Graph Out-of-Distribution Generalization via Causal Intervention

The Web Conference (WWW), 2024

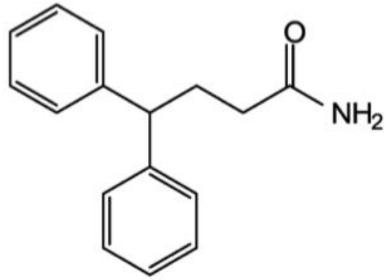
Qitian Wu, Fan Nie, Chenxiao Yang, Tianyi Bao, Junchi Yan
Shanghai Jiao Tong University

Paper: https://arxiv.org/pdf/2402.11494
Code: https://github.com/fannie1208/CaNet

Background: Graph-Structured Data

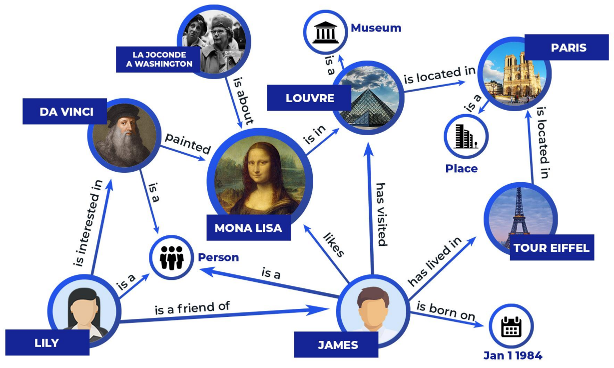
❑ **Graph-structured data are ubiquitous in various domains**

molecular

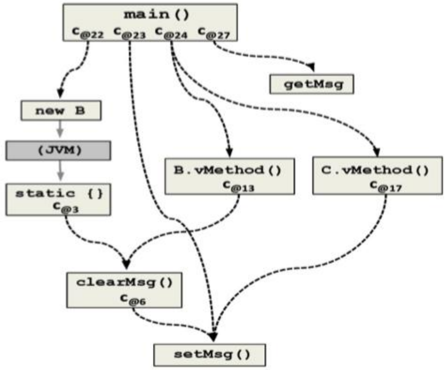


social network

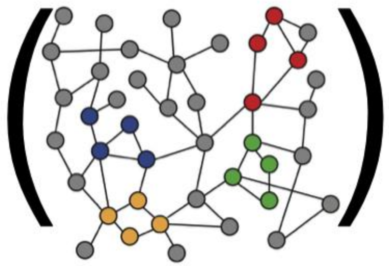
knowledge graph



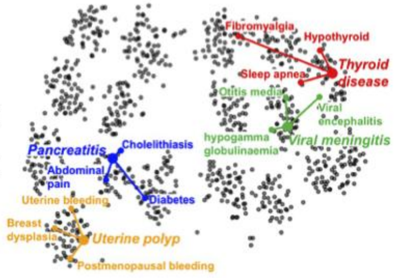
code



❑ **Graph representation learning:** find a functional map that converts nodes in a graph into embeddings in latent space

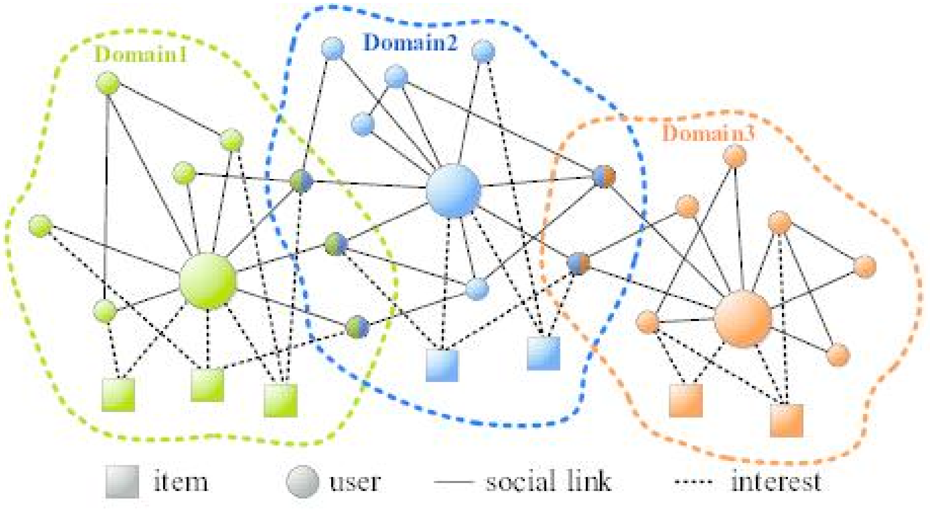
f() =

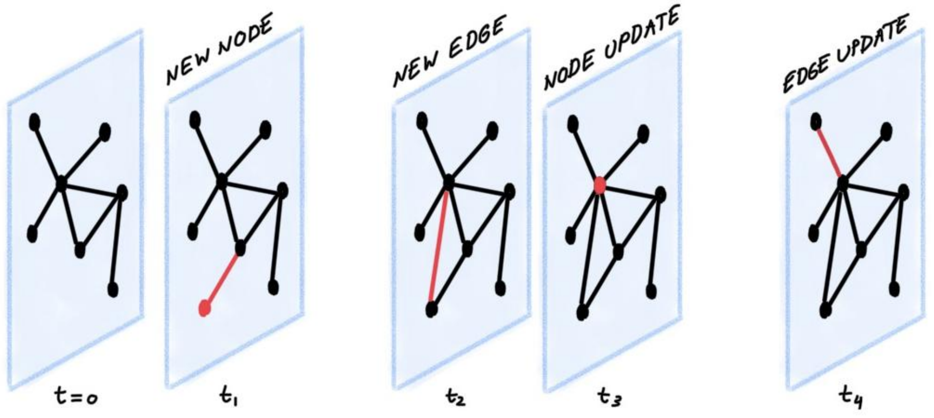
Input graph



2D node embeddings

Distribution Shifts on Graph Data



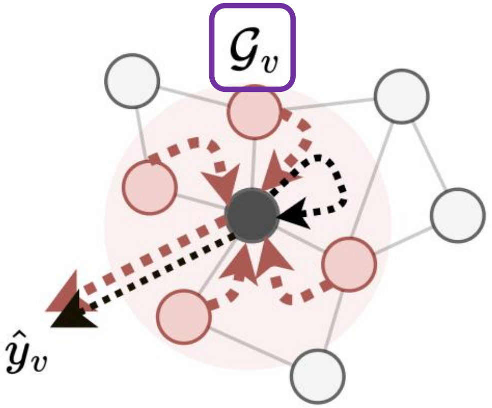


Graph data from multiple domains **Dynamic temporal networks**

- **Distribution shifts cause different data distributions** $P_{train}(\mathcal{D}) \neq P_{test}(\mathcal{D})$
- **Challenges:**
  - New data from unknown distribution are unseen by training
  - Distribution shifts involve structural information of non-Euclidean space

The Impact of Distribution Shifts

what is processed by graph neural networks as inputs



ego-graph feature $\mathcal{G}_v$		*node label* y_v	*environment* e_v
a user's friends are young a user's friends like sports	✓ (spurious) ✓ (causal)	the user likes playing basketball	university
a user's friends are young a user's friends like sports	✗ (spurious) ✓ (causal)	the user likes playing basketball	Linkedin

✓ *positive correlation* ✗ *no correlation*

┈┈▶ *spurious correlation: only hold in a few environments*

──▶ *causal relation: universally hold in all environments*

Observation: spurious correlation that only holds in training data is harmful for generalization, but the causal relation that universally hold is beneficial for generalization

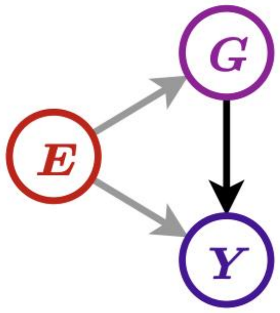
Out-of-Distribution Generalization

- **Graph notation:** a graph $G = (A, X)$, adjacency matrix $A = \{a_{uv} | v, u \in V\}$ node features $X = \{x_v | v \in V\}$, node labels $Y = \{y_v | v \in V\}$

$$p(G, Y|E) = p(G|E)p(Y|G, E)$$

where E denotes environment (that affects data generation)

- **Observation:** environment is a latent confounder in data generation



Distribution shifts cause varying environments from training to testing

$$p(G, Y|E = e_{tr}) \neq p(G, Y|E = e_{te})$$

- ✓ Social networks collected from different regions (environment)
- ✓ Citation networks formed at different times (environment)
- ✓ Protein interaction networks of different species (environment)

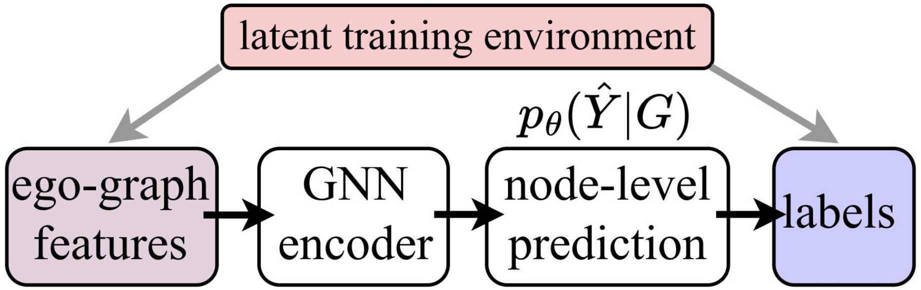
Causal Analysis of Graph Neural Networks

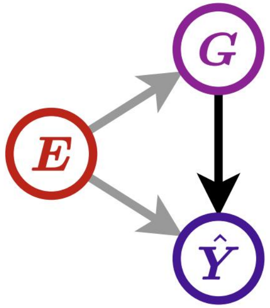
- **Graph neural networks (GNN) for node-level prediction:**

$$\mathbf{z}_v^{(1)} = \phi_{in}(\mathbf{x}_v) \qquad \mathbf{z}_v^{(l+1)} = \sigma\left(\mathrm{Conv}^{(l)}\left(\{\mathbf{z}_u^{(l)} | u \in \mathcal{N}_v \cup \{v\}\}\right)\right) \qquad \hat{\mathbf{y}}_v = \phi_{out}(\mathbf{z}_v^{(L+1)})$$

- **Maximum Likelihood Estimation (MLE) yields trained model parameters:**

$$\theta^* = \arg\min_{\theta} -\frac{1}{|\mathcal{V}_{tr}|} \sum_{v \in \mathcal{V}_{tr}} \mathbf{y}_v^\top \log f_\theta(\mathcal{G}_v)$$



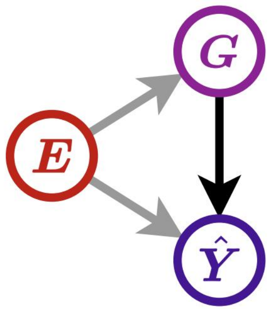


- $G \rightarrow \hat{Y}$: by predictive distribution of $p_\theta(\hat{Y}|G)$ GNN model $\hat{\mathbf{y}}_v = f_\theta(\mathcal{G}_v)$
- $E \rightarrow G$: by definition of data generation $p(G|E)$
- $E \rightarrow \hat{Y}$: by training process of Maximum Likelihood Estimation

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{e \sim p_{tr}(E), (\mathcal{G}_v, \mathbf{y}_v) \sim p(G, Y | E=e)}[-\mathbf{y}_v^\top \log f_\theta(\mathcal{G}_v)]$$

Causal Intervention via Backdoor Adjustment

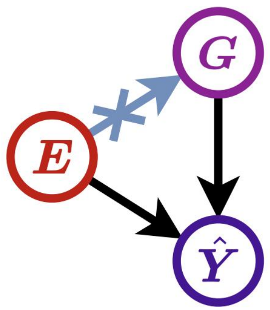
- Harmful effect: the confounding bias of latent environment



- E establishes a shortcut (spurious correlation) between G and Y
- Model training tends to exploit spurious correlation in training data

($\mathcal{G}_v$ "a user's friends are young" to y_v "the user likes playing basketball")

- Potential solution: cutting off the dependence between E and G



Key idea: replace $p_\theta(\hat{Y}|G)$ with $p_\theta(\hat{Y}|do(G))$

- According to Backdoor Adjustment in causal inference [Pearl et al., 2016]:

$$p_\theta(\hat{Y}|do(G)) = \mathbb{E}_{\boxed{p_0(E)}}[p_\theta(\hat{Y}|G,E)]$$

a model-free prior for E

Causal Intervention with Env. Inference

original intervention objective

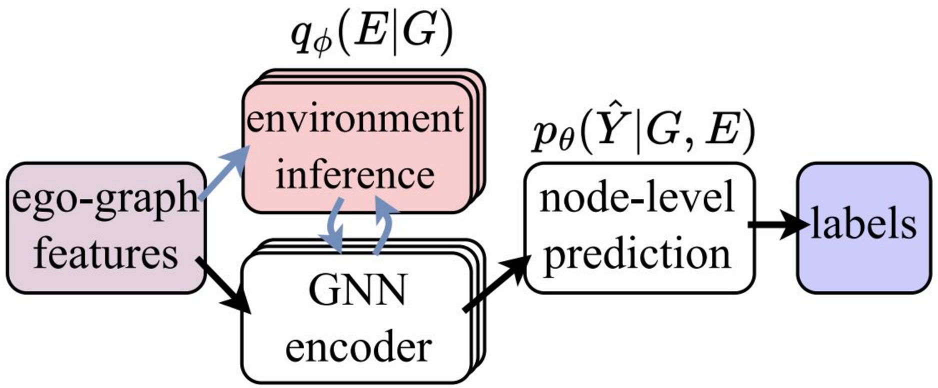
$$\log \sum_e p_\theta(\hat{Y}|G, E=e)P(E=e)$$

$$= \log \sum_e p_\theta(\hat{Y}|G, E=e)p_0(E=e)\frac{q_\phi(E=e|G)}{q_\phi(E=e|G)}$$

$$\geq \sum_e q_\phi(E=e|G)\log p_\theta(\hat{Y}|G, E=e)p_0(E=e)\frac{1}{q_\phi(E=e|G)}$$

variational lower bound of the objective

$$= \sum_e q_\phi(E=e|G)\log p_\theta(\hat{Y}|G, E=e) - \sum_e q_\phi(E=e|G)\log\frac{q_\phi(E=e|G)}{p_0(E=e)}$$



Model instantiation:

- $q_\phi(E|G)$: pseudo environment estimator
- $p_\theta(\hat{Y}|G, E)$: GNN predictor conditioned on E
- $p_0(E)$: a trivial prior distribution

Model Architecture Design

- **Pseudo Environment Estimator** $q_\phi(E|G)$

$$\boldsymbol{\pi}_v^{(l)} = \mathrm{Softmax}(\mathbf{W}_S^{(l)}\mathbf{z}_v^{(l)}) \qquad e_{vk}^{(l)} = \frac{\exp\left(\left(\pi_{vk}^{(l)} + g_k\right)/\tau\right)}{\sum_k \exp((\pi_{vk}^{(l)} + g_k)/\tau)}, \quad g_k \sim \mathrm{Gumbel}(0,1)$$

model env. as a latent discrete variable at each layer

use Gumbel reparameterization trick for enabling differentiable sampling

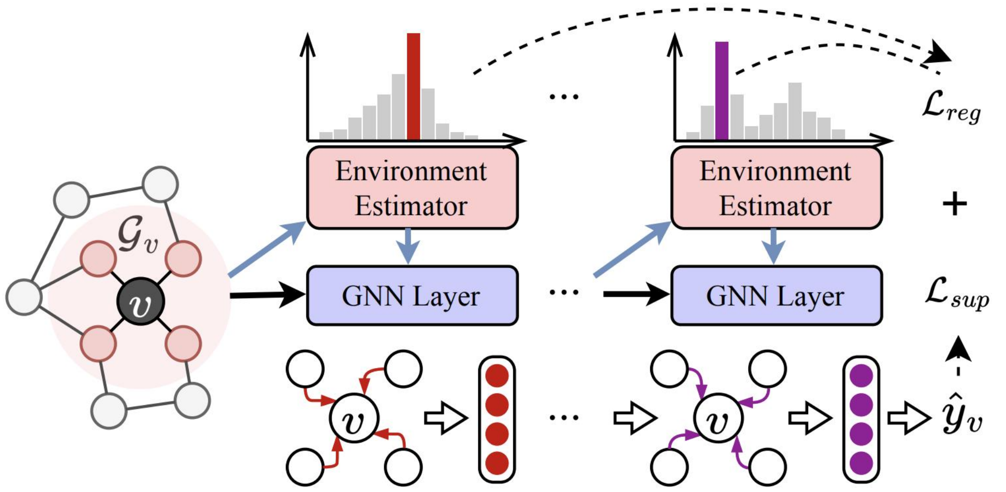
- **Mixture-of-expert GNN Predictor** $p_\theta(\hat{Y}|G,E)$
  - CaNet-GCN: use graph convolution unit

$$\mathbf{z}_u^{(l+1)} = \sigma\left(\sum_{k=1}^{K} e_{u,k}^{(l)} \sum_{v, a_{uv}=1} \frac{1}{\sqrt{d_u d_v}} \mathbf{W}_D^{(l,k)}\mathbf{z}_v^{(l)} + \mathbf{W}_S^{(l,k)}\mathbf{z}_u^{(l)}\right)$$

  - CaNet-GAT: use graph attention unit

$$\mathbf{z}_u^{(l+1)} = \sigma\left(\sum_{k=1}^{K} e_{u,k}^{(l)} \sum_{v, a_{uv}=1} w_{uv}^{(l,k)} \mathbf{W}_D^{(l,k)}\mathbf{z}_v^{(l)} + \mathbf{W}_S^{(l,k)}\mathbf{z}_u^{(l)}\right)$$

$$w_{uv}^{(l,k)} = \frac{\mathrm{LeakyReLU}((\mathbf{b}^{(l,k)})^\top[\mathbf{W}_A^{(l,k)}\mathbf{z}_u^{(l)} \| \mathbf{W}_A^{(l,k)}\mathbf{z}_v^{(l)}])}{\sum_{w=1}^{N} \mathrm{LeakyReLU}(\mathbf{b}^{(l,k)})^\top[\mathbf{W}_A^{(l,k)}\mathbf{z}_u^{(l)} \| \mathbf{W}_A^{(l,k)}\mathbf{z}_w^{(l)}])}$$

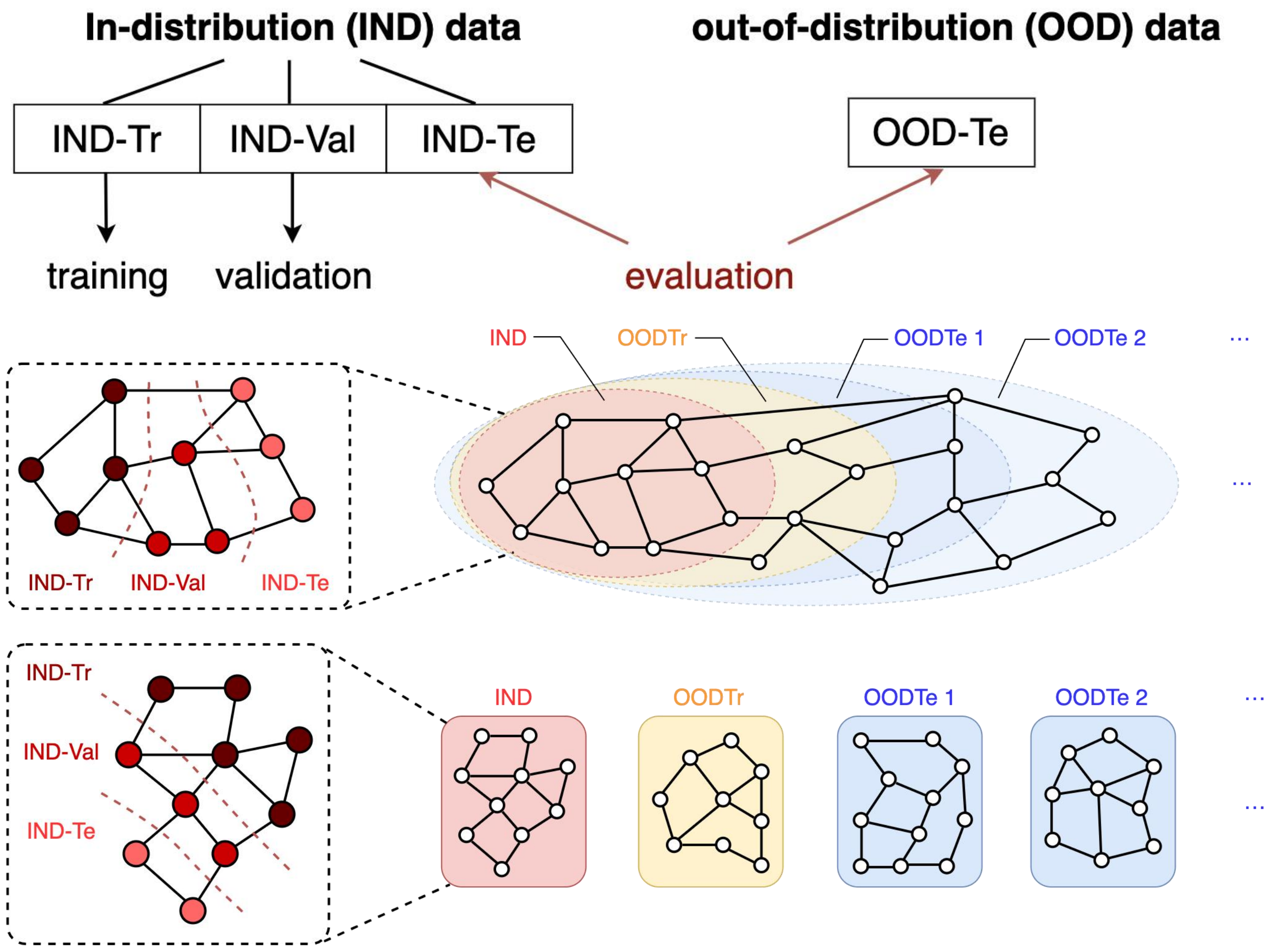


Experiment Protocols

- Split data into in-distribution and out-of-distribution portions; for IND data, randomly split into IND-Tr/IND-Val/IND-Te
- For temporal graph dataset: use time information for data split of IND and OOD
- For multi-graph dataset: use domain information for data split of IND and OOD

Qitian Wu, et al., Handling Distribution Shifts on Graphs: An Invariance Perspective, ICLR 2022

Qitian Wu, et al., Energy-based Out-of-Distribution Detection for Graph Neural Networks, ICLR 2023

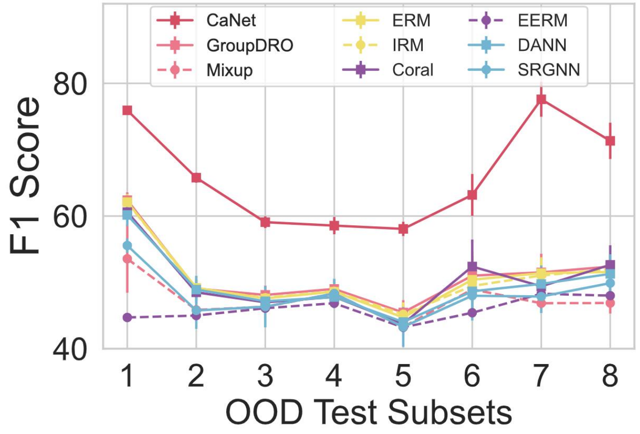


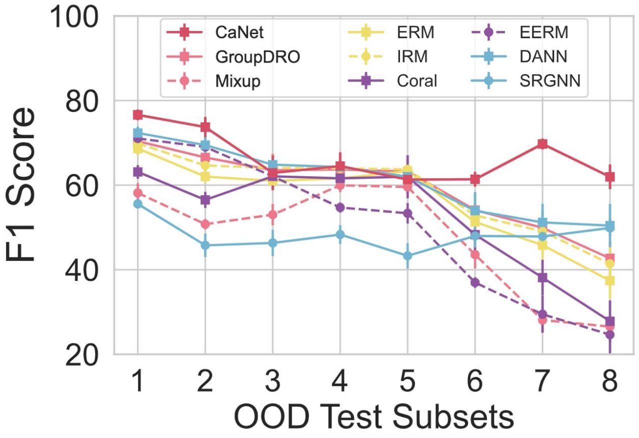
Experiment Results

Testing results (Accuracy for *Arxiv*, ROC-AUC for *Twitch*) on real-world datasets

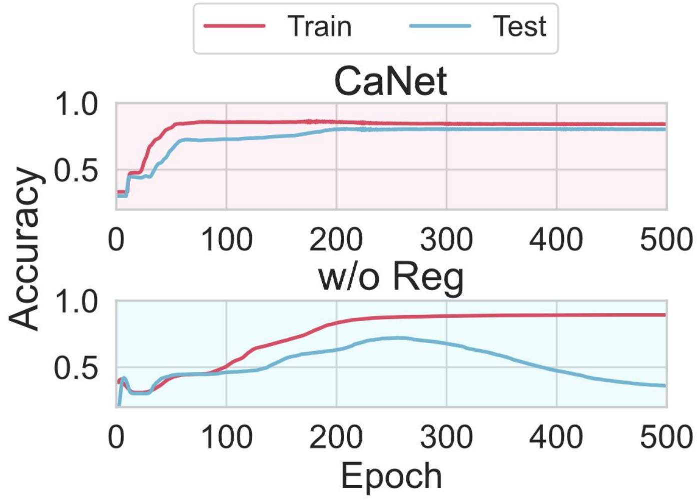
Method	Arxiv OOD 1	Arxiv OOD 2	Arxiv OOD 3	Arxiv ID	Twitch OOD 1	Twitch OOD 2	Twitch OOD 3	Twitch ID
ERM	56.33 ± 0.17	53.53 ± 0.44	45.83 ± 0.47	59.94 ± 0.45	66.07 ± 0.14	52.62 ± 0.01	63.15 ± 0.08	75.40 ± 0.01
IRM	55.92 ± 0.24	53.25 ± 0.49	45.66 ± 0.83	60.28 ± 0.23	66.95 ± 0.27	52.53 ± 0.02	62.91 ± 0.08	74.88 ± 0.02
Coral	56.42 ± 0.26	53.53 ± 0.54	45.92 ± 0.52	60.16 ± 0.12	66.15 ± 0.14	52.67 ± 0.02	**63.18** ± 0.03	75.40 ± 0.01
DANN	56.35 ± 0.11	53.81 ± 0.33	45.89 ± 0.37	60.22 ± 0.29	66.15 ± 0.13	52.66 ± 0.02	**63.20** ± 0.06	75.40 ± 0.02
GroupDRO	56.52 ± 0.27	53.40 ± 0.29	45.76 ± 0.59	60.35 ± 0.27	66.82 ± 0.26	**52.69** ± 0.02	62.95 ± 0.11	75.03 ± 0.01
Mixup	56.67 ± 0.46	**54.02** ± 0.51	46.09 ± 0.58	60.09 ± 0.15	65.76 ± 0.30	**52.78** ± 0.04	63.15 ± 0.08	75.47 ± 0.06
SRGNN	**56.79** ± 1.35	**54.33** ± 1.78	**46.24** ± 1.90	60.02 ± 0.52	65.83 ± 0.45	52.47 ± 0.06	62.74 ± 0.23	75.75 ± 0.09
EERM	OOM	OOM	OOM	OOM	**67.50** ± 0.74	51.88 ± 0.07	62.56 ± 0.02	74.85 ± 0.05
CaNet	**59.01** ± 0.30	**56.88** ± 0.70	**56.27** ± 1.21	61.42 ± 0.10	**67.47** ± 0.32	**53.59** ± 0.19	**64.24** ± 0.18	75.10 ± 0.08

Testing F1 score for *Elliptic* with GCN and GAT as the encoder backbone

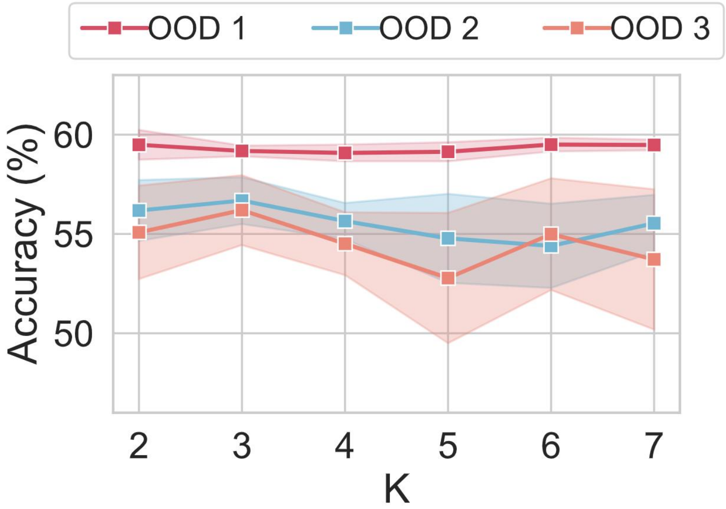




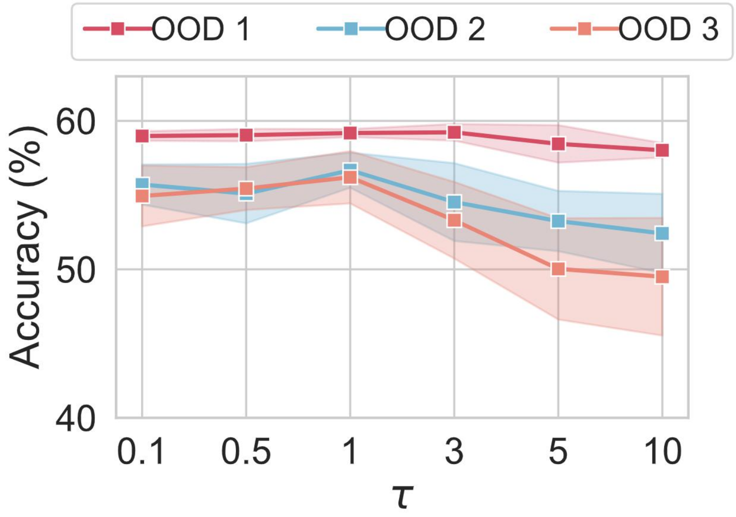
Ablation Study and Hyperparameters



Regularization loss in the new objective is effective for improving generalization



Model performance is stable for proper K (number of pseudo env.)



Small temperature (sharp results) can produce satisfactory performance

Conclusion

Main contributions of our work:

- ☐ We identify that the confounding bias of latent environments in graph data leads to poor generalization on out-of-distribution data
- ☐ We propose a new learning approach resorting to causal intervention and variational inference for improving out-of-distribution generalization
- ☐ We demonstrate the spuriority of the new model on diverse real-world datasets and achieve improvements over state-of-the-arts

Paper: https://arxiv.org/pdf/2402.11494
Code: https://github.com/fannie1208/CaNet

Qitian Wu, et al., Handling Distribution Shifts on Graphs: An Invariance Perspective, ICLR 2022
Qitian Wu, et al., Energy-based Out-of-Distribution Detection for Graph Neural Networks, ICLR 2023