

Dynamically Expandable Graph Convolution for Streaming Recommendation

Bowei He*

City University of Hong Kong
Hong Kong SAR
boweihe2-c@my.cityu.edu.hk

Xu He

Huawei Noah's Ark Lab
Shenzhen, China
hexu27@huawei.com

Yingxue Zhang

Huawei Noah's Ark Lab Montreal
Montreal, Canada
yingxue.zhang@huawei.comRuiming Tang
Huawei Noah's Ark Lab
Shenzhen, China
tangruiming@huawei.com

Chen Ma†

City University of Hong Kong
Hong Kong SAR
chenma@cityu.edu.hk

ABSTRACT

Personalized recommender systems have been widely studied and deployed to reduce information overload and satisfy users' diverse needs. However, conventional recommendation models solely conduct a one-time training-test fashion and can hardly adapt to evolving demands, considering user preference shifts and ever-increasing users and items in the real world. To tackle such challenges, the streaming recommendation is proposed and has attracted great attention recently. Among these, continual graph learning is widely regarded as a promising approach for the streaming recommendation by academia and industry. However, existing methods either rely on the historical data replay which is often not practical under increasingly strict data regulations, or can seldom solve the *over-stability* issue. To overcome these difficulties, we propose a novel **Dynamically Expandable Graph Convolution (DEGC)** algorithm from a *model isolation* perspective for the streaming recommendation which is orthogonal to previous methods. Based on the motivation of disentangling outdated short-term preferences from useful long-term preferences, we design a sequence of operations including graph convolution pruning, refining, and expanding to only preserve beneficial long-term preference-related parameters and extract fresh short-term preferences. Moreover, we model the temporal user preference, which is utilized as user embedding initialization, for better capturing the individual-level preference shifts. Extensive experiments on the three most representative GCN-based recommendation models and four industrial datasets demonstrate the effectiveness and robustness of our method.

CCS CONCEPTS

- Information systems → Recommender systems; • Computing methodologies → Online learning settings.

*Work done as an intern in Huawei Noah's Ark Lab, Hong Kong.

†Corresponding author

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

WWW '23, April 30–May 4, 2023, Austin, TX, USA

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9416-1/23/04...\$15.00

<https://doi.org/10.1145/3543507.3583237>

KEYWORDS

Continual graph learning, Graph neural network, Streaming recommendation

ACM Reference Format:

Bowei He, Xu He, Yingxue Zhang, Ruiming Tang, and Chen Ma. 2023. Dynamically Expandable Graph Convolution for Streaming Recommendation. In *Proceedings of the ACM Web Conference 2023 (WWW '23), April 30–May 4, 2023, Austin, TX, USA*. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3543507.3583237>

1 INTRODUCTION

The recommender system (RS), aiming to provide the personalized contents to different users precisely [11–13, 31, 32], has been deployed in many online Internet applications. However, traditional recommender systems trained on the offline static datasets face three challenges in the real online platforms: user preference shift, ever-increasing users and items, and intermittent user activities. In fact, these can cause the unacceptable performance degradation as shown in Figure 1, and lead to the necessity of dynamical model updating. Thus, how to update the model dynamically to tackle the above challenges has attracted great attention in real-world RS [5, 46]. This ensures the indispensable need of *streaming recommendation*, referring to updating and applying recommendation models dynamically over the data stream.

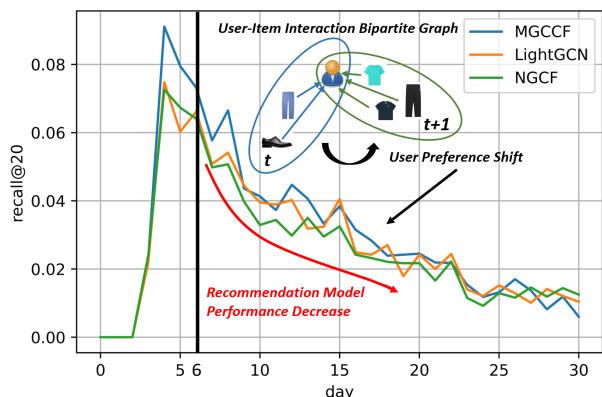


Figure 1: LightGCN, NGCF, and MGCF only trained in the first week and then tested in the following three weeks on Taobao2014 dataset.

Due to the strong capability of modeling user-item, user-user, and item-item relationships, graph convolution neural network (GCN) has been widely-used as recommendation models. Some recent works [1, 51, 52, 56, 61] develop continual learning methods for GCN-based recommendation methods to achieve the streaming recommendation, also known as continual graph learning for streaming recommendation.

To enable continual GCN-based recommendation, most works focus on two realizations: *experience replay* [1, 52, 68] and *knowledge distillation/weight regularization* [29, 51, 56, 61]. Although these methods have achieved acceptable results, there are still drawbacks which hinder them from being applied in the real-world systems. First, *experience replay* needs complete and accurate historical data when training the model on the newly coming data. Nevertheless, in the real-world system, the missing data [14, 55] and malicious data attack [67] issues are pretty common. What's worse, more and more strict data governance policies often make the user historical behavior inaccessible. Training with historical data replay also brings the huge memory cost and increasing time consumption. Second, knowledge distillation/model regularization can hardly capture the varying short-term preferences, especially for those users whose preferences change quickly and dramatically. Such issue is also known as *over-stability* [35] in continual learning. Third, previous continual graph learning methods for streaming recommendation do not model the user-level temporal preference changes explicitly and only relies on the graph convolution itself to capture the shift, which is greatly limited by the aforementioned two challenges: ever-increasing users and intermittent user activities. This is also far from the fine-grained user modeling for personalized recommendation. Besides, few existing *model isolation*-based methods [6, 66] (another main approach for continual learning) are designed specially for the typical continual learning paradigm with task boundaries and can hardly handle the growing dataset (data stream) on a single task. Such *model isolation* methods need to search the whole GCN architecture in each task which increases the method complexity greatly. In streaming recommendation setting, *model isolation*-based continual graph learning is less investigated though some apparent advantages like no longer needing to replay historical data and the potential to overcome the *over-stability* issue.

To tackle the above challenges, we propose a model-isolation continual graph learning method, namely **Dynamically Expandable Graph Convolution** (DEGC), to better model the user preference shift without the need of historical data replay. First, we design a graph convolution network-based sparsification training method to disentangle short-time preference-related parameters from long-time preference-related parameters. Then we remove outdated short-time preference-related filters and preserve long-term preference-related filters which are further refined with newly-collected data. Next, the graph convolution network is expanded by additional filters to extract the current short-term preference. The added filters will also be partly pruned to eliminate the redundant ones and prevent the network width explosion catastrophe. Moreover, inspired by the Kalman filter [57], a temporal attention model is utilized to explicitly encode the temporal user preference, which works as the user embedding initialization for training on new data.

In summary, the main contributions of this paper are:

- We propose a *model isolation*-based continual graph learning method, DEGC, for streaming recommendation. We design a sequence of graph convolution operations including pruning, refining, and expanding to overcome the *over-stability* challenge.
- To address the challenges of ever-increasing users and intermittent user activities, we model the temporal user preference as the user embedding initialization to help learn users' preferences on newly coming data.
- Experiments on three representative GCN-based recommendation models and four real-world datasets demonstrate the effectiveness and robustness of our method.

2 RELATED WORK

2.1 Streaming Recommendation

Due to the real-world dynamics like user preference continuous shift and ever-increasing users and items, conventional recommender systems trained on the static fixed datasets usually suffer from: predicting previous interactions and preferences, disregarding trends and shifting preferences, and ignoring real industrial constraints like few time and limited resources. To tackle these challenges, streaming recommendation is proposed in which data and recommendation model are both updated dynamically along the timeline [8, 10, 15, 16, 47, 48, 53]. Early works recommend items to users based on the popularity, recency, and trend analysis [7, 30, 49] but pay few attention to the collaborative signal distilling. To extracting such information, some other works [8, 16, 18, 44] introduce the classical recommendation algorithms like collaborative filtering and machine factorization into the streaming setting. In addition, there are also some recent works from the perspectives of online clustering of bandits and collaborative filtering bandits [2, 19, 20, 27, 28] to perform streaming recommendation. Thanks to the great success of graph neural network on complex relationship modeling, how to apply GCN-based recommendation models to the streaming recommendation is attracting more and more attention recently [1, 51, 52, 56, 61]. Besides, streaming recommendation algorithms have been successfully deployed to industrial online service platforms like Google, Huawei, and Tencent [5, 15, 46]. However, for a long time, there lacks a standardized definition to streaming recommendation, especially in the deep model-based recommendation setting. In this paper, we draw intuitions from previous research and most recent progress, and then summarize a definition of the streaming recommendation.

2.2 Continual Learning

Continual learning was originally paid great attention in computer vision and nature language process areas in which different tasks come in sequence. Various methods have been proposed to prevent catastrophic forgetting and effectively transfer knowledge. The mainstream continual learning algorithms can be classified into three categories: *experience replay* [9, 25, 34, 38, 41], *knowledge distillation/model regularization* [17, 23, 24, 40], and *model isolation* [21, 35, 39, 45, 59, 60, 65]. Continual learning is often regarded as a trade-off between knowledge retention (stability) and knowledge expansion (plasticity) [35], and *model isolation*-based methods provide an more explicit control over such trade-off. Considering that graph-based models have been widely studied

to model the complex data relationships, continual graph learning [6, 29, 33, 36, 37, 51, 52, 68] has also attracted more and more attentions recently. When it comes to the continual graph in recommendation setting [1, 51, 56, 61], we focus more on the data coming continuously in chronological order rather than the data with task boundaries. Different from the conventional continual learning, the continual graph learning for streaming recommendation which is studied in this work pays more attention to the effective knowledge transfer across the time segments rather than preventing catastrophic forgetting. This is because performance degradation on historical data makes no sense to a real recommender system.

3 PRELIMINARIES

In this section, we first formalize the continual graph learning for streaming recommendation. Then we briefly introduce three classical graph convolution based recommendation models used in this paper.

3.1 Definitions and Formulations

DEFINITION 1. *Streaming Recommendation.* Massive user-item interaction data $\tilde{\mathbf{D}}$ streams into industrial recommender system continuously. For convenience [4, 51, 52], the continuous data stream is split into consecutive data segments $D_1, \dots, D_t, \dots, D_T$ with the same time span. At each time segment t , the model needs to optimize the recommendation performance on D_t with the knowledge inherited from D_1, D_2, \dots, D_{t-1} . The recommendation performance is evaluated along the whole timeline.

DEFINITION 2. *Streaming Graph.* A streaming graph is represented as a sequence of graphs $\mathcal{G} = (G_1, G_2, \dots, G_t, \dots, G_T)$, where $G_t = G_{t-1} + \Delta G_t$. $G_t = (\mathbf{A}_t, \mathbf{X}_t)$ is an attributed graph at time t , where \mathbf{A}_t and \mathbf{X}_t are the adjacency matrix and node features of G_t , respectively. $\Delta G_t = (\Delta \mathbf{A}_t, \Delta \mathbf{X}_t)$ is the changes of graph structures and node attributes at t . The changes contain newly added nodes and newly built connections between different nodes.

DEFINITION 3. *Continual Graph Learning for Streaming Graph.* Given a streaming graph $\mathcal{G} = (G_1, G_2, \dots, G_t, \dots, G_T)$, the goal of **continual graph learning (CGL)** is to learn $\Delta G_t (D_t)$ sequentially while transferring historical knowledge to new graph segments effectively. Mathematically, the goal of **CGL** for streaming graph is to find the optimal GNN structure S_t and parameters \mathbf{W}_t at each segment t such that:

$$(S_t^*, \mathbf{W}_t^*) = \arg \min_{(S_t, \mathbf{W}_t)} \mathcal{L}_t(S_t, \mathbf{W}_t, \Delta G_t), \quad (1)$$

where $(S_t, \mathbf{W}_t) \in (\mathcal{S}, \mathcal{W})$. $\mathcal{L}_t(S_t, \mathbf{W}_t, \Delta G_t)$ is the loss function of current task defined on ΔG_t . The \mathcal{S} and \mathcal{W} are corresponding search spaces, respectively.

Since the user-item interaction data is actually a bipartite graph, the continual learning task for streaming recommendation is essentially the continual graph learning for streaming graph. For each segment t , the GNN structure S_t and parameters \mathbf{W}_t need to be adjusted and refined simultaneously to achieve a satisfying recommendation performance. We use the Bayesian Personalized Ranking (BPR) [42] loss as the loss function in this work, because it is effective and has broad applicability in top-K recommendation tasks. The major notations are summarized in Appendix A.

3.2 GCN-based Recommender Models

Many graph convolution-based recommender models [22, 50, 54, 64] have been developed recently to capture the collaborative signal, which is not encoded by the early matrix factorization and other deep learning based models. A general graph convolution process for such models can be summarized below: On the user-item bipartite graph, the layer- k embedding of user u is obtained via the following processing:

$$\begin{aligned} \mathbf{h}^{u,k} &= \sigma(\mathbf{W}^{u,k} \cdot [\mathbf{h}^{u,k-1}; \mathbf{h}^{\mathcal{N}(u),k-1}]), \mathbf{h}^{u,0} = \mathbf{e}^u, \\ \mathbf{h}^{\mathcal{N}(u),k-1} &= \text{AGGREGATOR}^u(\mathbf{h}^{i,k-1}, i \in \mathcal{N}(u)), \end{aligned} \quad (2)$$

where \mathbf{e}^u is the initial user embeddings, $\sigma(\cdot)$ is the activation function, $\mathbf{h}^{\mathcal{N}(u),k-1}$ is the learned neighborhood embedding, and $\mathbf{W}^{u,k}$ is the layer- k user transformation matrix shared among all users. The AGGREGATOR^u is the designed aggregation function in order to aggregate neighbor information for user nodes. Similarly, the layer- k embedding of item i is obtained via the following processing:

$$\begin{aligned} \mathbf{h}^{i,k} &= \sigma(\mathbf{W}^{i,k} \cdot [\mathbf{h}^{i,k-1}; \mathbf{h}^{\mathcal{N}(i),k-1}]), \mathbf{h}^{i,0} = \mathbf{e}^i, \\ \mathbf{h}^{\mathcal{N}(i),k-1} &= \text{AGGREGATOR}^i(\mathbf{h}^{u,k-1}, u \in \mathcal{N}(i)). \end{aligned} \quad (3)$$

In our setting, considering the timestamps of the above matrices, convolution parameters $\mathbf{W}_t = \{\mathbf{W}_t^{u,1}, \mathbf{W}_t^{u,2}, \dots, \mathbf{W}_t^{u,k}, \mathbf{W}_t^{i,1}, \dots, \mathbf{W}_t^{i,k}, \mathbf{W}_t^{u,K}, \mathbf{W}_t^{i,K}\}$, where $\mathbf{W}_t^{u,k}$ is the $\mathbf{W}^{u,k}$ on the time segment t . Each column in such matrices is a graph convolution filter. NGCF [54], LightGCN [22], and MGCCF [50] are three representative GCN-based recommendation models which will be utilized in our work. Some of their details are provided below.

NGCF [54]. One of the most widely used GCN-based recommendation model. NGCF exploits the user-item bipartite graph structure by propagating embeddings on it with graph convolution.

LightGCN [22]. An improved version of NGCF which still convolves on user-item graph. Compared with NGCF, LightGCN no more needs neighbor node feature transformation and nonlinear activation. Based on the original model setting, we add a dense matrix to each layer to align the dimensions of aggregated embeddings for better adapting to our approach.

MGCCF [50]. A graph convolution-based recommender framework which explicitly incorporates multiple graphs in the embedding learning process. Compared with the above two models, MGCCF adds a Multi-Graph Encoding(MGE) module to capture the inter-user and inter-item proximity information from user-user graph and item-item graph via homogeneous graph convolution, respectively.

4 METHODOLOGY

In this section, we introduce our **DEGC** method towards continual graph learning for streaming recommendation. We first model the temporal user preference to capture the preference shifts between segments, which will be utilized for the user embedding initialization of training. Then, we successively use two operations, *historical graph convolution pruning and refining* as well as *graph convolution expanding and pruning* (shown in Figure 2), in such a *model isolation* way to obtain the best structure and optimal parameters of the graph convolution.

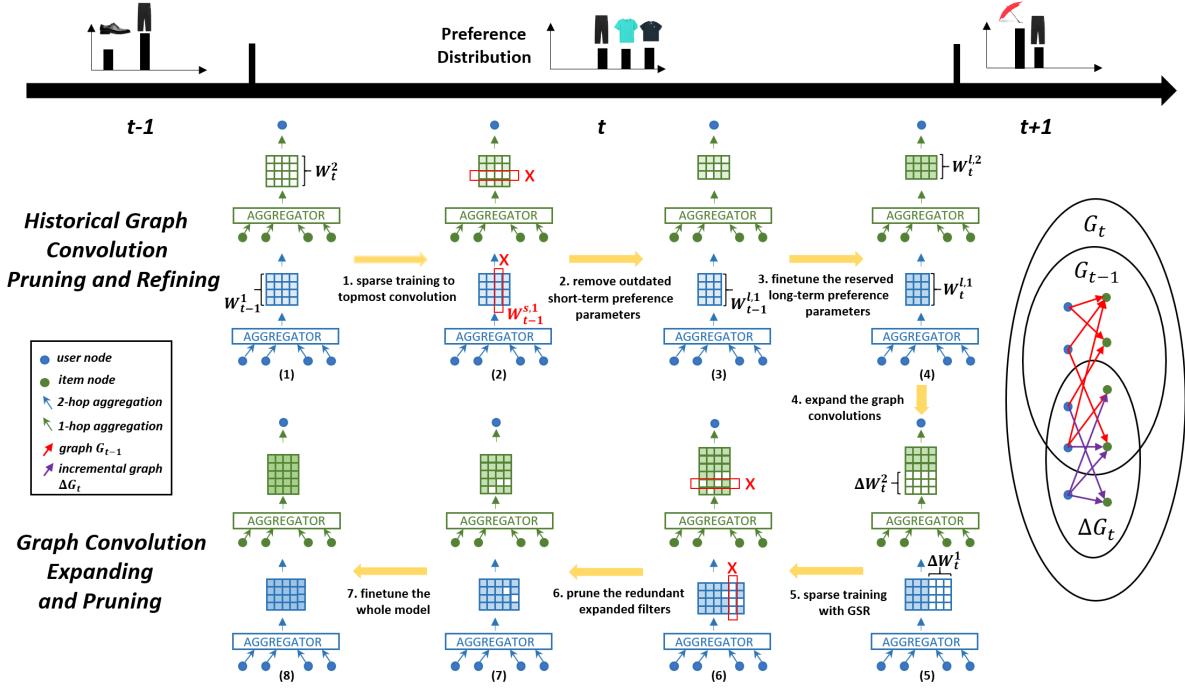


Figure 2: Overview to Dynamically Expandable Graph Convolution. Take 2-layer graph convolution as an example. Operations 1,2,3 correspond to the methods introduced in Sec. 4.2. And operations 4,5,6,7 illustrate the methods mentioned in Sec. 4.3.

4.1 Temporal User Preference Modeling as Initialization

Conventional continual graph learning algorithms directly inherit the user embeddings trained in the previous segments as the initialization without considering the user preference shift between the time segments. This type of methods totally rely on the graph convolution network to capture the user preference shift which is far from the fine-grained user modeling. To explicitly model the user-level preference shift and provide a better embedding initialization for graph convolution model training, we propose a temporal attention (TA) module to model the user temporal preference shift. This is motivated by the Kalman filter [3, 26, 57], which is an effective way to model the temporal state change. In recommender system, the user embedding e^u is utilized to represent the user u 's preference. At segment t , the preference shift of user u can be estimated with the Hadamard product of a time-scaled attention vector and the previous user embedding:

$$\Delta e_t^u = (\mathbf{W}_{TA} \Delta t) \odot e_{t^-}^u, \Delta t = t - t^-, \quad (4)$$

where t^- is the last time segment that user u appears and \mathbf{W}_{TA} is a learnable linear matrix. The larger the time interval Δt , the larger the user preference shift. Note that the users who are intermittently active on the platform (e.g., u_1 and u_2 in Figure 3) can also be modeled. Summing the user preference shift vector Δe_t^u and the user's previous preference vector $e_{t^-}^u$ gives the user's current preference e_t^u :

$$e_t^u = e_{t^-}^u + \Delta e_t^u. \quad (5)$$

As illustrated in Figure 3, the preferences of existing users (have appeared in the historical segments) at segment t can be estimated

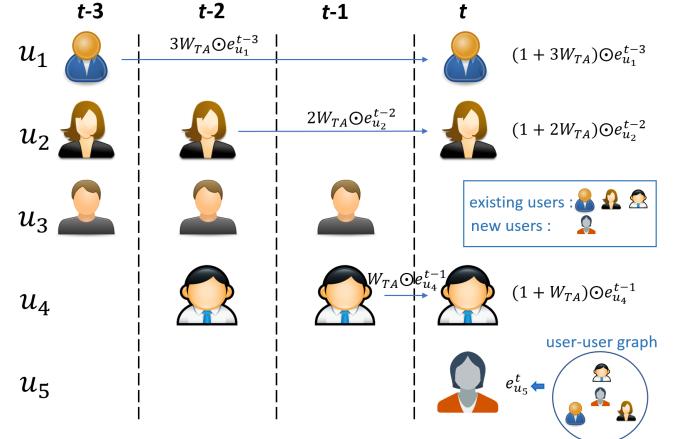


Figure 3: Temporal user preference modeling and new user initialization.

with the temporal attention module directly. As for new users (appears for the first time), their preference vectors are initialized with a one-hop neighbor aggregation from the user-user graph at segment t :

$$e_t^u = \frac{1}{|\mathcal{N}_t^u|} \sum_{j \in \mathcal{N}_t^u} e_t^j, \quad (6)$$

where j are existing users in user u 's one-hop neighbor on the user-user graph at segment t . The embeddings of existing users and new users will be utilized as the initial embeddings for graph convolution-based recommendation model training.

4.2 Historical Graph Convolution Pruning and Refining

In the recommender system, a user's preference is often regarded as the combination of her long-term preference (LTP) and short-term preference (STP) [31, 32, 43]. The long-term preference will not change drastically along the time and can take effect at most segments. This kind of preference is often determined by users' gender, occupation, education, and so on. By contrast, the short-term preference varies quickly and can only take effect in a certain time segment. STP is often influenced by the recommendation context information, like user emotion. And the recommendation model parameters store both users' long-term preferences and short-term preferences after learning on the interaction data. Previous continual graph learning methods for streaming recommendation inherit the parameters learned on the last segment indiscriminately and fine-tune them with the new data. However, these methods will preserve the outdated short-term preferences which only work for the last segment and hinder the model learning on the new segment, which corresponds to the *over-stability* issue mentioned above.

Based on the motivation of decoupling users' useful long-term preferences and outdated short-term preferences, we first design a sparsification training method to disentangle LTP-related graph convolution parameters \mathbf{W}_t^l and STP-related graph convolution parameters \mathbf{W}_t^s at segment t . As shown in subfigure (1) of Figure 2, we randomly initialize the topmost graph convolution layer \mathbf{W}_t^K and fix the rest graph convolution layers with parameters \mathbf{W}_{t-1} learned on ΔG_{t-1} and then only train the topmost graph convolution layer \mathbf{W}_t^K with the new incremental interaction graph ΔG_t :

$$\min_{\mathbf{W}_t^K} \mathcal{L}_t(\mathbf{W}_t^K; \mathbf{W}_{t-1}^{1:K-1}, \mathbf{e}, \Delta G_t) + \lambda_1 \|\mathbf{W}_t^K\|_1. \quad (7)$$

The \mathbf{e} denotes the embeddings of graph nodes. The added $L1$ regulation term is to encourage the sparse connection between layer K and layer $K - 1$ illustrated in subfigure (2) of Figure 2. Once obtaining sparse \mathbf{W}_t^K , we can identify the filters of \mathbf{W}_{t-1}^{K-1} that have no connection with layer K . Starting from this, we can find all the parameters in layer $1 : K - 1$ that have no connection with convolution layer K via breadth-first search. Because this sparsification effect is obtained by training on ΔG_t , such parameters are actually the users' outdated short-term preference-related parameters \mathbf{W}_{t-1}^s and cannot reflect users' current preferences at t . And the convolution parameters \mathbf{W}_{t-1}^l that have connection with \mathbf{W}_t^K represent another part of users' previous preferences that still takes effect at t , which is users' long-term preferences. Considering that the users' long-term preferences are also evolving along the time, we first remove the \mathbf{W}_{t-1}^s to avoid the negative knowledge transfer and then fine-tune the remaining LTP-related graph convolution parameters \mathbf{W}_t^l initialized with \mathbf{W}_{t-1}^l :

$$\min_{\mathbf{W}_t^l} \mathcal{L}_t(\mathbf{W}_t^l, \mathbf{e}, \Delta G_t) + \lambda_2 \|\mathbf{W}_t^l\|_2. \quad (8)$$

Here, we use $L2$ regularization to prevent the model overfitting. The operation of removing \mathbf{W}_{t-1}^s also corresponds to the best GNN structure search mentioned in Definition 3. Note that only fine-tuning the preserved long-term preference parameters can reduce

the computing overload which is of great significance in streaming recommendation setting.

Algorithm 1: DEGC

```

Input: A sequence of user-item interaction graphs  $\mathcal{G}$ .
Output: Graph convolution parameters  $\mathbf{W}_t$ , user
embeddings  $\mathbf{e}_t^u$ , item embeddings  $\mathbf{e}_t^i$ .
1 Process:
2 for each  $t = 1, 2, \dots, T$  do
3   | Initialize the user embeddings with Eqn. 5 and 6;
4   | Train the topmost graph convolution layer  $\mathbf{W}_t^K$  with
   | Eqn. 7;
5   | Obtain  $\mathbf{W}_{t-1}^l$  and  $\mathbf{W}_{t-1}^s$  with breadth-first search;
6   | Refine the  $\mathbf{W}_t^l$  initialized by  $\mathbf{W}_{t-1}^l$  with Eqn. 8;
7   | Expand the graph convolution layers and train the
   | expanded filters  $\Delta \mathbf{W}_t$  with Eqn. 9;
8   | Prune the expanded filters and finetune the whole
   | model with Eqn. 10;
9   | Update the  $\mathbf{W}_{TA}$  in Eqn. 4;
10 end

```

4.3 Graph Convolution Expanding and Pruning

Only the refined long-term preferences are not enough to reflect the users' comprehensive preferences. To extract users' current short-term preferences at segment t , we expand the graph convolution layers and train the expanded part from scratch independently with new data after obtaining fine-tuned \mathbf{W}_t^l via operations 1, 2, 3 in Figure 2. In detail, we add N filters to each graph convolution layer. Then, we initialize the expanded graph convolution parameters $\Delta \mathbf{W}_t$ randomly and train them with ΔG_t while fixing the \mathbf{W}_t^l :

$$\min_{\Delta \mathbf{W}_t} \mathcal{L}_t(\mathbf{W}_t^l; \Delta \mathbf{W}_t, \mathbf{e}, \Delta G_t) + \lambda_1 \|\Delta \mathbf{W}_t\|_1 + \lambda_g \sum_g \|\Delta \mathbf{W}_t^g\|_2. \quad (9)$$

Here, we use both $L1$ regularization and group sparse regularization (GSR) [58] to sparsify the expanded convolution parameters $\Delta \mathbf{W}_t$. g is the group consisting of the parameters of each newly added filter. The purpose of sparsification terms here is to prevent the convolution layer width explosion catastrophe if adding N filters to each layer constantly at each segment.

After obtaining sparsified $\Delta \mathbf{W}_t$ (operation 5 in Figure 2), we prune $\Delta \mathbf{W}_t^k$ at each layer k . Specifically, for each $\Delta \mathbf{W}_t^k$, we first search the filters whose weights are all zeros and remove such filters. Meanwhile, the corresponding parameters in $\Delta \mathbf{W}_t^{k+1}$ of layer $k + 1$ are also pruned. In such way, we not only extract the current short-term preferences at segment t but also eliminate the redundant expansion parameters. It needs to be mentioned that our expanding and pruning operations echo the GNN structure optimization in Definition 3 again. Finally, the fixed \mathbf{W}_t^l , the pruned $\Delta \mathbf{W}_t$, and the embeddings \mathbf{e} will be finetuned:

$$\min_{\mathbf{W}_t^l, \Delta \mathbf{W}_t, \mathbf{e}} \mathcal{L}_t(\mathbf{W}_t^l; \Delta \mathbf{W}_t, \mathbf{e}, \Delta G_t) + \lambda_1 (\|\mathbf{W}_t^l\|_1 + \|\Delta \mathbf{W}_t\|_1) + \lambda_2 \|\mathbf{e}\|_2. \quad (10)$$

The embeddings e include both user embeddings e^u and item embeddings e^i . Here, the $L1$ regularization is to sparsify the whole graph convolution structure to facilitate the historical convolution pruning on the next segment $t + 1$. The whole workflow of our method is illustrated in Algorithm 1. Because our method is actually orthogonal to previous methods like *experience replay* and *knowledge distillation*, we will combine our method with previous methods to demonstrate its advantages.

5 EXPERIMENTS

In this section, we conduct experiments on four real-world timestamped recommendation datasets to show the effectiveness of our method. We mainly focus on the following questions:

- **RQ1:** Whether our method can get better recommendation effects than the state-of-art methods on the most common recommendation metrics like $Recall@20$ and $NDCG@20$?
- **RQ2:** Whether our method is robust to different datasets and GCN-based recommendation models?
- **RQ3:** Whether pruning historical convolution parameters to forget the outdated short-term preferences is necessary for improving streaming recommendation?
- **RQ4:** Whether temporal user preference modeling as initialization help learn users' preferences in the GCN update phase?

5.1 Experiment Settings

5.1.1 Datasets.

- **Taobao2014¹:** This dataset contains real users-commodities behavior data collected from Alibaba's M-Commerce platforms, spanning 30 days. The rich background informations, like users' location information and behavior timestamp are also included. We filter out users and items with less than 10 interactions.
- **Taobao2015²:** This dataset contains user behavior data between July 1st, 2015 and Nov. 30th, 2015 accumulated on Taobao.com and the app Alipay. The online actions and timestamps are both recorded. In this work, we only use the first month data for analysis. Besides, We filter out users and items with less than 20 interactions.
- **Netflix³:** This movie rating dataset contains over 100 million ratings from 480 thousand randomly-chosen Netflix customers over 17 thousand movie titles. The data were collected between October, 1998 and December, 2005 and reflect the distribution of all ratings received during this period. A similar filtering operation is executed and the thresholds are both set as 30. We use the first 24 months data for analysis.
- **Foursquare [62, 63]:** This dataset includes long-term (about 22 months from Apr. 2012 to Jan. 2014) global-scale check-in data collected from Foursquare, a local search-and-discovery mobile APP. The check-in dataset contains 22,809,624 checkins by 114,324 users on 3,820,891 venues. For this dataset, we set the filtering thresholds as 20.

The data statistics after filtering of the above datasets are detailed in Appendix B. Average entity overlapping rate between adjacent segments (AER) is a metric to measure the stability of a data stream.

¹<https://tianchi.aliyun.com/dataset/dataDetail?dataId=46>

²<https://tianchi.aliyun.com/dataset/dataDetail?dataId=53>

³<https://academictorrents.com/details/9b13183dc4d60676b773c9e2cd6de5e5542cee9a>

The larger the AER, the more stable the data stream. The data on each segment is split into training, validation, and test sets using a ratio of 8:1:1. We repeat each experiment five times and report the average results to reduce the variance brought by the randomness.

5.1.2 Baselines.

- **Finetune:** Finetune first inherits the parameters from the previous segment and then fine-tune the model only with the data of the current segment.
- **Uniform Sampling (Uniform):** A kind of naive *experience replay* method which first sample the historical data uniformly and then combine the new data with it. The model is trained with the combined data from scratch at each segment.
- **Inverse Degree Sampling (Inverse) [1]:** A similar sampling-based *experience replay* method. However, the sampling probability of each interaction is proportional to its user's inverse degree on the interaction graph.
- **ContinualGNN [51]:** A continual graph learning method which combines the *experience replay* and *knowledge distillation* for existing pattern consolidation.
- **Topology-aware Weight Preserving (TWP) [29]:** A *knowledge distillation* method which explores the local structure of the interaction graph and stabilize the parameters playing pivotal roles in the topological aggregation.
- **GraphSAIL [61]:** A *knowledge distillation* method which preserves each node's local structure, global structure, and self-information, respectively at the new segment.
- **SGCT [56]:** A *knowledge distillation* method which uses contrastive distillation on each node's embedding where a single user-item graph is used to construct positive samples.
- **MGCT [56]:** A *knowledge distillation* method which uses contrastive distillation on each node's embedding where multiple graphs (user-item, user-user, item-item graphs) is used to construct positive samples.
- **LWC-KD [56]:** Based on MGCT, LWC-KD adds the intermediate layer distillation to inject layer-level supervision.

We compare our **DEGC** model with the above continual graph learning methods. Note that we will show the experiment results of **DEGC+Finetune** and **DEGC+LWC-KD** (the combined methods mentioned in Section 4.3) for a fair comparison. The implementation details are provided in Appendix D. The code is available at <https://github.com/BokwaiHo/DEGC>.

5.1.3 *Evaluation Metrics.* All the methods are evaluated in terms of $Recall@k$ and $NDCG@k$. For each user, our recommendation model will recommend an ordered list of items to her. $Recall@k$ (abbreviated as $R@k$) indicates the percentage of her rated items that appear in the top k items of the recommended list. The $NDCG@k$ (abbreviated as $N@k$) is the normalized discounted cumulative gain at a ranking position k to measure the ranking quality. Similar to previous related papers [1, 56, 61], we set the k as 20.

5.2 Results and Analysis

5.2.1 *Overall Performance (RQ1).* To answer **RQ1**, we evaluate our model performance from two perspectives: time-varying performance and average performance on the data stream. In Figure 4, we visualize the $Recall@20$ and $NDCG@20$ curves on the

Table 1: The average performance with MGCCF as our base model. * indicates the improvements over baselines are statistically significant (t -test, p -value ≤ 0.01).

Method	Taobao2014		Taobao2015		Netflix		Foursquare	
	Recall@20	NDCG@20	Recall@20	NDCG@20	Recall@20	NDCG@20	Recall@20	NDCG@20
Finetune	0.0412	0.0052	0.4256	0.0117	0.3359	0.0580	0.1154	0.0115
Uniform	0.0308	0.0038	0.4194	0.0113	0.3298	0.0533	0.1024	0.0101
Inverse	0.0323	0.0039	0.4217	0.0114	0.3321	0.0536	0.1063	0.0107
ContinualGNN	0.0311	0.0035	0.4203	0.0111	0.3089	0.0491	0.1056	0.0098
TWP	0.0398	0.0050	0.4320	0.0122	0.3428	0.0580	0.1048	0.0104
GraphSAIL	0.0395	0.0051	0.4371	0.0126	0.3470	0.0583	0.1086	0.0110
SGCT	0.0423	0.0054	0.4411	0.0129	0.3300	0.0572	0.1255	0.0137
MGCT	0.0421	0.0055	0.4446	0.0131	0.3252	0.0562	0.1192	0.0127
LWC-KD	<u>0.0440</u>	<u>0.0059</u>	<u>0.4512</u>	<u>0.0139</u>	0.2616	0.0482	<u>0.1280</u>	<u>0.0144</u>
DEGC+LWC-KD	0.1082* ($\uparrow 146\%$)	0.0142* ($\uparrow 141\%$)	0.4892* ($\uparrow 8.42\%$)	0.0167* * ($\uparrow 20.1\%$)	0.2949* ($\downarrow 15.0\%$)	0.0520* ($\downarrow 10.8\%$)	0.1425* ($\uparrow 11.3\%$)	0.0178* ($\uparrow 23.6\%$)
DEGC+Finetune	0.0825* ($\uparrow 87.5\%$)	0.0106* ($\uparrow 79.7\%$)	0.4563* ($\uparrow 1.13\%$)	0.0142* ($\uparrow 2.16\%$)	0.3583* ($\uparrow 3.26\%$)	0.0612* ($\uparrow 4.97\%$)	0.1324* ($\uparrow 3.44\%$)	0.0153* ($\uparrow 6.25\%$)

Table 2: The average performance with NGCF as our base model. * indicates the improvements over baselines are statistically significant (t -test, p -value ≤ 0.01).

Method	Taobao2014		Netflix	
	R@20	N@20	R@20	N@20
Finetune	0.0304	0.0040	0.3131	0.0541
Uniform	0.0340	0.0038	0.3263	0.0525
Inverse	0.0347	0.0039	0.3256	0.0518
ContinualGNN	0.0338	0.0036	0.3047	0.0479
TWP	0.0358	0.0047	0.3159	0.0531
GraphSAIL	0.0318	0.0042	0.3245	<u>0.0554</u>
SGCT	0.0350	0.0046	0.3044	0.0533
MGCT	0.0346	0.0045	0.2957	0.0511
LWC-KD	<u>0.0380</u>	<u>0.0050</u>	0.2496	0.0454
DEGC+LWC-KD	0.0961* ($\uparrow 153\%$)	0.0123* ($\uparrow 146\%$)	0.2713* ($\downarrow 16.9\%$)	0.0486* ($\downarrow 12.3\%$)
DEGC+Finetune	0.0816* ($\uparrow 115\%$)	0.0107* ($\uparrow 114\%$)	0.3454* ($\uparrow 5.85\%$)	0.0594* ($\uparrow 7.22\%$)

data streams of Taobao2014 dataset and Netflix dataset. Comparing the DEGC+Finetune and DEGC+LWC-KD with Finetune and LWC-KD, respectively, we can observe that our methods achieve significant time-varying performance gain over their corresponding base methods. Note that the zero performance in the first two days of Taobao2014 dataset is due to the limited amount of data and the model overfitting. The main reason that the observed sharp $NDCG@20$ decreases in the first five months of Netflix dataset is the rapidly increasing numbers of users and items. In Table 1, we show the average performance of different methods on four datasets while choosing MGCCF as the base GCN recommendation model. It can be observed that our DEGC+Finetune and DEGC+LWC-KD get better recommendation effects than all state-of-art methods on all four datasets, except the DEGC+LWC-KD on Netflix. We argue that this is because the poor performance of LWC-KD itself on Netflix. Comparing the DEGC+LWC-KD with LWC-KD independently, our

Table 3: The average performance with LightGCN as our base model. * indicates the improvements over baselines are statistically significant (t -test, p -value ≤ 0.01).

Method	Taobao2014		Netflix	
	R@20	N@20	R@20	N@20
Finetune	0.0339	0.0040	0.3179	0.0537
Uniform	0.0377	0.0041	0.3289	0.0533
Inverse	0.0386	0.0042	0.3275	0.0530
ContinualGNN	0.0382	0.0041	0.3035	0.0475
TWP	0.0338	0.0040	0.3204	0.0542
GraphSAIL	0.0342	0.0042	0.3282	<u>0.0544</u>
SGCT	0.0342	0.0043	0.3073	0.0519
MGCT	0.0357	0.0047	0.2983	0.0516
LWC-KD	0.0402	<u>0.0053</u>	0.2571	0.0461
DEGC+LWC-KD	0.0975* ($\uparrow 143\%$)	0.0125* ($\uparrow 136\%$)	0.2776* ($\downarrow 15.6\%$)	0.0491* ($\downarrow 9.74\%$)
DEGC+Finetune	0.0833* ($\uparrow 107\%$)	0.0109* ($\uparrow 106\%$)	0.3483* ($\uparrow 5.90\%$)	0.0596* ($\uparrow 9.56\%$)

method still improves the performance by 12.7% on $Recall@20$ and 7.9% on $NDCG@20$, which also demonstrate the effectiveness of our method. As for the poor performance of SGCT, MGCT, LWC-KD on Netflix, this is resulted by the *over-stability* issue. Such type of *knowledge distillation* methods can hardly accurately capture the user preferences' shifts when they change rapidly. Besides, it can be noticed that the performance of *experience replay* methods including Uniform, Inverse, and ContinualGNN are even worse than Finetune. Actually, in the streaming recommendation, such methods can replay previous data containing users' outdated short-term preferences, which negatively influences the model learning on new segments.

5.2.2 Method Robustness Analysis (RQ2). To answer RQ2, we conduct the experiments with NGCF and LightGCN as the base GCN models on both Taobao2014 and Netflix datasets. The

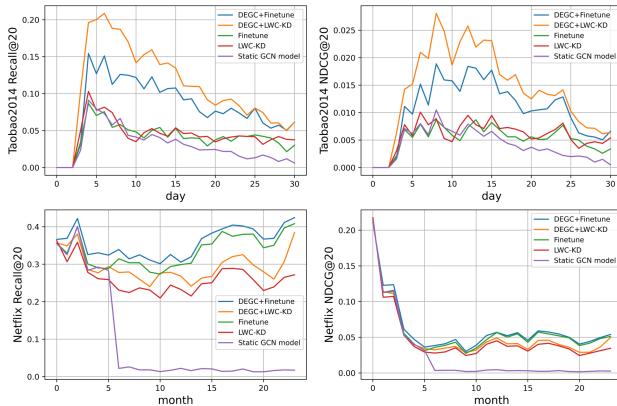


Figure 4: The time-varying model performance on the data stream of Taobao2014 dataset and Netflix dataset.

corresponding results are shown in Tables 2 and 3. For GCN models NGCF and LightGCN, the improvements of our methods on Taobao2014 are both significant. DEGC+Finetune and DEGC+LWC-KD both achieve the state-of-art recommendation performance. An interesting observation is that DEGC+Finetune in Table 3 even gets better performance than that in Table 1. This also shows the performance potential of DEGC on different kinds of base GCN models. As for the dataset Netflix, DEGC+Finetune improves the *Recall@20* by 10.3% and *NDCG@20* by 9.8% over Finetune when choosing NGCF as the GCN model. DEGC+LWC-KD improves the *Recall@20* by 8.7% and *NDCG@20* by 7.0% over LWC-KD, meanwhile. Besides, DEGC+Finetune achieves the best recommendation effect over all previous methods. Similar improvements can also be observed when taking LightGCN as the base GCN model. Such observations demonstrate the robustness of our methods to different datasets and GCN-based recommendation models.

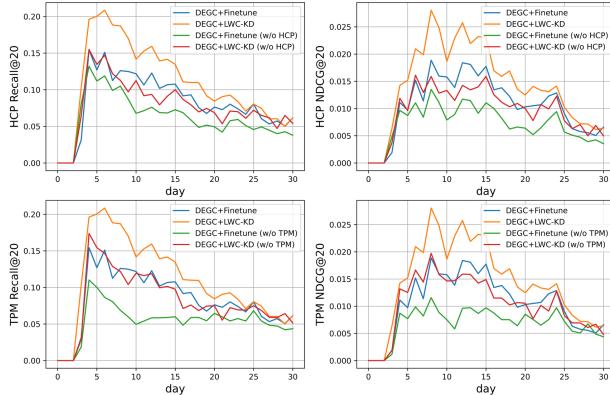


Figure 5: Ablation study to historical convolution pruning (HCP) and temporal preference modeling (TPM) on Taobao2014 dataset.

5.2.3 Ablation Study to Historical Convolution Pruning (RQ3). To answer RQ3, we conduct the experiments with DEGC+Finetune (w/o HCP) and DEGC+LWC-KD (w/o HCP) as the continual graph learning methods while taking the MGCCF as base GCN model on

Taobao2014 and Netflix datasets. The top two subfigures of Figure 5 illustrate the time-varying recommendation performance with two metrics: *Recall@20* and *NDCG@20*. Comparing the DEGC+Finetune with DEGC+Finetune (w/o HCP) and DEGC+LWC-KD with DEGC+LWC-KD (w/o HCP), respectively, we can find that historical convolution pruning is of great significance to our methods' effectiveness. These demonstrate that pruning historical convolution parameter to forget the outdated short-term preferences is necessary. It is also validated that conventional continual graph methods that inherit the parameters learned on the last segment indiscriminately and then finetune them with the new data hinders the model learning on the new segment. This can also be regarded as, at least, part of the reasons that lead to the 'over-stability' challenge in continual learning for streaming recommendation. We also quantitatively analyze the performance drop after removing the historical convolution pruning. We find that DEGC+Finetune decreases by 25.3% and DEGC+LWC-KD decreases by 28.4% on average. The more severe decrease effect of DEGC+LWC-KD is due to that LWC-KD preserves more historical knowledge which may contain users' outdated short-term preferences. We can also observe similar results regarding Netflix dataset in Figure 6 of Appendix C.

5.2.4 Ablation Study to Temporal Preference Modeling (RQ4).

To answer RQ4, we conduct the experiments with DEGC+Finetune (w/o TPM) and DEGC+LWC-KD (w/o TPM) on Taobao2014 and Netflix datasets while taking the MGCCF as the base GCN model. From the bottom two subfigures of Figure 5, we can observe that the *Recall@20* and *NDCG@20* metrics both decrease obviously after removing the temporal preference modeling. This proves that modeling temporal user preference as initialization does benefit the user preference learning in the GCN update phase. Actually, this also corresponds to the other two challenges except the continuous user preference shifts in streaming recommendation: ever-increasing users and intermittent user activities mentioned in Section 1. Traditional continual graph learning methods like Finetune and LWC-KD directly use the user embeddings learned in the last segment as the embeddings initialization in the current segment. So they can hardly provide an accurate embedding initialization to users whose active intervals on the online platform are longer than a time segment. Also, they cannot provide a warm embedding initialization for newly coming users. Our temporal preference modeling as user embedding initialization solves such two challenges to some extent and improves the *Recall@20* by 34.3% and 24.9% on average, over DEGC+Finetune (w/o TPM) and DEGC+LWC-KD (w/o TPM), respectively. The similar trends on Netflix dataset can also be observed in Figure 6 of Appendix C.

6 CONCLUSION

Streaming recommendation has attracted great attention due to the dynamics of the real world. In this paper, we first propose the temporal preference modeling as the user embedding initialization of each time segment. Then, we start from the *model isolation* perspective and propose the *historical graph convolution pruning and refining* and *graph convolution expanding and pruning* operations, in such ways to only preserve useful long-term preferences and further extract current short-term preferences. Extensive experiments

on four real-world datasets and three most representative GCN-based recommendation models also demonstrate the effectiveness and robustness of our method.

ACKNOWLEDGMENTS

This work was supported by the Start-up Grant (No. 9610564) and the Strategic Research Grant (No. 7005847) of City University of Hong Kong.

REFERENCES

- [1] Kian Ahrabian, Yishi Xu, Yingxue Zhang, Jiapeng Wu, Yuening Wang, and Mark Coates. 2021. Structure Aware Experience Replay for Incremental Learning in Graph-based Recommender Systems. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2832–2836.
- [2] Yikun Ban and Jingrui He. 2021. Local clustering in contextual multi-armed bandits. In *Proceedings of the Web Conference 2021*. 2335–2346.
- [3] Alex Beutel, Paul Covington, Sagar Jain, Can Xu, Jia Li, Vince Gatto, and Ed H Chi. 2018. Latent cross: Making use of context in recurrent recommender systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 46–54.
- [4] Lucas Caccia, Jing Xu, Myle Ott, Marc'Aurelio Ranzato, and Ludovic Denoyer. 2021. On anytime learning at macroscale. *arXiv preprint arXiv:2106.09563* (2021).
- [5] Guohao Cai, Jieming Zhu, Quanyu Dai, Zhenhua Dong, Xiuqiang He, Ruiming Tang, and Rui Zhang. 2022. ReLoop: A Self-Correction Continual Learning Loop for Recommender Systems. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*. ACM, 2692–2697.
- [6] Jie Cai, Xin Wang, Chaoyu Guan, Yateng Tang, Jin Xu, Bin Zhong, and Wenwu Zhu. 2022. Multimodal Continual Graph Learning with Neural Architecture Search. In *Proceedings of the ACM Web Conference 2022*. 1292–1300.
- [7] Badrish Chandramouli, Justin J Levandoski, Ahmed Eldawy, and Mohamed F Mokbel. 2011. Streamrec: a real-time recommender system. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*. 1243–1246.
- [8] Shiyu Chang, Yang Zhang, Jiliang Tang, Dawei Yin, Yi Chang, Mark A Hasegawa-Johnson, and Thomas S Huang. 2017. Streaming recommender systems. In *Proceedings of the 26th international conference on world wide web*. 381–389.
- [9] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and M Ranzato. 2019. Continual learning with tiny episodic memories. (2019).
- [10] Chen Chen, Hongzhi Yin, Junjie Yao, and Bin Cui. 2013. Terec: A temporal recommender system over tweet stream. *Proceedings of the VLDB Endowment* 6, 12 (2013), 1254–1257.
- [11] Yankai Chen, Huifeng Guo, Yingxue Zhang, Chen Ma, Ruiming Tang, Jingjie Li, and Irwin King. 2022. Learning binarized graph representations with multi-faceted quantization reinforcement for top-k recommendation. In *SIGKDD*.
- [12] Yankai Chen, Menglin Yang, Yingxue Zhang, Mengchen Zhao, Ziqiao Meng, Jianye Hao, and Irwin King. 2022. Modeling Scale-free Graphs with Hyperbolic Geometry for Knowledge-aware Recommendation. In *WSDM*. 94–102.
- [13] Yankai Chen, Yaming Yang, Yujing Wang, Jing Bai, Xiangchen Song, and Irwin King. 2022. Attentive Knowledge-aware Graph Convolutional Networks with Collaborative Guidance for Personalized Recommendation. In *ICDE*.
- [14] Ed H Chi. 2020. From Missing Data to Boltzmann Distributions and Time Dynamics: The Statistical Physics of Recommendation. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 1–2.
- [15] Abhinandan S Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. 2007. Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th international conference on World Wide Web*. 271–280.
- [16] Robin Devooght, Nicolas Kourtellis, and Amin Mantrach. 2015. Dynamic matrix factorization with priors on unknown values. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 189–198.
- [17] Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. 2019. Learning without memorizing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5138–5146.
- [18] Ernesto Diaz-Aviles, Lucas Drumond, Lars Schmidt-Thieme, and Wolfgang Nejdl. 2012. Real-time top-n recommendation in social streams. In *Proceedings of the sixth ACM conference on Recommender systems*. 59–66.
- [19] Claudio Gentile, Shuai Li, Purushottam Kar, Alexandros Karatzoglou, Giovanni Zappella, and Evans Etrue. 2017. On context-dependent clustering of bandits. In *International Conference on machine learning*. PMLR, 1253–1262.
- [20] Claudio Gentile, Shuai Li, and Giovanni Zappella. 2014. Online clustering of bandits. In *International Conference on Machine Learning*. PMLR, 757–765.
- [21] Siavash Golkar, Michael Kagan, and Kyunghyun Cho. 2019. Continual learning via neural pruning. *arXiv preprint arXiv:1903.04476* (2019).
- [22] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 639–648.
- [23] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* 2, 7 (2015).
- [24] Shaihu Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. 2019. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 831–839.
- [25] David Isele and Akansel Cosgun. 2018. Selective experience replay for lifelong learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
- [26] Srijan Kumar, Xikun Zhang, and Jure Leskovec. 2019. Predicting dynamic embedding trajectory in temporal interaction networks. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 1269–1278.
- [27] Shuai Li, Wei Chen, Shuai Li, and Kwong-Sak Leung. 2019. Improved Algorithm on Online Clustering of Bandits. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, Sarit Kraus (Ed.). ijcai.org, 2923–2929.
- [28] Shuai Li, Alexandros Karatzoglou, and Claudio Gentile. 2016. Collaborative filtering bandits. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 539–548.
- [29] Huihui Liu, Yiding Yang, and Xinchao Wang. 2021. Overcoming Catastrophic Forgetting in Graph Neural Networks. In *AAAI*. AAAI Press, 8653–8661.
- [30] Andreas Lommatzsch and Sahin Albayrak. 2015. Real-time recommendations for user-item streams. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*. 1039–1046.
- [31] Chen Ma, Peng Kang, and Xue Liu. 2019. Hierarchical gating networks for sequential recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 825–833.
- [32] Chen Ma, Liheng Ma, Yingxue Zhang, Ruiming Tang, Xue Liu, and Mark Coates. 2020. Probabilistic metric learning with adaptive margin for top-K Recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1036–1044.
- [33] Yao Ma, Ziyi Guo, Zhaocun Ren, Jiliang Tang, and Dawei Yin. 2020. Streaming graph neural networks. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 719–728.
- [34] Fei Mi, Xiaoyu Lin, and Boi Faltings. 2020. Ader: Adaptively distilled exemplar replay towards continual learning for session-based recommendation. In *Fourteenth ACM Conference on Recommender Systems*. 408–413.
- [35] Oleksiy Ostapenko, Pau Rodriguez, Massimo Caccia, and Laurent Charlin. 2021. Continual learning via local module composition. *Advances in Neural Information Processing Systems* 34 (2021), 30298–30312.
- [36] Aldo Pareja, Giacomo Domeniconi, Jie Chen, Tengfei Ma, Toyotaro Suzumura, Hiroki Kanezashi, Tim Kaler, Tao B. Schardl, and Charles E. Leiserson. 2020. EvolveGCN: Evolving Graph Convolutional Networks for Dynamic Graphs. In *AAAI*. AAAI Press, 5363–5370.
- [37] Massimo Perini, Giorgia Ramponi, Paris Carbone, and Vasiliki Kalavri. 2022. Learning on streaming graphs with experience replay. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*. 470–478.
- [38] Ameya Prabhu, Philip HS Torr, and Puneet K Dokania. 2020. Gdumb: A simple approach that questions our progress in continual learning. In *European conference on computer vision*. Springer, 524–540.
- [39] Qi Qin, Wenpeng Hu, Han Peng, Dongyan Zhao, and Bing Liu. 2021. BNS: Building Network Structures Dynamically for Continual Learning. *Advances in Neural Information Processing Systems* 34 (2021), 20608–20620.
- [40] Amal Rannen, Rahaf Aljundi, Matthew B Blaschko, and Tinne Tuytelaars. 2017. Encoder based lifelong learning. In *Proceedings of the IEEE International Conference on Computer Vision*. 1320–1328.
- [41] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2001–2010.
- [42] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *UAI*. AUAI Press, 452–461.
- [43] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*. 811–820.
- [44] Steffen Rendle and Lars Schmidt-Thieme. 2008. Online-updating regularized kernel matrix factorization models for large-scale recommender systems. In *Proceedings of the 2008 ACM conference on Recommender systems*. 251–258.
- [45] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. 2016. Progressive neural networks. *arXiv preprint arXiv:1606.04671* (2016).
- [46] Chijun Sima, Yao Fu, Man-Kit Sit, Liyi Guo, Xuri Gong, Feng Lin, Junyu Wu, Yongsheng Li, Haidong Rong, Pierre-Louis Aublin, et al. 2022. Ekko: A {Large-Scale} Deep Learning Recommender System with {Low-Latency} Model Update.

- In 16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22). 821–839.
- [47] Qingquan Song, Xiao Huang, Hancheng Ge, James Caverlee, and Xia Hu. 2017. Multi-aspect streaming tensor completion. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 435–443.
- [48] Yang Song, Ziming Zhuang, Huajing Li, Qiankun Zhao, Jia Li, Wang-Chien Lee, and C Lee Giles. 2008. Real-time automatic tag recommendation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. 515–522.
- [49] Karthik Subbian, Charu Aggarwal, and Kshiteesh Hegde. 2016. Recommendations for streaming data. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*. 2185–2190.
- [50] Jianing Sun, Yingxue Zhang, Chen Ma, Mark Coates, Hufeng Guo, Ruiming Tang, and Xiuqiang He. 2019. Multi-graph convolution collaborative filtering. In *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1306–1311.
- [51] Junshan Wang, Guojie Song, Yi Wu, and Liang Wang. 2020. Streaming graph neural networks via continual learning. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 1515–1524.
- [52] Junshan Wang, Wenhao Zhu, Guojie Song, and Liang Wang. 2022. Streaming Graph Neural Networks with Generative Replay. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Washington, DC, USA, August 14 - 18, 2022. 1878–1888.
- [53] Weiqing Wang, Hongzhi Yin, Zi Huang, Qinyong Wang, Xingzhong Du, and Quoc Viet Hung Nguyen. 2018. Streaming ranking based recommender systems. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 525–534.
- [54] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural graph collaborative filtering. In *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval*. 165–174.
- [55] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. 2019. Doubly robust joint learning for recommendation on data missing not at random. In *International Conference on Machine Learning*. PMLR, 6638–6647.
- [56] Yuening Wang, Yingxue Zhang, and Mark Coates. 2021. Graph Structure Aware Contrastive Knowledge Distillation for Incremental Learning in Recommender Systems. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 3518–3522.
- [57] Greg Welch, Gary Bishop, et al. 1995. An introduction to the Kalman filter. (1995).
- [58] Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. 2016. Learning structured sparsity in deep neural networks. *Advances in neural information processing systems* 29 (2016).
- [59] Lemeng Wu, Bo Liu, Peter Stone, and Qiang Liu. 2020. Firefly neural architecture descent: general approach for growing neural networks. *Advances in Neural Information Processing Systems* 33 (2020), 22373–22383.
- [60] Ju Xu and Zhanxing Zhu. 2018. Reinforced continual learning. *Advances in Neural Information Processing Systems* 31 (2018).
- [61] Yishi Xu, Yingxue Zhang, Wei Guo, Hufeng Guo, Ruiming Tang, and Mark Coates. 2020. Graphsail: Graph structure aware incremental learning for recommender systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2861–2868.
- [62] Dingqi Yang, Bingqing Qu, Jie Yang, and Philippe Cudré-Mauroux. 2019. Revisiting user mobility and social relationships in lsns: a hypergraph embedding approach. In *The world wide web conference*. 2147–2157.
- [63] Dingqi Yang, Bingqing Qu, Jie Yang, and Philippe Cudré-Mauroux. 2020. Lsns2vec++: Heterogeneous hypergraph embedding for location-based social networks. *IEEE Transactions on Knowledge and Data Engineering* (2020).
- [64] 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 *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 974–983.
- [65] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. 2018. Lifelong Learning with Dynamically Expandable Networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.
- [66] Fajie Yuan, Guoxia Zhang, Alexandros Karatzoglou, Joemon Jose, Beibei Kong, and Yudong Li. 2021. One person, one model, one world: Learning continual user representation without forgetting. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 696–705.
- [67] Hengtong Zhang, Changxin Tian, Yaliang Li, Lu Su, Nan Yang, Wayne Xin Zhao, and Jing Gao. 2021. Data Poisoning Attack against Recommender System Using Incomplete and Perturbed Data. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2154–2164.
- [68] Fan Zhou and Chengtai Cao. 2021. Overcoming Catastrophic Forgetting in Graph Neural Networks with Experience Replay. In *AAAI*. AAAI Press, 4714–4722.

A NOTATIONS

We summarize the main notations used in this paper in Table 4.

Table 4: Major notations.

T	The total number of data/time segments
K	The total number of graph convolution layers
\tilde{D}	The user-item interaction data stream
D_t	The data streaming into the system at time segment t
ΔG_t	The graph structure of interaction data D_t
G_t	The graph structure of the union of D_1, D_2, \dots, D_t
S_t	The graph convolution structure at segment t
W_t	The graph convolution parameters at segment t
W^K_t	The topmost graph convolution layer parameters at segment t
W^S_t	The short-term preference-related parameters of W_t
W^L_t	The long-term preference-related parameters of W_t
ΔW_t	The expansion part of graph convolution at segment t

Table 5: Data statistics of filtered datasets.

Dataset	Tb2014	Tb2015	Netflix	Foursquare
user #	8K	192K	301K	52K
item #	39K	10K	9K	37K
interaction #	749K	9M	49M	2M
time span	31 days	123 days	74 months	22 months
AER	35.5%	26.0%	58.4%	60.0%

B DATA STATISTICS OF FILTERED DATASETS

The data statistics of fours filter datasets used in this work are summarized in Table 5.

C SUPPLEMENTARY ABLATION STUDY

The ablation study results on Netflix dataset are present in Figure 6.

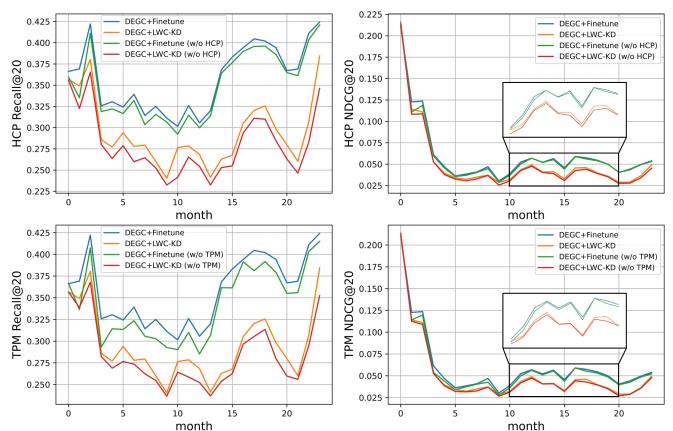


Figure 6: Ablation study to historical convolution pruning (HCP) and temporal preference modeling (TPM) on Netflix dataset.

D IMPLEMENTATION DETAILS

We report our implementation details here. We use the Adam optimizer with an initial learning rate as 0.001. The embedding size and

the width of each graph convolution layer are set to 128. The $L1$ regularization coefficient λ_1 is set to 0.001. The $L2$ regularization coefficient λ_2 and the GSR regularization coefficient λ_g are set to 0.01. We set the batch size as 1,000 when training the GCN models.

The number N of the expansion filters at each layer is set as 30. Without specifications, the hyper-parameters are set same as the original papers. We implement our algorithm with Tensorflow and test it on the NVIDIA GeForce RTX 3090 GPU with 24 GB memory.