leq \frac{1}{\sqrt{k}}\|\mathbf{P}\mathbf{x}\|_2^2 \leq (1+\varepsilon)\|\mathbf{x}\|_2^2) \geq 1 - e^{-\varepsilon^2 k/8}. \quad (18)$$

Suppose $\mathbf{X}^\top \mathbf{X} - \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \succeq 0$ and \mathbf{x} is the eigenvector of $\mathbf{X}^\top \mathbf{X} - \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}$ corresponding to its largest eigenvalue σ_{\max} , then we have:

$$\|\mathbf{X}^\top \mathbf{X} - \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}\|_2 = \mathbf{x}^\top (\mathbf{X}^\top \mathbf{X} - \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}) \mathbf{x} = \sigma_{\max} \quad (19)$$

$$= \mathbf{x}^\top \mathbf{X}^\top \mathbf{X} \mathbf{x} - \mathbf{x}^\top \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \mathbf{x} \quad (20)$$

$$= \|\mathbf{X}\mathbf{x}\|_2^2 - \|\tilde{\mathbf{X}}\mathbf{x}\|_2^2 \quad (21)$$

$$= \|\mathbf{X}\mathbf{x}\|_2^2 - \|\mathbf{P}\mathbf{X}\mathbf{x}\|_2^2. \quad (22)$$

Applying Eq. (13), there is at least probability $1 - e^{-\varepsilon^2 k/8}$ such that:

$$\begin{aligned} \|\mathbf{X}^\top \mathbf{X} - \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}\|_2 &\leq \|\mathbf{X}\mathbf{x}\|_2^2 + (\varepsilon - 1)\|\mathbf{X}\mathbf{x}\|_2^2 \\ &= \varepsilon \|\mathbf{X}\mathbf{x}\|_2^2 \\ &\leq \varepsilon \|\mathbf{X}\|_2^2 \|\mathbf{x}\|_2^2 (\|\mathbf{x}\|_2^2 = 1) \\ &\leq \varepsilon \text{Tr}(\mathbf{X}^\top \mathbf{X}) \end{aligned} \quad (23)$$

In similar way, it is easy to show that above equation also holds for $\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} - \mathbf{X}^\top \mathbf{X} \succeq 0$. Thus, we have:

$$\mathcal{P}(\|\mathbf{X}^\top \mathbf{X} - \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}\|_2 \leq \varepsilon \text{Tr}(\mathbf{X}^\top \mathbf{X})) \geq 1 - e^{-\varepsilon^2 k/8}, \quad (24)$$

which completes the proof.

B Accelerating the Feature Augmentation with a Very Sparse Random Projection

To accelerate the random projection, approach [20] presented a sparse random projection as an improvement over the Gaussian random projection, in which entries of \mathbf{P} are i.i.d. sampled from:

$$p_{ij} = \begin{cases} \sqrt{s} & \text{with probability } \frac{1}{2s}, \\ 0 & \text{with probability } 1 - \frac{1}{s}, \\ -\sqrt{s} & \text{with probability } \frac{1}{2s}, \end{cases} \quad (25)$$

where $\frac{1}{s}$ denotes the density of matrix \mathbf{P} .

The computation cost of the random projection is related to the sparsity of the sparse random matrix (SRP). The low density of the SRP reduces the computational cost while it may affect the performance at the same time. To explore the relationship between the density and performance, we vary the density of the random matrix in the range of [0.001, 0.002, 0.004, 0.008, 0.016, 0.32, 0.064]. We plot the relationship between the performance and density on three datasets in Figure 9. It is apparent that the performance improves as the density increases. However, the performance reaches a relatively high peak at a low density ($\frac{1}{s} < 0.01$). This suggests that one could enjoy the efficiency provided by the SPMM without sacrificing the performance.

C Statistics of Datasets

Below we describe datasets from Table 5:

- **Cora, CiteSeer, Pubmed, DBLP.** These are well-known citation network datasets, in which nodes represent publications and edges indicate their citations. All nodes are labeled according to paper subjects [17].

Dataset	#Nodes	#Edges	#Features	#Classes
Wiki-CS [23]	11,701	216,123	300	10
Amazon-Computers[22]	13,752	245,861	767	10
Amazon-Photo[22]	7,650	119,081	745	8
Coauthor-CS [29]	18,333	81,894	6,805	15
Coauthor-Physics [29]	34,493	247,962	8,415	5
Cora [17]	2,708	5,429	1,433	7
Citeseer [17]	3,327	4,732	3,703	6
Pubmed [17]	19,717	44,338	500	3
DBLP [17]	17,716	105,734	1,639	4

Table 5: Statistics of datasets used in our experiments.

- **WikiCS.** It is a network of Wikipedia pages related to the computer science, with edges showing cross-references. Each article is assigned to one of 10 subfields (classes), with characteristics computed using the content’s averaged GloVe embeddings. We make no changes to the 20 train/val/test data splits provided by [23].
- **Am-Computer, AM-Photo.** Both of these networks are based on Amazon’s co-purchase data. Nodes represent products, while edges show how frequently they were purchased together. Each product is described using a Bag-of-Words representation based on the reviews (node features). There are ten node classes (product categories) and eight node classes (product categories), respectively [22].
- **Coauthor-CS, Coauthor-Physics.** These are two networks that were extracted from the Microsoft Academic Graph dataset. The edges reflect a collaboration between two authors, while the nodes represent writers. The keywords that each author uses in their articles are utilized to categorize them (Bag-of-Words representation; node features). According to [29], there are 15 author research fields (node classes) and 5 author research fields (node classes).