Causality-Inspired Spatial-Temporal Explanations for Dynamic Graph Neural Networks

Kesen Zhao
City University of Hong Kong
kesenzhao2-c@my.cityu.edu.hk

Liang Zhang
Shenzhen Research Institute of Big Datao
zhangliang@sribd.cn

Background & Motivation

- Dynamic Graph Neural Networks (DyGNNs)
  - Spatial interpretability
  - Temporal interpretability

Background & Motivation

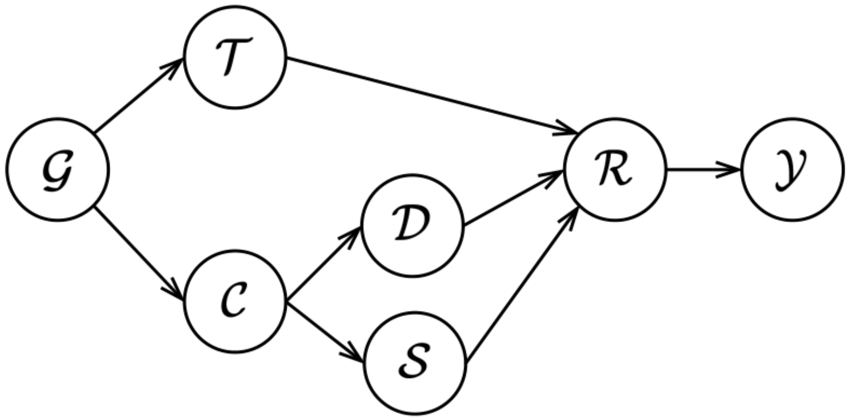
- Dynamic Graph Neural Networks (DyGNNs)
  - Spatial interpretability
  - Temporal interpretability
- DyGNNExplainer
  - Disentangle the trivial relationship and the causal relationship
  - Disentangle the dynamic relationship and the static relationship

A causal view on DyGNNs

- Backdoor path
  - Between causal and trivial

    $\mathcal{C} \leftarrow \mathcal{G} \rightarrow \mathcal{T} \rightarrow \mathcal{R} \rightarrow \mathcal{Y}$
  - Between dynamic and static

    $\mathcal{D} \leftarrow \mathcal{C} \rightarrow \mathcal{S} \rightarrow \mathcal{R} \rightarrow \mathcal{Y}$



$\mathcal{G}$: graph data
$\mathcal{C}$: causal factor
$\mathcal{S}$: static factor
$\mathcal{Y}$: prediction
$\mathcal{T}$: trivial factor
$\mathcal{D}$: dynamic factor
$\mathcal{R}$: representation

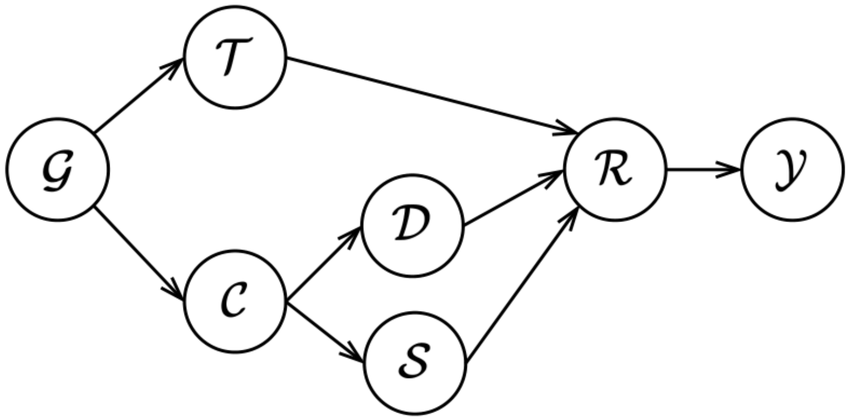
A causal view on DyGNNs

- Backdoor path
  - Between causal and trivial

    $\mathcal{C} \leftarrow \mathcal{G} \rightarrow \mathcal{T} \rightarrow \mathcal{R} \rightarrow \mathcal{Y}$
  - Between dynamic and static

    $\mathcal{D} \leftarrow \mathcal{C} \rightarrow \mathcal{S} \rightarrow \mathcal{R} \rightarrow \mathcal{Y}$
- Backdoor adjustment

$$
\begin{aligned}
P(\mathcal{Y}|do(\mathcal{D})) &= \sum P(\mathcal{Y}|do(\mathcal{D}),\mathcal{S})P(\mathcal{S}|do(\mathcal{D})) \\
&= \sum P(\mathcal{Y}|do(\mathcal{C}))P(\mathcal{S}) \\
&= \sum P(\mathcal{S}) \sum P(\mathcal{Y}|do(\mathcal{C}),\mathcal{T})P(\mathcal{T}|do(\mathcal{C})) \\
&= \sum P(\mathcal{S}) \sum P(\mathcal{Y}|\mathcal{G})P(\mathcal{T}).
\end{aligned}
$$



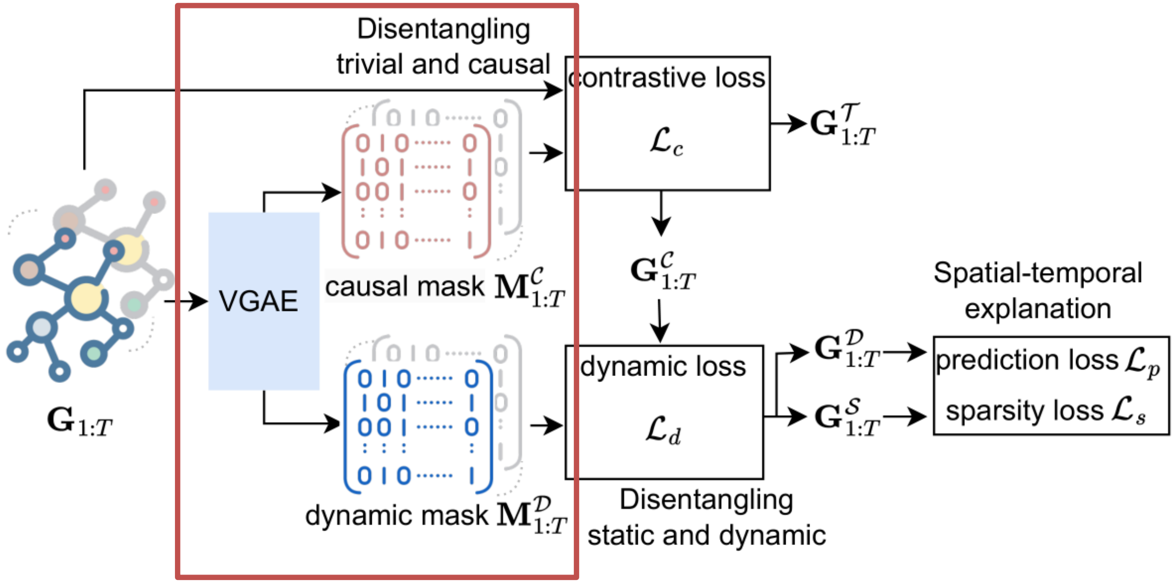
$\mathcal{G}$: graph data
$\mathcal{C}$: causal factor
$\mathcal{S}$: static factor
$\mathcal{Y}$: prediction
$\mathcal{T}$: trivial factor
$\mathcal{D}$: dynamic factor
$\mathcal{R}$: representation

Disentangling Complex Causal Relationships

- Estimating soft mask
  - VGAE-based dynamic encoder-decoder

$$q(\mathbf{H}_t \mid \mathbf{G}_{1:t}) = \Pi_{i=1}^{N} q\left(\mathbf{h}_{t,i} \mid \mathbf{G}_{1:t}\right), q\left(\mathbf{h}_{t,i} \mid \mathbf{G}_{1:t}\right) = \mathcal{N}\left(\mathbf{h}_{t,i} \mid \boldsymbol{\mu}_{t,i}, \operatorname{diag}\left(\boldsymbol{\sigma}_{t,i}^{2}\right)\right)$$

$$p(\mathbf{M}_t^{\mathcal{C}} \mid \boldsymbol{H}_t) = \prod_{i=1}^{N}\prod_{j=1}^{N} p\left(M_{t,ij}^{\mathcal{C}} \mid \mathbf{h}_{t,i}, \mathbf{h}_{t,j}\right), p\left(M_{t,ij}^{\mathcal{C}} = 1 \mid \mathbf{h}_{t,i}, \mathbf{h}_{t,j}\right) = g\left(\mathbf{h}_{t,i}, \mathbf{h}_{t,j}\right)$$



Disentangling Complex Causal Relationships

- Estimating soft mask
  - VGAE-based dynamic encoder-decoder
  - Casual, dynamic, static factor

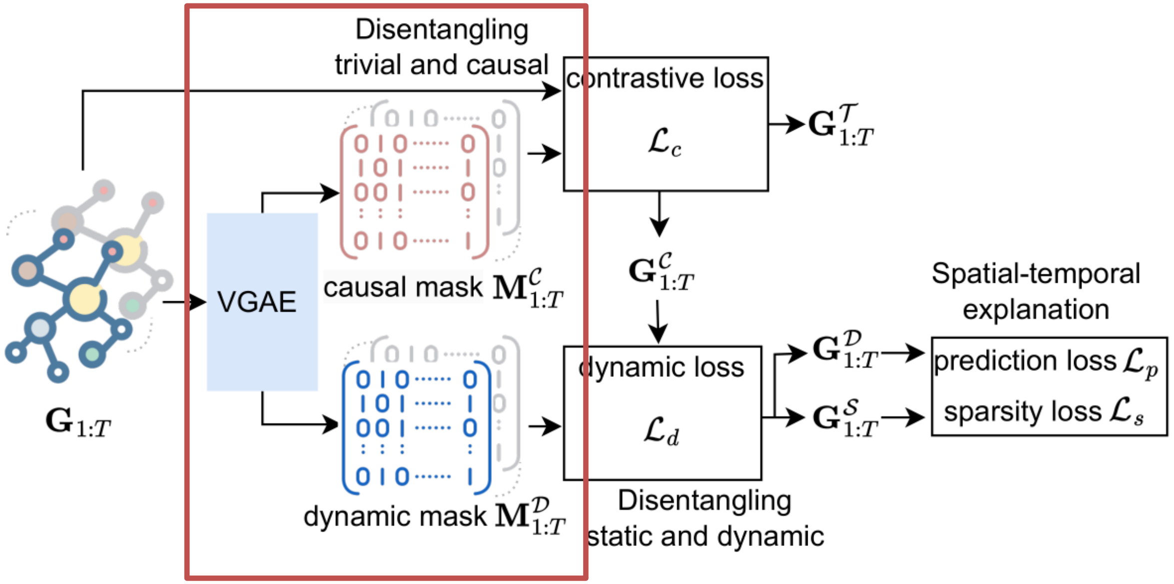
$$\mathbf{M}_t^{\mathcal{C}} = f_v\left(\mathbf{X}_{1:t}, \mathbf{A}_{1:t}; \Theta_{\mathcal{C}}\right) = p(\mathbf{M}_t^{\mathcal{C}} \mid \mathbf{H}_t) q(\mathbf{H_t} \mid \mathbf{G}_{1:t})$$

$$\mathbf{M}_t^{\mathcal{D}} = f_v\left(\mathbf{X}_{1:t}, \mathbf{A}_{1:t} \oplus \mathbf{M}_{1:t}^{\mathcal{C}}; \Theta_{\mathcal{D}}\right)$$

$$\mathbf{A}_{1:T}^{\mathcal{C}} = \mathbf{A}_{1:T} \oplus \mathbf{M}_{1:T}^{\mathcal{C}}$$

$$\mathbf{A}_{1:T}^{\mathcal{S}} = \mathbf{A}_{1:T} \oplus \mathbf{M}_{1:T}^{\mathcal{C}} \oplus \overline{\mathbf{M}}_{1:T}^{\mathcal{D}}$$

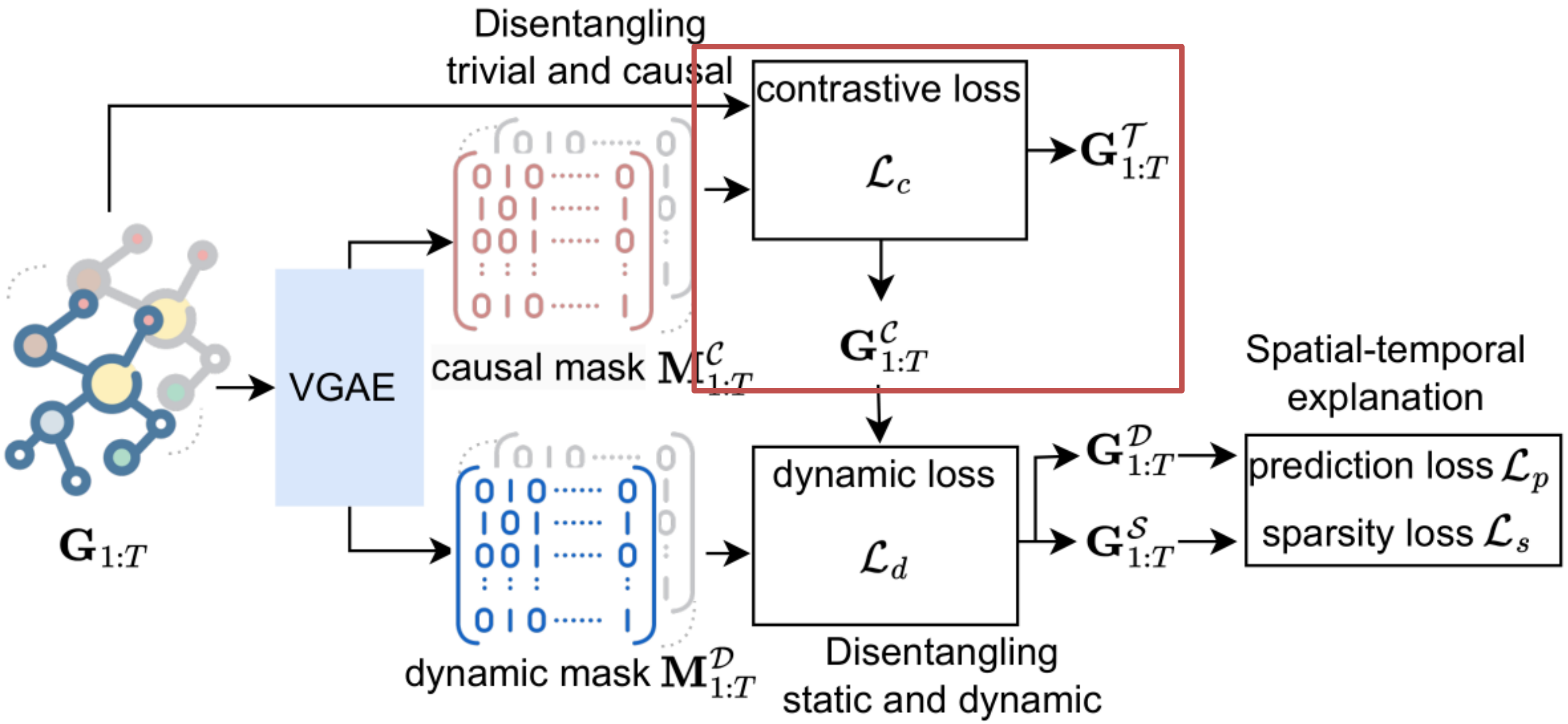
$$\mathbf{A}_{1:T}^{\mathcal{D}} = \mathbf{A}_{1:T} \oplus \mathbf{M}_{1:T}^{\mathcal{C}} \oplus \mathbf{M}_{1:T}^{\mathcal{D}}$$



Disentangling Complex Causal Relationships

- Disentangling trivial and causal

$$\mathcal{L}_c = \frac{1}{T}\sum_{t=1}^{T} \log \frac{\exp\left(s(\mathbf{e}_t, \mathbf{e}_t^{\mathcal{C}})/\tau\right)}{\exp\left(s(\mathbf{e}_t, \mathbf{e}_t^{\mathcal{C}})/\tau\right) + \alpha_1 \exp\left(s(\mathbf{e}_t^{\mathcal{T}}, \mathbf{e}_t^{\mathcal{C}})/\tau\right) + \alpha_2 \sum_{k\neq t} \exp\left(s(\mathbf{e}_t^{\mathcal{T}}, \mathbf{e}_k^{\mathcal{C}})/\tau\right)}$$



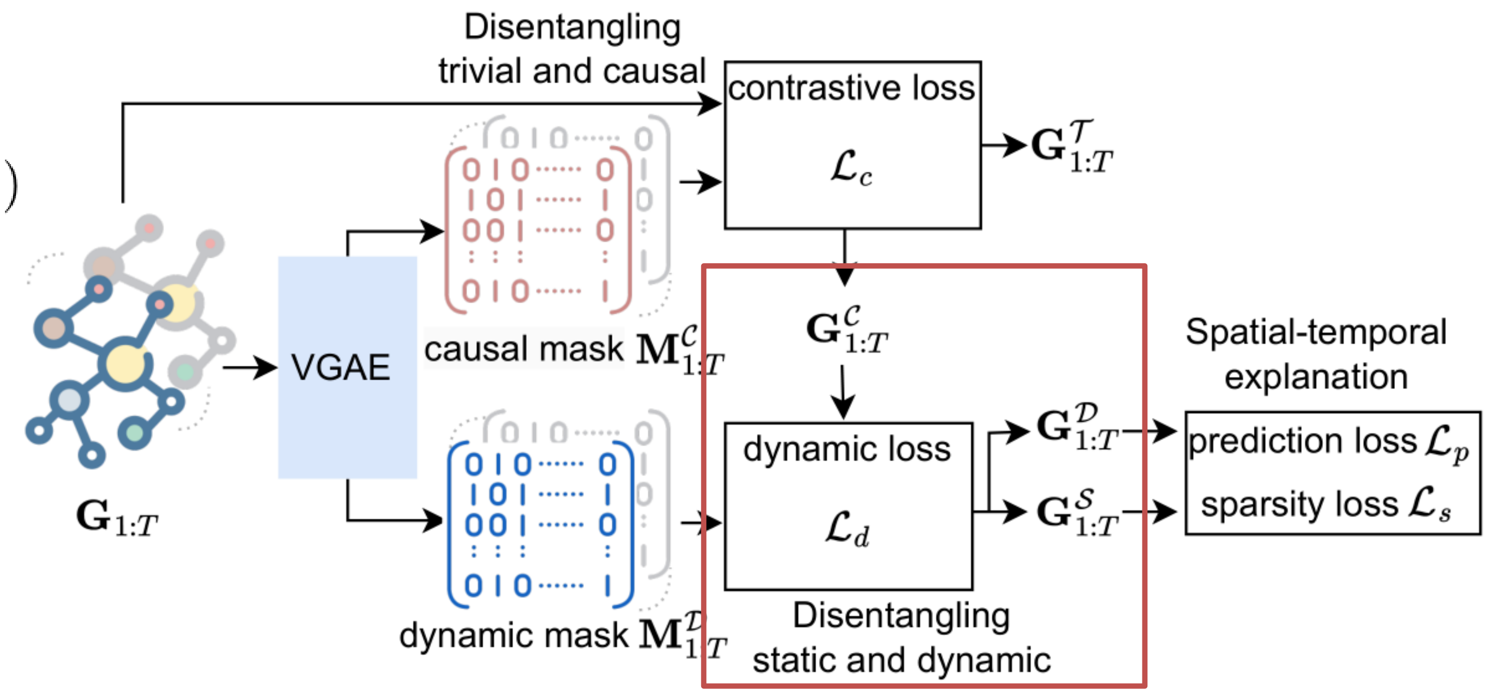
Disentangling Complex Causal Relationships

- Disentangling trivial and causal
- Disentangling static and dynamic

$$\boldsymbol{H}_t^{\mathcal{D}} = GCN(\boldsymbol{A}_t^{\mathcal{D}}, \boldsymbol{X}_t; \Psi_{\mathcal{D}}), \boldsymbol{H}_t^{\mathcal{S}} = GCN(\boldsymbol{A}_t^{\mathcal{S}}, \boldsymbol{X}_t; \Psi_{\mathcal{S}})$$

$$\boldsymbol{H}_{1:(t-1)}^{\mathcal{D}} \longrightarrow \boldsymbol{H}_t^{\mathcal{D}}, \quad \boldsymbol{H}_{1:(t-1)}^{\mathcal{S}} \perp \boldsymbol{H}_t^{\mathcal{S}}$$

$$\mathcal{L}_d = \frac{1}{T-1}\sum_{t=2}^{T} d\left(f_a\left(\mathbf{G}_{\mathbf{1:(t-1)}}^{\mathcal{D}}\right), \boldsymbol{H}_t^{\mathcal{D}}\right)$$



Disentangling Complex Causal Relationships

- Disentangling trivial and causal
- Disentangling static and dynamic
- Spatial-temporal explanation

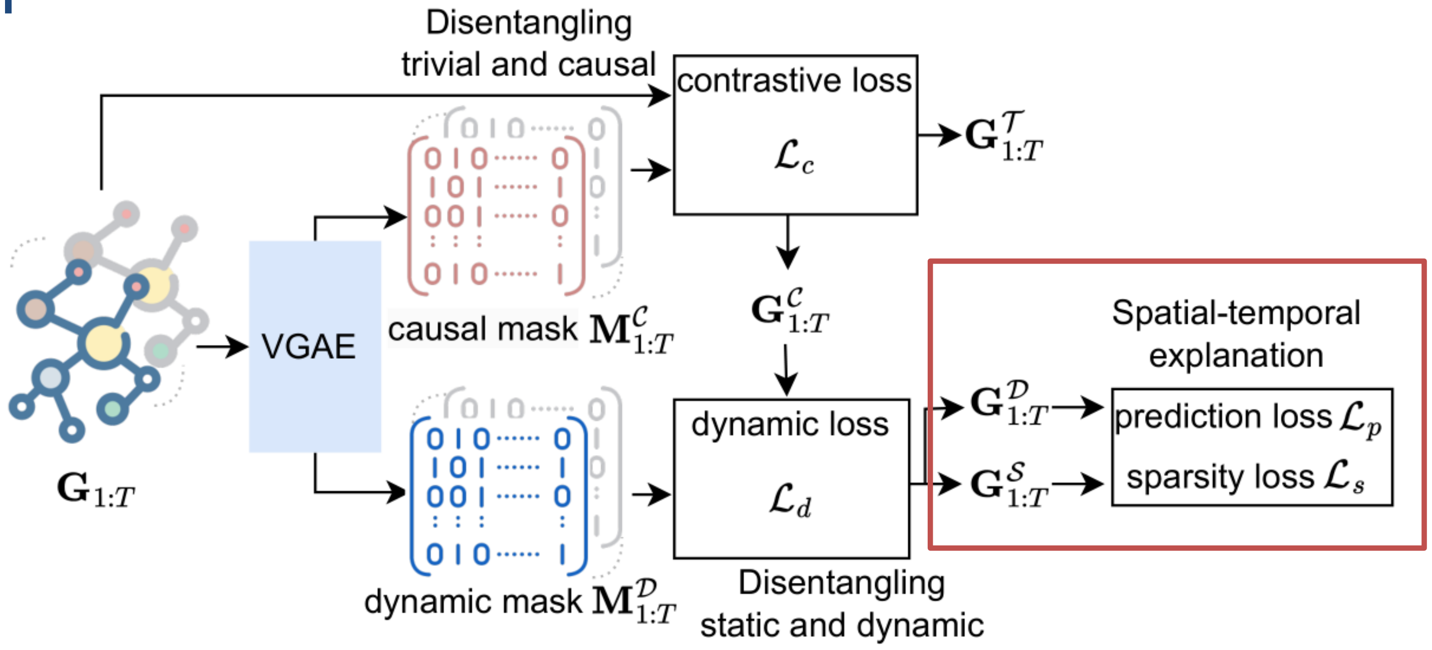
$$\Delta \boldsymbol{H}_t^{\mathcal{D}} = f_a\left(\mathbf{G}_{1:t}^{\mathcal{D}}\right) - f_a\left(\mathbf{G}_{1:(t-1)}^{\mathcal{D}}\right)$$

$$\boldsymbol{H}_T = \sum^{T} t_p(\Delta \boldsymbol{H}_t^{\mathcal{D}} \oplus \boldsymbol{H}_t^{\mathcal{S}}) \Delta \boldsymbol{H}_t^{\mathcal{D}} \oplus \boldsymbol{H}_t^{\mathcal{S}}$$

$$t_p(\mathbf{H}) = Softmax(\Psi_{\mathcal{P}} \mathbf{H} / \|\Psi_{\mathcal{P}}\|)$$

$$\mathcal{L}_p = l(f_d(\boldsymbol{H}_T), \mathcal{Y}))$$

$$\mathcal{L}_s = \sum_{t=1}^{T} \frac{\left\|\mathbf{A}_t^{\mathcal{C}}\right\|_1 + \left\|\mathbf{A}_t^{\mathcal{D}}\right\|_1}{\|\mathbf{A}_t\|_1}$$



Explanation fidelity

Table 2: Explanation accuracy of different models (%). Where best performances are bold.

Task	Dataset	GNNExplainer	PGExplainer	Gem	OrphicX	DyGNNExplainer
Node cls.	DBA-Shapes	92.1	92.9	93.6	94.3	**97.8***
	DTree-Cycles	92.8	93.7	94.4	96	**98.2***
	DTree-Grid	85.2	85.9	87.1	90.5	**94.2***
	Elliptic	92.4	94.1	94.6	96.1	**98.7***
Graph cls.	DBA-2motifs	86.5	88.0	90.7	91.4	**96.3***
	MemeTracker	88.2	89.2	91.0	91.9	**97.4***

"*" indicates the statistically significant improvements (i.e., two-sided t-test with $p < 0.05$) over the best baseline. 'cls.' is short for classification.

- Observations
  - DyGNNExplainer surpasses all other baselines

Explanation fidelity

Table 2: Explanation accuracy of different models (%). Where best performances are bold.

Task	Dataset	GNNExplainer	PGExplainer	Gem	OrphicX	DyGNNExplainer
Node cls.	DBA-Shapes	92.1	92.9	93.6	94.3	**97.8***
	DTree-Cycles	92.8	93.7	94.4	96	**98.2***
	DTree-Grid	85.2	85.9	87.1	90.5	**94.2***
	Elliptic	92.4	94.1	94.6	96.1	**98.7***
Graph cls.	DBA-2motifs	86.5	88.0	90.7	91.4	**96.3***
	MemeTracker	88.2	89.2	91.0	91.9	**97.4***

"*" indicates the statistically significant improvements (i.e., two-sided t-test with $p < 0.05$) over the best baseline. 'cls.' is short for classification.

- Observations
  - Causal-based methods OrphicX and Gem also outperform other baselines

Explanation interpretability analysis

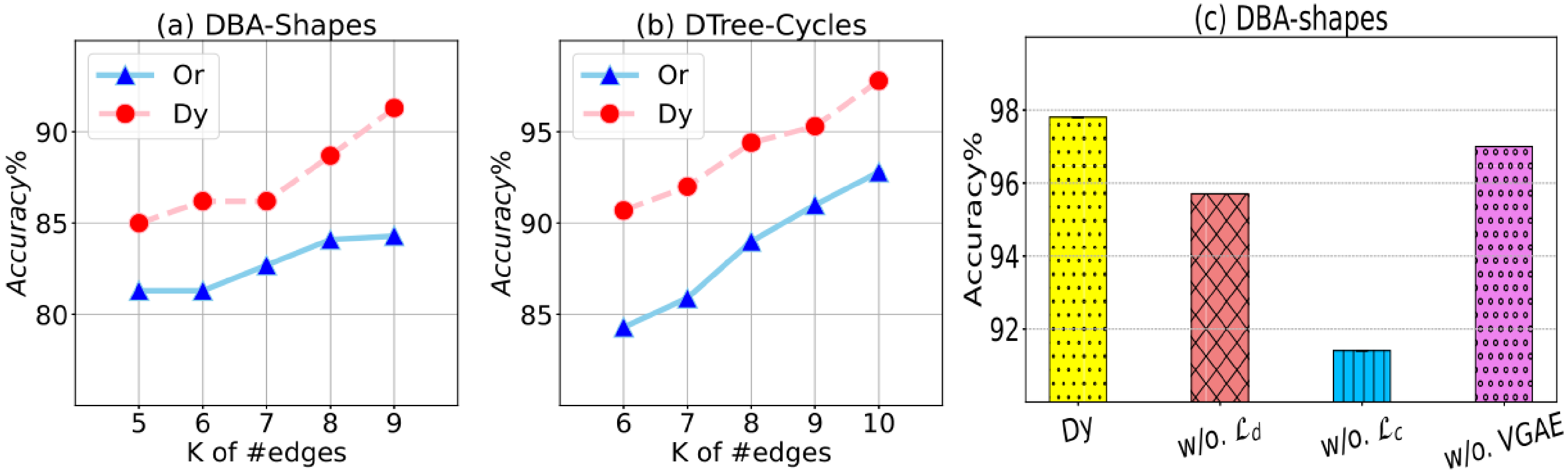


Figure 2: Interpretability analysis and ablation study. (a) Sparsity analysis on DBA-Shapes dataset (b) Sparsity analysis on DTree-Cycles dataset. (c) Ablation study on DBA-shapes. K is the edge number of each explanation subgraph. 'Or' is the OrphicX model, and 'Dy' is our DyGNNExplainer. 'w/o. $\mathcal{L}_d$', 'w/o. $\mathcal{L}_c$', and 'w/o. VGAE' are DyGNNExplainer without dynamic loss, contrastive loss, and VGAE, respectively.

- Sparsity
  - DyGNNExplainer outperforms OrphicX with fewer edges in the subgraphs.

Conclusion

DyGNNExplainer has addressed the critical challenges associated with interpretability in Dynamic Graph Neural Networks :

- Pioneering the development of DyGNN explanation
- Generating synthetic dynamic datasets tailored for dynamic graph interpretability tasks
- Demonstrating the superior performance of DyGNNExplainer in both explanation tasks and real predictions

kesenzhao2-c@my.cityu.edu.hk

香港城市大學
City University of Hong Kong