CAUSALITY-INSPIRED SPATIAL-TEMPORAL EXPLANATIONS FOR DYNAMIC GRAPH NEURAL NETWORKS

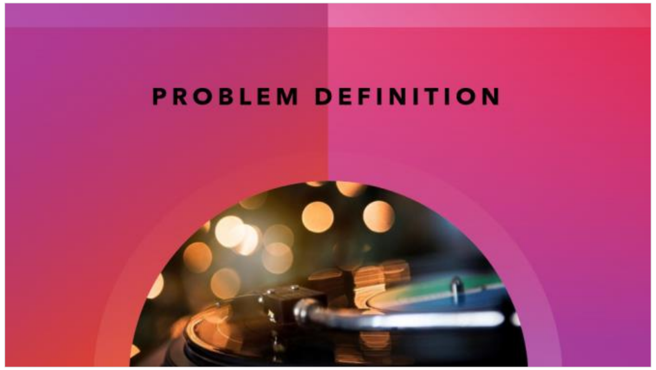
Hooman Zolfaghari

AGENDA

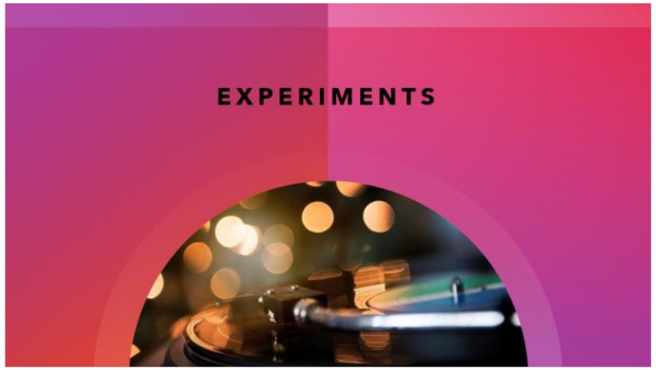
- Introduction
- Problem Definition
- Proposed Method
- Experiments
- Related Work
- Conclusion

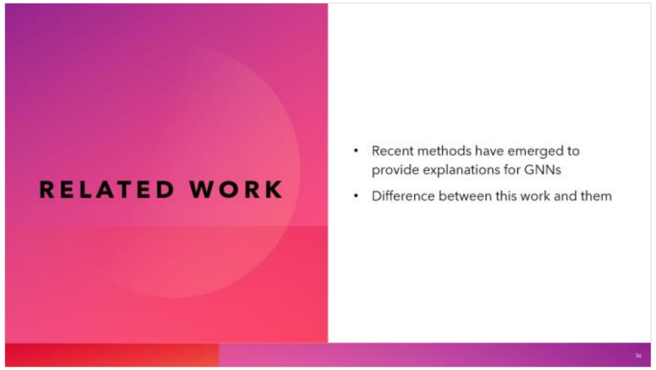
ZOOM

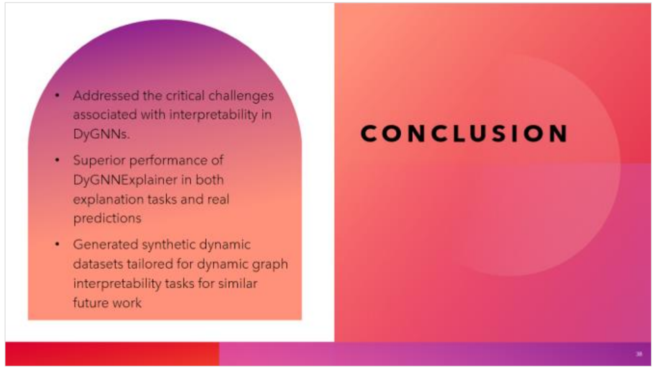

INTRODUCTION


PROBLEM DEFINITION


Framework Overview
A Causal View On DyGNNs
Backdoor Adjustments
Disentangling Complex Causal Relationships
PROPOSED METHOD
Causality-Inspired Spatial-Temporal DyGNN Explainer


EXPERIMENTS


RELATED WORK
Recent methods have emerged to provide explanations for GNNs
Difference between this work and them


Addressed the critical challenges associated with interpretability in DyGNNs.
Superior performance of DyGNNExplainer in both explanation tasks and real predictions
Generated synthetic dynamic datasets tailored for dynamic graph interpretability tasks for similar future work
CONCLUSION

INTRODUCTION

PROBLEM DEFINITION

OVERVIEW OF THE TARGET MODEL

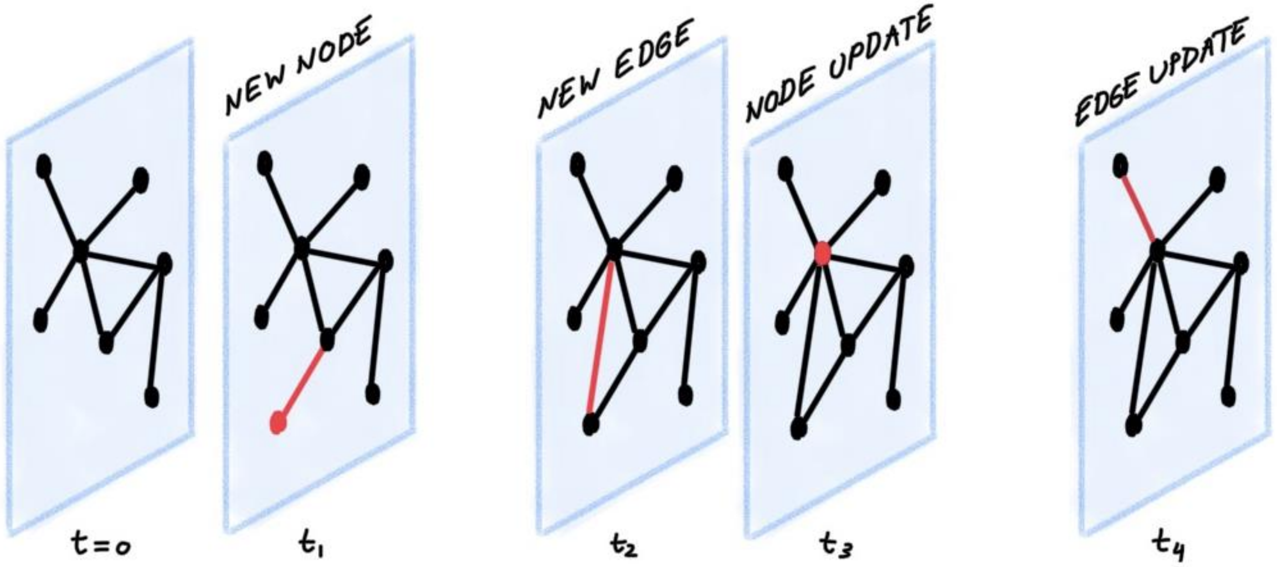
Dynamic Graph Neural Network (DyGNN)

- $f = f_d \circ f_a$
- $f_a: \mathcal{G}_{1:T} \rightarrow \mathcal{R}$
  - **Aggregation Function:** Captures temporal structures and feature patterns
  - **Input:** Dynamic graphs $\mathcal{G}_{1:T}$ over T time steps.
  - **Output:** High-dimensional graph representation in $\mathcal{R}$
- Downstream Task: $f_d: \mathcal{R} \rightarrow \mathcal{Y}$
  - **Downstream Task Function:** Transforms graph representation to label space.
  - **Output:** Final label prediction in $\mathcal{Y}$

OVERVIEW OF THE TARGET MODEL
(CONT.)

Dynamic Graph Structure

- Input at Each Time Step t: $G_t = (X_t, A_t)$
- Node Attribute Matrix: $X_t \in \mathbb{R}^{|V| \times D}$
  - $|V|$: Number of nodes.
  - D: Dimension of node attributes.
- Adjacency Matrix:$A_t \in \mathbb{R}^{|V| \times |V}$



OBJECTIVE

Explanation Objective Critical Criteria for DyGNNs

1. Fidelity
   - Explanations (dynamic subgraphs) should accurately reflect the model’s behavior around the input graphs.
   - Explanatory subgraphs, when fed back into the model, should produce similar predictions as the original dynamic graphs.
2. Interpretability
   - Explanations should highlight the most important parts of the input while ignoring irrelevant components.
   - Spatial Interpretability: Identifies key subgraphs within each graph
   - Temporal Interpretability: Identifies key time steps contributing to the prediction

OBJECTIVE (CONT.)

Model-Agnostic Explainer. Generative Model $\mathcal{F}$

- Identifies aspects of the input that contribute to DyGNN's predictions.
- Meets both fidelity and interpretability criteria.
- Model-Agnostic: Can explain any black-box DyGNN without needing access to its internal mechanisms or ground-truth labels.
- Focus: providing spatial-temporal explanations (dynamic subgraph set) for dynamic graph structures.

- Framework Overview
- A Causal View On DyGNNs
- Backdoor Adjustments
- Disentangling Complex Causal Relationships

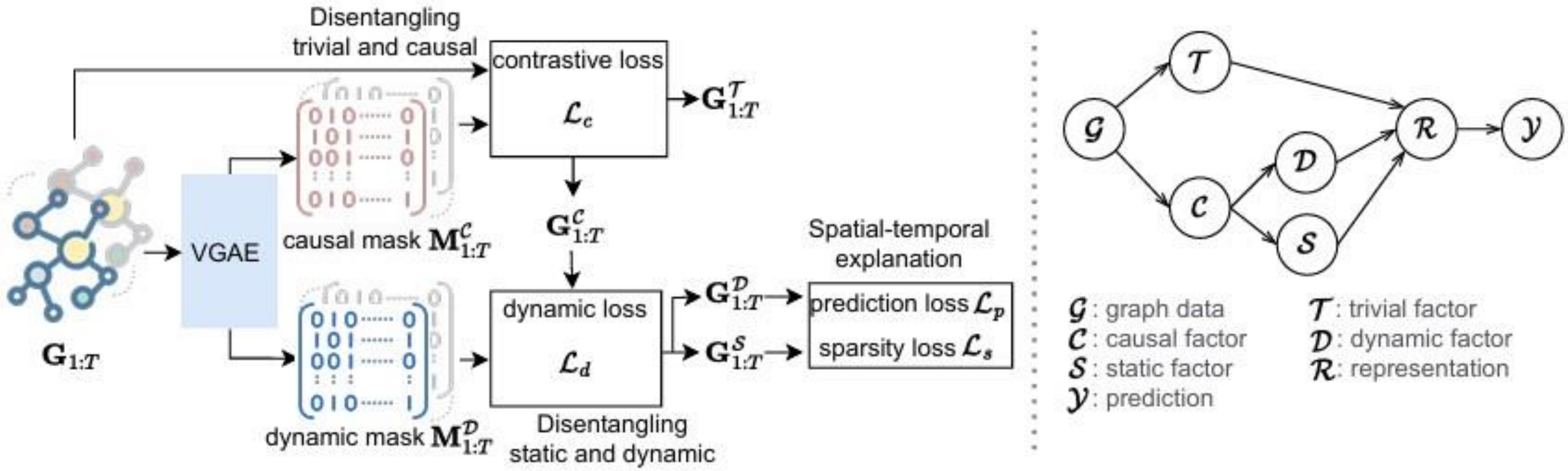
PROPOSED METHOD

Causality-Inspired Spatial-Temporal DyGNN Explainer

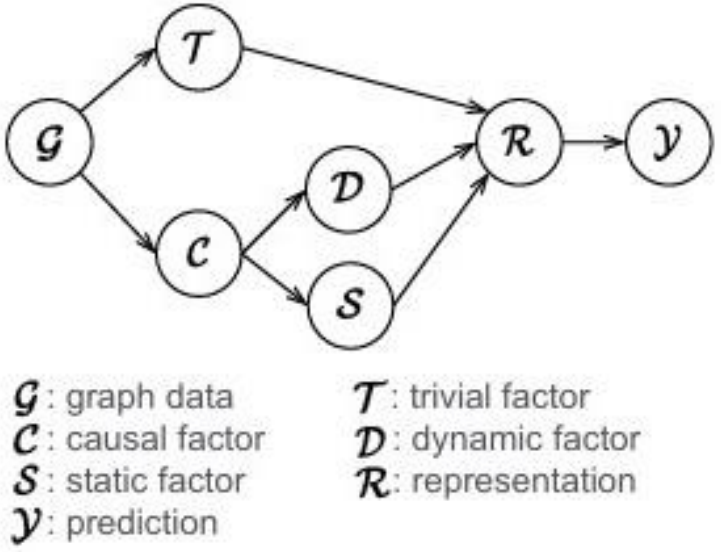
FRAMEWORK OVERVIEW

1. Construct a Structural Causal Model (SCM) , enabling a comprehensive understanding of dynamic graphs
2. generate causal and dynamic soft masks to enable backdoor adjustment, to intervene in the targeting causal and dynamic factors
3. Use:
   - Contrastive Loss: Separates trivial and causal relationships
   - Dynamic Loss: Disentangles static and dynamic relationships
4. Use Prediction & Sparsity Loss: Enhance prediction accuracy and interpretability

FRAMEWORK OVERVIEW



CAUSAL VIEW



- $\mathcal{T} \leftarrow \mathcal{G} \rightarrow \mathcal{C}$:
  - $\mathcal{C}$: Genuine causal relationships
  - $\mathcal{T}$: Trivial relationships (data biases/spurious patterns)
- $\mathcal{T} \rightarrow \mathcal{R} \leftarrow \mathcal{C}$
  - $\mathcal{R}$: high−dimensional representation of dynamic graph node data.
  - Uses both $\mathcal{T}$ and $\mathcal{C}$ to extract discriminative information.
- $\mathcal{D} \rightarrow \mathcal{R} \leftarrow \mathcal{S}$: Causal relationships consist of dynamic relationship and static relationship
- $\mathcal{R} \rightarrow \mathcal{Y}$: Ultimate aim of dynamic graph representation learning to predict graph properties

CAUSAL VIEW

Backdoor Paths from the SCM:

- Path 1: $\mathcal{C} \leftarrow \mathcal{G} \rightarrow \mathcal{T} \rightarrow \mathcal{R} \rightarrow \mathcal{Y}$
  - Issue: T acts as a confounder between G and Y
  - Consequence: Misleading correlation between C and Y leading to incorrect predictions
- Path 2: $\mathcal{D} \leftarrow \mathcal{C} \rightarrow \mathcal{S} \rightarrow \mathcal{R} \rightarrow \mathcal{Y}$
  - Issue: S acts as a confounder between C and Y
- Thus:
  - It is important to <u>block backdoor paths</u> to ensure DyGNNs <u>utilize genuine causal relationships</u> without interference from confounders

BACKDOOR ADJUSTMENT

- Safeguard DyGNNs against confounding factors and distinguish between dynamic and static relationships
- Focus: representation learning that eliminates the backdoor path
- Do-Calculus on $\mathcal{T}$:
    - Estimate $P(\mathcal{Y}|do(\mathcal{C}))$ without interference from $\mathcal{T}$
- Do-Calculus on $\mathcal{D}$:
    - Estimate $P(\mathcal{Y}|do(\mathcal{D}))$ eliminating backdoor from $\mathcal{S}$

BACKDOOR ADJUSTMENT

Challenge:

- Can't directly employ the standard backdoor adjustment method due to confounder $\mathcal{T}$
- Merge the estimation of $P(\mathcal{Y}|do(\mathcal{C}))$ with that of $P(\mathcal{Y}|do(\mathcal{D}))$:

$$P(\mathcal{Y}|do(\mathcal{D})) = \sum P(\mathcal{Y}|do(\mathcal{C}))P(\mathcal{S}) = \sum P(\mathcal{S}) \sum P(\mathcal{Y}|\mathcal{G})P(\mathcal{T})$$

- No explicit information is available for the identification of trivial, dynamic, and static relationships.

DISENTANGLING COMPLEX CAUSAL RELATIONSHIPS

- Casual soft masks for Casual relationship at t^{th} time step as $M_t^{\mathcal{C}} \in \mathbb{R}^{|V| \times |V|}$.
- Each element represents an attention score typically in $[0,1]$.
- Complementary Mask: $\bar{M} = \mathbf{1} - M$
- Partition dynamic graph set as:
    - Casual set $G_{1:T}^{\mathcal{C}} = (X_{1:T}, A_{1:T} \oplus M_{1:T}^{\mathcal{C}})$
    - Trivial set $G_{1:T}^{\mathcal{T}} = (X_{1:T}, A_{1:T} \oplus \bar{M}_{1:T}^{\mathcal{C}})$
        - Elementwise dot product : $\oplus$

DISENTANGLING COMPLEX CAUSAL RELATIONSHIPS

Disentangling Complex Causal Relationships

- Similarly, denote the dynamic soft masks as $M_t^{\mathcal{D}} \in \mathbb{R}^{|V|\times|V|}$, to extract dynamic relationships and its complementary to extract the static relationships.
- Dynamic Casual set $G_{1:T}^{\mathcal{D}} = (X_{1:T}, A_{1:T} \oplus M_{1:T}^{\mathcal{C}} \oplus M_{1:T}^{\mathcal{D}})$
- Static Casual set $G_{1:T}^{\mathcal{S}} = (X_{1:T}, A_{1:T} \oplus M_{1:T}^{\mathcal{C}} \oplus \bar{M}_{1:T}^{\mathcal{D}})$
- The ground-truth trivial set, dynamic causal set, and static causal set are unavailable in real world applications.
- So, we aim to capture the trivial, dynamic, and static relationships from the full graph by learning the masks.

DISENTANGLING COMPLEX CAUSAL RELATIONSHIPS

Estimating soft mask

- dynamic VGAE-based encoder-decoder to estimate the soft masks of explainable subgraphs.
- At the t-th time step, the causal soft mask matrix can be calculated as

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

- Latent representation calculation

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

- Generating explainable subgraphs:

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

DISENTANGLING COMPLEX CAUSAL RELATIONSHIPS

- static relationship S can also be treated as the co-founder between D and Y, just like the trivial relationship T respect to C and Y.
- The same VGAE-based encoder-decoder framework, different parameters Θ_D:

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

- Now we have adjacency matrix of Casual Set, Dynamic Causal Set, Static Causal Set
- We need to disentangle the trivial, dynamic, and static relationships

DISENTANGLING COMPLEX CAUSAL RELATIONSHIPS

Disentangling trivial and causal

- Objective: Ensure Explanation Fidelity
  - Causal Subgraph Set: Target for explanations, representing essential information.
  - Trivial Subgraph Set: Treated as noise, serving as negative examples.
- Approach:
  - Fidelity Criterion: Explanations with causal subgraphs should mimic the original graph's behavior.
  - Negative Treatment: Trivial subgraphs should not influence the model's predictions.

DISENTANGLING COMPLEX CAUSAL RELATIONSHIPS

Disentangling trivial and causal

- Methodology: Embedding Extraction:
- Utilize the pre-trained aggregation function from the DyGNN.
- During each time step t:
  - f_a would generate the embeddings via extracting the essential information until time t, for:
  - Original Graph Set $G_{1:t}$, Causal Subgraph Set $G_{1:t}^{\mathcal{C}}$, Trivial Subgraph Set $G_{1:t}^{\mathcal{T}}$.

$$\mathbf{R}_t = f_a(\mathbf{G}_{1:t}), \mathbf{R}_t^{\mathcal{C}} = f_a(\mathbf{G}_{1:t}^{\mathcal{C}}), \mathbf{R}_t^{\mathcal{T}} = f_a(\mathbf{G}_{1:t}^{\mathcal{T}})$$

  - The above outputs can be the node (graph) embedding for downstream tasks.

DISENTANGLING COMPLEX CAUSAL RELATIONSHIPS

Disentangling trivial and causal

- Utilize contrastive learning
- Ensure the semantic similarity between the causal embedding $\boldsymbol{e}_t^{\mathcal{C}}$ and the original embedding $\boldsymbol{e}_t$
- Enlarging the semantic distance between the causal embedding $\boldsymbol{e}_t^{\mathcal{C}}$ and the trivial embedding $\boldsymbol{e}_t^{\mathcal{T}}$.

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

DISENTANGLING COMPLEX CAUSAL RELATIONSHIPS

Disentangling static and dynamic

- Extract the dynamic relationship and static relationship
- From the dynamic causal set $G_{1:T}^{D}$ and static causal set $G_{1:T}^{S}$
- Utilize GCN with learn-able parameters Ψ_D and Ψ_S.

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

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

DISENTANGLING COMPLEX CAUSAL RELATIONSHIPS

Disentangling static and dynamic

- we can use the pre-trained aggregation function $f_a(\cdot)$ again, which extracts the dynamic relationship from the original graph set.
- Fidelity ⇒ Generated dynamic graph set should guarantee that f_a can extract the dynamic relationship from it
- Dynamic loss:

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

DISENTANGLING COMPLEX CAUSAL RELATIONSHIPS

Spatial-temporal explanation

- Due to the highly temporal correlation for dynamic relationships, it would be difficult to disentangle the dynamic relationship.
- Treat the dynamic relationship at time t as an invention
- Define the causal effect at time t as follows:

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

DISENTANGLING COMPLEX CAUSAL RELATIONSHIPS

Spatial-temporal explanation

Combine the causal effect for the dynamic relationship and the static relationship at time t as the key causal information for Gt

Propose the learn-able weight pooling method to aggregate all the information across all time slots as follows

$$\boldsymbol{H}_T = \sum_{t=1}^{T} t_p(\Delta \boldsymbol{H}_t^{\mathcal{D}} \oplus \boldsymbol{H}_t^{\mathcal{S}}) \Delta \boldsymbol{H}_t^{\mathcal{D}} \oplus \boldsymbol{H}_t^{\mathcal{S}}, \quad t_p(\mathbf{H}) = Softmax(\Psi_{\mathcal{P}} \mathbf{H} / \|\Psi_{\mathcal{P}}\|)$$

DISENTANGLING COMPLEX CAUSAL RELATIONSHIPS

Spatial-temporal explanation

Use aggregated embedding to explain the ground-truth label via prediction loss

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

Take the sparsity requirement for both the causal graph set and the dynamic causal graph set via the sparsity loss

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

DISENTANGLING COMPLEX CAUSAL RELATIONSHIPS

Summary:

learn the optimal explainable causal subgraphs, dynamic subgraphs, and temporal importance by solving the following optimization problems

$$\min_{\Theta, \Psi} \mathcal{L}(\Theta, \Psi) = \lambda_1 \mathcal{L}_c + \lambda_2 \mathcal{L}_s + \lambda_3 \mathcal{L}_p + \lambda_4 \mathcal{L}_d$$

EXPERIMENTS

EXPERIMENTS

- Use 4 synthetic datasets and 2 real-world datasets
- Node classification task and Graph classification task

Dataset	Node classification				Graph classification	
	DBA-Shapes	DTree-Cycles	DTree-Grid	Elliptic	DBA-2motifs	MemeTracker
#nodes	700	871	1,231	203,769	25,000	3.3 mil.
#edges	4,110	1,950	3410	234,355	51,392	27.6 mil.
#labels	7	3	3	2	3	2

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***

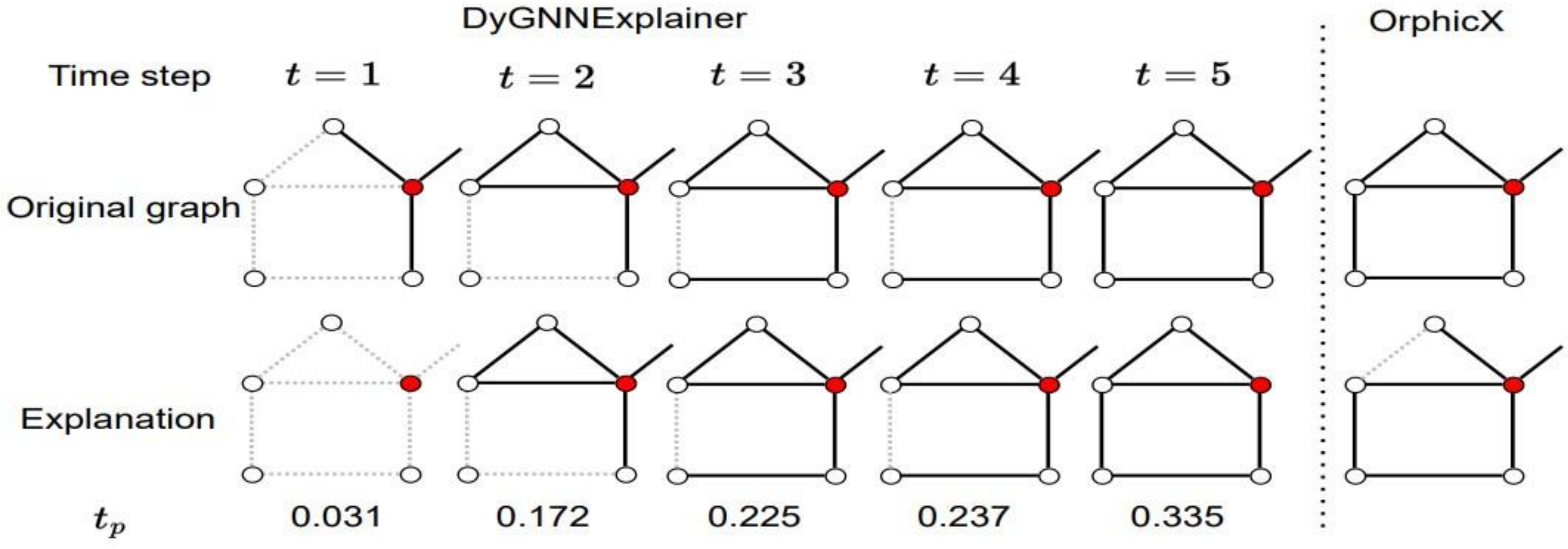
'*' indicates the statistically significant improvements (i.e., two-sided t-test with $p < 0.05$) over the best baseline. 'cls.' is short for classification.

EXPERIMENTS

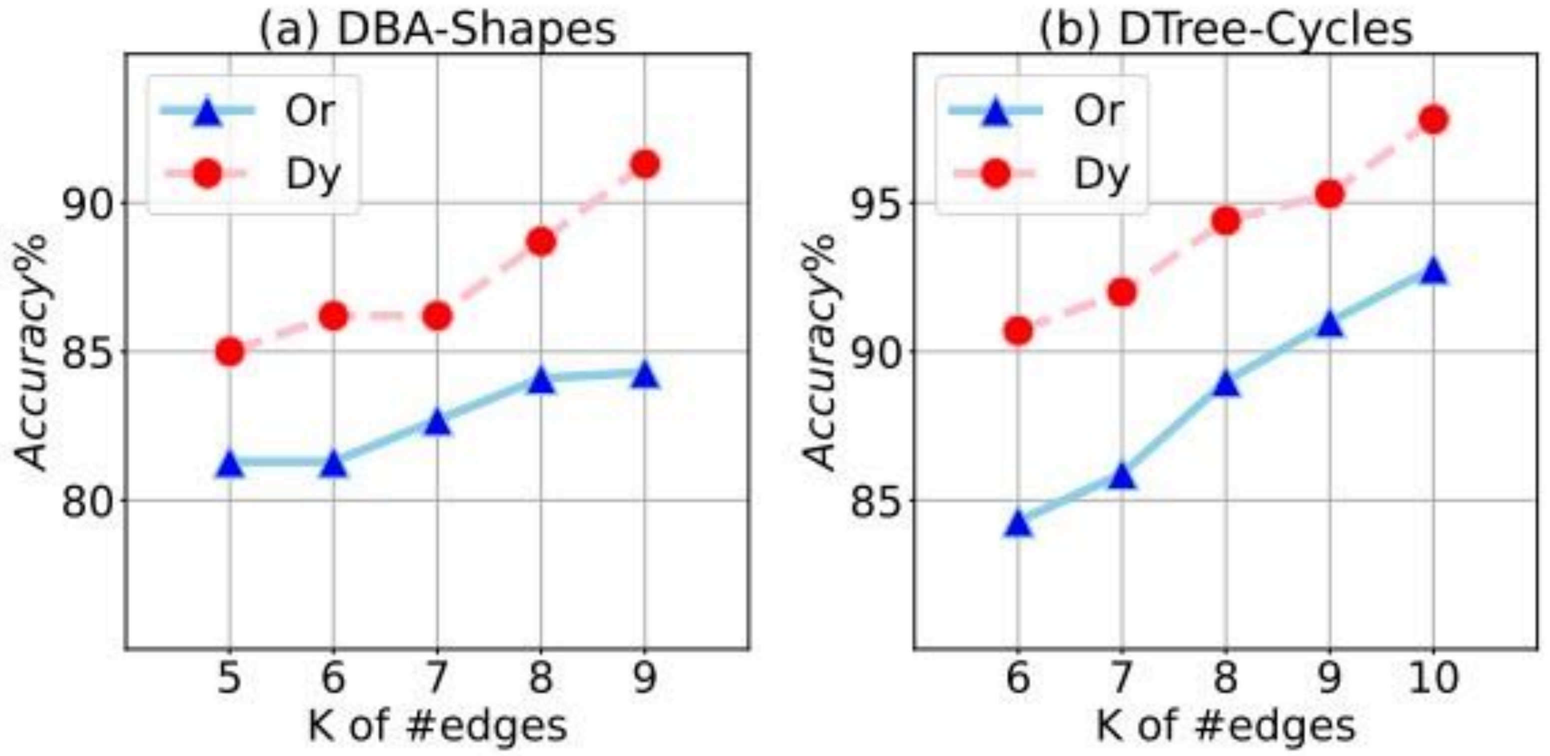
- Explanation fidelity: We compare the predicted labels of explanatory subgraphs with the predicted labels of the original graphs as generated by the target model.
- Explanation interpretability analysis: The explanation subgraphs should exhibit a high degree of sparsity. Measure the number of subgraph edges

EXPERIMENTS

Case Study



EXPERIMENTS



EXPERIMENTS

- Prediction accuracy analysis

Table 3: Prediction accuracy of different models (%). Where best performances are bold.

Dataset	GNNExplainer	PGExplainer	Gem	OrphicX	Target	DyGNNExplainer
DBA-Shapes	35.5	36.3	38.5	38.7	40.2	**44.6***
Elliptic	39.7	45.6	43.5	47.8	84.3	**89.2***

*" indicates the statistically significant improvements (i.e., two-sided t-test with $p < 0.05$) over the best baseline.

RELATED WORK

- Recent methods have emerged to provide explanations for GNNs
- Difference between this work and them

RELATED WORK

- Methods predominantly aim to generate input-dependent explanations
- GNNExplainer (Ying et al., 2019) seeks soft masks for edges and node features through mask optimization to explain predictions
- Typically explain each instance individually and lack the ability to generalize graphs
- PGExplainer (Luo et al., 2020) proposes learning a mask predictor for edge masks to provide explanations.
- XGNN (Yuan et al., 2020) focuses on investigating graph patterns leading to specific classes
- In contrast to these approaches, this work leverages **causality** to achieve faithful explanations

CONCLUSION

- Addressed the critical challenges associated with interpretability in DyGNNs.
- Superior performance of DyGNNExplainer in both explanation tasks and real predictions
- Generated synthetic dynamic datasets tailored for dynamic graph interpretability tasks for similar future work

THANK YOU

Hooman Zolfaghari