

MultiBiSage: A Web-Scale Recommendation System Using Multiple Bipartite Graphs at Pinterest

Saket Gurukar[†], Nikil Pancha[‡], Andrew Zhai[‡], Eric Kim[‡], Samson Hu[‡],
Srinivasan Parthasarathy[†], Charles Rosenberg[‡], Jure Leskovec[‡]

[†] The Ohio State University, [‡] Pinterest
United States

gurukar.1@osu.edu, {npancha, andrew, erickim, sansonhu}@pinterest.com
srini@cse.ohio-state.edu, {crosenberg, jure}@pinterest.com,

ABSTRACT

Graph Convolutional Networks (GCN) can efficiently integrate graph structure and node features to learn high-quality node embeddings. These embeddings can then be used for several tasks such as recommendation and search. At Pinterest, we have developed and deployed PinSage, a data-efficient GCN that learns pin embeddings from the Pin-Board graph. The Pin-Board graph contains pin and board entities and the graph captures the pin belongs to a board interaction. However, there exist several entities at Pinterest such as users, idea pins, creators, and there exist heterogeneous interactions among these entities such as add-to-cart, follow, long-click.

In this work, we show that training deep learning models on graphs that captures these diverse interactions would result in learning higher-quality pin embeddings than training PinSage on only the Pin-Board graph. To that end, we model the diverse entities and their diverse interactions through multiple bipartite graphs and propose a novel data-efficient MultiBiSage model. MultiBiSage can capture the graph structure of multiple bipartite graphs to learn high-quality pin embeddings. We take this pragmatic approach as it allows us to utilize the existing infrastructure developed at Pinterest – such as Pixie system [7] that can perform optimized random-walks on billion node graphs, along with existing training and deployment workflows. We train MultiBiSage on six bipartite graphs including our Pin-Board graph. Our offline metrics show that MultiBiSage significantly outperforms the deployed latest version of PinSage on multiple user engagement metrics.

KEYWORDS

Recommendation Systems, Graph based Recommendation Systems, Bipartite Graphs

ACM Reference Format:

Saket Gurukar[†], Nikil Pancha[‡], Andrew Zhai[‡], Eric Kim[‡], Samson Hu[‡], Srinivasan Parthasarathy[†], Charles Rosenberg[‡], Jure Leskovec[‡]. 2022. MultiBiSage: A Web-Scale Recommendation System Using Multiple Bipartite

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

© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06...\$15.00
<https://doi.org/10.1145/1122445.1122456>

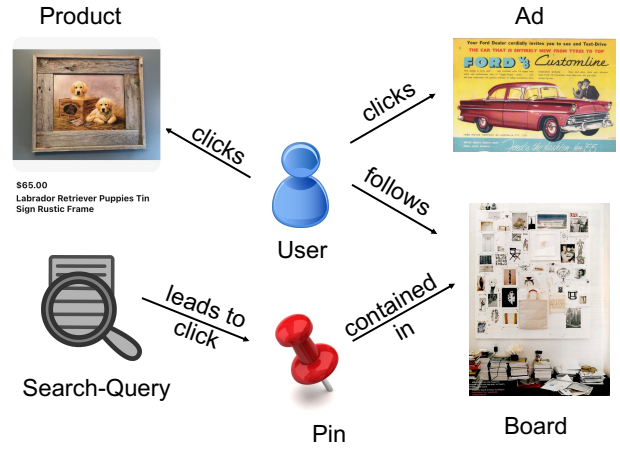


Figure 1: Pinterest's Heterogeneous Graph

Graphs at Pinterest. In . ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/1122445.1122456>

1 INTRODUCTION

Pinterest helps its users discover things they love by using 100+ billion pins present on its platform. The users discover things by exploring the personalized recommendations that are powered by Graph Convolutional Network (GCN) based recommendation system, PinSage[36]. PinSage aggregates the visual and textual features of pins along with their local graph neighborhood from the Pin-Board graph to generate the pin embeddings. The learned pin embeddings are then fed as input to several machine learning models at Pinterest for personalized recommendation, spam classification, ads recommendation, and search.

The PinSage [36] model currently relies on only the Pin-Board bipartite graph to identify the local graph neighborhood of the pin. However, at Pinterest, we have multiple entities such as users, idea-pins, creators, products, ads, search-queries, boards, along with other entities. Moreover, these entities can have diverse interactions on Pinterest. Figure 1 shows an example heterogeneous graph at Pinterest. Here, a user can click on a product or an ad, a user can follow a board, a pin can belong to a board, a user can enter a search-query and then click on a pin shown in the search results.

We hypothesize that these interactions among entities contain rich signals that can be utilized to learn high-quality pin embeddings.

Recently, several extensions of Graph Convolutional Networks for heterogeneous graphs have been proposed [3, 16, 19, 28, 38]. These proposed models often operate on moderate-size to large-scale graphs. One such heterogeneous graph model, HGT [16], showed promising results on Open Academic Graph [39] consisting of 179 million nodes and 2 billion edges. However, the proposed version of HGT is trained on a single machine and requires all the training data (heterogeneous graph and node features) to be present on that single system. At Pinterest, in an arbitrary version of our graph, we have 2+ billion pins, 2+ billion boards, 7+ billion pin-board edges and 400+ million users [23]. The storage of graph and node features itself requires more than 3TB of space. Hence, given the scale of the data at Pinterest, training existing heterogeneous graph models on a single machine is not feasible. There also exist a few frameworks such as Distributed Deep Graph Library (DistDGL) [40] and Aligraph [41] that support distributed training of heterogeneous GNNs models. DistDGL and Aligraph trained distributed GCNs models on Ogbn-papers100m [15] (~111 million nodes) and Taobao-large (~483 million nodes) graphs, respectively. Both DistDGL and Aligraph rely on multi-processing for "on-the-fly" sampling of node's neighbors. On a 4+ billion node graph, the "on-the-fly" sampling of neighbors in every epoch or after a few epochs, results in a high training time of the model. The running time of sampling can be decreased by increasing the number of workers but that results in additional costs. Hence, it is challenging to train the heterogeneous graph models on web-scale graphs.

Present work: In this work, we adopt a pragmatic approach that allows us to utilize the existing infrastructure developed at Pinterest to train heterogeneous graph models. The existing infrastructure includes the random-walker system, Pixie [7], that loads the graph (without node features) in a single AWS x1.32xlarge (2 TB RAM, 128 cpus) instance and utilizes C++ routines to identify the local graph neighborhood of the nodes. We also utilize the existing model training and feature extraction workflows. The approach can be summarized as follows. Given a heterogeneous graph we first decompose it into multiple bipartite graphs. We then compute the local graph neighborhood of pins with the help of Pixie. To efficiently aggregate the node features and node neighbor features from multiple bipartite graphs, we propose a transformer based MultiBiSage model. Given k bipartite graphs, MultiBiSage learns k pin embeddings with the help of transformer [29] model where each pin embedding corresponds to each input bipartite graphs. These pin embeddings are then aggregated with another transformer layer that generates the final pin embedding. The contributions of our work are summarized as follows

- (1) To the best of our knowledge, this is the largest-ever application of heterogeneous graph models on web-scale graphs with over 4.8+ billion nodes and 9.7+ billion edges.
- (2) We show that our proposed MultiBiSage model significantly outperforms the currently deployed PinSage model.
- (3) We present several ablations studies and case study that helped us arrive at the design of MultiBiSage.

2 PRELIMINARIES AND RELATED WORK

In this section, we present the necessary definitions and provide a review of related work that utilize graph convolutional networks for recommendations. We also review distributed training frameworks for heterogeneous graph models.

2.1 Definitions

Definition 2.1. Heterogeneous Graph: A heterogeneous graph $H = (V, E, T, R, X, \phi, \psi, \varphi)$ is a graph where V, E, T, R are the set of nodes, set of edges, set of node types, and set of relations, respectively. A node $v \in V$ has node type $\phi(v) : V \rightarrow T$ and an edge $e(u, v) \in E$ between two nodes u, v has edge type $\psi(e(u, v)) : E \rightarrow R$. A node v can have node features denoted by $\varphi(v) : V \rightarrow X$.

Heterogeneous graphs can naturally capture multi-modal interactions in the real-world. For instance, nodes can represent diverse entities such as users, pins, boards, products, and ads; while edges can represent diverse interactions types such as click, add-to-cart, and purchase.

Definition 2.2. Bipartite Graph: A bipartite graph $B = (V_1, V_2, E, X, \varphi)$ is a graph consisting of two sets of nodes V_1 and V_2 and an edge $e(u, v) \in E$ between two nodes $u \in V_1, v \in V_2$ has one edge type. A node v can have node features denoted by $\varphi(v) : V \rightarrow X$.

Definition 2.3. Heterogeneous Graph Embedding: Given a heterogeneous graph H , a heterogeneous graph embedding method learns a function $f(H) : V \rightarrow \mathbb{R}^d$ such that $\Theta = f(H)$ is a matrix consisting of d -dimensional node embeddings. The function f is learned such that similarity between nodes in the heterogeneous graph is approximated by the closeness between nodes in the embedding space.

In this work, we aim to learn a embedding function f' such that

$$f(H) \approx f'(\cup_{V_i} \cup_{V_j} \cup_{E_k} B(V_i, V_j, E_k, X, \varphi)) \quad (1)$$

as computing $f(H)$ might not always be feasible.

2.2 Related Work

2.2.1 Proximity-based heterogeneous graph models. Given a heterogeneous graph, the proximity preserving heterogeneous graph models rely on the notion of similarity between nodes in the graph space to learn the node embedding. The similarity in the graph space can be defined through random-walks visits [11] or first/second-order proximity between nodes [27]. Metapath2vec [6] is a popular random-walk-based model that learns the embeddings of nodes in the heterogeneous graph by performing random walks through meta-path schema. Metapath2vec inspired several meta-path schema-based heterogeneous graph models. For instance, HIN2Vec [8] learns node embeddings by estimating the presence of meta-path between two nodes. Meta-path schema weighting techniques are also proposed [4, 26] to learn efficient node embeddings. Other random-walk-based approaches, proposed models that do not require specifying meta-paths. For instance, JUST [17] proposed a Jump and Stay random-walk strategy while Hao et al.[13] proposed an attention mechanism for performing random-walk on heterogeneous graphs. The first/second-order proximity-based methods define proximity based on the edge-weights or the number of times

the nodes are part of a subgraph or meta-graphs [34]. PTE [28] decomposes a heterogeneous graph into multiple bipartite graphs and learns node embeddings by preserving the first-order and second-order proximity between nodes in each bipartite graph. The PTE model is trained through the proposed joint training framework. PTE is considered as a shallow network embedding method as it is implicitly factorizing a form of Laplacian matrix [24]. The interested readers can refer to the surveys of heterogeneous graph models [31, 34].

2.2.2 Graph Convolutional Networks for Recommendation: In recent years, we have seen the superior performance of graph convolutional networks [12, 21, 30] on several machine learning tasks on graphs. Relational GCN [25] extends GCN to graph with different edge types by proposing a learnable weight matrix for each relation to improve link prediction performance. Neural Graph Collaborative Filtering (NGCF) [32] operates on user-item bipartite graph and explicitly incorporates the high-order connectivity between users and items to learn their embeddings. LightGCN [14] simplifies GCN for recommendation by proposing linear embedding propagation between user and items.

2.2.3 Heterogeneous Graph Neural Networks: HAN [33] proposed a hierarchical, node-level and semantic-level attention based heterogeneous graph neural network. HAN requires performing random-walks based on meta-paths. HetGNN [37] proposed a random-walk with restart strategy that first samples heterogeneous neighbor nodes and then aggregates them based on neighbor features with Bi-LSTM. GATNE[2] utilizes heterogeneous skip-gram objective function to learn node embeddings of attributed multiplex heterogeneous graphs. The training samples for skip-gram are curated by performing random-walks on nodes with same relations. HGNCN [42] proposed a heterogeneous graph convolutional network that can learn fine-grained relational features of the heterogeneous graph. Heterogeneous Graph Transformer (HGT) [16] proposed a scalable model that can handle dynamic heterogeneous graphs. HGT introduced heterogeneous mutual attention and heterogeneous message passing to perform convolutions on heterogeneous graphs. Note that, these heterogeneous graph models are trained on single-machine system and training these models on a distributed setup requires solving many challenges which are discussed in the next section.

2.2.4 Distributed Frameworks for Heterogeneous Graph Models: The distributed training frameworks for heterogeneous graphs in general addresses three main scalability challenges related to storage, sampling and GNN operations. The AliGraph [37] system address these challenges by proposing three layers: i) Storage layer partitions the graphs using four graph partition algorithms and introduces indexing and caching operations for retrieval of node and its neighbors features, ii) Sampling layer identify the node's neighbors which are required by GNN. Sampling is efficient as each worker sample neighbors on the assigned graph partition, and iii) Operations layers that implement aggregation and combination operation of GNN in an optimized manner. DistDGL [40] stores the graphs on multiple devices by partitioning the graph using METIS [20] algorithm. The node features are stored using a distributed key-value store. For sampling, DistDGL introduced a distributed

Algorithm 1 Training Pipeline of MultiBiSage

- 1: Construct the bipartite graphs from the logs at Pinterest.
 - 2: Get the local graph structure of the nodes from the bipartite graphs.
 - a: Perform random-walks on each bipartite graph with the help of in-house random-walker Pixie system.
 - b: Select top-k highly visited neighbors of the nodes from the random-walks.
 - 3: Collect training data.
 - a: Curate query-pin and engaged-pin from multiple surfaces at Pinterest.
 - b: Collect the visual and textual features of query pin, engaged pin and their neighbors.
 - c: Collect a random sample of million pins and treat these pins as negative samples.
 - 4: Train MultiBiSage model with a modified Sampled Softmax [18].
-

sampler that follows client-server model to samples the node neighbors in parallel. DistDGL perform random-walks on each partition and a single random-walk does not move across different partitions. We estimate that the migration to these new frameworks in the Pinterest software ecosystem would also incur significant infrastructure costs. P3 [9] proposed a novel distributed training approach that eliminates partitioning overheads and high communication among workers. P3 introduced a pipelined push-pull parallelism based execution strategy for the distributed training of GNNs.

3 METHODOLOGY

Given the scale of Pinterest data, we require an optimized data curation and ingestion pipeline to feed the bipartite graph features to the machine learning model. Moreover, we also require a novel model architecture to better capture the bipartite graph features in order to learn high-quality pin embeddings. We describe these details in this section.

3.1 Training Pipeline

Algorithm 1 shows the training pipeline of MultiBiSage. The first step is the construction of bipartite graphs from the user's interaction logs at Pinterest. However constructing these graphs on a single machine is time-consuming and not practical. Hence, we develop a spark-based graph generator, which given the two entity-types and the interaction type, curates the corresponding bipartite graph by parsing the logs. The spark-based graph generator also includes a degree-based graph pruning algorithm that can reduce the number of the edges in the graph. The algorithm can be described as follows: let a and b be the specified minimum and maximum node degrees and let p be the pruning factor between 0 and 1. The pruning algorithm selects node (say u) with node degree greater than a and then randomly removes the edges incident on node u such that new node degree of node u is $\min(a * p, b)$.

Once the bipartite graphs are created, our next goal is to identify the neighbors of pin nodes from the bipartite graphs. As mentioned earlier, the "on-the-fly" sampling of neighbors during model training from billion-nodes and billion-edges graph is extremely inefficient and not practical. Hence, we identify the node neighbors before training MultiBiSage using Pixie [7] system. Pixie system loads the entire graph in 2 TB RAM machine and performs random-walks using optimized C++ threads. For each pin node, we select top-k


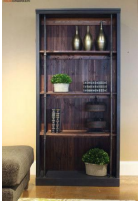




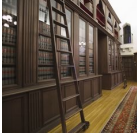










Candidate	Graph	Rank-1	Rank-2	Rank-3	Rank-4	Rank-5
	Pin-Board					
	User-Product					
	User-Ad					
	SearchQuery-Pin					

Table 1: Candidate pin and its neighbors from diverse bipartite graphs.

highly visited pin neighbors. The pseudo-code of Pixie’s random-walker is shared in the Appendix A. Table 1 shows a sample of the top-5 highly visited pins of a candidate pin.

Next, we collect the training samples and the features of the training samples with the help of spark jobs. The training sample is constructed as follows. A user interacts with a pin (referred to as query pin) and is shown recommendations, if the user interacts with one of the recommended pin (referred to as engaged pin) we treat the query and engaged pin as positive pair. The goal of MultiBiSage model is to learn pin embeddings such that distance between query and engaged pin is close to each other. We describe the model architecture next.

3.2 MultiBiSage model architecture

Figure 2 shows the model architecture of MultiBiSage. Let p and x_p be the pin and embedding of pin p , respectively. Here $x_p \in \mathbb{R}^d$ is l_2 normalized d -dimensional vector. Let $v_p \in \mathbb{R}^{d_v}$ and $t_p \in \mathbb{R}^{d_t}$ be the visual and textual features of pin p , respectively. Assume

there are k bipartite graphs with G_i being the i^{th} bipartite graph. Let n be the number of neighbors from each bipartite graph. Let $\vartheta_{i,p} \in \mathbb{R}^{n \times d_v}$ and $\tau_{i,p} \in \mathbb{R}^{n \times d_t}$ be the visual and textual features of the neighbors of pin p from G_i , respectively.

Mathematically, the MultiBiSage model can be described as follows. First, we pass the visual and textual features of pin and its neighbors through a feedforward neural network (FFN) to learn the intermediate representation of dimension size d_h .

$$\begin{aligned}
 x_{p,v_i} &= FFN(v_p) = ReLU(v_p W_{v_i} + b_{v_i}) \\
 x_{p,t_i} &= FFN(t_p) = ReLU(t_p W_{t_i} + b_{t_i}) \\
 x_{p,\vartheta_i} &= FFN(\vartheta_{i,p}) = ReLU(\vartheta_{i,p} W_{\vartheta_i} + b_{\vartheta_i}) \\
 x_{p,\tau_i} &= FFN(\tau_{i,p}) = ReLU(\tau_{i,p} W_{\tau_i} + b_{\tau_i})
 \end{aligned} \tag{2}$$

where $W_{v_i} \in \mathbb{R}^{d_v \times d_h}$, $W_{t_i} \in \mathbb{R}^{d_t \times d_h}$, $W_{\vartheta_i} \in \mathbb{R}^{d_v \times d_h}$, and $W_{\tau_i} \in \mathbb{R}^{d_t \times d_h}$ be the learnable weight matrices. Here, $b_{v_i} \in \mathbb{R}^{d_h}$, $b_{t_i} \in \mathbb{R}^{d_h}$, $b_{\vartheta_i} \in \mathbb{R}^{d_h}$ and $b_{\tau_i} \in \mathbb{R}^{d_h}$ are the biases parameters.

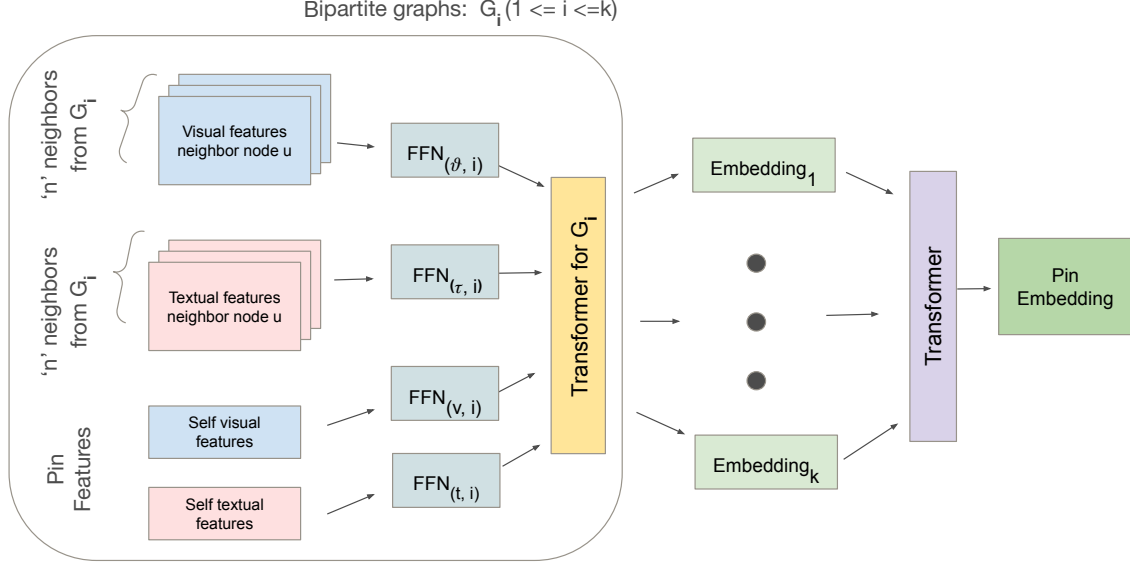


Figure 2: The model architecture of MultiBiSage.

The intermediate representations $x_{v,i}, x_{t,i}, x_{\theta,i}, x_{\tau,i}$ and a global token $x_{g_i} \in \mathbb{R}^{d_h}$ are then concatenated and passed to the transformer layer.

$$s_{p,i} = \text{Concat}(x_{g_i}, x_{v,i}, x_{t,i}, x_{\theta,i}, x_{\tau,i}) \quad (3)$$

where $s_{p,i} \in \mathbb{R}^{(1+2 \times (1+n)) \times d_h}$. Here, $s_{p,i}$ can be considered as a sequence with $1 + 2 \times (1 + n)$ number of tokens and each token is represented in d_h dimensional space.

We pass the sequence $s_{p,i}$ through the multihead attention layer [29] where each input token has three intermediate representations referred to as query (Q), key (K), and value (V). The attention (A) between all tokens is computed using the below equation.

$$A(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

where d_k is dimension size of queries and keys. We apply attention with H number of heads as shown below:

$$\begin{aligned} \text{MultiHead}_i(Q, K, V) &= \text{Concat}(\text{head}_1, \dots, \text{head}_H)W^O \\ \text{head}_j &= A(QW_j^Q, KW_j^K, VW_j^V) \end{aligned} \quad (5)$$

where $W_j^Q \in \mathbb{R}^{d_h \times d_k}$, $W_j^K \in \mathbb{R}^{d_h \times d_k}$, $W_j^V \in \mathbb{R}^{d_h \times d_v}$ and $W^O \in \mathbb{R}^{Hd_v \times d}$ are trainable parameters and $d_k = d_v = d_h/H$.

The representation of global token x_{g_i} after passing it through multihead attention is treated as pin embedding $x_{p,i} \in \mathbb{R}^d$ of pin p from bipartite graph G_i . The pin embeddings $x_{p,i}$ are then passed to second transformer along with global token $x_g \in \mathbb{R}^d$ as shown below:

$$s = \text{Concat}(x_g, x_{p,1}, x_{p,2}, \dots, x_{p,k}) \quad (6)$$

where $s \in \mathbb{R}^{(1+k) \times d}$. The representation of global token x_g after passing it through transformer layer is L_2 normalized and considered as pin embedding $x_p \in \mathbb{R}^d$.

3.2.1 Training objective. Let q_i , and e_i be the i^{th} query pin and engaged pin in a batch \mathcal{B} , respectively. Let $x_{q_i} \in \mathbb{R}^d$ and $x_{e_i} \in \mathbb{R}^d$ be the embedding of these pins computed by MultiBiSage. Also, let $C_{\mathcal{B}} = \{e_1, \dots, e_{|\mathcal{B}|}\}$ be the set of all the engaged pins in the batch. Then, we minimize the sampled softmax loss function [18] as shown below:

$$-\frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \log \frac{e^{\langle x_{q_i}, x_{e_i} \rangle} - \log Q_p(e_i | q_i)}{\sum_{e' \in C_{\mathcal{B}}} e^{\langle x_{q_i}, x_{e'} \rangle} - \log Q_p(e' | q_i)} \quad (7)$$

where $\langle \cdot, \cdot \rangle$ represents the dot product between learned embeddings. Here, we are performing softmax over sampled classes $C_{\mathcal{B}}$ instead of performing softmax over all the classes. The softmax over-sampled classes introduce sampling bias in the full softmax computation. This bias is corrected through the proposal distribution $Q_p(e_i | q_i)$ which is the probability of e_i being included as a positive sample in the training batch. We utilize count-min sketch [5] to estimate $Q_p(e_i | q_i)$ in a streaming manner.

Notice that, for query-pin q_i , we are treating all the other engaged pins e_j ($e_j \neq e_i$) in the batch as negative samples. If a pin is popular, it would frequently appear as an engaged pin (say e_k) in the batch. The above loss function would then unfairly penalize engaged pin e_k , as they are more likely to be selected as negative pin during training. To address this issue, we adopt the mixed negative sampling approach [35], in which we select a set of random negatives \mathcal{M} where $|\mathcal{M}| = |\mathcal{B}|$ and compute the below loss function.

Graphs	Num. Entity 1	Num. Entity 2	Num. Edges
Pin-Board-ContainedIn	1,995,423,350	1,927,181,284	7,144,399,779
User-Product-LC	74,130,469	40,198,339	472,307,057
User-AD-LC	114,059,486	9,392,321	710,736,288
SearchQuery-Pin-LC	397,290,595	189,846,404	1,307,227,847
Creator-IdeaPin-Created	1,508,535	12,349,297	13,984,765
IdeaPin-User-FollowedBy	2,502,566	41,002,754	98,570,789

Table 2: The statistics of Pinterest’s bipartite graphs. Graph nomenclature convention: Entity 1 - Entity 2 - EdgeType. The graphs represent a random subset of users, pins, boards, search-queries, ads, products, creators, idea pins, and video pins.

Statistic	Count
Number of distinct Query Pins	158,489,849
Number of distinct Engaged Pins	150,706,623
Number of distinct Query-Engaged Pins	573,056,337

Table 3: The training data volume collected over a certain temporal period.

$$-\frac{1}{|\mathcal{M}|} \sum_{i=1}^{|\mathcal{M}|} \log \frac{e^{\langle x_{q_i}, x_{e_i} \rangle} - \log Q_n(e_i)}{\sum_{e' \in \mathcal{M}} e^{\langle x_{q_i}, x_{e'} \rangle} - \log Q_n(e')} \quad (8)$$

where $Q_n(e)$ represents the probability of sampling pin e . We also utilize count-min sketch [5] to estimate $Q_n(e)$ in a streaming manner. We optimize both the loss function in Equation 7 and Equation 8 by summing them with equal weightage.

4 EXPERIMENTS

4.1 Experiment Details

4.1.1 Dataset. We perform experiments on the six bipartite graphs curated from the logs present at Pinterest. The statistics of these graphs are present in Table 2. All the graphs are collected over the period of one year and represent a random subset of users, pins, boards, search-queries, ads, products, creators, idea pins, and video pins. The edge type LC refers to the long click. A long click refers to the interaction where a user clicks on a pin and does not return to the Pinterest website in 10 secs. LC often acts as a signal that the user found the item that he/she was searching for. The raw Pin-Board graph consists of 100+ billion edges. We perform the degree-based pruning to reduce the number of edges from 100+ billion to 7 billion by setting the parameters as follows: minimum node degree=10, maximum node degree=10000 and prune factor=0.86. We apply the degree-based pruning to only pin-board graph. In the case of the SearchQuery-Pin-LC graph, each search-query text is considered as a node, and no text processing is performed on the search-query text. We select 50 neighbors from every bipartite graph. Table 3 contains the training data statistics on which we train all the models. The training data is collected over a certain temporal period.

4.1.2 Experimental Setting: We train the models with Adam optimizer with a learning rate of 0.002 and batch size of 8032. We gradually increase learning rate in optimizer [10] using Cosine Annealing [22].

Surface/Entity Type	Engagement Type	PinSage	MultiBiSage
Organic Surface	Add-to-cart	0.893	0.907 (+1.57%)
	Checkout	0.893	0.907 (+1.57%)
Ads	Good-click-through	0.651	0.684 (+5.07%)
Related Products	Long-click	0.733	0.749 (+2.18%)
Idea Pin	Close-ups	0.724	0.776 (+7.18%)
	Repin	0.797	0.849 (+6.52%)
Video Pin	Close-ups	0.446	0.47 (+5.38%)
	Repin	0.571	0.603 (+5.60%)

Table 4: Recommendation performance: Recall@10

4.1.3 Baselines. We primarily focus on the comparison of the state-of-the-art production model PinSage currently deployed at Pinterest. We also include advanced deep learning models and a few of their variants. The baselines we consider are

- **Transformer** [29]: We pass the visual and textual embeddings of the neighbors from the bipartite graphs to the model. The five additional bipartite graphs increase the sequence length passed to the transformer by 5x. This model can be considered as an extension of the deployed PinSage model to the six bipartite graphs.
- **SharedTransformer**: Instead of having a transformer for every bipartite graph, we have one transformer model that is shared across multiple bipartite graph features.
- **NFFNTransformer**: Here, the bipartite graph embeddings are passed through a feed forward neural network layer which aggregates the embeddings from all the bipartite graphs.
- **NSumTransformer**: We perform an element-wise sum of all the bipartite graph embeddings and then pass the resultant embeddings to the transformer.
- **NHadamardTransformer**: We perform the Hadamard product of all the bipartite graph embeddings and then pass the resultant embeddings to the transformer.
- **PinFeatToLastTransformer**: Here, we modify the Multi-BiSage model and pass the pin-features to only the last transformer layer instead of repeatedly passing them to every bipartite graph transformer.
- **AggregateByFFN**: Here, we modify the MultiBiSage model and replace the last transformer layer with a feed forward neural network that aggregates the bipartite graph pin embeddings $(x_{p,i})$.

4.1.4 Metrics. We measure the performance of our recommendations on multiple types of engagement. These engagements happen on different surfaces at Pinterest. The organic surface refers to pins present on the home feed (the landing page with an endless set of recommended pins) surface. The related products surface corresponds to the page where a user clicks on a pin and is shown pin recommendations. The entities such as Ads, Idea Pins, and Video Pins can be present on both organic and related products surface. The users can have various types of engagement over these surfaces and entities. The Add-to-Cart and Checkout engagement corresponds to the user’s actions related to purchasing. The Good-click-through engagement occurs when the user is displayed an ad and after clicking the ad, the user does not return to the website

in less than 30 secs. In Long click engagement, the user clicks on a pin and does not return back to the website in less than 10 secs. Close-ups refer to the action where the user zooms in into the pin. Repin refers to the action where the user saves the pin to his/her board. The users often perform these engagements on our set of recommended pins.

Our goal is to improve the quality of user engagement by providing better recommendations. Here, we measure the quality of the recommendations using the recall@10 metric. Given a pair of query pin q and engaged pin e , we first generate the query pin embedding x_q and engaged pin embedding x_e . We additionally generate the embeddings of one million randomly sampled pins. We then find the ten nearest neighbors pins of query pin q using x_q in the embedding space. The metric Recall@10 computes the fraction of times the ground-truth engaged pin e appears in the top-10 nearest neighbors of query pin.

4.2 Results

4.2.1 Comparison with Baselines: In Table 4, we compare our proposed model MultiBiSage against the currently deployed PinSage version at Pinterest. The percentages in Table 4 in column MultiBiSage refers to the percentage improvement over PinSage model. We observe that MultiBiSage outperforms PinSage on all the performance metrics. The difference in performance between MultiBiSage and PinSage is statistically significant with standard paired t-test at significance level 0.01. On Ads, Idea Pins, and Video Pins, we observe more than 5% improvement over PinSage. These results shows that training MultiBiSage on graphs that captures diverse interactions result in learning higher-quality pin embeddings than training PinSage on only Pin-Board graph.

We compare the performance of MultiBiSage with the baselines in Table 5. The percentages shown in parenthesis refer to the percentage improvement over the deployed PinSage model. The difference in performance between MultiBiSage and the baselines is statistically significant with standard paired t-test at significance level 0.01. We also observe that most of the baselines result in the improvement of recommendation performance over PinSage. This is partly due to the fact that these models can utilize the additional neighborhood context of pins from diverse bipartite graphs. Next, we observe that the distribution of the node’s neighbors from each bipartite graph is different. This can be inferred from the performance of the Shared-Transformer model where the transformer is shared across multiple bipartite graph features. The models NSum-Transformer, NHadarnard-Transformer, and NFFN-Transformer recommendation is similar to that of Transformer. The difference between them is not statistically significant with standard paired t-test at significance level 0.01. The MultiBiSage model variants PinFeatToLastTransformer and AggregateByFFN recommendation performance is poorer than that of MultiBiSage.

4.2.2 Impact of the number of neighbors: We study the impact of the number of neighbors of nodes on the MultiBiSage performance. If the lower number of neighbors from different bipartite graphs can achieve similar performance, then the model inference would be faster. Table 6 shows the model performance on 10, 20 and 50 number of neighbors. We observe that the recommendation performance of MultiBiSage with 10 and 20 number of neighbors is

similar. The difference in performance between them is not statistically significant with paired t-test at significance level 0.01. However, the difference in the performance of MultiBiSage with 50 neighbors is significantly better than that of MultiBiSage with 10 and 20 neighbors.

4.2.3 Ablation Study: Next, we study the impact of incorporating each bipartite graph on the MultiBiSage performance. The result of this ablation study is in shown in Table 7 and Table 8. In Table 7, we train the MultiBiSage model with each introduced bipartite graph along with Pin-Board graph, while in Table 8 we train MultiBiSage with only the introduced bipartite graphs. We also have an experiment where we train MultiBiSage on all the five bipartite graphs except the Pin-Board graph in Table 8. From Table 7, we observe MultiBiSage+UserAd+PB achieves over 4.3% improvement on Ads over deployed PinSage model. At the same time, both MultiBiSage+CCI+PB and MultiBiSage+IUF+PB that are trained on graphs containing Idea pins show from 3.18% to 4.42% improvement on Idea pins over the PinSage model. Since Idea pins can also contain videos, we observe that these two models show improvement over videos pin recommendation tasks. The model MultiBiSage+UserProd+PB and MultiBiSage+QueryPin+PB show improvement in the recommendation performance across multiple metrics. However, the best performance on all the metrics is achieved by MultiBiSage trained with all six bipartite graphs. The difference in the performance of MultiBiSage with all the models is statistically significant with pair t-test at significance level 0.01. In Table 8, we observe that training MultiBiSage on each bipartite graph alone results in poorer performance as compared to the PinSage model. This result is less surprising as the Pin-Board graph is orders of magnitude large than the other bipartite graphs. If we train MultiBiSage with all bipartite graphs except Pin-Board, we see improvement over PinSage on Ads and Idea Pins.

4.2.4 Case Study: We present a case study in Table 9 where a user first interacts with the shown query pin (a foldable electric scooter) and then immediately interacts with the engaged pin. The query pin is contained in a board titled "Electric Vehicles" and that board also contains cars. As a result, the top-5 recommendations provided by the PinSage model contain vehicles including cars. However, the engaged pin’s embedding learned by PinSage is not too similar to that of query pin’s embedding, as the rank of engaged-pin – computed based on the similarity between the query and engaged pin’s embeddings – is 364. On the other hand, MultiBiSage learns the engaged pin’s embedding closer to that of query-pin. As a result, engaged pin’s rank is 5 with MultiBiSage – a positive recall with ten nearest neighbors. In addition, we see that the other 3 nearest pins retrieved by MultiBiSage are also related to electric scooters.

5 CONCLUSION

We develop a novel web-scale deep learning model, called MultiBiSage, that leverages multiple bipartite graphs (each capturing diverse entities and different types of interactions) to learn high-quality pin embeddings on a production system at Pinterest. We train MultiBiSage on six bipartite graphs that include the billion nodes Pin-Board graph. Our experiments show that our proposed

Surface/Entity Type	Engagement Type	Transformer	Shared Transformer	NSum Transformer	NHadamard Transformer	NFFN Transformer	PinFeatToLast Transformer	Aggregate ByFFN
Organic Surface	Add-to-cart Checkout	0.900 (+0.8%)	0.902 (+1.0%)	0.901 (+0.9%)	0.899 (+0.7%)	0.900 (+0.8%)	0.874 (-2.1%)	0.885 (-0.9%)
		0.898 (+0.6%)	0.900 (+0.8%)	0.900 (+0.8%)	0.901 (+0.9%)	0.902 (+1.0%)	0.875 (-2.0%)	0.885 (-0.9%)
Ads	Good-click-through	0.670 (+2.9%)	0.676 (+3.8%)	0.669 (+2.8%)	0.668 (+2.6%)	0.672 (+3.2%)	0.629 (-3.4%)	0.661 (+1.5%)
Related Products	Long-click	0.740 (+1.0%)	0.738 (+0.7%)	0.738 (+0.7%)	0.737 (+0.5%)	0.741 (+1.1%)	0.708 (-3.4%)	0.741 (+1.1%)
Idea Pin	Close-ups Repin	0.749 (+3.5%)	0.757 (+4.6%)	0.747 (+3.2%)	0.746 (+3.0%)	0.752 (+3.9%)	0.729 (+0.7%)	0.760 (+5.0%)
		0.827 (+3.8%)	0.833 (+4.5%)	0.829 (+4.0%)	0.824 (+3.4%)	0.827 (+3.8%)	0.808 (+1.4%)	0.837 (+5.0%)
Video	Close-ups Repin	0.449 (+0.7%)	0.454 (+1.8%)	0.451 (+1.1%)	0.453 (+1.6%)	0.451 (+1.1%)	0.432 (-3.1%)	0.463 (+3.8%)
		0.578 (+1.2%)	0.581 (+1.8%)	0.575 (+0.7%)	0.573 (+0.4%)	0.578 (+1.2%)	0.553 (-3.2%)	0.591 (+3.5%)

Table 5: Baselines. The percentages refer to the percentage improvement over deployed PinSage model.

Surface/Entity Type	Engagement Type	10	20	50
Organic	Add-to-cart Checkout	0.899	0.903	0.907
		0.900	0.904	0.907
Ads	Good-click-through	0.681	0.68	0.684
Related Products	Long-click	0.738	0.743	0.749
Idea Pin	Close-ups Repin	0.771	0.772	0.776
		0.842	0.846	0.849
Video	Close-ups Repin	0.469	0.468	0.470
		0.593	0.597	0.603

Table 6: Impact of the number of selected neighbors.

MultiBiSage model can significantly outperform the currently deployed PinSage production system on multiple user engagement metrics. A key novelty is how we aggregate the pin features along with the local graph structure of nodes from multiple bipartite graphs within our proposed MultiBiSage model. We also rely on several key optimizations to ensure the model can accommodate Pinterest scale data.

REFERENCES

- [1] Josh Beal et al. 2022. Billion-Scale Pretraining with Vision Transformers for Multi-Task Visual Representations. In *WACV*.
- [2] Yukuo Cen, Xu Zou, Jianwei Zhang, and et al. 2019. Representation Learning for Attributed Multiplex Heterogeneous Network. In *KDD*.
- [3] Shiyu Chang et al. 2015. Heterogeneous network embedding via deep architectures. In *KDD*. 119–128.
- [4] Ting Chen and Yizhou Sun. 2017. Task-guided and path-augmented heterogeneous network embedding for author identification. In *WSDM*. 295–304.
- [5] Graham Cormode and Shan Muthukrishnan. 2005. An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms*.
- [6] Yuxiao Dong et al. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *KDD*. 135–144.
- [7] Chantat Eksombatchai, Pranav Jindal, Jerry Zitao Liu, Yuchen Liu, Rahul Sharma, Charles Sugnet, Mark Ulrich, and Jure Leskovec. 2018. Pixie: A system for recommending 3+ billion items to 200+ million users in real-time. In *TheWeb*.
- [8] Tao-yang Fu, Wang-Chien Lee, and Zhen Lei. 2017. Hin2vec: Explore meta-paths in heterogeneous information networks for representation learning. In *CIKM*.
- [9] Swapnil Gandhi and Anand Padmanabha Iyer. 2021. P3: Distributed deep graph learning at scale. In *USENIX*. 551–568.
- [10] Priya Goyal et al. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677* (2017).
- [11] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *KDD*.
- [12] Will Hamilton, Zhitaoying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *NeurIPS*.
- [13] Yunzhi Hao et al. 2021. Walking with Attention: Self-guided Walking for Heterogeneous Graph Embedding. *TKDE* (2021).
- [14] Xiangnan He et al. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *SIGIR*. 639–648.
- [15] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2020. Open Graph Benchmark: Datasets for Machine Learning on Graphs. *arXiv preprint arXiv:2005.00687* (2020).
- [16] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020. Heterogeneous graph transformer. In *TheWeb*. 2704–2710.
- [17] Rana Hussein, Dingqi Yang, and Philippe Cudré-Mauroux. 2018. Are meta-paths necessary? Revisiting heterogeneous graph embeddings. In *CIKM*. 437–446.
- [18] Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine translation. *ACL* (2015).
- [19] Zhuoren Jiang et al. 2020. Task-oriented genetic activation for large-scale complex heterogeneous graph embedding. In *TheWeb*. 1581–1591.
- [20] George Karypis and Vipin Kumar. 1998. A fast and high quality multilevel scheme for partitioning irregular graphs. *SDM* 20, 1 (1998), 359–392.
- [21] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. In *arXiv preprint arXiv:1609.02907*.
- [22] Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* (2016).
- [23] Pinterest. 2022. Your audience is here. <https://business.pinterest.com/en/audience/>.
- [24] Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Kuansan Wang, and Jie Tang. 2018. Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec. In *WSDM*.
- [25] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European semantic web conference*. Springer, 593–607.
- [26] Yizhou Sun, Brandon Norick, Jiawei Han, Xifeng Yan, Philip S Yu, and Xiao Yu. 2013. Pathselclus: Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. *TKDD* 7, 3 (2013), 1–23.
- [27] Jian Tang et al. 2015. Line: Large-scale information network embedding. In *TheWeb*.
- [28] Jian Tang, Meng Qu, and Qiaozhu Mei. 2015. Pte: Predictive text embedding through large-scale heterogeneous text networks. In *KDD*. 1165–1174.
- [29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.
- [30] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. *ICLR*.
- [31] Xiao Wang, Deyu Bo, Chuan Shi, Shaohua Fan, Yanfang Ye, and Philip S Yu. 2020. A survey on heterogeneous graph embedding: methods, techniques, applications and sources. *arXiv preprint arXiv:2011.14867* (2020).
- [32] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural graph collaborative filtering. In *SIGIR*. 165–174.
- [33] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. 2019. Heterogeneous graph attention network. In *TheWeb*. 2022–2032.
- [34] Carl Yang et al. 2020. Heterogeneous network representation learning: A unified framework with survey and benchmark. *TKDE* (2020).
- [35] Ji Yang et al. 2020. Mixed Negative Sampling for Learning Two-tower Neural Networks in Recommendations. In *TheWeb*. 441–447.
- [36] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec. 2018. Graph convolutional neural networks for web-scale recommender systems. In *KDD*.
- [37] Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla. 2019. Heterogeneous graph neural network. In *KDD*. 793–803.

Surface/Entity Type	Engagement Type	MultiBiSage+ UserProd+PB	MultiBiSage+ UserAd+PB	MultiBiSage+ QueryPin+PB	MultiBiSage+ CCI+PB	MultiBiSage+ IUF+PB
Organic Surface	Add-to-cart Checkout	0.897 (+0.45%) 0.898 (+0.56%)	0.899 (+0.67%) 0.896 (+0.34%)	0.899 (+0.67%) 0.897 (+0.45%)	0.897 (+0.45%) 0.897 (+0.45%)	0.897 (+0.45%) 0.897 (+0.45%)
Ads	Good-click-through	0.660 (+1.38%)	0.679 (+4.30%)	0.677 (+3.99%)	0.662 (+1.69%)	0.658 (+1.08%)
Related Products	Long-click	0.737 (+0.55%)	0.739 (+0.82%)	0.737 (+0.55%)	0.738 (+0.68%)	0.735 (+0.27%)
Idea Pin	Close-ups Repin	0.742 (+2.49%) 0.814 (+2.13%)	0.751 (+3.73%) 0.822 (+3.14%)	0.746 (+3.04%) 0.815 (+2.26%)	0.747 (+3.18%) 0.824 (+3.39%)	0.756 (+4.42%) 0.833 (+4.52%)
Video	Close-ups Repin	0.459 (+2.91%) 0.586 (+2.63%)	0.463 (+3.81%) 0.591 (+3.50%)	0.461 (+3.36%) 0.588 (+2.98%)	0.465 (+4.26%) 0.587 (+2.80%)	0.461 (+3.36%) 0.589 (+3.15%)

Table 7: Ablation Study on Bipartite Graphs. PB, UserProd, UserAd, QueryPin, CCI, and IUF refers to Pin-Board-Contained, User-Product-LC, User-Ad-LC, SearchQuery-Pin-LC, Creator-IdeaPin-Created, IdeaPin-User-Follow bipartite graphs, respectively.

Surface/Entity Type	Engagement Type	MultiBiSage+ UserProd	MultiBiSage+ UserAd	MultiBiSage+ QueryPin	MultiBiSage+ CCI	MultiBiSage+ IUF	MultiBiSage+ UserProd+ UserAd+ QueryPin + CCI+IUF
Organic Surface	Add-to-cart Checkout	0.810 (-9.29%) 0.811 (-9.18%)	0.822 (-7.95%) 0.819 (-8.29%)	0.845 (-5.38%) 0.844 (-5.49%)	0.807 (-9.63%) 0.807 (-9.63%)	0.811 (-9.18%) 0.814 (-8.85%)	0.874 (-2.13%) 0.874 (-2.13%)
Ads	Good-click-through	0.606 (-6.91%)	0.630 (-3.23%)	0.627 (-3.69%)	0.605 (-7.07%)	0.609 (-6.45%)	0.670 (+2.92%)
Related Products	Long-click	0.666 (-9.14%)	0.665 (-9.28%)	0.685 (-6.55%)	0.660 (-9.96%)	0.663 (-9.55%)	0.711 (-3.00%)
Idea Pin	Close-ups Repin	0.637 (-12.02%) 0.722 (-9.41%)	0.637 (-12.02%) 0.723 (-9.28%)	0.648 (-10.50%) 0.733 (-8.03%)	0.654 (-9.67%) 0.738 (-7.40%)	0.690 (-4.70%) 0.780 (-2.13%)	0.748 (+3.31%) 0.828 (+3.89%)
Video	Close-ups Repin	0.397 (-10.99%) 0.511 (-10.51%)	0.407 (-8.74%) 0.510 (-10.68%)	0.404 (-9.42%) 0.521 (-8.76%)	0.404 (-9.42%) 0.511 (-10.51%)	0.406 (-8.97%) 0.516 (-9.63%)	0.450 (+0.90%) 0.563 (-1.40%)

Table 8: Ablation Study on Bipartite Graphs. UserProd, UserAd, QueryPin, CCI and IUF refers to User-Product-LC, User-Ad-LC, SearchQuery-Pin-LC, Creator-IdeaPin-Created, and IdeaPin-User-Follow bipartite graphs, respectively.














Query-Pin	Model	Engaged-Pin	Rank-1	Rank-2	Rank-3	Rank-4	Rank-5
	PinSage	 Rank: 364					
	MultiBiSage	 Rank: 5					

Table 9: Case study

[38] Chuxu Zhang, Ananthram Swami, and Nitesh V Chawla. 2019. Shne: Representation learning for semantic-associated heterogeneous networks. In *WSDM*.

[39] Fanjin Zhang, Xiao Liu, Jie Tang, Yuxiao Dong, Peiran Yao, Jie Zhang, Xiaotao Gu, Yan Wang, Bin Shao, Rui Li, et al. 2019. Oag: Toward linking large-scale

- heterogeneous entity graphs. In *KDD*.
- [40] Da Zheng et al. 2020. Distsgl: distributed graph neural network training for billion-scale graphs. In *2020 IEEE/ACM 10th Workshop on Irregular Applications: Architectures and Algorithms (IA3)*. IEEE, 36–44.
 - [41] Rong Zhu, Kun Zhao, Hongxia Yang, Wei Lin, Chang Zhou, Baole Ai, Yong Li, and Jingren Zhou. 2019. Aligraph: a comprehensive graph neural network platform. *arXiv preprint arXiv:1902.08730* (2019).
 - [42] Zhihua Zhu et al. 2020. Hgcn: A heterogeneous graph convolutional network-based deep learning model toward collective classification. In *KDD*.
 - [43] Jinfeng Zhuang and Yu Liu. 2019. PinText: A multitask text embedding system in pinterest. In *KDD*. 2653–2661.

Algorithm 2 Random-walk pseudocode of Pixie

Input: Bipartite Graph G , number of steps per node nw , random-walk reset probability α

Output: top-k highly visited nodes

```

1: Initialize Visits={}
2: for each pin  $p$  in  $G$  do
3:   current_node=p
4:   for each step  $i$  in  $nw$  do
5:     while true do
6:       Sample a neighbor node from the current_node's neighbors
       by performing uniform random sampling.
7:       Visits[p][current_node]+=1 if current_node of type pin.
8:       Flip a random coin with heads probability= $\alpha$ . Break if heads.
9:     end while
10:  end for
11: end for
12: Return top-k neighbors of each pin from Visits.
```

A PIXIE

The algorithm 2 outlines the random-walk code of Pixie. We set the number of steps per node $nw = 10,000$ and random-walk reset probability $\alpha = 0.9$.

B EXPERIMENTAL SETUP DETAILS

The node features details of Pinterest dataset is summarized below

- Visual: A 1024-dimensional unified visual embedding computed through a large-scale Transformer-based pretraining[1].
- Textual: A 64-dimensional embedding learned by Pintext [43] system.

Training details: For the model parameters, we use three-layer FeedForward Networks for visual features: $1024 \rightarrow 2048 \rightarrow 512$. The visual features FFN have dropout set to 0.25 and has ReLU activations. The textual features FFN consist of 1 linear layer: $64 \rightarrow 512$. We set the batch size to 8032 and the number of epochs to 100k. The output of the MultiBiSage model is L2-normalized 256-dimensional pin embeddings. The graph construction and feature extraction jobs execute on Spark cluster with 900 executors each with 40g executor and 40g driver memory. Pixie random-walks are performed on a single AWS x1.32xlarge (2 TB RAM, 128 cpus) instance. Model training is done in a distributed manner with two p3dn.24xlarge instances. Each p3dn.24xlarge instance has 96 custom vCPUs, 8 NVIDIA V100 Tensor Core GPUs with 32 GB of memory each.