Dynamic Graph Information Bottleneck

Haonan Yuan
School of Computer Science and Engineering, BDBC,
Beihang University
Beijing, China
yuanhn@buaa.edu.cn

Qingyun Sun*
School of Computer Science and Engineering, BDBC,
Beihang University
Beijing, China
sunqy@buaa.edu.cn

Xingcheng Fu
Key Lab of Education Blockchain and Intelligent Technology,
Guangxi Normal University
Guilin, Guangxi, China
fuxc@gxnu.edu.cn

Cheng Ji
School of Computer Science and Engineering, BDBC,
Beihang University
Beijing, China
jicheng@buaa.edu.cn

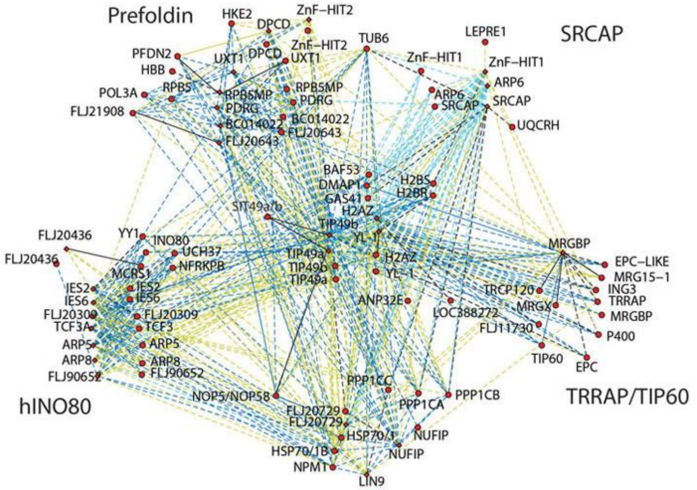
Jianxin Li
School of Computer Science and Engineering, BDBC,
Beihang University
Beijing, China
lijx@buaa.edu.cn

*Corresponding author.

Paper Code Scholar

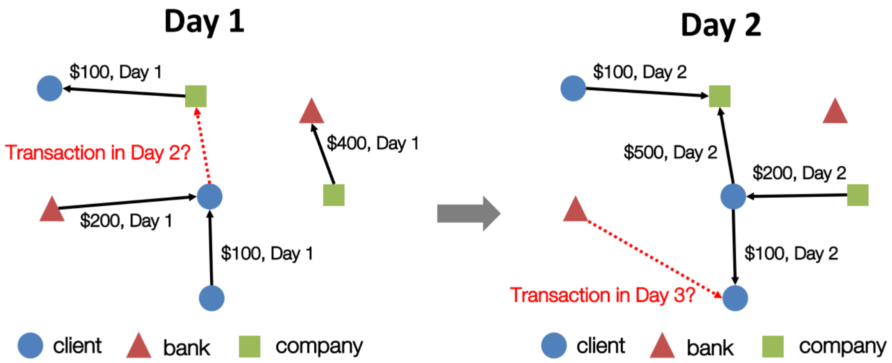
■ (Dynamic) Graphs / Networks are everywhere v.s. temporal graph?

Social Web Network



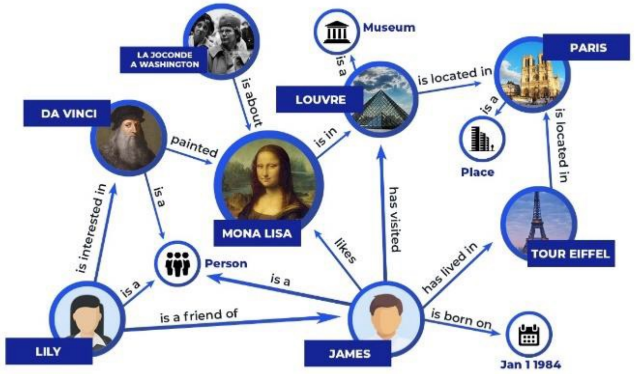
Biology Network

Traffic Network



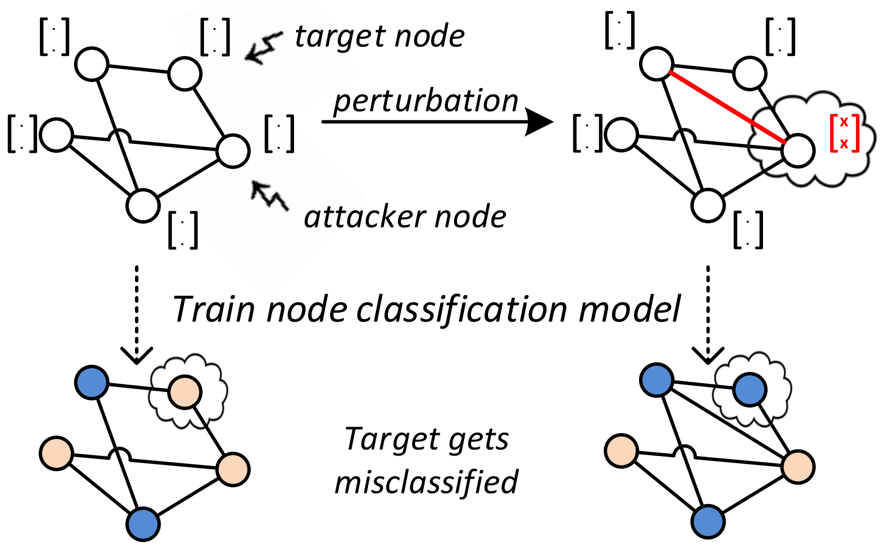
Transaction Network

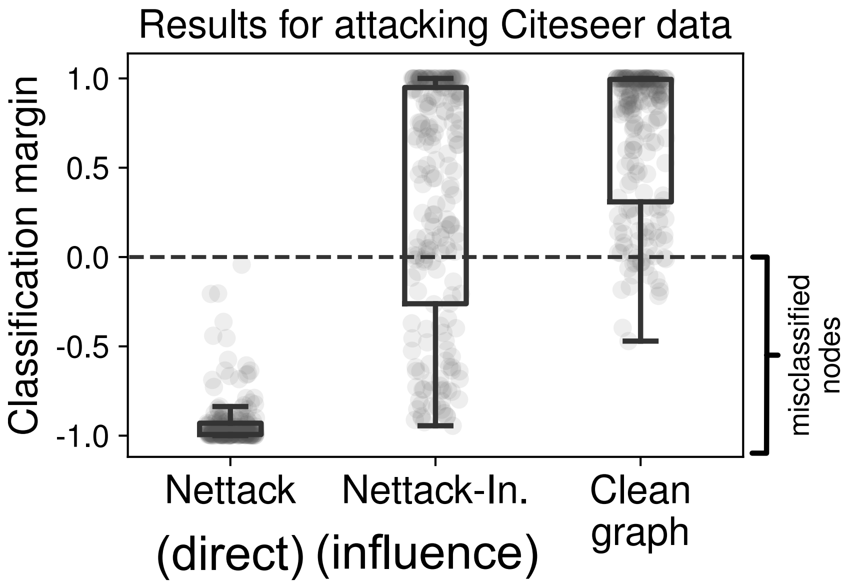
Internet of Things



Knowledge Graph

■ However, GNNs are susceptible to adversarial attacks



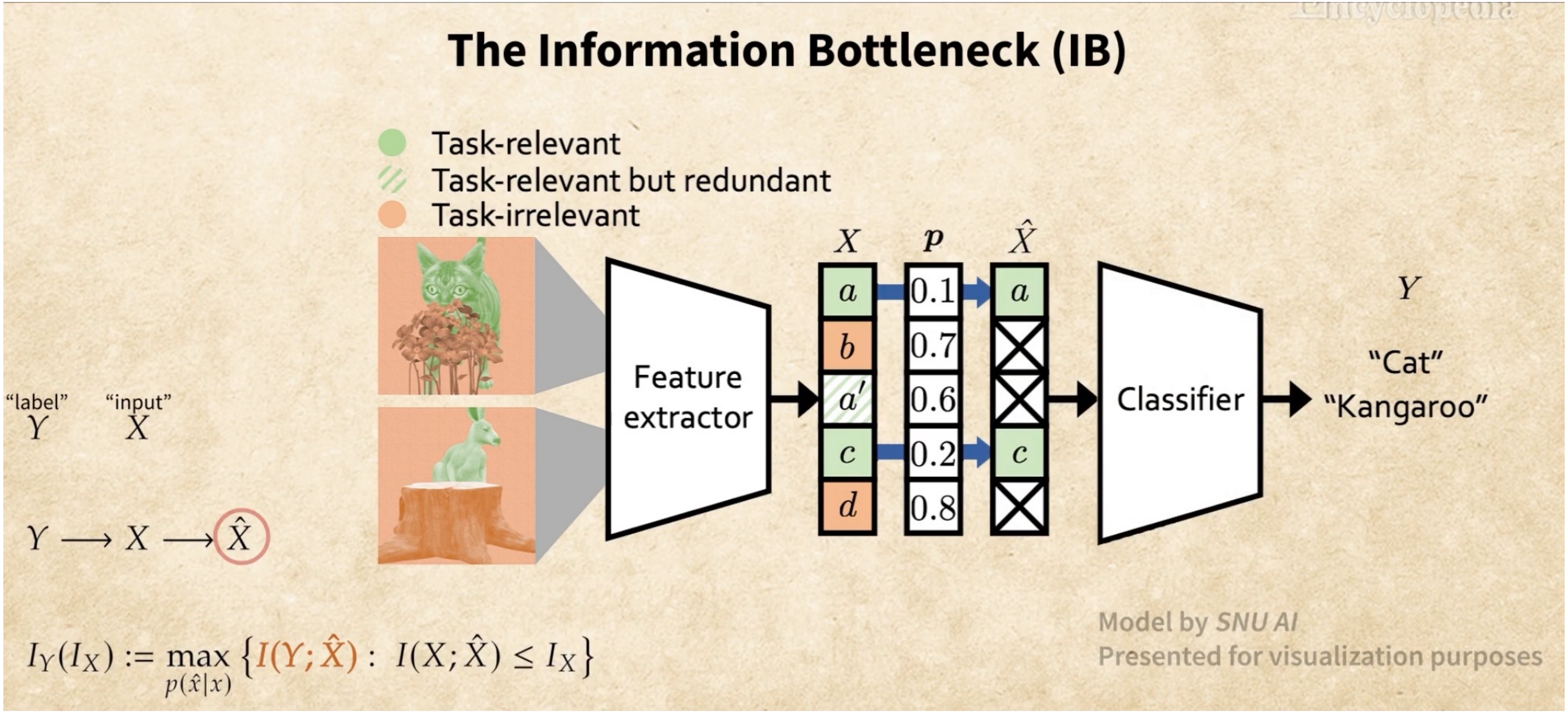


- ■ Small perturbations of the graph structure and node features lead to misclassification of the target [1].
- ■ **Dynamic** GNNs are more sensitive to random noise and targeted adversarial attacks!

■ Question: What is a good representation for graphs?

[1] Acknowledgement from Zügner *et al*. Adversarial attacks on neural networks for graph data. In KDD 2018.

Good graph representation: in the perspective of the Information Bottleneck

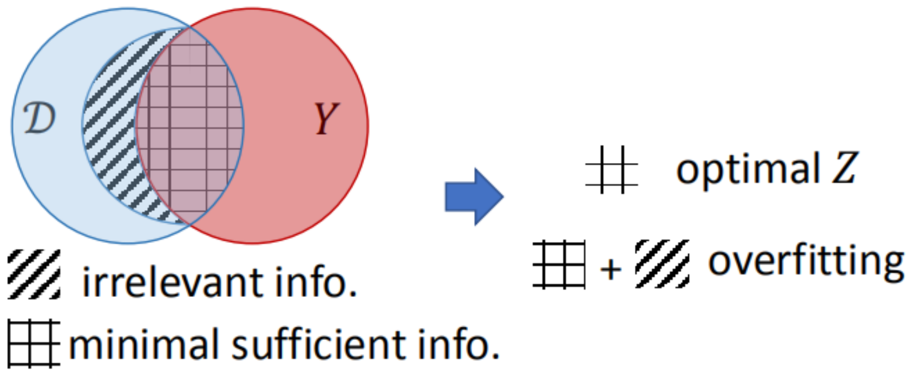


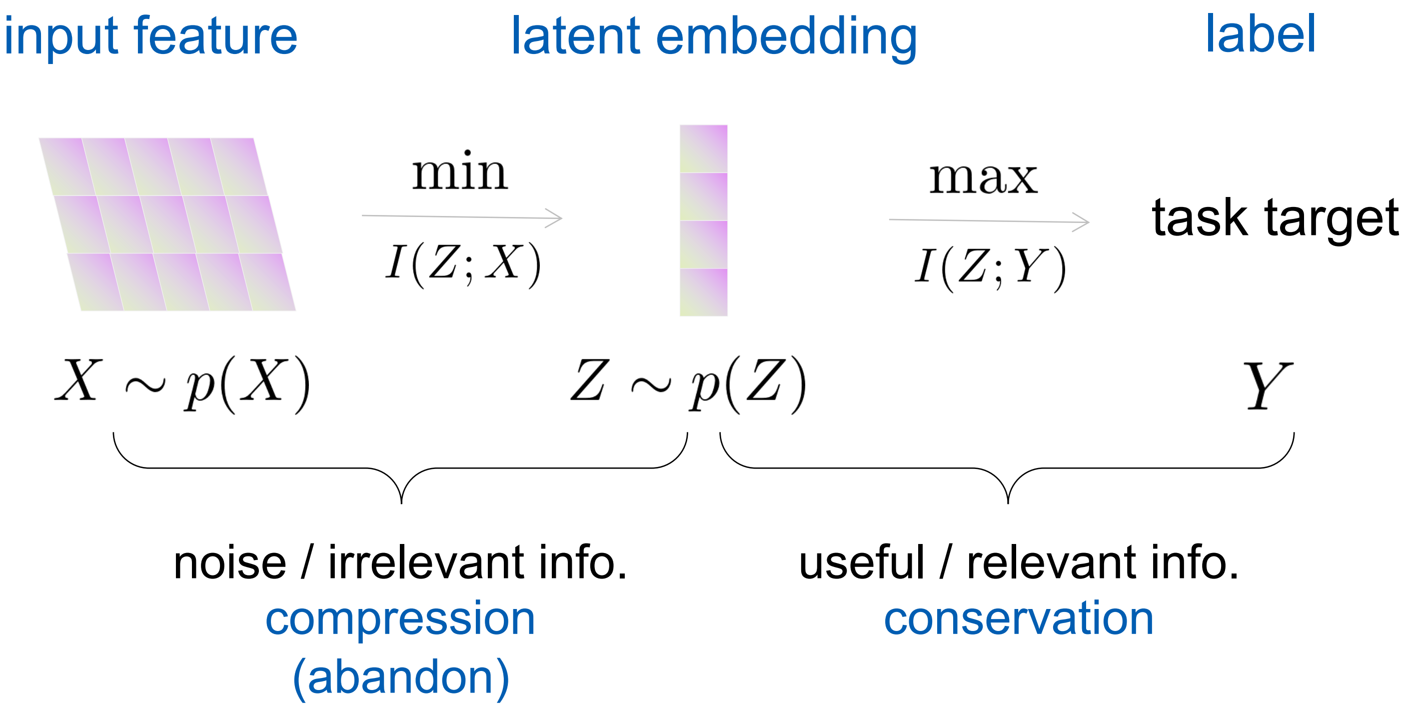
Acknowledgement from Agmon, S. The Information Bottleneck's Ordinary Differential Equation. Encyclopedia.

Good graph representation: in the perspective of the Information Bottleneck

- The **Information Bottleneck (IB)** method is a technique in information theory, designed for finding the best tradeoff between accuracy and compression [1].
- **(a) Vanilla IB.**

$$\max_{Z} I(Z;Y) - \beta I(Z;X)$$



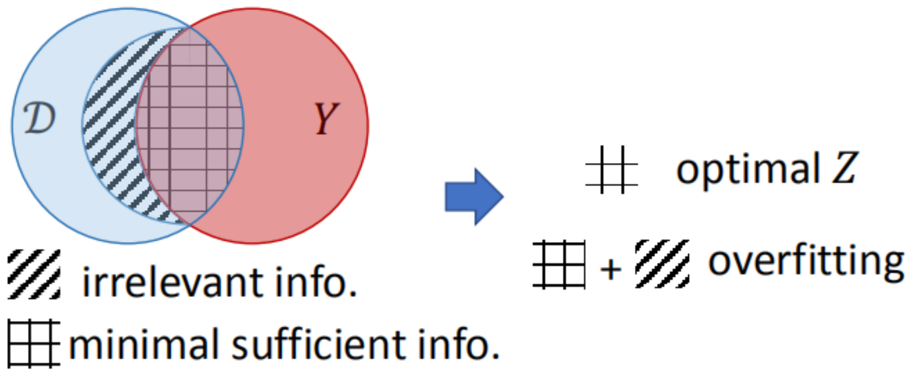


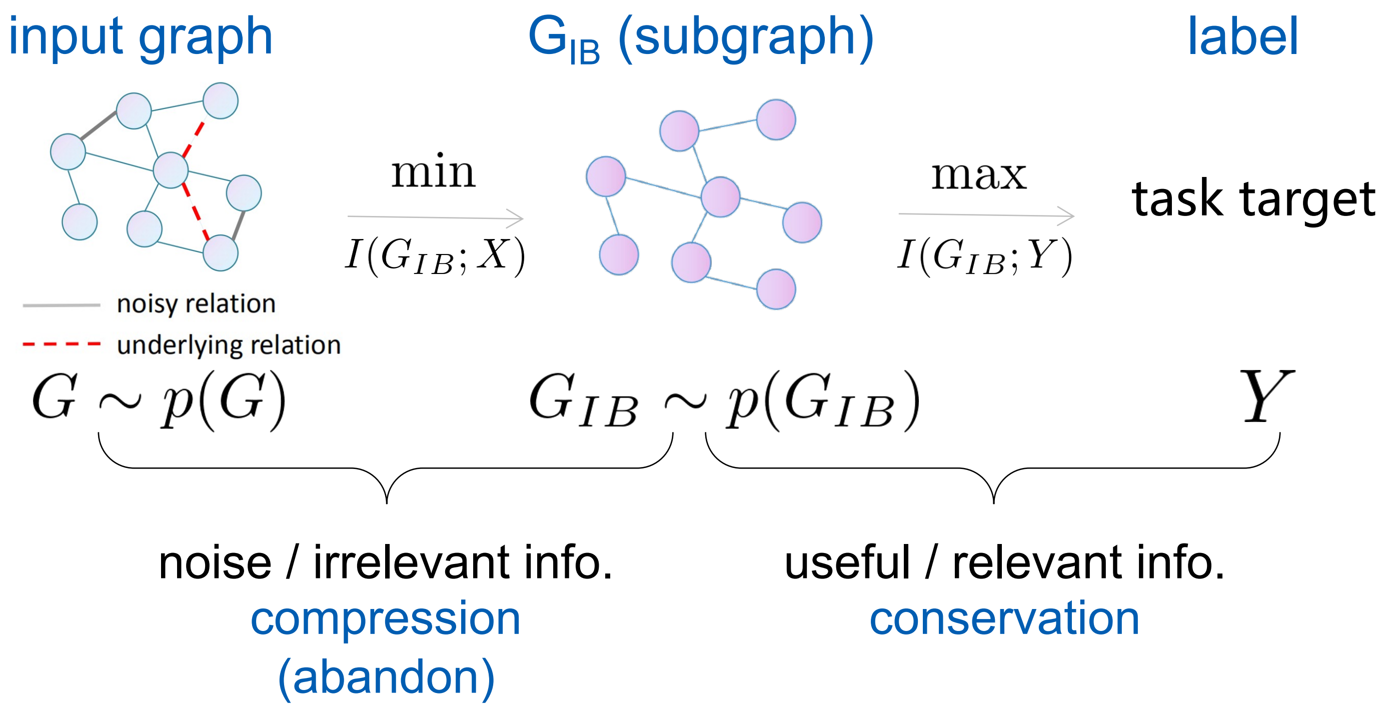
[1] Acknowledgement from Tishby N, Pereira F C, Bialek W. The information bottleneck method, 2000.

Good graph representation: in the perspective of the Information Bottleneck

- The **Information Bottleneck (IB)** method is a technique in information theory, designed for finding the best tradeoff between accuracy and compression [1].
- **(b) Non-structure-involved GIB (Graph IB).**

$$\max_{G_{IB}} I(G_{IB};Y) - \beta I(G_{IB};G)$$





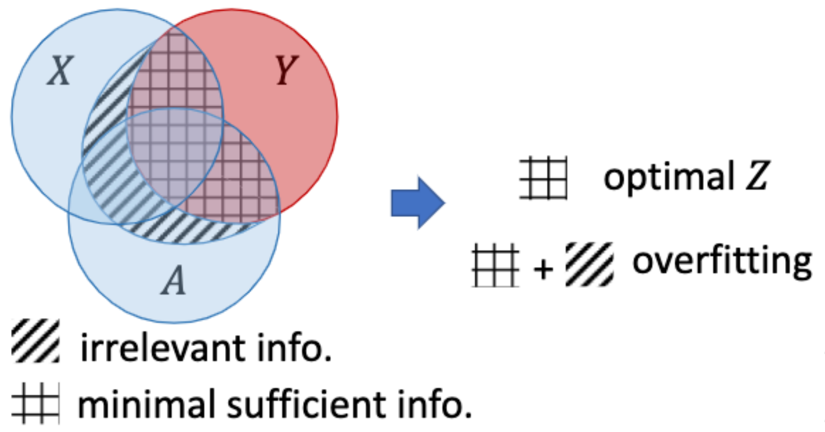
[1] Acknowledgement from Tishby N, Pereira F C, Bialek W. The information bottleneck method, 2000.

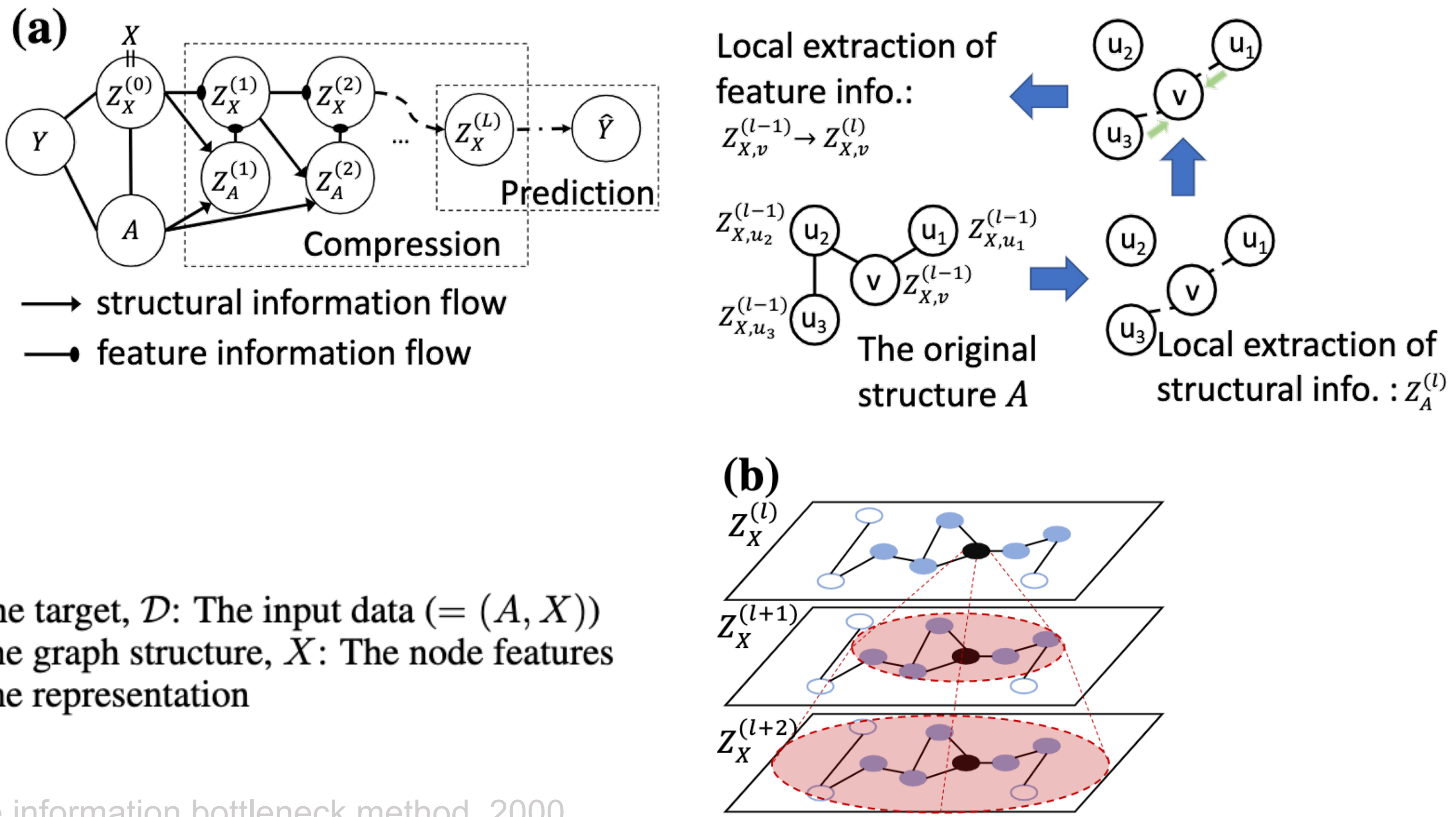
■ Good graph representation: in the perspective of the Information Bottleneck

- The **Information Bottleneck (IB)** method is a technique in information theory, designed for finding the best tradeoff between accuracy and compression [1].
- **(c) Structure-involved GIB (Graph IB) [2].**

Graph Information Bottleneck:

$$\min_{\mathbb{P}(Z|\mathcal{D})\in\Omega} \mathrm{GIB}_\beta(\mathcal{D}, Y; Z) \triangleq [-I(Y;Z) + \beta I(\mathcal{D};Z)]$$

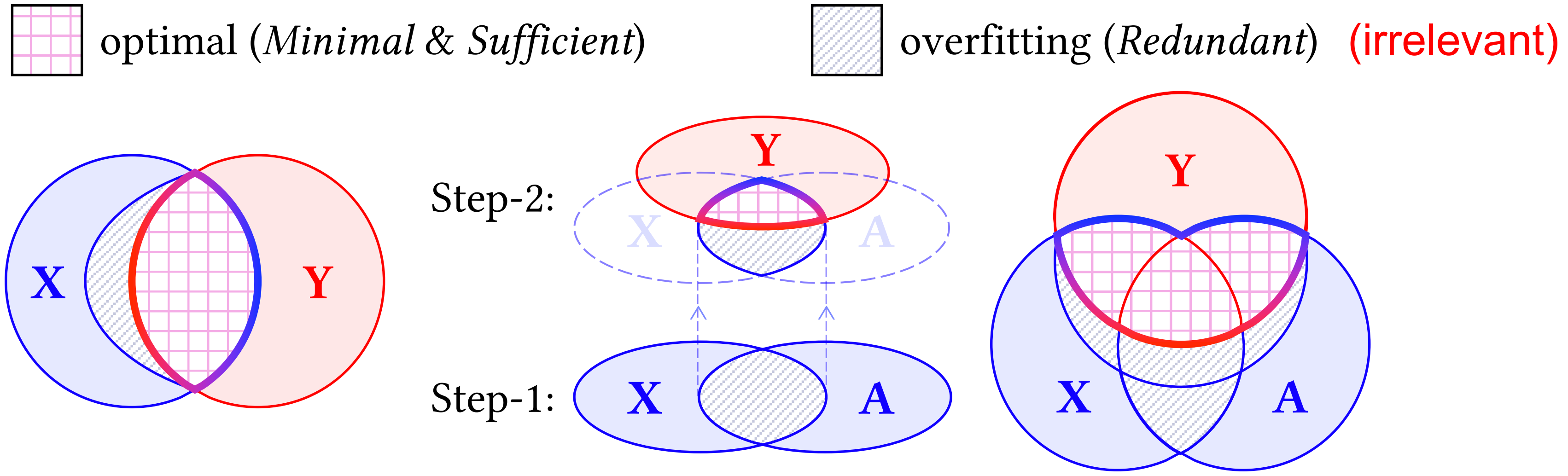




Y: The target, $\mathcal{D}$: The input data $(= (A, X))$
A: The graph structure, X: The node features
Z: The representation

[1] Acknowledgement from Tishby N, Pereira F C, Bialek W. The information bottleneck method, 2000.
[2] Acknowledgement from Tailin Wu *et al*. Graph Information Bottleneck. In NeurIPS 2020.

Good graph representation: in the perspective of the Information Bottleneck



(a) Vanilla IB. (b) Non-struct.-invol. GIB. (c) Struct.-invol. GIB.

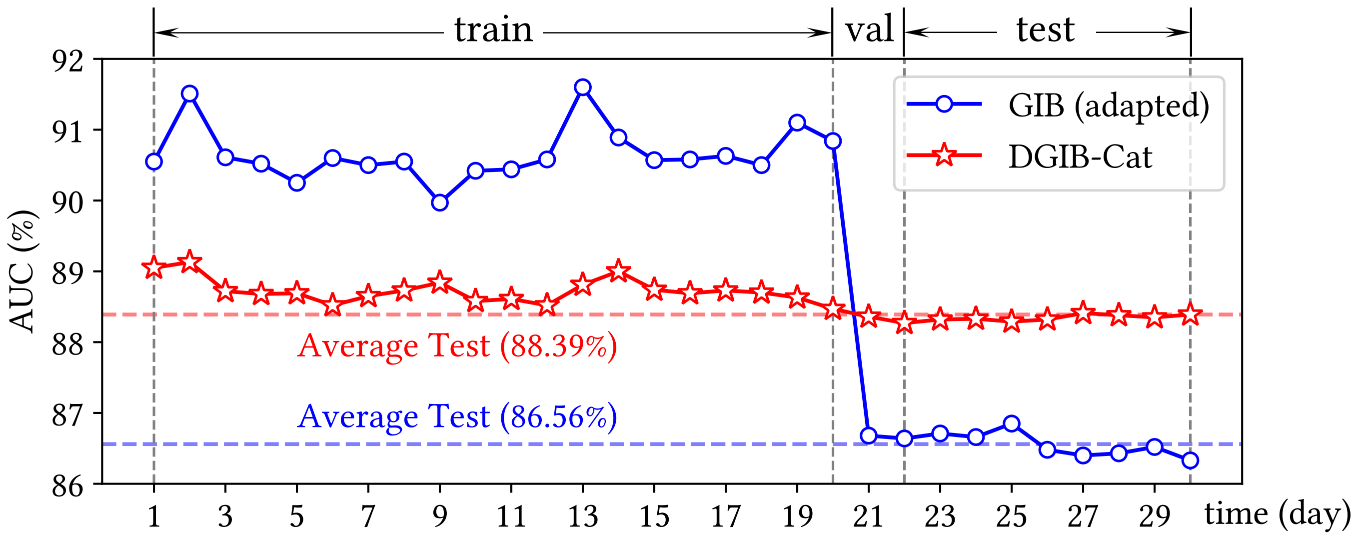
- The high sensitivity of the model to adversarial attacks is because the information that is irrelevant to the target is encoded in the representation [1].
- Past works: A good graph representation contains the **minimal and sufficient** information to achieve good prediction!

Question Further: What is a good representation for dynamic graphs?

[1] Acknowledgement from Tailin Wu *et al*. Graph Information Bottleneck. In NeurIPS 2020.

Good dynamic graph representation: "minimal and sufficient" is inadequate

- **Case study:** directly adapt GIB [1] to dynamic graph representation learning for link prediction



- **A noteworthy finding:** the prediction performance for the next time-step graph during training significantly surpasses that for the target graph during validating and testing.

We argue: A good representation for dynamic graphs should be

Minimal, ***Sufficient***, and ***Consensual*** (*MSC* Condition), but How?

[1] Acknowledgement from Tailin Wu *et al*. Graph Information Bottleneck. In NeurIPS 2020.

Major Challenges

- **How to** understand what constitutes the optimal representation that is both discriminative and robust for downstream task prediction under the dynamic scenario?
- **How to** appropriately compress the input dynamic graph features by optimizing the information flow across graph snapshots with structures straightforwardly involved?
- **How to** optimize the intractable IB objectives, which are incalculable on the non-Euclidean dynamic graphs?

Minimal-Sufficient-Consensual (MSC) Condition

- **Assumption 1 (*Minimal-Sufficient-Consensual (MSC)* Condition)**

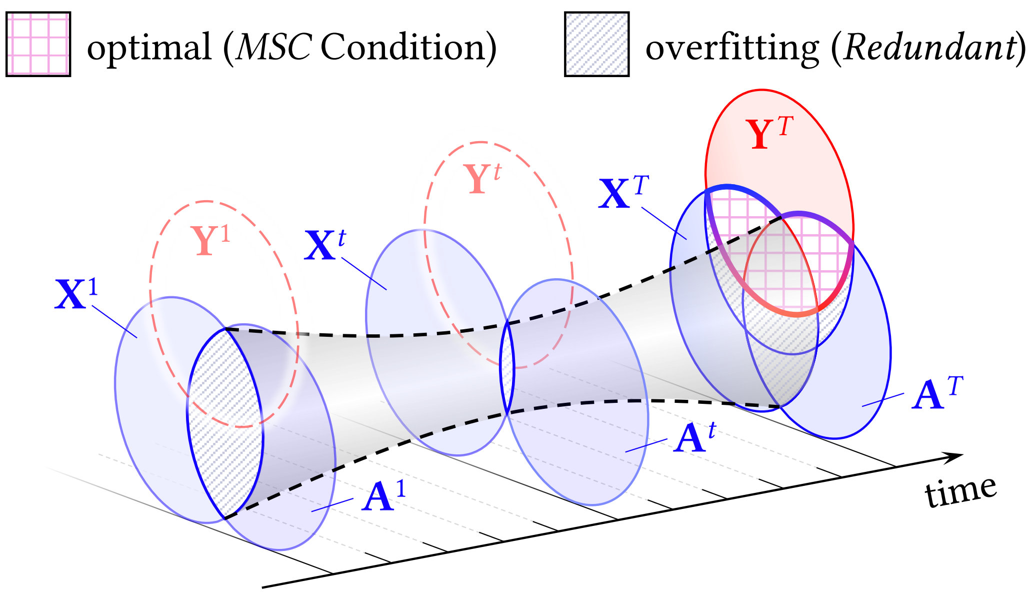
  Given $\mathrm{DG} = \{\mathcal{G}^t\}_{t=1}^{T}$, the optimal representation Z^{T+1} for the robust future link prediction should satisfy the *Minimal-Sufficient-Consensual (MSC)* Condition, such that:

$$\mathrm{Z}^{T+1} = \arg \min_{S(\mathcal{D}):I(S(\mathcal{D});\mathrm{Y}^{T+1})=I(\mathcal{D};\mathrm{Y}^{T+1})} I(S(\mathcal{D});\mathcal{D}),$$

  where $\mathcal{D}$ is the training data containing previous graphs $\mathcal{G}^{1:T}$ and node feature X^{T+1} at the next time-step, and $S(\cdot)$ is the partitions of $\mathcal{D}$, which is implemented by a stochastic encoder $\mathbb{P}(\mathrm{Z}^{T+1} \mid \mathcal{D}, C(\theta))$ where $C(\theta)$ satisfies the *Consensual* Constraint.

- Assumption 1 declares the optimal representation Z^{T+1} for the robust future link prediction task of dynamic graphs should be minimal, sufficient and consensual (*MSC*).
- How to understand “Consensual”? **==> The “baton” for compression process.**

Dynamic Graph Information Bottleneck (DGIB) Principle



Definition 1 (Dynamic Graph Information Bottleneck)

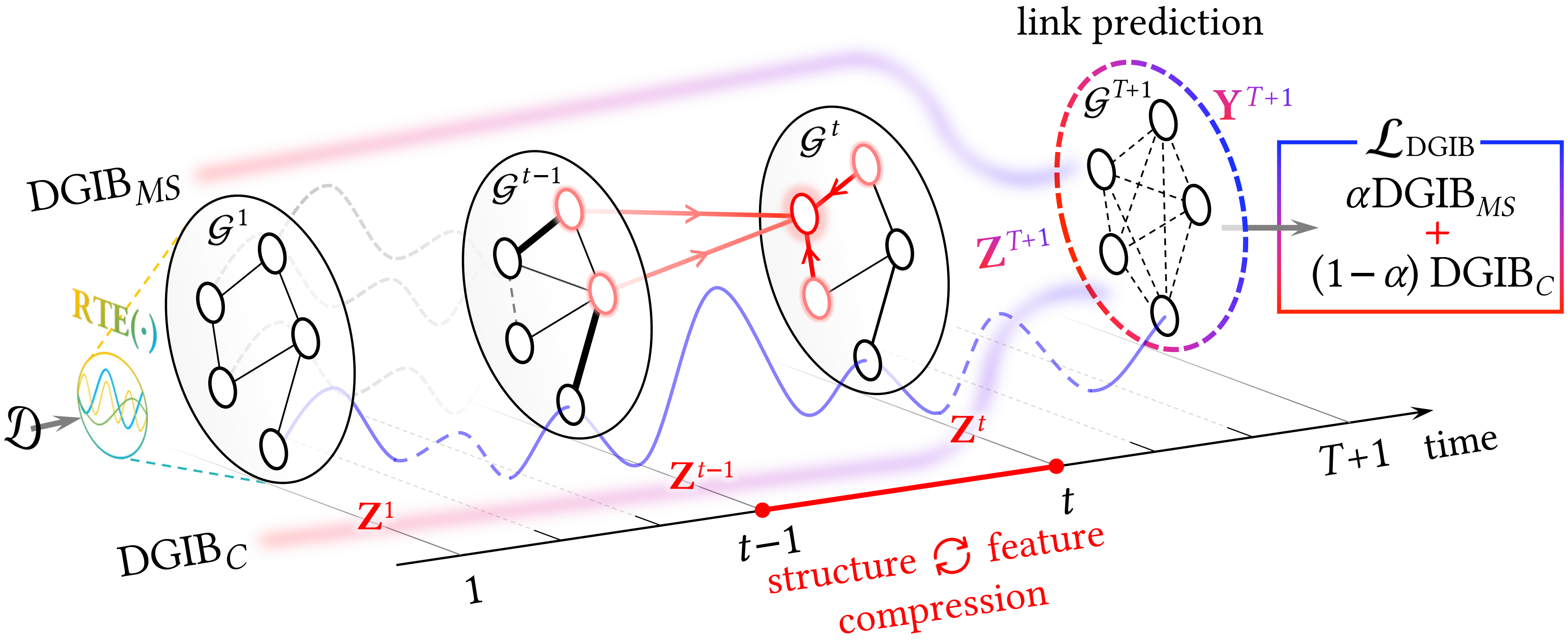
Given DG $= \{\mathcal{G}^t\}_{t=1}^T$, and the nodes feature $\mathbf{X}^{T+1}$ at the next time-step, the Dynamic Graph Information Bottleneck (DGIB) is to learn the optimal representation $\mathbf{Z}^{T+1}$ that satisfies *MSC* Condition by:

$$\mathbf{Z}^{T+1} = \arg \min_{\mathbb{P}(\mathbf{Z}^{T+1} \mid \mathcal{D}, C(\boldsymbol{\theta})) \in \Omega} \text{DGIB}(\mathcal{D}, \mathbf{Y}^{T+1}; \mathbf{Z}^{T+1})$$

$$\triangleq \left[-I(\mathbf{Y}^{T+1}; \mathbf{Z}^{T+1}) + \beta I(\mathcal{D}; \mathbf{Z}^{T+1}) \right].$$

- How to implement and optimize the DGIB objective? Decompose it into DGIB_{MS} and DGIB_C.
  - **DGIB_{MS} channel**: optimize $\mathbf{Z}^{T+1}$ under $\mathbb{P}(\mathbf{Z}^{T+1} \mid \mathcal{D}, \boldsymbol{\theta})$
  - **DGIB_C channel**: encourages $\mathbf{Z}^{1:T}$ and $\mathbf{Z}^{T+1}$ share consensual predictive pattern for predicting $\mathbf{Y}^{T+1}$ under $\mathbb{P}(\mathbf{Z}^{T+1} \mid \mathbf{Z}^{1:T}, C(\boldsymbol{\theta}))$

■ Dynamic Graph Information Bottleneck (DGIB) Framework



- The DGIB framework iteratively compresses structures and node features between graphs.
- The overall $\mathcal{L}_{\mathrm{DGIB}}$ is decomposed to DGIB_{MS} and DGIB_{C} channels, which act jointly to satisfy the *MSC* Condition.

DGIB Principle Derivation and Implementation

- For DGIB$_{MS}$:

- **Assumption 2 (Spatio-Temporal Local Dependence)**

  Given $\mathbf{DG} = \{\mathcal{G}^t\}_{t=1}^{T}$, let $\mathcal{N}_{ST}(v, k, t)$ be the spatio-temporal k-hop neighbors of any node $v \in \mathcal{V}$. The rest of the $\mathbf{DG}$ will be independent of node v and its spatio-temporal k-hop neighbors, *i.e.*:

$$\mathbb{P}\left(\mathbf{X}_v^t \mid \mathcal{G}^{t-1:t}_{\mathcal{N}_{ST}(v,k,t)}, \overline{\mathcal{G}^{1:T}_{\{v\}}}\right) = \mathbb{P}\left(\mathbf{X}_v^t \mid \mathcal{G}^{t-1:t}_{\mathcal{N}_{ST}(v,k,t)}\right),$$

  where $\mathcal{G}^{t-1:t}_{\mathcal{N}_{ST}(v,k,t)}$ denotes $\mathcal{N}_{ST}(v, k, t)$-related subgraphs, and $\overline{\mathcal{G}^{1:T}_{\{v\}}}$ denotes complement graphs in terms of node v and associated edges.

- Assumption 2 is applied to constrain the search space Ω as $\mathbb{P}(\mathbf{Z}^{T+1} \mid \mathcal{D}, \theta)$, which leads to a more feasible DGIB$_{MS}$.

DGIB Principle Derivation and Implementation

- For DGIB$_{MS}$:

$$\mathrm{Z}^{T+1} = \arg \min_{\mathbb{P}(\mathrm{Z}^{T+1}|\mathcal{D},\boldsymbol{\theta})\in\Omega} \mathrm{DGIB}_{MS}(\mathcal{D}, \mathrm{Y}^{T+1}; \mathrm{Z}^{T+1})$$
$$\triangleq \left[- I(\mathrm{Y}^{T+1}; \mathrm{Z}^{T+1}) + \beta_1 I(\mathcal{D}; \mathrm{Z}^{T+1}) \right].$$

Proposition 1 (Lower Bound of $I(\mathrm{Y}^{T+1}; \mathrm{Z}^{T+1})$).

$$I(\mathrm{Y}^{T+1}; \mathrm{Z}^{T+1}) \geq 1 + \mathbb{E}_{\mathbb{P}(\mathrm{Y}^{T+1},\mathrm{Z}^{T+1})}\left[\log \frac{\mathbb{Q}_1(\mathrm{Y}^{T+1} \mid \mathrm{Z}^{T+1})}{\mathbb{Q}_2(\mathrm{Y}^{T+1})}\right] - \mathbb{E}_{\mathbb{P}(\mathrm{Y}^{T+1})\mathbb{P}(\mathrm{Z}^{T+1})}\left[\frac{\mathbb{Q}_1(\mathrm{Y}^{T+1} \mid \mathrm{Z}^{T+1})}{\mathbb{Q}_2(\mathrm{Y}^{T+1})}\right].$$

converges to constants

Proposition 2 (Upper Bound of $I(\mathcal{D}; \mathrm{Z}^{T+1})$). Let $\mathcal{I}_{\mathrm{A}}, \mathcal{I}_{\mathrm{Z}} \subset [T+1]$ be the random time indices. Based on the Markov property $\mathcal{D} \perp\!\!\!\perp \mathrm{Z}^{T+1} \mid (\{\hat{\mathrm{A}}^t\}_{t\in\mathcal{I}_{\mathrm{A}}} \cup \{\mathrm{Z}^t\}_{t\in\mathcal{I}_{\mathrm{Z}}})$, for any $\mathbb{Q}(\hat{\mathrm{A}}^t)$ and $\mathbb{Q}(\mathrm{Z}^t)$:

$$I(\mathcal{D}; \mathrm{Z}^{T+1}) \leq I(\mathcal{D}; \{\hat{\mathrm{A}}^t\}_{t\in\mathcal{I}_{\mathrm{A}}} \cup \{\mathrm{Z}^t\}_{t\in\mathcal{I}_{\mathrm{Z}}}) \leq \sum_{t\in\mathcal{I}_{\mathrm{A}}} \mathcal{A}^t + \sum_{t\in\mathcal{I}_{\mathrm{Z}}} \mathcal{Z}^t,$$

where
$$\mathcal{A}^t = \mathbb{E}\left[\log \frac{\mathbb{P}(\hat{\mathrm{A}}^t \mid \hat{\mathrm{Z}}^t, \mathrm{Z}^{t-1}, \mathrm{A}^t)}{\mathbb{Q}(\hat{\mathrm{A}}^t)}\right],$$
$$\mathcal{Z}^t = \mathbb{E}\left[\log \frac{\mathbb{P}(\mathrm{Z}^t \mid \hat{\mathrm{Z}}^t, \mathrm{Z}^{t-1}, \hat{\mathrm{A}}^t)}{\mathbb{Q}(\mathrm{Z}^t)}\right].$$

DGIB Principle Derivation and Implementation

- For DGIB_{MS}:

$$\mathrm{Z}^{T+1} = \arg \min_{\mathbb{P}(\mathrm{Z}^{T+1}|\mathcal{D},\theta)\in\Omega} \text{DGIB}_{MS}(\mathcal{D}, \mathrm{Y}^{T+1}; \mathrm{Z}^{T+1})$$
$$\triangleq \left[- I(\mathrm{Y}^{T+1}; \mathrm{Z}^{T+1}) + \beta_1 I(\mathcal{D}; \mathrm{Z}^{T+1}) \right].$$

Proposition 1 (Lower Bound of $I(\mathrm{Y}^{T+1}; \mathrm{Z}^{T+1})$).

$$I(\mathrm{Y}^{T+1}; \mathrm{Z}^{T+1}) \geq 1 + \mathbb{E}_{\mathbb{P}(\mathrm{Y}^{T+1},\mathrm{Z}^{T+1})}\left[\log \frac{\mathbb{Q}_1(\mathrm{Y}^{T+1} \mid \mathrm{Z}^{T+1})}{\mathbb{Q}_2(\mathrm{Y}^{T+1})}\right]$$

$$-\mathcal{L}_{\text{CE}}(g(\mathrm{Z}^{T+1}), \mathrm{Y}^{T+1}).$$

Proposition 2 (Upper Bound of $I(\mathcal{D}; \mathrm{Z}^{T+1})$). Let $\mathcal{I}_{\mathrm{A}}, \mathcal{I}_{\mathrm{Z}} \subset [T+1]$ be the random time indices. Based on the Markov property $\mathcal{D} \perp\!\!\!\perp \mathrm{Z}^{T+1} \mid (\{\hat{\mathrm{A}}^t\}_{t\in\mathcal{I}_{\mathrm{A}}} \cup \{\mathrm{Z}^t\}_{t\in\mathcal{I}_{\mathrm{Z}}})$, for any $\mathbb{Q}(\hat{\mathrm{A}}^t)$ and $\mathbb{Q}(\mathrm{Z}^t)$:

$$I(\mathcal{D}; \mathrm{Z}^{T+1}) \leq I(\mathcal{D}; \{\hat{\mathrm{A}}^t\}_{t\in\mathcal{I}_{\mathrm{A}}} \cup \{\mathrm{Z}^t\}_{t\in\mathcal{I}_{\mathrm{Z}}}) \leq \sum_{t\in\mathcal{I}_{\mathrm{A}}} \mathcal{A}^t + \sum_{t\in\mathcal{I}_{\mathrm{Z}}} \mathcal{Z}^t,$$

where
$$\mathcal{A}^t = \mathbb{E}_{\mathbb{P}(\hat{\mathrm{A}}^t|\hat{\mathrm{Z}}^t,\mathrm{Z}^{t-1},\mathrm{A}^t)}\left[\log \frac{\mathbb{P}(\hat{\mathrm{A}}^t \mid \hat{\mathrm{Z}}^t, \mathrm{Z}^{t-1}, \mathrm{A}^t)}{\mathbb{Q}(\hat{\mathrm{A}}^t)}\right]$$

$$\mathcal{Z}^t = \sum_{v\in\mathcal{V}^{t\prime}} \left[\log \Phi(\mathrm{Z}_v^t; \mu_{\mathbb{P}}, \sigma_{\mathbb{P}}^2) - \log \Phi(\mathrm{Z}_v^t; \mu_{\mathbb{Q}}, \sigma_{\mathbb{Q}}^2)\right]$$

$$\mathcal{A}_{\mathrm{B}}^t \doteq \sum_{v\in\mathcal{V}^t} \mathrm{KL}\left[\mathrm{Bern}(\phi_{v,k}^t) \middle\| \mathrm{Bern}(|\mathcal{N}_{ST}(v,k,t)|^{-1})\right]$$
$$\mathcal{A}_{\mathrm{C}}^t \doteq \sum_{v\in\mathcal{V}^t} \mathrm{KL}\left[\mathrm{Cat}(\phi_{v,k}^t) \middle\| \mathrm{Cat}(|\mathcal{N}_{ST}(v,k,t)|^{-1})\right]$$

DGIB Principle Derivation and Implementation

- For DGIB$_C$:

$$\mathrm{Z}^{T+1} = \arg \min_{\mathbb{P}(\mathrm{Z}^{T+1}|\mathrm{Z}^{1:T}, C(\boldsymbol{\theta}))\in\Omega} \mathrm{DGIB}_C\left(\mathrm{Z}^{1:T}, \mathrm{Y}^{T+1}; \mathrm{Z}^{T+1}\right)$$
$$\triangleq \left[- I\left(\mathrm{Y}^{T+1}; \mathrm{Z}^{T+1}\right) + \beta_2 \, I\left(\mathrm{Z}^{1:T}; \mathrm{Z}^{T+1}\right) \right].$$

Proposition 1 (**Lower Bound of** $I(\mathrm{Y}^{T+1}; \mathrm{Z}^{T+1})$).

$$I\left(\mathrm{Y}^{T+1}; \mathrm{Z}^{T+1}\right) \geq 1 + \mathbb{E}_{\mathbb{P}(\mathrm{Y}^{T+1}, \mathrm{Z}^{T+1})}\left[\log \frac{\mathbb{Q}_1\left(\mathrm{Y}^{T+1} \mid \mathrm{Z}^{T+1}\right)}{\mathbb{Q}_2\left(\mathrm{Y}^{T+1}\right)}\right] - \mathbb{E}_{\mathbb{P}(\mathrm{Y}^{T+1})\mathbb{P}(\mathrm{Z}^{T+1})}\left[\frac{\mathbb{Q}_1\left(\mathrm{Y}^{T+1} \mid \mathrm{Z}^{T+1}\right)}{\mathbb{Q}_2\left(\mathrm{Y}^{T+1}\right)}\right].$$

converges to constants

Proposition 3 (Upper Bound of $I(\mathrm{Z}^{1:T}; \mathrm{Z}^{T+1})$).

$$I\left(\mathrm{Z}^{1:T}; \mathrm{Z}^{T+1}\right) \leq \mathbb{E}\left[\log \frac{\mathbb{P}\left(\mathrm{Z}^{T+1} \mid \mathrm{Z}^{1:T}\right)}{\mathbb{Q}\left(\mathrm{Z}^{T+1}\right)}\right].$$

DGIB Principle Derivation and Implementation

- For DGIB$_C$:

$$\mathrm{Z}^{T+1} = \arg \min_{\mathbb{P}(\mathrm{Z}^{T+1}|\mathrm{Z}^{1:T}, C(\theta)) \in \Omega} \mathrm{DGIB}_C(\mathrm{Z}^{1:T}, \mathrm{Y}^{T+1}; \mathrm{Z}^{T+1})$$

$$\triangleq \left[- I(\mathrm{Y}^{T+1}; \mathrm{Z}^{T+1}) + \beta_2 I(\mathrm{Z}^{1:T}; \mathrm{Z}^{T+1}) \right].$$

Proposition 1 (Lower Bound of $I(\mathrm{Y}^{T+1}; \mathrm{Z}^{T+1})$).

$$I(\mathrm{Y}^{T+1}; \mathrm{Z}^{T+1}) \geq 1 + \mathbb{E}_{\mathbb{P}(\mathrm{Y}^{T+1}, \mathrm{Z}^{T+1})} \left[\log \frac{\mathbb{Q}_1(\mathrm{Y}^{T+1} \mid \mathrm{Z}^{T+1})}{\mathbb{Q}_2(\mathrm{Y}^{T+1})} \right]$$

$$-\mathcal{L}_{\mathrm{CE}}(g(\mathrm{Z}^{T+1}), \mathrm{Y}^{T+1})$$

Proposition 3 (Upper Bound of $I(\mathrm{Z}^{1:T}; \mathrm{Z}^{T+1})$).

$$I(\mathrm{Z}^{1:T}; \mathrm{Z}^{T+1}) \leq$$

$$\sum_{v \in \mathcal{V}^{(T+1)\prime}} \left[\log \Phi(\mathrm{Z}_v^{T+1}; \boldsymbol{\mu}_{\mathbb{P}}, \boldsymbol{\sigma}_{\mathbb{P}}^2) - \log \Phi(\mathrm{Z}_v^{T+1}; \boldsymbol{\mu}_{\mathbb{Q}}, \boldsymbol{\sigma}_{\mathbb{Q}}^2) \right]$$

DGIB Principle Derivation and Implementation

- For DGIB_{MS}:

$$\mathbf{Z}^{T+1} = \arg \min_{\mathbb{P}(\mathbf{Z}^{T+1}|\mathcal{D},\boldsymbol{\theta})\in\Omega} \text{DGIB}_{MS}(\mathcal{D}, \mathbf{Y}^{T+1}; \mathbf{Z}^{T+1}) \triangleq \left[- I(\mathbf{Y}^{T+1}; \mathbf{Z}^{T+1}) + \beta_1 I(\mathcal{D}; \mathbf{Z}^{T+1}) \right].$$

Minimal & Sufficient

- For DGIB_{C}:

$$\mathbf{Z}^{T+1} = \arg \min_{\mathbb{P}(\mathbf{Z}^{T+1}|\mathbf{Z}^{1:T},C(\boldsymbol{\theta}))\in\Omega} \text{DGIB}_{C}(\mathbf{Z}^{1:T}, \mathbf{Y}^{T+1}; \mathbf{Z}^{T+1}) \triangleq \left[- I(\mathbf{Y}^{T+1}; \mathbf{Z}^{T+1}) + \beta_2 I(\mathbf{Z}^{1:T}; \mathbf{Z}^{T+1}) \right].$$

Consensual

- **Overall:**

$$\mathcal{L}_{\text{DGIB}} = \alpha \text{DGIB}_{MS} + (1-\alpha) \text{DGIB}_{C}$$

DGIB Overall Training Pipeline

Algorithm 1: Overall training pipeline of DGIB framework.

Input: Dynamic graph DG $= \{\mathcal{G}^t\}_{t=1}^{T}$; Node features $\mathbf{X}^{1:T+1}$; Labels $\mathbf{Y}^{1:T}$ of link occurrence; Number of layers L; Relative time encoding function RTE$(\cdot)$; Non-linear rectifier τ; Activation function σ; Hyperparameters k, α, β_1 and β_2.

Output: The optimized robust model $f_{\theta}^{\star} = w \circ g$; Predicted label $\hat{\mathbf{Y}}^{T+1}$ of link occurrence at time $T+1$.

```
Initialize weights W and learnable parameters θ randomly;
Z^{1:T+1,(0)} ← RTE(X^{1:T+1}), attentions w ← GAT(Z^{1:T+1,(0)});
Construct N_ST(v,k,t) ← {u ∈ V^{t-1:t} | d(u,v) = k};
for l = 1, 2, ..., L and v ∈ V^{1:T+1} do
    for t in range [T + 1] do
        Ẑ^{t(l-1)} ← τ(Z^{t(l-1)}) W^{(l)};
        φ^{t(l)}_{v,k} ← σ{(Ẑ^{t(l-1)}_v || Ẑ^{t-1:t(l-1)}_u) w^T}_{u∈N_ST(v,k,t)};
        Â^{t(l)} ← ∪_{v∈V^t} {u ∈ N_ST(v,k,t) | u ~ Bern(φ^{t(l)}_{v,k})}
            or ∪_{v∈V^t} {u ∈ N_ST(v,k,t) | u ~ Cat(φ^{t(l)}_{v,k})};
        Z^{t(l)} ← Σ_{(u,v)∈Â^{t(l)}} {Ẑ^{t(l-1)}_v}_{v∈V^t};
    end
    Ŷ^{T+1} = g(Z^{T+1(L)});
    DGIB_MS ← Eq. (8), DGIB_C ← Eq. (13);
    Calculate the overall loss, as L_DGIB ← Eq. (21);
    Update θ by minimizing L_DGIB and back-propagation.
end
```

Dynamic Graph Datasets

Dataset	# Node	# Link	# Link Type	Length (Split)	Temporal Granularity
COLLAB	23,035	151,790	5	16 (10/1/5)	year
Yelp	13,095	65,375	5	24 (15/1/8)	month
ACT	20,408	202,339	5	30 (20/2/8)	day

Baselines

- **Static GNNs:** GAE, VGAE, GAT
- **Dynamic GNNs:** GCRN, EvolveGCN, DySAT
- **Robust (D)GNNs and Regularizer:** IRM, V-REx, GroupDRO, RGCN, DIDA, GIB

Adversarial Attack Settings

- **Non-targeted:** for structures, randomly remove links; for node features, add random noise
- **Targeted:** apply NETTACK library to perform both evasion attack and poisoning attack
  - evasion: train on clean datasets and perform attacking on each graph snapshot in testing.
  - poisoning: attack the whole dataset before model training and testing.

■ Against Non-targeted Adversarial Attacks

Table 1: AUC score (% ± standard deviation) of future link prediction task on real-world datasets against *non-targeted* adversarial attacks. The best results are shown in bold type and the runner-ups are underlined.

Dataset	COLLAB					Yelp					ACT				
Model	Clean	Structure Attack	Feature Attack			Clean	Structure Attack	Feature Attack			Clean	Structure Attack	Feature Attack		
			$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 1.5$			$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 1.5$			$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 1.5$
GAE [20]	77.15±0.5	74.04±0.8	50.59±0.8	44.66±0.8	43.12±0.8	70.67±1.1	64.45±5.0	51.05±0.6	45.41±0.6	41.56±0.9	72.31±0.5	60.27±0.4	56.56±0.5	52.52±0.6	50.36±0.9
VGAE [20]	86.47±0.0	74.95±1.2	56.75±0.6	50.39±0.7	48.68±0.7	76.54±0.5	65.33±1.4	55.53±0.7	49.88±0.8	45.08±0.6	79.18±0.5	66.29±1.3	60.67±0.7	57.39±0.8	55.27±1.0
GAT [44]	88.26±0.4	77.29±1.8	58.13±0.9	51.41±0.9	49.77±0.9	77.93±0.1	69.35±1.6	56.72±0.3	52.51±0.5	46.21±0.5	85.07±0.3	77.55±1.2	66.05±0.4	61.85±0.3	59.05±0.3
GCRN [35]	82.78±0.5	69.72±0.5	54.07±0.9	47.78±0.8	46.18±0.9	68.59±1.0	54.68±7.6	52.68±0.6	46.85±0.6	40.45±0.6	76.28±0.5	64.35±1.2	59.48±0.7	54.16±0.6	53.88±0.7
EvolveGCN [30]	86.62±1.0	76.15±0.9	56.82±1.2	50.33±1.0	48.55±1.0	78.21±0.0	53.82±2.0	57.91±0.5	51.82±0.3	45.32±1.0	74.55±0.3	63.17±1.0	61.02±0.5	53.34±0.5	51.62±0.7
DySAT [33]	88.77±0.2	76.59±0.2	58.28±0.3	51.52±0.3	49.32±0.5	78.87±0.6	66.09±1.4	58.46±0.4	52.33±0.7	46.24±0.7	78.52±0.4	66.55±1.2	61.94±0.8	56.98±0.8	54.14±0.7
IRM [2]	87.96±0.9	75.42±0.9	60.51±1.3	53.89±1.1	52.17±0.9	66.49±10.8	56.02±16.0	50.96±3.3	48.58±5.2	45.32±3.3	80.02±0.6	69.19±1.4	62.84±0.1	57.28±0.2	56.04±0.2
V-REx [20]	88.31±0.3	76.24±0.8	61.23±1.5	54.51±1.0	52.24±1.1	79.04±0.2	66.41±1.9	61.49±0.5	53.72±1.0	51.32±0.9	83.11±0.3	70.15±1.1	65.59±0.1	60.03±0.3	58.79±0.2
GroupDRO [32]	88.76±0.1	76.33±0.3	61.10±1.3	54.62±1.0	52.33±0.8	79.38±0.4	66.97±0.6	61.78±0.8	55.37±0.9	52.18±0.7	85.19±0.5	74.35±1.6	66.05±0.5	61.85±0.4	59.05±0.3
RGCN [58]	88.21±0.1	78.66±0.7	61.29±0.5	54.29±0.6	52.99±0.6	77.28±0.3	74.29±0.4	59.72±0.3	52.88±0.3	50.40±0.2	87.22±0.2	82.66±0.4	68.51±0.2	62.67±0.2	61.31±0.2
DIDA [56]	91.97±0.0	80.87±0.4	61.32±0.8	55.77±0.9	54.91±0.9	78.22±0.4	75.92±0.9	60.83±0.6	54.11±0.6	50.21±0.6	89.84±0.8	78.64±1.0	70.97±0.2	64.49±0.4	62.57±0.2
GIB [46]	91.36±0.2	80.89±0.1	61.88±0.8	55.15±0.8	54.65±0.9	77.52±0.4	75.03±0.3	61.94±0.9	56.15±0.3	52.21±0.8	92.33±0.3	86.99±0.3	72.16±0.5	66.72±0.2	64.96±0.5
DGIB-Bern	92.17±0.2	83.58±0.1	63.54±0.9	56.92±1.0	**56.24±1.0**	76.88±0.2	75.61±0.0	**63.91±0.9**	**59.28±0.9**	**54.77±1.0**	94.49±0.2	87.75±0.1	73.05±0.9	68.49±0.9	**66.27±0.9**
DGIB-Cat	**92.68±0.1**	**84.16±0.1**	**63.99±0.5**	**57.76±0.8**	55.63±1.0	**79.53±0.2**	**77.72±0.1**	61.42±0.9	55.12±0.7	51.90±0.9	**94.89±0.2**	**88.27±0.2**	**73.92±0.8**	**68.88±0.9**	65.99±0.7

■ Against Targeted Adversarial Attacks

Table 2: AUC score (% ± standard deviation) of future link prediction task on real-world datasets against *targeted* adversarial attacks. The best results are shown in bold type and the runner-ups are underlined.

Dataset	Model	Clean	Evasion Attack $n=1$	Evasion Attack $n=2$	Evasion Attack $n=3$	Evasion Attack $n=4$	Evasion Avg. Decrease	Poisoning Attack $n=1$	Poisoning Attack $n=2$	Poisoning Attack $n=3$	Poisoning Attack $n=4$	Poisoning Avg. Decrease
COLLAB	VGAE [20]	86.47±0.0	73.39±0.1	62.18±0.1	51.72±0.1	46.97±0.1	↓ 32.27	63.42±0.3	52.63±0.3	50.98±0.4	45.64±0.3	↓ 38.51
	GAT [44]	88.26±0.4	76.21±0.1	66.56±0.1	57.92±0.1	50.96±0.1	↓ 28.71	66.59±0.5	55.31±0.6	51.34±0.7	48.99±0.9	↓ 37.05
	DySAT [33]	88.77±0.2	77.91±0.1	68.22±0.1	58.82±0.1	51.39±0.1	↓ 27.80	69.02±0.3	57.62±0.3	52.76±0.3	50.07±0.8	↓ 35.37
	RGCN [58]	88.21±0.1	77.65±0.1	67.11±0.1	59.06±0.1	52.02±0.1	↓ 27.49	69.48±0.2	58.39±0.3	52.48±0.6	50.62±0.9	↓ 34.53
	GIB [46]	91.36±0.2	78.95±0.0	69.63±0.1	60.98±0.0	54.48±0.2	↓ 27.74	71.47±0.3	61.03±0.4	54.97±0.7	52.09±1.0	↓ 34.44
	DGIB-Bern	92.17±0.2	**81.36±0.0**	**72.79±0.0**	**63.25±0.1**	**57.22±0.1**	**↓ 25.51**	**74.06±0.3**	**61.93±0.2**	**56.57±0.2**	52.62±0.3	**↓ 33.49**
	DGIB-Cat	**92.68±0.1**	81.29±0.0	71.32±0.1	62.03±0.1	55.08±0.1	↓ 27.24	72.55±0.2	60.99±0.3	55.62±0.4	**53.08±0.3**	↓ 34.65
Yelp	VGAE [20]	76.54±0.5	65.86±0.1	54.82±0.2	48.08±0.1	46.25±0.1	↓ 29.77	62.73±0.6	52.61±0.4	47.72±0.4	45.43±0.5	↓ 31.90
	GAT [44]	77.93±0.1	67.96±0.1	59.47±0.1	50.27±0.1	48.62±0.1	↓ 27.39	65.34±0.5	54.51±0.2	50.24±0.4	48.96±0.4	↓ 29.72
	DySAT [33]	78.87±0.6	69.77±0.1	60.66±0.1	52.16±0.1	50.15±0.1	↓ 26.22	66.87±0.6	56.31±0.3	50.44±0.6	50.49±0.5	↓ 28.96
	RGCN [58]	77.28±0.3	68.54±0.1	60.69±0.1	51.51±0.1	49.72±0.1	↓ 25.44	65.55±0.4	55.47±0.3	49.08±0.6	49.09±0.6	↓ 29.09
	GIB [46]	77.52±0.4	68.59±0.1	61.22±0.1	51.26±0.1	49.58±0.1	↓ 25.61	65.59±0.3	56.79±0.3	50.92±0.4	49.55±0.4	↓ 28.13
	DGIB-Bern	76.88±0.2	**72.27±0.1**	60.96±0.0	**54.32±0.1**	**51.73±0.1**	**↓ 22.19**	**68.64±0.2**	56.73±0.2	**53.18±0.3**	50.21±0.2	**↓ 25.61**
	DGIB-Cat	**79.53±0.2**	70.17±0.0	**62.25±0.1**	52.69±0.1	50.87±0.1	↓ 25.82	67.38±0.3	**57.02±0.2**	51.39±0.2	**50.53±0.2**	↓ 28.85
ACT	VGAE [20]	79.18±0.5	67.59±0.1	62.98±0.1	54.33±0.1	52.26±0.0	↓ 25.11	62.55±1.6	55.15±1.7	51.02±1.8	50.11±1.9	↓ 30.90
	GAT [44]	85.07±0.3	75.14±0.1	67.25±0.1	59.75±0.1	58.51±0.1	↓ 23.40	71.26±0.9	61.43±1.1	57.35±1.1	58.53±1.0	↓ 26.95
	DySAT [33]	78.52±0.4	70.64±0.1	63.35±0.0	56.36±0.0	55.12±0.1	↓ 21.84	66.21±0.9	56.28±0.9	53.45±1.1	54.43±1.0	↓ 26.65
	RGCN [58]	87.22±0.2	78.64±0.1	70.11±0.1	62.99±0.1	61.31±0.1	↓ 21.73	73.71±0.8	63.43±0.9	59.97±1.3	60.41±0.8	**↓ 26.18**
	GIB [46]	92.33±0.3	85.61±0.1	74.08±0.1	65.44±0.1	64.04±0.1	↓ 21.70	80.01±0.7	67.04±0.8	63.85±0.6	60.95±0.7	↓ 26.39
	DGIB-Bern	94.49±0.2	**89.83±0.1**	**85.81±0.1**	**79.95±0.1**	**78.01±0.1**	**↓ 11.73**	**80.92±0.3**	**70.76±0.4**	**65.27±0.6**	61.93±0.9	↓ 26.21
	DGIB-Cat	**94.89±0.2**	84.98±0.1	76.78±0.1	67.69±0.1	66.68±0.1	↓ 21.98	80.16±0.4	68.71±0.5	64.38±0.6	**65.43±0.9**	↓ 26.57

Ablation Study

- **DGIB (w/o *Cons*):** remove the *Consensual* channel DGIB_C in the overall training objective, and optimizing only with the *Minimal* and *Sufficient* channel DGIB_{MS}.
- **DGIB (w/o $\mathcal{A}$):** remove the structure sampling term ($\mathcal{A}$) in the upper bound of $I(\mathcal{D}; \mathbf{Z}^{T+1})$.
- **DGIB (w/o $\mathcal{Z}$):** remove the feature sampling term ($\mathcal{Z}$) in the upper bound of $I(\mathcal{D}; \mathbf{Z}^{T+1})$.

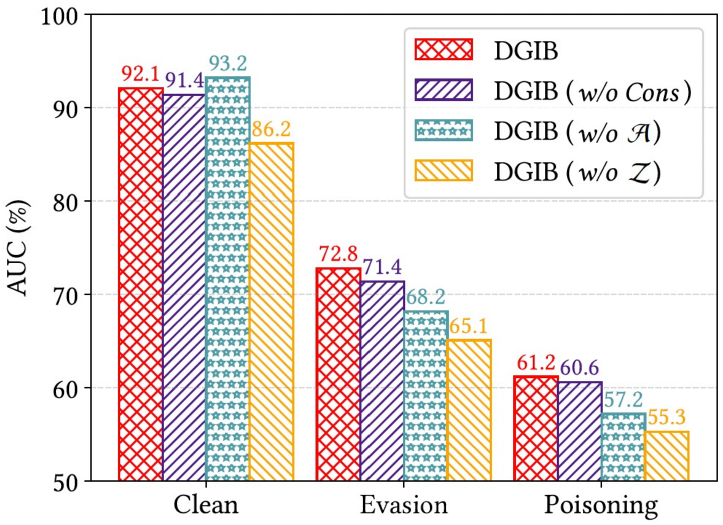


Figure 4: Ablation study results.
(COLLAB)

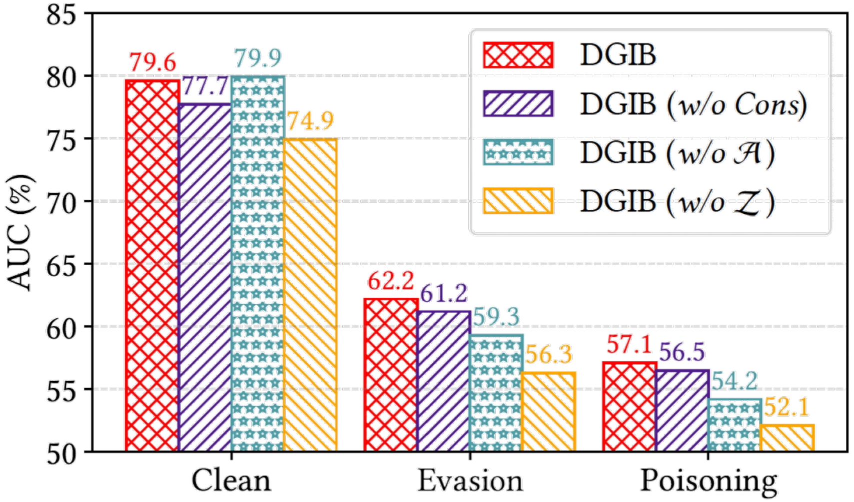


Figure B.1: Results of ablation study on Yelp.

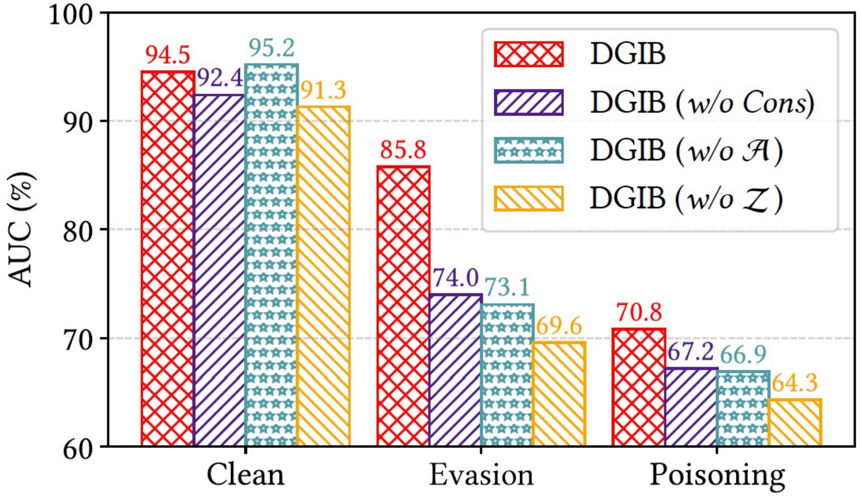
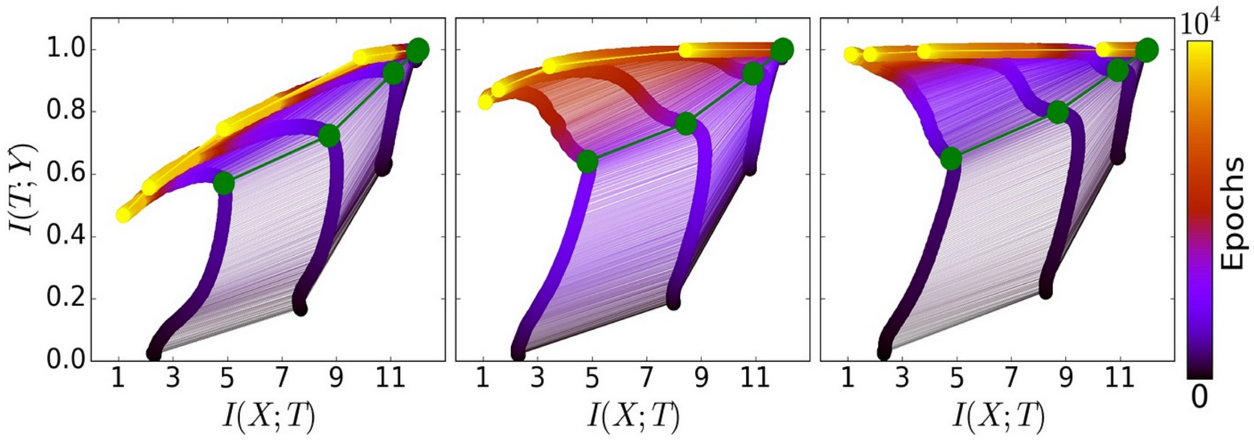


Figure B.2: Results of ablation study on ACT.

Information Plane Analysis

- **Information Plane:** observe the evolution of the IB compression process, including the changes in the mutual information between input, latent representations, and output during training.
- Given the Markov Chain $<\mathrm{X} \rightarrow \mathrm{Y} \rightarrow \mathrm{Z}>$, the latent representation is uniquely mapped to a point in the Information Plane with coordinates $(I(\mathrm{X};\mathrm{Z}), I(\mathrm{Y},\mathrm{Z}))$.

$$
\begin{aligned}
\mathrm{Z}^{T+1} &= \arg \min_{\mathbb{P}(\mathrm{Z}^{T+1} \mid \mathcal{D}, C(\theta)) \in \Omega} \mathrm{DGIB}(\mathcal{D}, \mathrm{Y}^{T+1}; \mathrm{Z}^{T+1}) \\
&\triangleq \left[-I(\mathrm{Y}^{T+1}; \mathrm{Z}^{T+1}) + \beta I(\mathcal{D}; \mathrm{Z}^{T+1}) \right],
\end{aligned}
$$



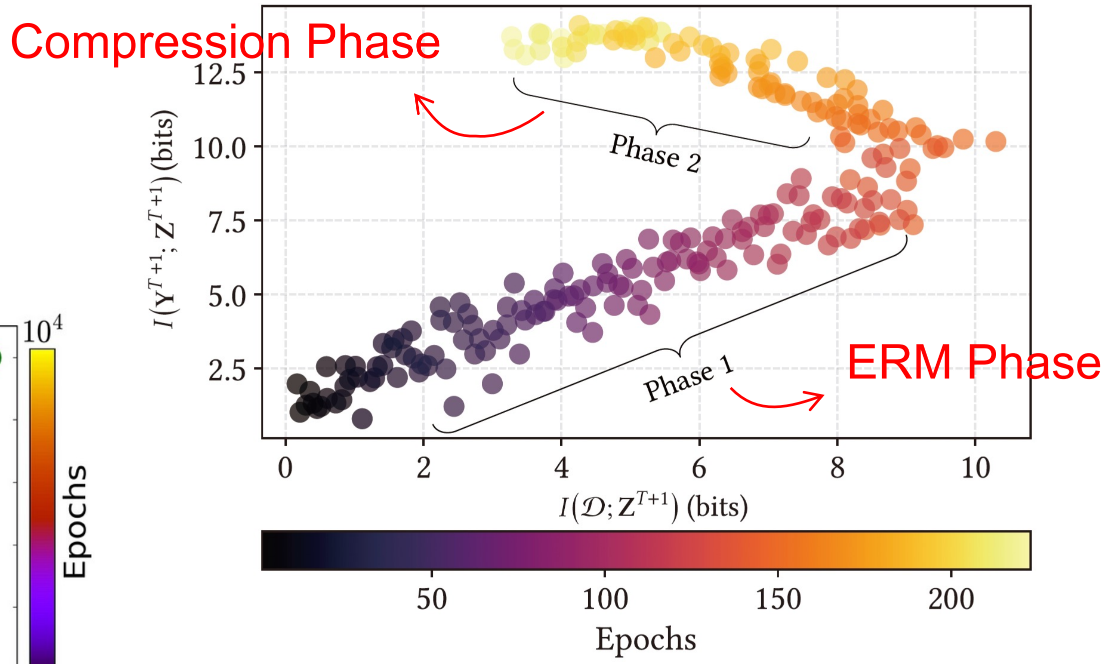


Figure 5: Information Plane analysis.

Hyperparameter Analysis

- We analyze the sensitivity of hyperparameters α, β_1 and β_2.
- We analyze the impact of the compression parameters β_1 and β_2 on the trade-off between the performance of prediction and robustness.
- We conduct experiments based on DGIB-Bern with different ratios of MSE term and compression term $(1/\beta_1, 1/\beta_2)$ on the clean, evasion attacked $(n = 2)$ and poisoning attacked $(n = 2)$ COLLAB.

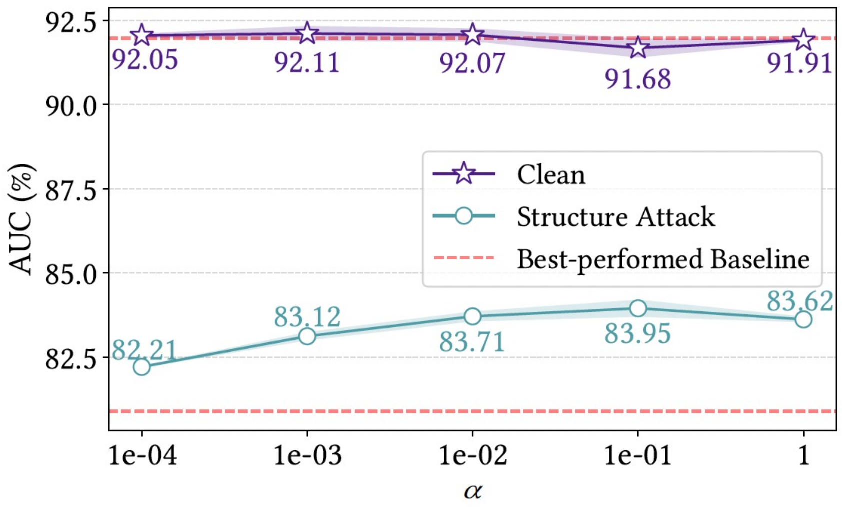


Figure B.3: Sensitivity analysis of α on COLLAB.

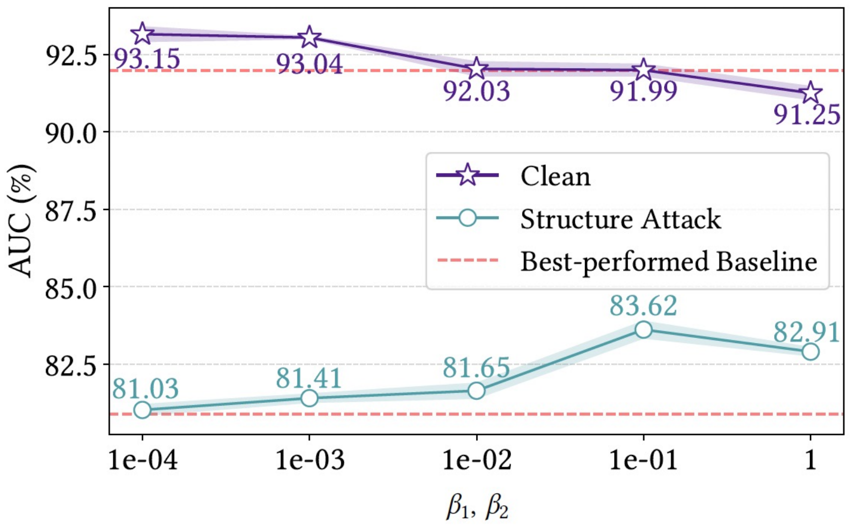


Figure B.4: Sensitivity analysis of β_1, β_2 on COLLAB.

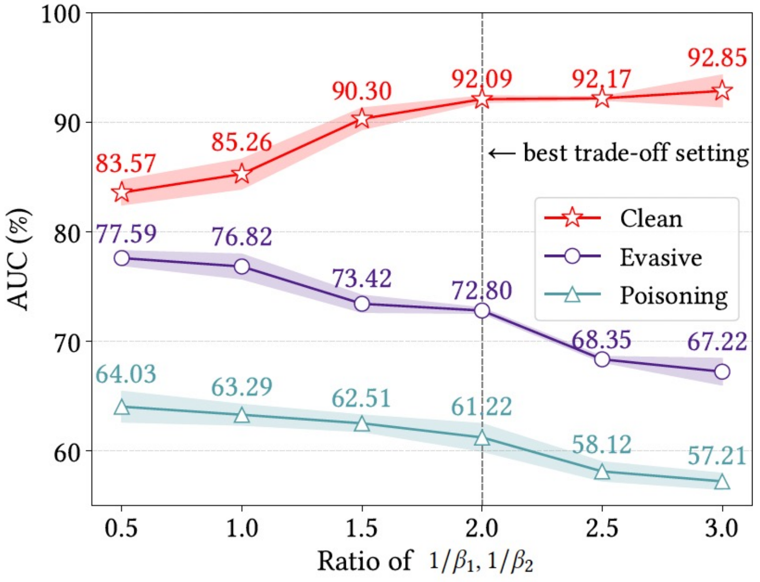


Figure 6: Performance trade-off analysis.

Paper Highlights

- We propose the novel DGIB framework for robust dynamic graph representation learning. To the best of our knowledge, this is the first exploration to extend IB on dynamic graphs with structures directly involved in the IB optimization.

- We investigate a new insight and propose the *Minimal-Sufficient-Consensual (MSC)* Condition, which can be satisfied by the cooperation of both DGIB_{MS} and DGIB_{C} channels to refine the spatio-temporal information flow for feature compression. We further introduce their variational bounds for optimization.

- Extensive experiments on both real-world and synthetic dynamic graph datasets demonstrate the superior robustness of our DGIB against targeted and non-targeted adversarial attacks compared with state-of-the-art baselines.

Future Works

- Extend DGIB to more downstream graph tasks, such as node classification, graph classification, *etc.*
- Adapt DGIB to continuous dynamic graph scenarios for wider applications, such as in-time traffic prediction, anomaly detections, *etc.*
- Investigate the possibility of combining DGIB with up-to-date GNN backbones and architectures, such as graph Transformers, *etc.*
- Explore more efficiently-estimated IB variational bounds to decrease complexity.
- Construct a uniform framework to handle both static and dynamic graph robust representation learning.

THE WEB CONFERENCE ACM | 北京航空航天大學 BEIHANG UNIVERSITY WITH

Dynamic Graph Information Bottleneck

Haonan Yuan
School of Computer Science and Engineering, BDBC,
Beihang University
Beijing, China
yuanhn@buaa.edu.cn

Qingyun Sun*
School of Computer Science and Engineering, BDBC,
Beihang University
Beijing, China
sunqy@buaa.edu.cn

Xingcheng Fu
Key Lab of Education Blockchain and Intelligent Technology,
Guangxi Normal University
Guilin, Guangxi, China
fuxc@gxnu.edu.cn

Cheng Ji
School of Computer Science and Engineering, BDBC,
Beihang University
Beijing, China
jicheng@buaa.edu.cn

Jianxin Li
School of Computer Science and Engineering, BDBC,
Beihang University
Beijing, China
lijx@buaa.edu.cn

*Corresponding author.

Paper Code Scholar