Spectral Invariant Learning for Dynamic Graphs under Distribution Shifts

Zeyang Zhang†, Xin Wang‡, Ziwei Zhang†, Zhou Qin‡, Weigao Wen‡, Hui Xue‡, Haoyang Li‡, Wenwu Zhu‡

1Department of Computer Science and Technology, BNRist, Tsinghua University, 2Alibaba Group
zy-zhang20@mails.tsinghua.edu.cn, {xin_wang, zwzhang}@tsinghua.edu.cn,
{qinzhou.qinzhou, weigao.wen, hui.xueh}@alibaba-inc.com,
lihy18@mails.tsinghua.edu.cn, wuwzhu@tsinghua.edu.cn

Abstract

Dynamic graph neural networks (DyGNNs) currently struggle with handling distribution shifts that are inherent in dynamic graphs. Existing work on DyGNNs with out-of-distribution settings only focuses on the time domain, failing to handle cases involving distribution shifts in the spectral domain. In this paper, we discover that there exist cases with distribution shifts unobservable in the time domain while observable in the spectral domain, and propose to study distribution shifts on dynamic graphs in the spectral domain for the first time. However, this investigation poses two key challenges: i) it is non-trivial to capture different graph patterns that are driven by various frequency components entangled in the spectral domain; and ii) it remains unclear how to handle distribution shifts with the discovered spectral patterns. To address these challenges, we propose Spectral Invariant Learning for Dynamic Graphs under Distribution Shifts (SILD), which can handle distribution shifts on dynamic graphs by capturing and utilizing invariant and variant spectral patterns. Specifically, we first design a DyGNN with Fourier transform to obtain the ego-graph trajectory spectrums, allowing the mixed dynamic graph patterns to be transformed into separate frequency components. We then develop a disentangled spectrum mask to filter graph dynamics from various frequency components and discover the invariant and variant spectral patterns. Finally, we propose invariant spectral filtering, which encourages the model to rely on invariant patterns for generalization under distribution shifts. Experimental results on synthetic and real-world dynamic graph datasets demonstrate the superiority of our method for both node classification and link prediction tasks under distribution shifts.

1 Introduction

Dynamic graph neural networks (DyGNNs) have achieved remarkable success in many predictive tasks over dynamic graphs [1, 2]. Existing DyGNNs exhibit limitations in handling distribution shifts, which naturally exist in dynamic graphs due to multiple uncontrollable factors [3, 4, 5, 6]. Existing work on out-of-distribution generalized DyGNNs focuses on handling distribution shifts in the time domain. For example, DIDA [7] utilizes dynamic graph attention to mask the graph trajectories to capture the invariant patterns on dynamic graphs, which assumes that in the time domain, the distribution shift is observable and the invariant and variant patterns can be easily disentangled. However, there exist cases that the distribution shift is unobservable in the time domain while observable in the spectral domain, as shown in Figure 1. The shift in frequency components can

---

*This work was done during the author’s internship at Alibaba Group
†Corresponding authors

Figure 1: An illustration example: the graph dynamics from different frequency components are entangled in the temporal domain, while it is much easier to distinguish different frequency components by masking the spectrums in the spectral domain. In this case, the frequency components in the invariant spectrums determine the node labels, while the relationship between the variant spectrums and labels is not stable under distribution shifts.

be clearly observed in the spectral domain, while these components are indistinguishable in the time domain. Moreover, in real-world applications, the observed dynamic graphs usually consist of multiple mixed graph structural and featural dynamics from various frequency components [8, 9, 10].

To address this problem, in this paper, we study the problem of handling distribution shifts on dynamic graphs in the spectral domain for the first time, which poses the following two key challenges: i) it is non-trivial to capture different graph patterns that are driven by various frequency components entangled in the spectral domain, and ii) it remains unclear how to handle distribution shifts with the discovered spectral patterns.

To tackle these challenges, we propose Spectral Invariant Learning for Dynamic Graphs under Distribution Shifts (SILD). Our proposed SILD model can effectively handle distribution shifts on dynamic graphs by discovering and utilizing invariant and variant spectral patterns. Specifically, we first design a DyGNN with Fourier transform to obtain the ego-graph trajectory spectrums so that the mixed graph dynamics can be transformed into separate frequency components. Then we develop a disentangled spectrum mask that leverages the amplitude and phase information of the ego-graph trajectory spectrums to obtain invariant and variant spectrum masks so that graph dynamics from various frequency components can be filtered. Finally, we propose an invariant spectral filtering that discovers the invariant and variant patterns via the disentangled spectrum masks, and minimize the variance of predictions with exposure to various variant patterns. As such, SILD is able to exploit invariant patterns to make predictions under distribution shifts. Experimental results on several synthetic and real-world datasets, including both node classification and link prediction tasks, demonstrate the superior performance of our SILD model compared to state-of-the-art baselines under distribution shifts. To summarize, we make the following contributions:

• We propose to study distribution shifts on dynamic graphs in the spectral domain, to the best of our knowledge, for the first time.

• We propose Spectral Invariant Learning for Dynamic Graphs under Distribution Shifts (SILD), which can handle distribution shifts on dynamic graphs in the spectral domain.

• We employ DyGNN with Fourier transform to obtain the node spectrums, design a disentangled spectrum mask to obtain invariant and variant spectrum masks in the spectral domain, and propose the invariant spectral filtering mechanism so that SILD is able to handle distribution shifts.

• We conduct extensive experiments on several synthetic and real-world datasets, including both node classification and link prediction tasks, to demonstrate the superior performance of our method compared to state-of-the-art baselines under distribution shifts.

3The codes are available at Github.
2 Problem Formulation and Notations

Dynamic Graphs A dynamic graph can be represented as $G = \{(G^t)_{t=1}^T\}$, where $T$ represents the total number of time stamps, and each $G^t = (V^t, E^t)$ corresponds to a graph snapshot at time stamp $t$ with the node set $V^t$ and the edge set $E^t$. For simplicity, we also represent a graph snapshot as $G^t = (X^t, A^t)$, which includes the node feature matrix $X^t$ and the adjacency matrix $A^t$. We further denote a random variable of $G^t$ as $\hat{G}^t$. The prediction task on dynamic graphs aims to utilize past graph snapshots to make predictions, i.e., $p(Y^t | \hat{G}^{1:t})$, where $G^{1:t} = \{G^1, G^2, \ldots, G^t\}$ denotes the graph trajectory, and the label $Y^t$ represents the node properties or the links at time $t + 1$. For brevity, we take node-level prediction tasks as an example in this paper. Following [7], the distribution of graph trajectory can be factorized into ego-graph trajectories, such that $p(Y^t | G^{1:t}) = \prod_v p(y^t_v | G^{1:t}_v)$.

Distribution Shifts on Dynamic Graphs The common optimization objective for prediction tasks on dynamic graphs is to learn an optimal predictor with empirical risk minimization (ERM), i.e. $\min_{f_\theta} \mathbb{E}_{y^t \sim p_{\text{tr}}(Y^t, G^{1:t})} \mathcal{L}(f_\theta(G^{1:t}_v), y^t)$. where $f_\theta$ is a learnable dynamic graph neural network. Under distribution shifts, however, the optimal predictor trained with ERM and the training distribution may not generalize well to the test distribution, since the risk minimization objectives under two distributions are different due to $p_{\text{tr}}(Y^t, G^{1:t}) \neq p_{\text{te}}(Y^t, G^{1:t})$. The distribution shift on dynamic graphs is complex that may originate from temporal distribution shifts [11, 6, 12, 13, 14] as well as structural distribution shifts [15, 16, 17]. For example, trends or community structures can affect interaction patterns in co-author networks [18] and recommendation networks [19], i.e., the distribution of ego-graph trajectories may vary through time and structures.

Following out-of-distribution (OOD) generalization literature [7, 15, 11, 20, 21, 22], we make the following assumptions of distribution shifts on dynamic graphs:

Assumption 1 For a given task, there exists a predictor $f(\cdot)$, for samples $(G^{1:t}_v, y^t_v)$ from any distribution, there exists an invariant pattern $P^1_f(v)$ and a variant pattern $P^2_f(v)$ such that the following conditions are satisfied: 1) the invariant patterns are sufficient to predict the labels, $y^t_v = f(P^1_f(v)) + \epsilon$, where $\epsilon$ is a random noise, 2) the observed data is composed of invariant and variant patterns, $P^1_f(v) = G^{1:t}_v, P^2_f(v)$. 3) the influence of the variant patterns on labels is shielded by the invariant patterns, $y^t_v \perp P^2_f(v) | P^1_f(v)$.

In the next section, inspired by [16], we give a motivation example to provide some high-level intuition before going to our formal method.

3 Motivation Example

Here we introduce a toy dynamic graph example to motivate learning invariant patterns in the spectral domain. We assume that the invariant and variant patterns lie in the 1-hop neighbors, i.e., each node has an invariant subgraph and a variant subgraph. For simplicity, we focus on the number of neighbors, i.e., each node $v$ has an invariant subgraph related degree $d_{v,1} \in \mathbb{R}^{T \times 1}$ and a variant subgraph related degree $d_{v,2} \in \mathbb{R}^{T \times 1}$. Only the former determines the node label, i.e., $y_v = g^T d_{v,1}$. Note that invariant and variant subgraphs are not observed in the data. We further assume a one-dimensional constant feature for each node, which is set as 1 without loss of generality.

For the model in the spatial-temporal domain, we adopt sum pooling as one-layer graph convolution, i.e., the message passing for each node and time is $h_v = \sum_{u \in N_v} 1 \ast d_{v,1} + d_{v,2}$. We further adopt a mask $m \in \mathbb{R}^{T \times 1}$ to filter patterns in the temporal domain and make predictions by a linear classifier, i.e., $\hat{y}_v = w^T (m \odot h_v)$, where $w \in \mathbb{R}^{T \times 1}$ denotes the learnable parameters. Then, the empirical risk in the training dataset $D_{tr}$ is $R_{tr}(w) = \frac{1}{|D_{tr}|} \sum_{v \in D_{tr}} (\hat{y}_v - y_v)^2$. We have the following proposition.

Proposition 1 For any mask $m \in \mathbb{R}^{T \times 1}$, for the optimal classifier in the training data $w^* = \arg \min_{w} R_{tr}(w)$, if $\|m \odot w^*\|_2 \neq 0$, there exist OOD nodes with unbounded error; i.e., $\exists v \text{ s.t. } \lim_{\|d_{v,2}\| \to \infty} (\hat{y}_v - y_v)^2 = \infty$.

The proposition [1] shows that a classifier trained with masks and empirical risk minimization has unbounded risks in testing data under distribution shifts as the classifier uses variant patterns to make predictions. Next, we show that under mild conditions, an invariant linear classifier in the spectral
domain can solve this problem. Denote $\Phi \in \mathbb{C}^{T \times T}$ as the Fourier bases, where $\Phi_{k,t} = \frac{1}{\sqrt{T}} e^{-j2\pi kt}$. Denote $z_v = \Phi \sum_{u \in N_v} 1$ as the spectral representation after a linear message-passing. The prediction is $\hat{y}_v = w^H(m \odot z_v)$, where $m \in \mathbb{C}^{T \times 1}$ is the mask to filter the spectral patterns, $w \in \mathbb{C}^{T \times 1}$ is a linear classifier, and $(\cdot)^H$ denotes Hermitian transpose. We have the following proposition.

**Proposition 2** If $\left( \Phi d_{u,1} \odot \Phi d_{v,1} \right) \odot \left( \Phi d_{u,2} \odot \Phi d_{v,2} \right) = 0, \forall d_{u,1}, d_{v,2}$, then $\exists m \in \mathbb{C}^{T \times 1}$ such that the optimal spectral classifier in the training data has bounded error, i.e., for $w^* = \arg \min_w R_{tr}(w)$, $\exists \epsilon > 0, \forall v, \lim_{||d_{v,2}|| \to \infty} (\hat{y}_v - y_v)^2 < \epsilon$.

The proposition shows that if the frequency bandwidths of invariant and variant patterns do not have any overlap, there exists a spectral mask such that a linear classifier trained with empirical risk minimization in the spectral domain will have bounded risk in any testing data distribution. This example motivates us to capture invariant and variant patterns in the spectral domain, which is not feasible in the spatial-temporal domain.

4 Method

In this section, we introduce our method named Spectral Invariant Learning for Dynamic Graphs under Distribution Shifts (SILD) to handle distribution shifts in dynamic graphs, including three modules, dynamic graph neural networks with Fourier transform, disentangled spectrum mask, and invariant spectral filtering. The framework of our method is shown in Figure 2.

4.1 Dynamic Graph Neural Network with Spectral Transform

**Dynamic Graph Trajectories Modeling** Each node on the dynamic graph has its ego-graph trajectory evolving through time that may determine the node properties or the occurrence of future links. Following [23][24][25], we adopt a message-passing network for each graph snapshot to aggregate the neighborhood information at the current time, i.e.,

$$m_{u \to v}^{t} \leftarrow \text{MSG}(h_u^{t}, h_v^{t}), h_v^{t} \leftarrow \text{AGG}([m_{u \to v}^{t} | u \in N^t(v)]),$$

where ‘MSG’ and ‘AGG’ denote message and aggregation functions, $m_{u \to v}^{t}$ is the message from node $u$ to node $v$, $h_u^{t}$ is the node embedding for node $u$ at time $t$, $N^t(v) = \{u | (u, v) \in E^t\}$ is node $v$’s neighborhood at time $t$. To model the high-order neighborhood information, we can stack multiple message-passing layers. In this way, the node embedding along time $h_u^{T} = \{h_u^{t} | t = 1\}$ summarizes the evolution of node $u$’s ego-graph trajectories. We denote $H \in \mathbb{R}^{T \times N \times d}$ as the ego-graph trajectory signals for all nodes on the dynamic graph, where $T$ denotes the total time length, $N$ denotes the number of nodes and $d$ denotes the hidden dimensionality.

**Spectral Transform** As some patterns on dynamic graphs are unobservable in the time domain, while observable in the spectral domain, we transform the summarized ego-graph trajectory signals $H$ into the spectral domain via Fourier transform for each node and hidden dimension, i.e.,

$$\Phi_{k,t} = \frac{1}{\sqrt{T}} e^{-j2\pi kt}, Z = \Phi H,$$

where $\Phi \in \mathbb{C}^{K \times T}$ denotes the Fourier bases, $K$ denotes the number of frequency components, and $Z \in \mathbb{C}^{K \times N \times d}$ denote the node embeddings along frequency components in the spectral domain, and $Z_{k,n,m} = \sum_{t=1}^{T} \Phi_{k,t} H_{n,m}^{t}$. By choosing the Fourier bases, our spectral transform has the following advantages: 1) we can use fast Fourier transform (FFT) [26] to accelerate the computation. The computation complexity of Eq. (2) can be reduced from $O(NdT^2)$ to $O(NdT logT)$. 2) Each basis has clear semantics, e.g., $Z_k$ denotes the node embeddings at the $k$-th frequency component in the spectral domain. In this way, we can observe how the nodes on the dynamic graph evolve in different frequency bandwidths. 3) Fourier transform is able to capture global and periodic patterns [27][28], which are real common in real-world dynamic graphs, e.g., the interactions on e-commerce networks may result from seasonal sales or product service cycles.
Dynamic Graph Neural Networks with Spectral Transform

Disentangled Spectrum Mask

Invariant Spectral Filtering

Figure 2: The framework of our proposed method SILD. Given a dynamic graph evolving through time, the dynamic graph neural networks with spectral transform first obtain the ego-graph trajectory spectrums in the spectral domain. Then the disentangled spectrum mask leverages the amplitude and phase information of the ego-graph trajectory spectrums to obtain invariant and variant spectrum masks. Last, invariant spectral filtering discovers the invariant and variant patterns via the disentangled spectrum masks, and minimizes the variance of predictions with exposure to various variant patterns, to help the model exploit invariant patterns to make predictions under distribution shifts.

4.2 Disentangled Spectrum Mask

To capture invariant patterns in the spectral domain, we propose to explicitly learn spectrum masks to disentangle the invariant and variant patterns. The embeddings in the spectral domain contain the amplitude information as well as the phase information for each node

$$\text{Amp}(Z) = \left(\text{Imag}^2(Z) + \text{Real}^2(Z)\right)^{\frac{1}{2}}, \quad \phi(Z) = \arctan \frac{\text{Imag}(Z)}{\text{Real}(Z)},$$

where $\text{Real}(\cdot)$ and $\text{Imag}(\cdot)$ denote the real and imaginary part of the complex number, i.e., $Z = \text{Real}(Z) + j \text{Imag}(Z)$, $j$ denotes the imaginary unit, $\text{Amp}(Z) \in \mathbb{R}^{K \times N \times d}$ and $\phi(Z) \in \mathbb{R}^{K \times N \times d}$ denote the amplitude and phase information respectively. For brevity, the tensor operators in Eq. (3) are all element-wise, e.g., $(\text{Imag}^2(Z))_{i,j,k} = (\text{Imag}(Z_{i,j,k}))^2$. Then, we obtain the spectrum masks by leveraging both the amplitude and phase information

$$M = \text{MLP}(\text{Real}(Z)||\text{Imag}(Z)), \quad M_I = \text{sigmoid}(M_I/\tau), \quad M_V = \text{sigmoid}(-M_I/\tau),$$

where MLP denotes multi-layer perceptrons, $\tau$ is the temperature, $M_I \in [0, 1]^K$ and $M_V \in [0, 1]^K$ denote the spectrum mask for invariant and variant patterns, and $||$ represents the concatenation of the embeddings. In this way, the invariant and variant masks have a negative relationship, and each node can have its own spectrum mask. As the phase information includes high-level semantics in the original signals, we keep the phase information unchanged to reduce harm in the fine-grained semantic information for the graph trajectories, and filter the spectrums by the learned disentangled masks in terms of amplitudes,

$$Z_I = \left(M_I \odot \text{Amp}(Z)\right) \odot (\cos \phi(Z) + j \sin \phi(Z)), \quad Z_V = \left(M_V \odot \text{Amp}(Z)\right)(\cos \phi(Z) + j \sin \phi(Z)), \quad (5)$$

where $Z_I$ and $Z_V$ denote the summarized invariant and variant patterns in the spectral domain. For node classification tasks, we can directly adopt the spectrums for the classifier to predict classes. For link prediction tasks, we can utilize inverse fast Fourier transform (IFFT) to transform the embeddings into the temporal domain for future link prediction

$$H'_I = \Phi^H Z_I, \quad H'_V = \Phi^H Z_V,$$
where $(\cdot)^H$ is Hermitian transpose, $H'_I$ and $H'_V$ denote the filtered invariant and variant patterns that are transformed back into the temporal domain respectively.

### 4.3 Invariant Spectral Filtering

Under distribution shifts, the variant patterns on dynamic graphs have varying relationships with labels, while the invariant patterns have sufficient predictive abilities with regard to labels. We propose invariant spectral filtering to capture the invariant and variant patterns in the spectral domain, and help the model focus on invariant patterns to make predictions, thus handling distribution shifts. We take node classification tasks for an example as follows.

Let $Z_I \in \mathbb{C}^{K \times N \times d}$ and $Z_V \in \mathbb{C}^{K \times N \times d}$ be the filtered invariant and variant spectrums in the spectral domain. Then we can utilize the invariant and variant node spectrums to calculate the task loss

$$L_I = l(f_I(Z_I), Y), \quad L_V = l(f_V(Z_V), Y),$$

where $f_I(\cdot)$ and $f_V(\cdot)$ are the classifiers for invariant and variant patterns respectively, $Y$ is the labels, and $l$ is the loss function. The task loss is utilized to capture the patterns with the predictive abilities of labels. Recall in Assumption 1 the influence of variant patterns on labels is shielded given invariant patterns as the invariant patterns have sufficient predictive abilities w.r.t labels, and thus the model’s predictions should not change when being exposed to different variant patterns and the original invariant patterns. Inspired by [15, 7], we calculate the invariance loss by

$$L_{INV} = \text{Var}(\{L_m | \tilde{z} : \tilde{z} \in \mathcal{S}\}),$$

where $L_m | \tilde{z}$ denotes the mixed loss to measure the model’s prediction ability with exposure to the specific variant pattern $\tilde{z} \in \mathbb{C}^{K \times d}$ that is sampled from a set of variant patterns $\mathcal{S}$. We adopt all the node embeddings in $Z_V$ to construct the set of variant patterns $\mathcal{S}$. Inspired by [34, 15], we calculate the mixed loss as

$$L_m | \tilde{z} = l(f_I(Z_I) \odot \sigma(f_V(\tilde{z})), Y),$$

where $\sigma$ denotes the sigmoid function. Then, the final training objective is

$$\min_{\theta, f_I} L_I + \lambda L_{INV} + \min_{f_V} L_V,$$

where $\theta$ is the parameters that encompass all the model parameters except the classifiers, $\lambda$ is a hyperparameter to balance the trade-off between the model’s predictive ability and invariance properties. A larger $\lambda$ encourages the model to capture patterns with better invariance under distribution shifts, with the potential risk of lower predictive ability during training, as the shortcuts brought by the variant patterns might be discarded in the training process. After training, we only adopt invariant patterns to make predictions in the inference stage. The overall algorithm for training on node classification datasets is summarized in Algo. 1.

#### Algorithm 1: Training pipeline for SILD on node classification datasets

**Require:** Training epochs $L$, sample number $S$, hyperparameter $\lambda$

1. for $l = 1, \ldots, L$
2. Obtain the node embeddings $H$ with snapshot-wise message passing as Eq. (1)
3. Transform the node embeddings into the spectral domain with FFT as Eq. (2)
4. Calculate the disentangled spectrum masks as Eq. (4)
5. Filter spectrums into invariant and variant patterns as Eq. (5)
6. Calculate the task loss as Eq. (7)
7. Sample $S$ variant patterns from collections of $Z_V$ and calculate the invariance loss as Eq. (8)
8. Update the model according to Eq. (10)
9. end for

### 5 Experiments

In this section, we conduct extensive experiments to verify that our proposed method can handle distribution shifts on dynamic graphs by discovering and utilizing invariant patterns in the spectral domain. More details of the settings and other results can be found in Appendix.

**Baselines.** We adopt several representative dynamic GNNs and Out-of-Distribution (OOD) generalization methods as our baselines.
Aminer [42, 43]. Following [7], we adopt the challenging inductive future link prediction task on

Table 2: Results of different methods on synthetic link prediction and node classification datasets. The best results are in bold and the second-best results are underlined. A larger ‘shift’ denotes a higher distribution shift level.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Link-Synthetic (AUC%)</th>
<th>Node-Synthetic (ACC%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>GCRN</td>
<td>72.57±0.72</td>
<td>72.29±0.47</td>
</tr>
<tr>
<td>EGCN</td>
<td>69.00±0.53</td>
<td>62.70±1.14</td>
</tr>
<tr>
<td>DySAT</td>
<td>70.24±1.26</td>
<td>64.01±0.19</td>
</tr>
<tr>
<td>IRM</td>
<td>69.40±0.09</td>
<td>63.97±0.37</td>
</tr>
<tr>
<td>VReX</td>
<td>70.44±1.08</td>
<td>63.99±0.21</td>
</tr>
<tr>
<td>GroupDRO</td>
<td>70.30±1.23</td>
<td>64.05±0.21</td>
</tr>
<tr>
<td>DIDA</td>
<td>85.20±0.84</td>
<td>82.89±0.23</td>
</tr>
<tr>
<td>SILD</td>
<td>85.95±0.18</td>
<td>84.69±1.18</td>
</tr>
</tbody>
</table>

- Dynamic GNNs: GCRN [35] is a representative dynamic GNN that first adopts a GCN [36] to obtain node embeddings and then a GRU [37] to model the network evolution. EGCN [25] adopts an LSTM [38] or GRU [37] to flexibly evolve the GCN [36] parameters through time. DySAT [24] aggregates neighborhood information at each graph snapshot using structural attention and models network dynamics with temporal self-attention.

- OOD generalization methods: IRM [20] aims at learning an invariant predictor which minimizes the empirical risks for all training domains. GroupDRO [39] puts more weight on training domains with larger errors to minimize the worst-group risks across training domains. V-REx [40] reduces the differences in the risks across training domains to reduce the model’s sensitivity to distributional shifts. As these methods are not specifically designed for dynamic graphs, we adopt the best-performed dynamic GNNs as their backbones on each dataset.

- OOD generalization methods for dynamic graphs: DIDA [7] utilizes disentangled attention to capture invariant and variant patterns in the spatial-temporal domain, and conducts spatial-temporal intervention mechanism to let the model focus on invariant patterns to make predictions.

5.1 Real-world Datasets

**Settings** We use 3 real-world dynamic graph datasets, including Collab [41][7], Yelp [24][2] and Aminer [42][53]. Following [7], we adopt the challenging inductive future link prediction task on Collab and Yelp, where the model should exploit historical graphs to predict the occurrence of links in the next time step. To measure the model’s performance under distribution shifts, the model is tested on another dynamic graph with different fields, which is unseen during training. For node classification, we adopt Aminer, a citation network, where nodes represent papers, and edges from u to v with timestamp t denote the paper u published at year t cites the paper v. The task is to predict the venues of the papers. We train on papers published between 2001 - 2011, validate on
those published in 2012 - 2014, and test on those published since 2015. On this dataset, the model is tested to exploit the invariant patterns and make stable predictions under distribution shifts, where the patterns on the dynamic graph may vary in different years.

**Results**  Based on the results in Table 1, we have the following observations: 1) Under distribution shifts, the general OOD generalization baselines have limited improvements over the dynamic GNNs, e.g., GroupDRO improves over DySAT with 0.9% in Yelp and 0.3% in Aminer15 respectively. A plausible reason is that they are not specially designed to handle distribution shifts on dynamic graphs, and may not consider the graph structural and temporal dynamics to capture invariant patterns. Another reason might be that they strongly rely on high-quality environment labels to capture invariant patterns, which are almost unavailable on real-world dynamic graphs. 2) Our method can better handle distribution shifts than the baselines. The datasets have strong distribution shifts, e.g., COVID-19 happens midway and has considerable influence on the consumer behavior on Yelp, and the citation patterns may shift with the outbreak of deep neural networks on Aminer. Nevertheless, our method SILD has significant improvements over the state-of-the-art OOD generalization baseline for dynamic graphs DIDA on all datasets, e.g., 2% on average for most datasets, which verifies that our method can better capture the invariant and variant patterns in the spectral domain, and thus handling distribution shifts on dynamic graphs.

5.2 Synthetic Datasets

**Settings**  To evaluate the model’s generalization ability under distribution shifts, we conduct experiments on synthetic link prediction and node classification datasets, which are constructed by introducing manually-designed distribution shifts. For link prediction datasets, we follow [7] to generate additional varying features for each node and timestamps on the original dataset Collab, where these additional features are constructed with spurious correlations w.r.t the labels, i.e., the links in the next timestamps. The spurious correlation degree is determined by a shift level parameter. On this dataset, to have better generalization ability, the model should not rely on variant patterns that exploit the additional features with spurious correlations. For node classification, we briefly introduce the construction of the synthetic dataset as follows. We generate the dynamic graph with stochastic block model [44], where the link probability between nodes at each graph snapshot is determined by two frequency factors. The correlation of one of the factors with class labels is always 1, while the other factor has a variant relationship with labels, where the relationship is also controlled by a shift level parameter. The model should discover and focus on the invariant frequency factors whose relationship with labels is invariant under distribution shifts. For both datasets, we set the shift level parameters as 0.4, 0.6, 0.8 for training and validation splits, and 0 for test splits.

**Results**  Based on the results in Table 2, we have the following observations: 1) Our method can better handle distribution shifts than the baselines, especially under stronger distribution shifts. SILD consistently outperforms DyGNN and general OOD generalization baselines by a significantly large margin, which can credit to our special design to handle distribution shifts on dynamic graphs in the spectral domain. Our method also has a significant improvement over the best-performed baseline under the strongest distribution shift, e.g., with absolute improvements of 5% in Link-Synthetic(0.8) and 7% in Node-Synthetic(0.8) respectively. 2) Our method can exploit invariant patterns to consistently alleviate the harmful effects of variant patterns under different distribution shift levels. As the distribution shift level increases, almost all methods decline in performance since the relationship between variant patterns and labels goes stronger, so that the variant patterns are much easier to be exploited by the model, misleading the training process. However, the performance drop of SILD is significantly lower than baselines, which demonstrates that our method can alleviate the harmful effects of variant patterns under distribution shifts by exploiting invariant patterns in the spectral domain.

5.3 Ablation Studies

We conduct ablation studies to verify the effectiveness of the proposed disentangled spectrum mask and invariant spectral filtering in SILD. The ablated version ‘SILD w/o I’ removes invariant spectral filtering in SILD by setting λ = 0, and ‘SILD w/o M’ is trained without the disentangled spectrum masks. From Figure 3, we have the following observations. First, our proposed SILD outperforms all the variants as well as the best-performed baseline on all datasets, demonstrating the effectiveness of
each component of our proposed method. Second, ‘SILD w/o I’ and ‘SILD w/o M’ drop drastically in performance on all datasets compared to the full version, which verifies that our proposed disentangled spectrum mask and spectral invariant learning can help the model to focus on invariant patterns to make predictions and significantly improve the performance under distribution shifts.

6 Related Works

Dynamic Graph Neural Networks  Dynamic graphs ubiquitously exist in real-world applications such as event forecasting, recommendation, etc. In comparison with static graphs, dynamic graphs contain rich temporal information. Considerable research attention has been devoted to dynamic graph neural networks (DyGNNs) to model the complex graph dynamics that include structures and features evolving through time. Some works adopt GNN to aggregate neighborhood information for each graph snapshot, and then utilize a sequence module to model the temporal information. Some others utilize time-encoding techniques to encode the temporal links into time-aware embeddings and adopt a GNN or memory module to process structural information. Some other related works leverage spectral graph neural networks, global graph framelet convolution, and graph wavelets to obtain better dynamic graph representations. However, distribution shifts remain largely unexplored in dynamic graph neural networks literature. The sole prior work DIDA handles spatial-temporal distribution shifts on dynamic graphs in the temporal domain. To the best of our knowledge, this is the first study of handling distribution shifts on dynamic graphs in the spectral domain.

Out-of-Distribution Generalization  A significant proportion of existing machine learning methodologies operate on the assumption that training and testing data are independent and identically distributed (i.i.d.). However, this assumption may not always hold true, especially in the context of complex real-world scenarios, and the uncontrollable distribution shifts between training and testing data distribution may lead to a significant decline in the model performance. Out-of-Distribution (OOD) generalization problem has recently drawn great attention in various areas. Some works handle structural distribution shifts on static graphs and temporal distribution shifts on time-series data. However, how to handle distribution shifts on dynamic graphs in the spectral domain remains unexplored.

Spectral Methods in Neural Networks  The applications of spectral methods in neural networks have been broadly explored in many areas, including static graph data, time-series data, etc., for their advantages of modeling global patterns, powerful expressiveness and interpretability. Some work proposes to reconstruct the image in the spectral domain to obtain robust image representations. Some work proposes to augment the image data by perturbing the amplitude information in the spectral domain. Some work proposes a multiwavelet-based method for compressing operator kernels. However, these methods are not applicable to dynamic graphs, not to mention the more complex scenarios under distribution shifts.
7 Conclusion

In this paper, we propose a novel model named Spectral Invariant Learning for Dynamic Graphs under Distribution Shifts (SILD), which can handle distribution shifts on dynamic graphs in the spectral domain. We design a DyGNN with Fourier transform to obtain the ego-graph trajectory spectrums. Then we propose a disentangled spectrum mask and invariant spectral filtering to discover the invariant and variant patterns in the spectral domain, and help the model rely on invariant spectral patterns to make predictions. Extensive experimental results on several synthetic and real-world datasets, including both node classification and link prediction tasks, demonstrate the superior performance of our method compared to state-of-the-art baselines under distribution shifts. One limitation is that in this paper we mainly focus on dynamic graphs in scenarios of discrete snapshots, and we leave extending our methods to continuous dynamic graphs for further explorations.

Acknowledgements

This work was supported by the National Key Research and Development Program of China No. 2020AAA0106300, National Natural Science Foundation of China (No. 62222209, 62250008, 62102222, 62206149), Beijing National Research Center for Information Science and Technology under Grant No. BNR2023RC01003, BNR2023TD03006, China National Postdoctoral Program for Innovative Talents No. BX20220185, China Postdoctoral Science Foundation No. 2022M711813, and Beijing Key Lab of Networked Multimedia. All opinions, findings, conclusions, and recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

References


