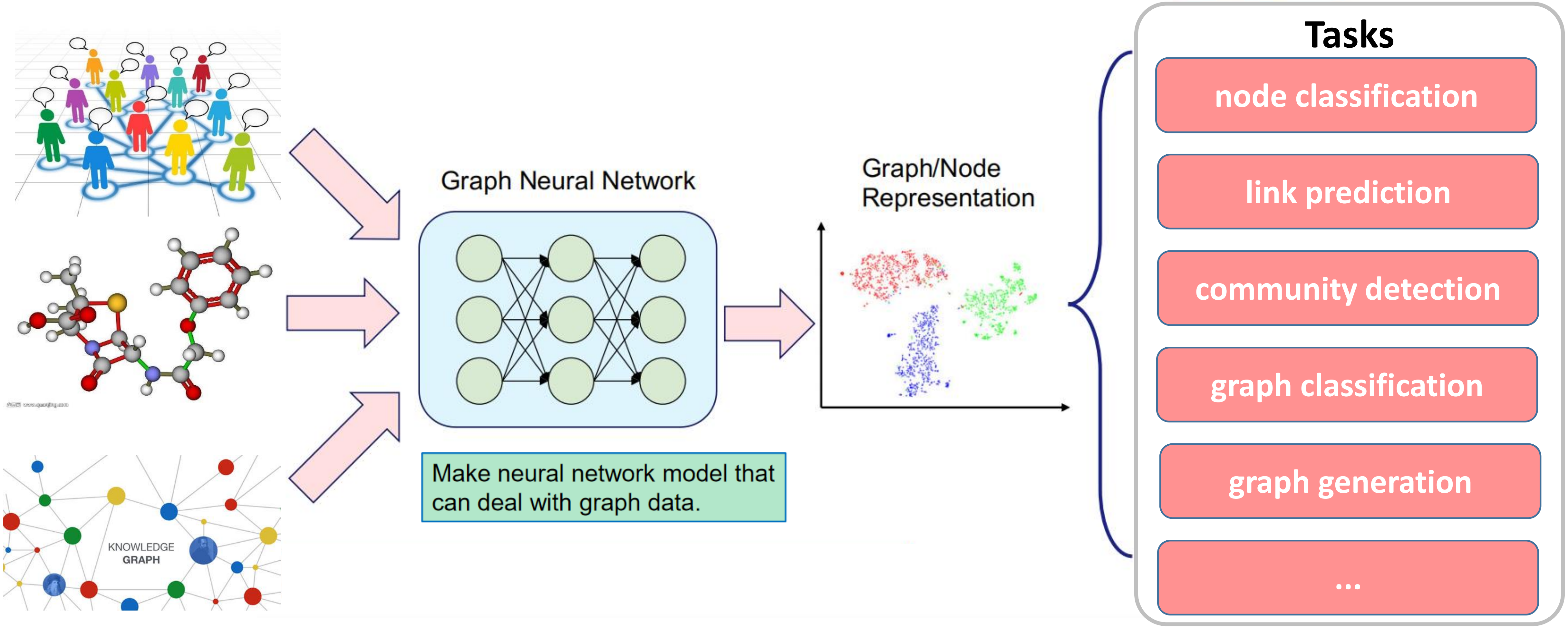
Graph Structure Learning with Variational Information Bottleneck

Qingyun Sun, Jianxin Li, Hao Peng, Jia Wu, Xingcheng Fu, Cheng Ji, Phillip S. Yu

Email: sunqy@act.buaa.edu.cn

Paper: https://arxiv.org/abs/2112.08903

Graph Neural Network



Acknowledgement from https://ai.tencent.com/ailab/ml/WWW-Deep-Graph-Learning.html

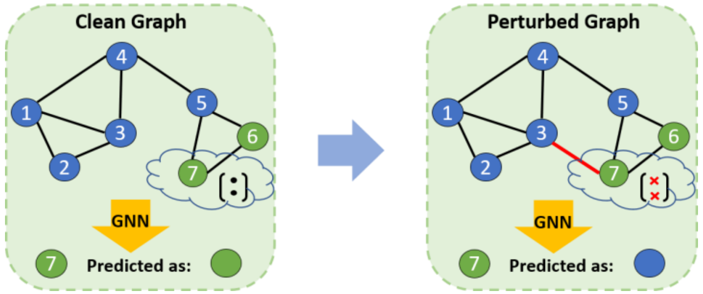
Why Graph Structure Learning?

One fundamental assumption of GNN: the observed topology is ground-truth information and consistent with the properties of GNNs.

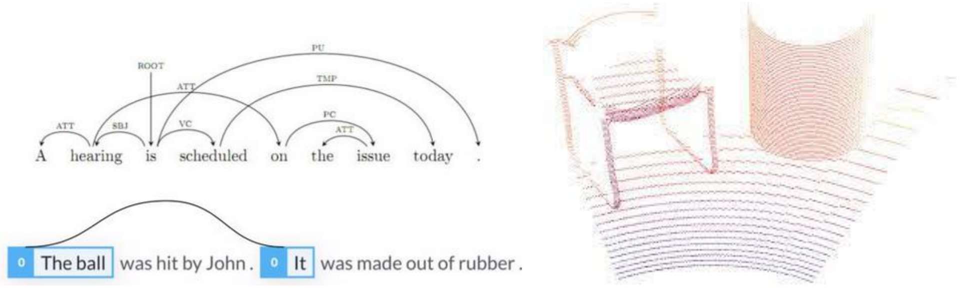
However,

- Questionable if the given intrinsic graph-structures are optimal (i.e., noisy, adversarial perturbation, incomplete) for the downstream tasks
- Many applications (e.g., NLP tasks) may only have non-graph structured data or even just the original feature matrix.

noisy social network



graph adverarial attack

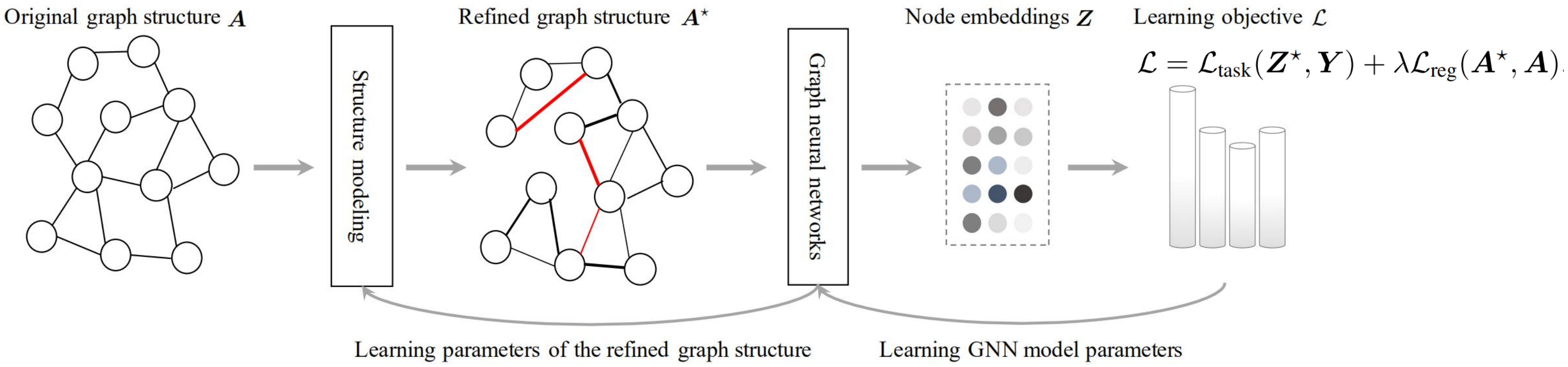


non-graph structure data

Graph Structure Learning

Graph structure learning targets jointly learning an optimized graph structure and corresponding representations to improving the robustness of GNN models.

Input: a raw graph **Output:** a refined graph structure

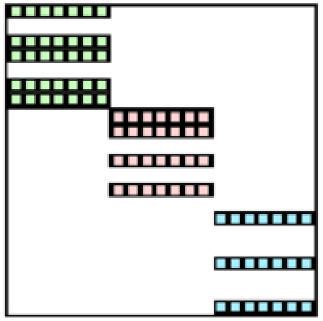


General Paradiam of GSL: Structure Modeling → Message Passing → Learning Objective

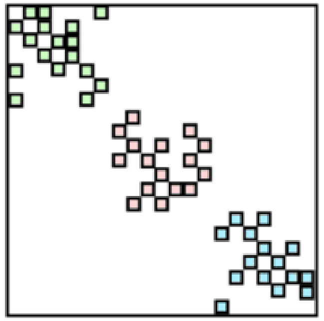
Acknowledgement from https://arxiv.org/abs/2103.03036

What is a “good” structure?

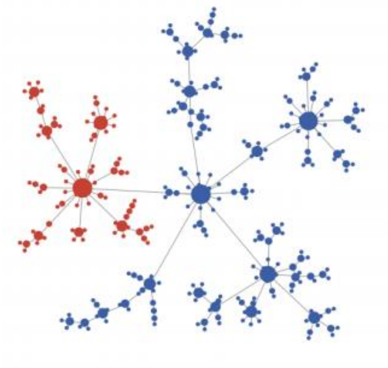
Previous works: basd on **assumptions** (e.g., homophily) or **certain constraints** (low-rank, sparse, conectted, feature-smoothing)



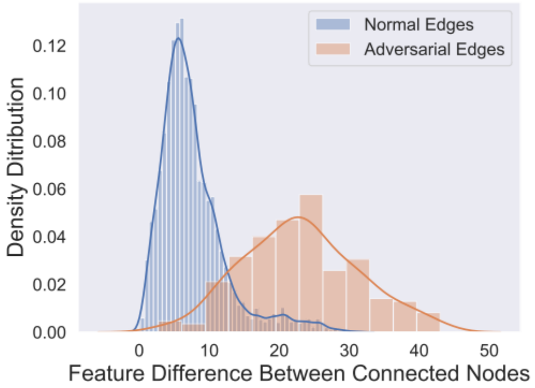
(b) low-rank



(c) sparse



absolute homophily:



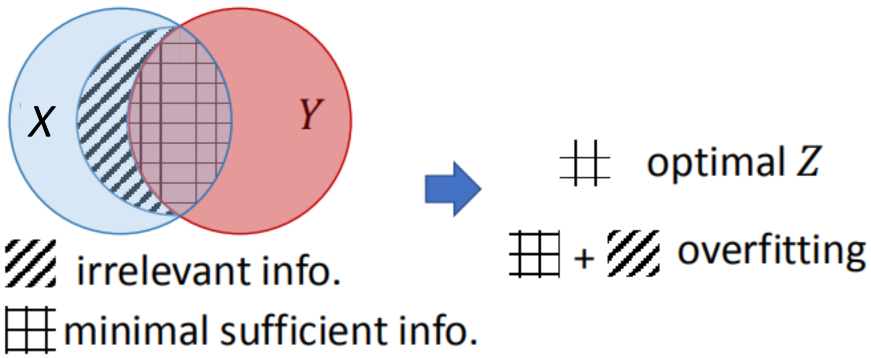
(d) Feature Smoothness

There is still a lack of a general framework that can mine underlying relations from the essence of representation learning

Acknowledgement from https://dl.acm.org/doi/abs/10.1145/3394486.3403049

Information Bottleneck (IB)

Information Bootleneck: an optimal representation **Z** should contain the **minimal** **sufficient** information for the downstream prediction task



X: data *Y: task* *Z: representation*

IB Objective: $\arg\max_{Z} I(Y,Z) - \beta I(X,Z)$

prediction term: make the prediction accurate

compression term: discourages acquiring irrelevant information

Acknowledgement from https://arxiv.org/abs/2010.12811

Our Work: Graph Structure Learning with IB

Advance IB principle for graph structure learning:

We focus on learning an optimal **IB-Graph** for G, which is compressed with minimum information loss in terms of G's properties for the downstream prediction task

Objective:

$$\underset{G_{\mathrm{IB}}}{\operatorname{argmin}} \underbrace{-I(G_{IB};Y)}_{\text{prediction term}} + \underbrace{\beta I(G_{IB};G)}_{\text{compression term}}$$

Proposition 1 (Upper bound of $-I(G_{\mathrm{IB}};Y)$). *For graph $G \in \mathbb{G}$ with label $Y \in \mathbb{Y}$ and IB-Graph G_{IB} learned from G, we have*

$$-I(Y;G_{\mathrm{IB}}) \leq -\iint p(Y,G_{\mathrm{IB}}) \log q_\theta(Y|G_{\mathrm{IB}}) dY dG_{\mathrm{IB}} + H(Y), \tag{5}$$

where $q_\theta(Y|G_{\mathrm{IB}})$ is the variational approximation of the true posterior $p(Y|G_{\mathrm{IB}})$.

Proposition 2 (Upper bound of $I(G_{\mathrm{IB}};G)$). *For graph $G \in \mathcal{G}$ and IB-Graph G_{IB} learned from G, we have*

$$I(G_{\mathrm{IB}};G) \leq \iint p(G_{\mathrm{IB}},G) \log \frac{p(G_{\mathrm{IB}}|G)}{r(G_{\mathrm{IB}})} dG_{\mathrm{IB}} dG, \tag{6}$$

where $r(G_{\mathrm{IB}})$ is the variational approximation to the prior distribution $p(G_{\mathrm{IB}})$ of G_{IB}.

Our Work: Graph Structure Learning with IB

Advance IB principle for graph structure learning:

We focus on learning an optimal **IB-Graph** for G, which is compressed with minimum information loss in terms of G's properties for the downstream prediction task

Objective:

$$\underset{G_{\mathrm{IB}}}{\operatorname{argmin}} \underbrace{-I(G_{IB};Y)}_{\text{prediction term}} + \underbrace{\beta I(G_{IB};G)}_{\text{compression term}}$$

$$\begin{aligned} & -I(G_{\mathrm{IB}};Y)+\beta I(G_{\mathrm{IB}};G) \\ \approx & -I(Z_{\mathrm{IB}};Y)+\beta I(Z_{\mathrm{IB}};G) \\ \leq & \frac{1}{N}\sum_{i=1}^{N}\left\{-\log q_\theta(Y_i|Z_{\mathrm{IB}i})+\beta p(Z_{\mathrm{IB}i}|G_i)\log\frac{p(Z_{\mathrm{IB}i}|G_i)}{r(Z_{\mathrm{IB}})}\right\}. \end{aligned} \qquad (8)$$

Graph Structure Learning with Variational Information Bottleneck, AAAI 2022

Our Work: Graph Structure Learning with IB

Step 1: Generate IB-Graph

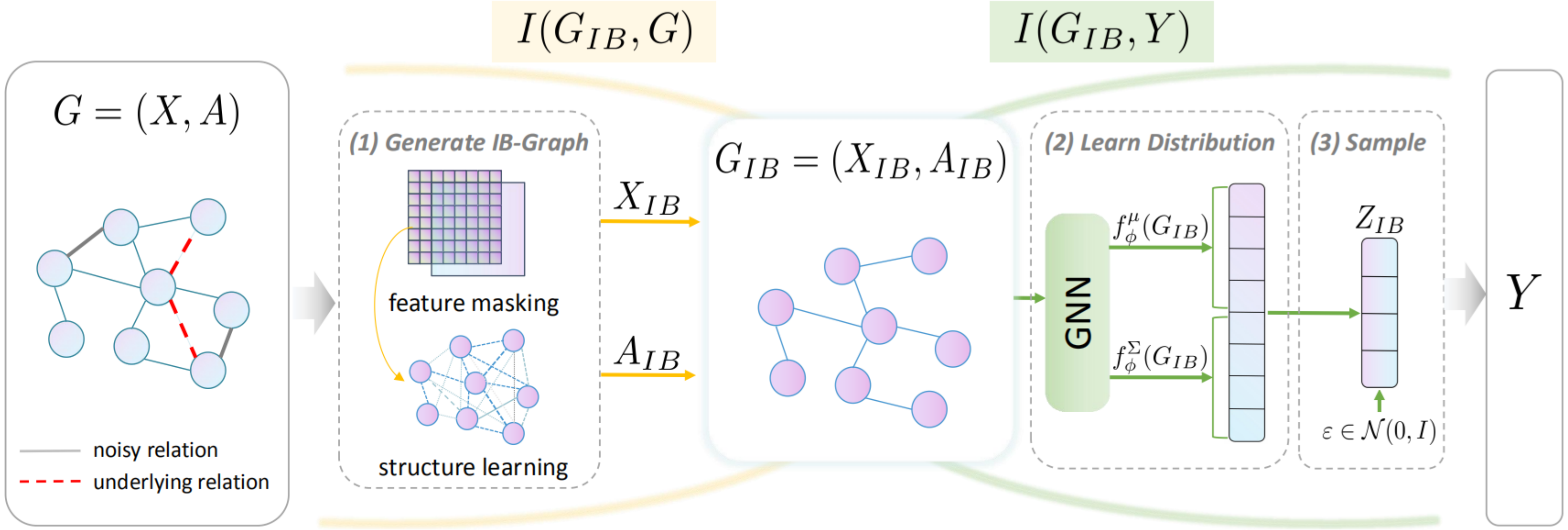
- **Feature Masking**: discretely drop features that are irrelevant to the task
- **Structure Learning**: model all edges as a set of mutually independent Bernoulli random variables parameterized by the learned attention weights

$$X_{\mathrm{IB}} = X_r + (X - X_r) \odot M,$$

$$A_{\mathrm{IB}} = \bigcup_{u,v \in V} \{a_{u,v} \sim \mathrm{Ber}(\pi_{u,v})\}$$

$$Z(u) = \mathbf{NN}(X_{\mathrm{IB}}(u)),$$

$$\pi_{u,v} = \mathrm{sigmoid}\left(Z(u)Z(v)^{\mathrm{T}}\right),$$



$$
\begin{aligned}
& -I(G_{\mathrm{IB}};Y) + \beta I(G_{\mathrm{IB}};G) \\
\approx & -I(Z_{\mathrm{IB}};Y) + \beta I(Z_{\mathrm{IB}};G) \\
\leq & \frac{1}{N}\sum_{i=1}^{N}\left\{-\log q_\theta(Y_i|Z_{\mathrm{IB}i}) + \beta p(Z_{\mathrm{IB}i}|G_i)\log\frac{p(Z_{\mathrm{IB}i}|G_i)}{r(Z_{\mathrm{IB}})}\right\}.
\end{aligned} \tag{8}
$$

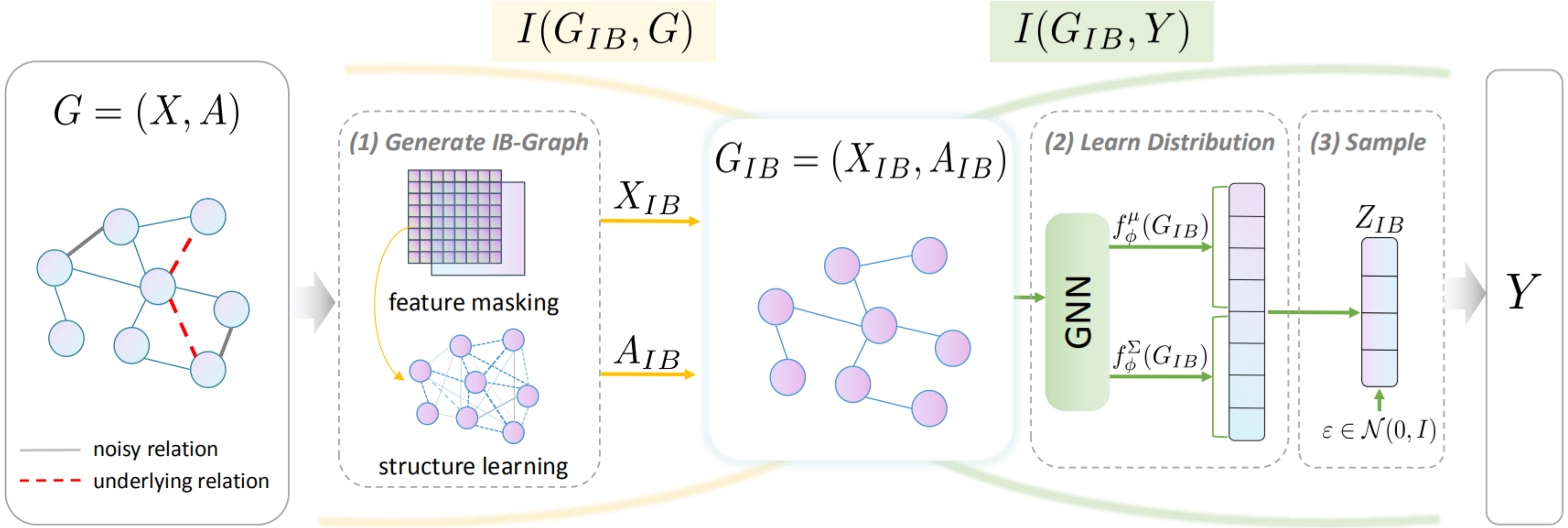
Graph Structure Learning with Variational Information Bottleneck, AAAI 2022

Our Work: Graph Structure Learning with IB

Step 2: Learn Distribution of IB-Graph Representation

- We consider a parametric Gaussian distribution as prior $r(Z_{\mathrm{IB}})$ and $p(Z_{\mathrm{IB}}|G)$i)
- We model the $f_\phi(G_{\mathrm{IB}})$ as a GNN, where $f_\phi^\mu(G_{\mathrm{IB}})$ and $f_\phi^\Sigma(G_{\mathrm{IB}})$ are the 2K-dimensional output value

$$r(Z_{\mathrm{IB}}) = \mathcal{N}(\mu_0, \Sigma_0), \qquad p(Z_{\mathrm{IB}}|G) = \mathcal{N}\left(f_\phi^\mu(G_{\mathrm{IB}}), f_\phi^\Sigma(G_{\mathrm{IB}})\right)$$



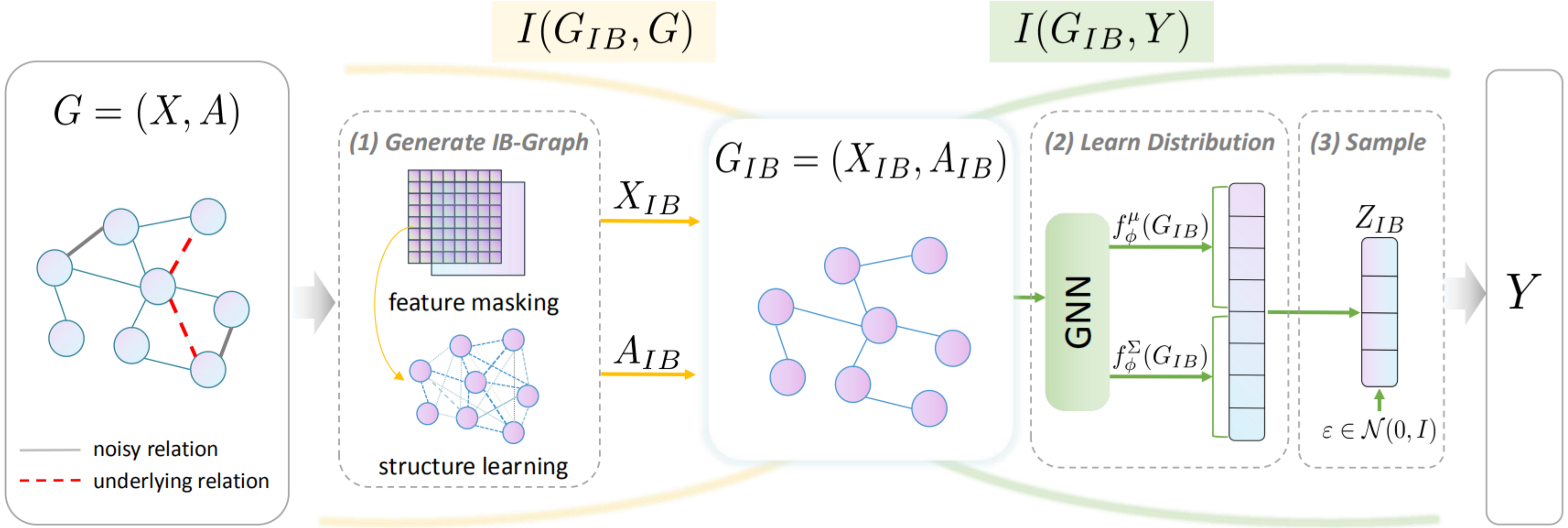
$$\begin{aligned}
&-I(G_{\mathrm{IB}};Y)+\beta I(G_{\mathrm{IB}};G)\\
\approx\ &-I(Z_{\mathrm{IB}};Y)+\beta I(Z_{\mathrm{IB}};G)\\
\leq\ &\frac{1}{N}\sum_{i=1}^{N}\left\{-\log q_\theta(Y_i|Z_{\mathrm{IB}i})+\beta p(Z_{\mathrm{IB}i}|G_i)\log\frac{p(Z_{\mathrm{IB}i}|G_i)}{r(Z_{\mathrm{IB}})}\right\}.
\end{aligned} \tag{8}$$

Our Work: Graph Structure Learning with IB

Step 3: Sample IB-Graph Representation

- We can use the reparameterization trick for gradients estimation

$$Z_{\mathrm{IB}} = f_\phi^\mu(G_{\mathrm{IB}}) + f_\phi^\Sigma(G_{\mathrm{IB}}) \odot \varepsilon,$$



$$\begin{aligned} & -I(G_{\mathrm{IB}};Y) + \beta I(G_{\mathrm{IB}};G) \\ \approx & -I(Z_{\mathrm{IB}};Y) + \beta I(Z_{\mathrm{IB}};G) \\ \leq & \frac{1}{N}\sum_{i=1}^{N}\left\{-\log q_\theta(Y_i|Z_{\mathrm{IB}i}) + \beta p(Z_{\mathrm{IB}i}|G_i)\log\frac{p(Z_{\mathrm{IB}i}|G_i)}{r(Z_{\mathrm{IB}})}\right\}. \end{aligned} \tag{8}$$

Evalauation on Graph Classification Task

- **Datasets:** Four social network datasets
- **Baselines:** Graph structure learners with different GNN backbones

Table 1: Summary of graph classification results: "average accuracy ± standard deviation" and "improvements" (%). Underlined: best performance of specific backbones, **bold**: best results of each dataset.

Structure Learner	Backbone	IMDB-B Accuracy	IMDB-B Δ	IMDB-M Accuracy	IMDB-M Δ	REDDIT-B Accuracy	REDDIT-B Δ	COLLAB Accuracy	COLLAB Δ
N/A	GCN	70.7±3.7	-	49.7±2.1	-	73.6±4.5	-	77.6±2.6	-
	GAT	71.3±3.5	-	50.9±2.7	-	73.1±2.6	-	75.4±2.4	-
	GIN	72.1±3.8	-	49.7±0.4	-	85.4±3.0	-	78.8±1.4	-
NeuralSparse	GCN	72.0±2.6	↑1.3	50.1±3.1	↑0.4	72.1±5.2	↓1.5	76.0±2.0	↓1.6
	GAT	73.4±2.2	↑2.1	53.7±3.1	↑2.8	74.3±3.1	↑1.2	75.4±5.8	0.0
	GIN	73.8±1.6	↑1.7	54.2±5.4	↑4.5	86.2±2.7	↑0.8	76.6±2.1	↓2.2
Subgraph-IB	GCN	72.2±3.9	↑1.5	51.8±3.9	↑2.1	76.7±3.0	↑3.1	76.3±2.3	↓1.3
	GAT	72.9±4.6	↑1.6	51.3±2.4	↑0.4	75.3±4.7	↑2.2	77.3±1.9	↑1.9
	GIN	73.7±7.0	↑1.6	51.6±4.8	↑1.9	85.7±3.5	↑0.3	77.2±2.3	↓1.6
IDGL	GCN	72.2±4.2	↑1.5	52.1±2.4	↑2.4	75.1±1.4	↑1.5	78.1±2.1	↑0.5
	GAT	71.5±4.6	↑0.2	51.8±2.4	↑0.9	76.2±2.5	↑3.1	76.8±4.4	↑1.4
	GIN	74.1±3.2	↑2.0	51.1±2.1	↑1.4	85.7±3.5	↑0.3	76.7±3.8	↓2.1
VIB-GSL	GCN	74.1±3.3	↑3.4	54.3±1.7	↑4.6	77.5±2.4	↑3.9	78.3±1.4	↑0.7
	GAT	75.2±2.7	↑3.9	54.1±2.7	↑3.2	78.1±2.5	↑5.0	79.1±1.2	↑3.7
	GIN	**77.1±1.4**	**↑5.0**	**55.6±2.0**	**↑5.9**	**88.5±1.8**	**↑3.1**	**79.3±2.1**	**↑0.5**

VIB-GSL can learn better graph structure to improve the representation quality

Graph Denosing and Paramete Sensitivity

- How does VIB-GSL perform on graph data with structure noise?
- How does the trade off between prediction and compression influence the performance of VIB-GSL?

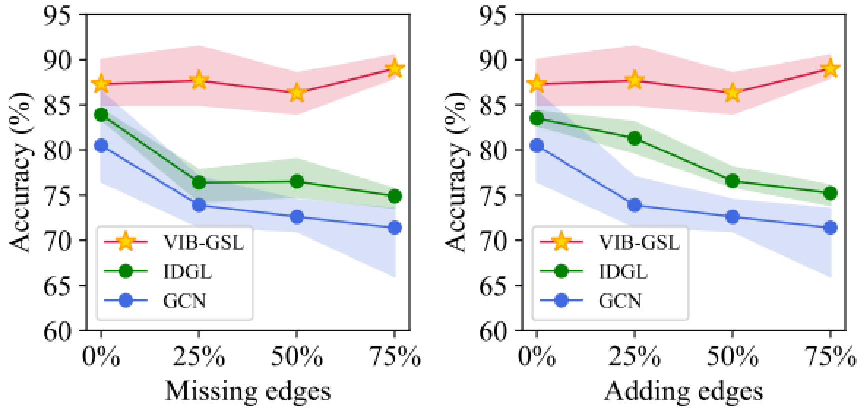


Figure 2: Test accuracy (± standard deviation) in percent for the edge attack scenarios on REDDIT-B (left: edge deletion, right: edge addition).

VIB-GSL is extremely robust to structure perturbations

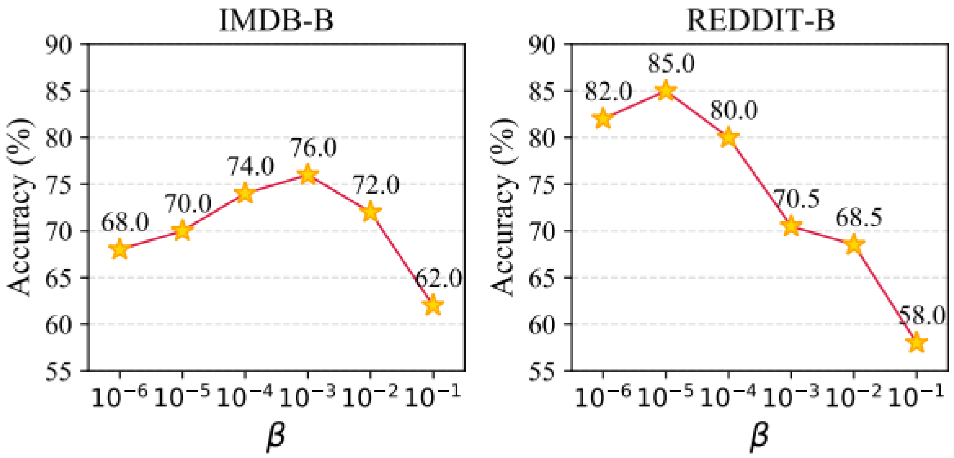


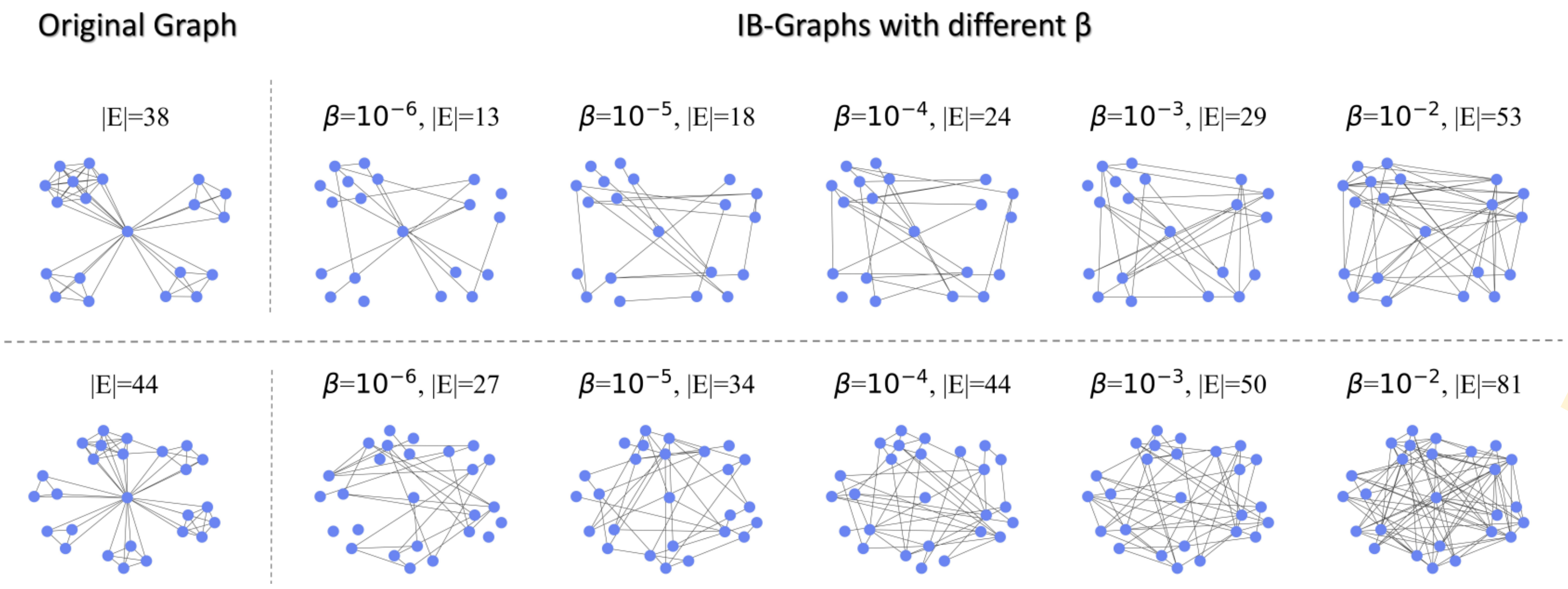
Figure 3: Impact of β on IMDB-B and REDDIT-B.

The accuracies of VIB-GSL variation across different β collapsed onto a hunchback shape

IB-Graph Visualization

How does the trade off between prediction and com-pression influence the learned IB-Graph?

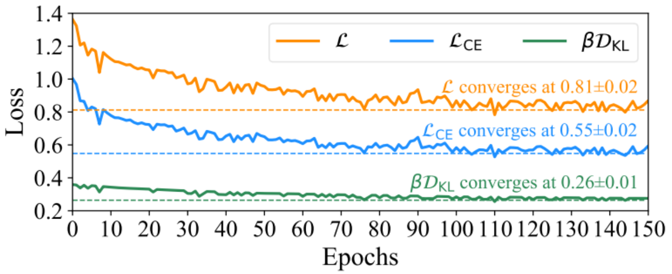
- We show original graph and IB-Graphs with different βwhen VIB-GSL achieves the same testing performance



- VIB-GSL tends to generate edges that connect nodes playing the same structure roles.
- VIB-GSL with larger β will generate a more dense graph structure.

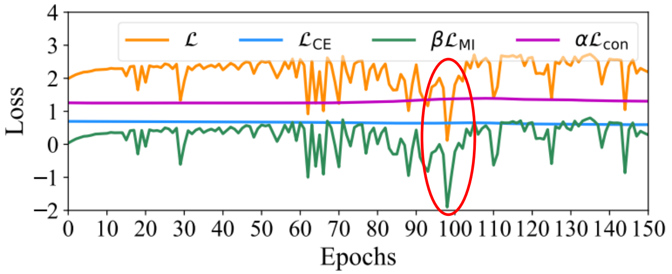
Training stability and efficiency

- **Training stability:** The tractable variational approximation for the IB objective facilitates the training stability
- **Efficiency:** Graph structure learners with different GNN backbones



(a) VIB-GSL.

VIB-GSL deduces a tractable variational approximation for the IB objective, which facilitates the training stability.



(b) Subgraph-IB.

Figure 5: Training dynamics of VIB-GSL and Subgraph-IB.

Subgraph-IB uses a bi-level optimization scheme for MI esitimation, leading to an unstable andinefficient training process

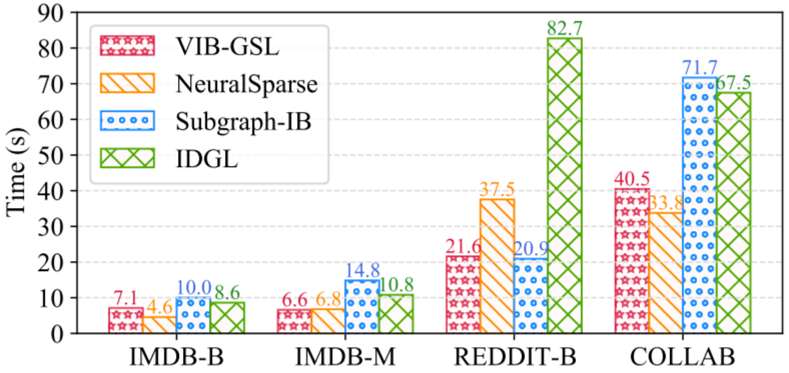


Figure 6: Training time of one epoch on various datasets.

VIB-GSL shows comparable efficiency with other methodswhen achieving the best performance

Conclusion and Future Works

- We advance the Information Bottleneck principle for graph structure learning and propose a framework named VIB-GSL, which jointly optimizes the graph structure and graph representations.
- VIB-GSL deduces a variational approximation to form a tractable IB objectivefunction that facilitates training stability and efficiency.
- Future works: A general, unified and scalable IB guided GSL framework for dfferent graph learning levels.

Graph Structure Learning with Variational Information Bottleneck

Qingyun Sun, Jianxin Li, Hao Peng, Jia Wu, Xingcheng Fu, Cheng Ji, Phillip S. Yu

Email: sunqy@act.buaa.edu.cn

Paper: https://arxiv.org/abs/2112.08903