Low-Distortion Graph Representation Learning: An Information-Theoretic Perspective

Qingyun Sun

sunqy@buaa.edu.cn

Beihang University, Beijing, China

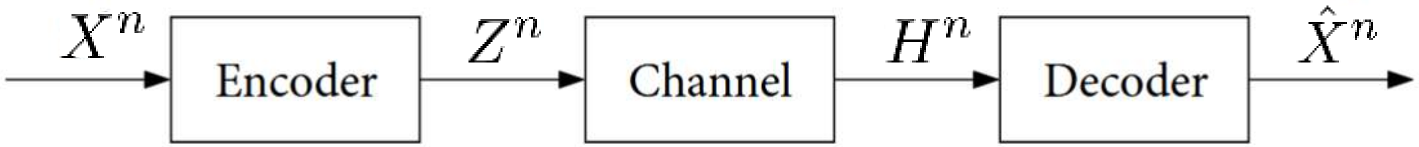
Outline:

- **Why Information Theory for Graph Learning?**
- How to Capture and Leverage Information in Graph?
- What’s Next? Future Directions of Information-Theoretic GRL

Information Theory

Claude Elwood Shannon
(1916-2001)

Shannon's Model of a Communication System (1948)



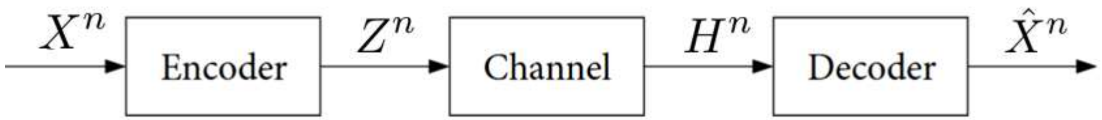
- A *k*-symbol sequence X is mapped by anencoder into an n-symbol input sequence Z
- The received channel output sequence H ismapped by a decoder into an estimate (reconstruction) sequence $\hat{X}$

Information can be efficiently compressed and transmitted using codes, laying the foundation of Information Theory

1. Shannon C E. A mathematical theory of communication[J]. The Bell system technical journal, 1948, 27(3): 379-423.

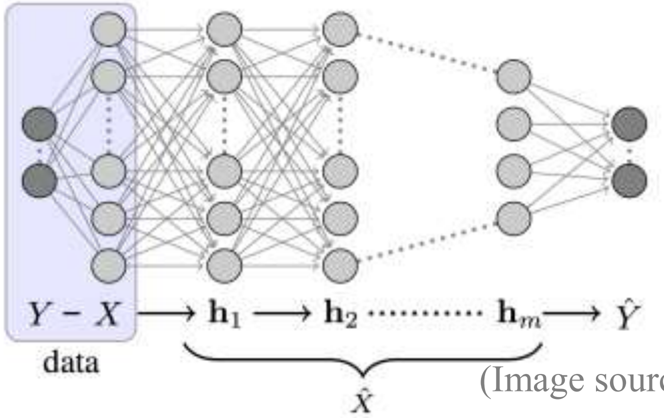
A Shared Goal: Minimal Distortion

Shannon's Model of a Communication System



Information Theory seeks to encode and decode signals with **minimal distortion.**

Model of Deep Learning



(Image source: Tishby and Zaslavsky, 2015)

Deep Learning similarly aims to extract information from data while preserving it with **minimal distortion**.

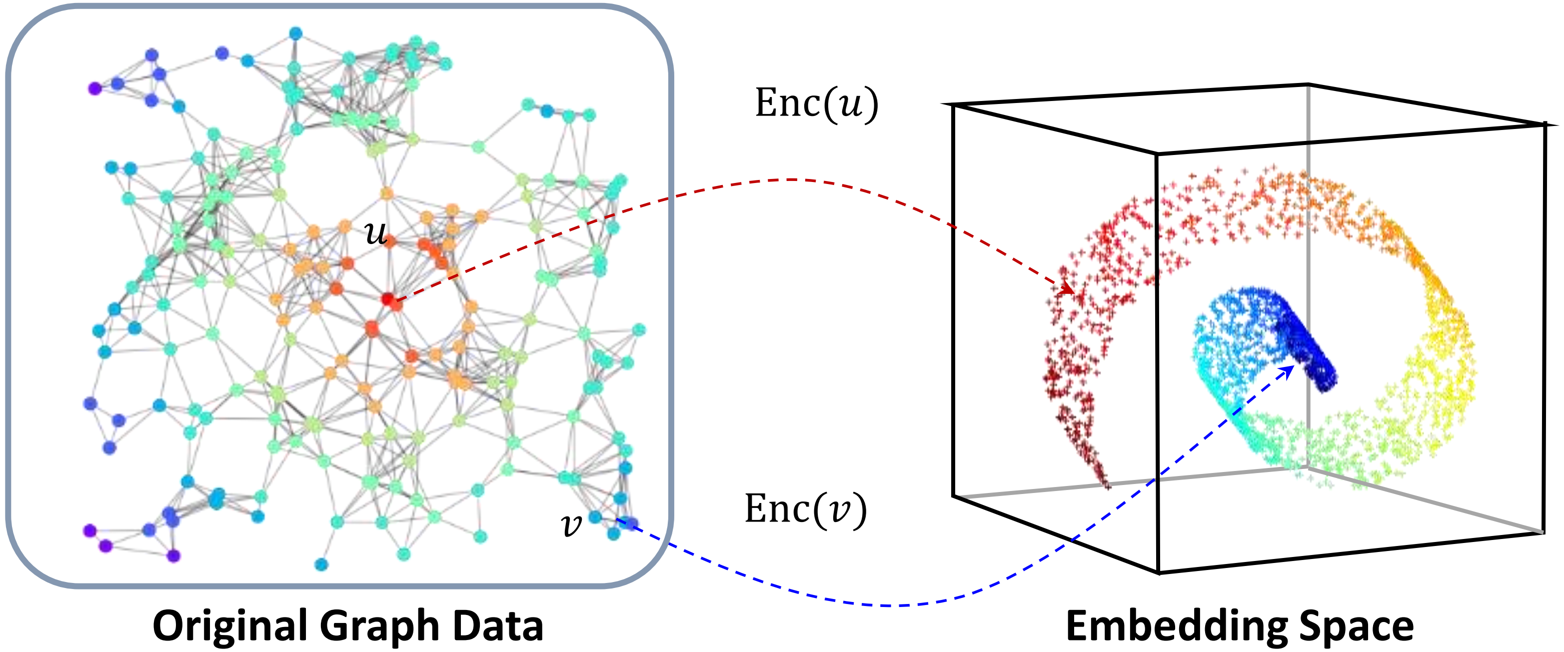
Shared Goal:

Compress the input, preserve the core information, and minimize distortion

1. Shannon C E. A mathematical theory of communication[J]. The Bell system technical journal, 1948, 27(3): 379-423.

Information Theory Meets Graph Learning

- Compressing graph data with rich structure and dependencies into embedding vectors **inevitably introduces distortion**.
- Information Theory provides a principled way to **measure, compress, and preserve information** in graph data.



How can we achieve low-distortion graph learning?

Let’s **explore** patterns and **draw inspiration from an Information-Theoretic Perspective**!

Outline:

- Why Information Theory for Graph Learning?
- **How to Capture and Leverage Information in Graph?**
- What's Next? Future Directions of Information-Theoretic GRL

How to Capture and Utilize Information?

- Entropy

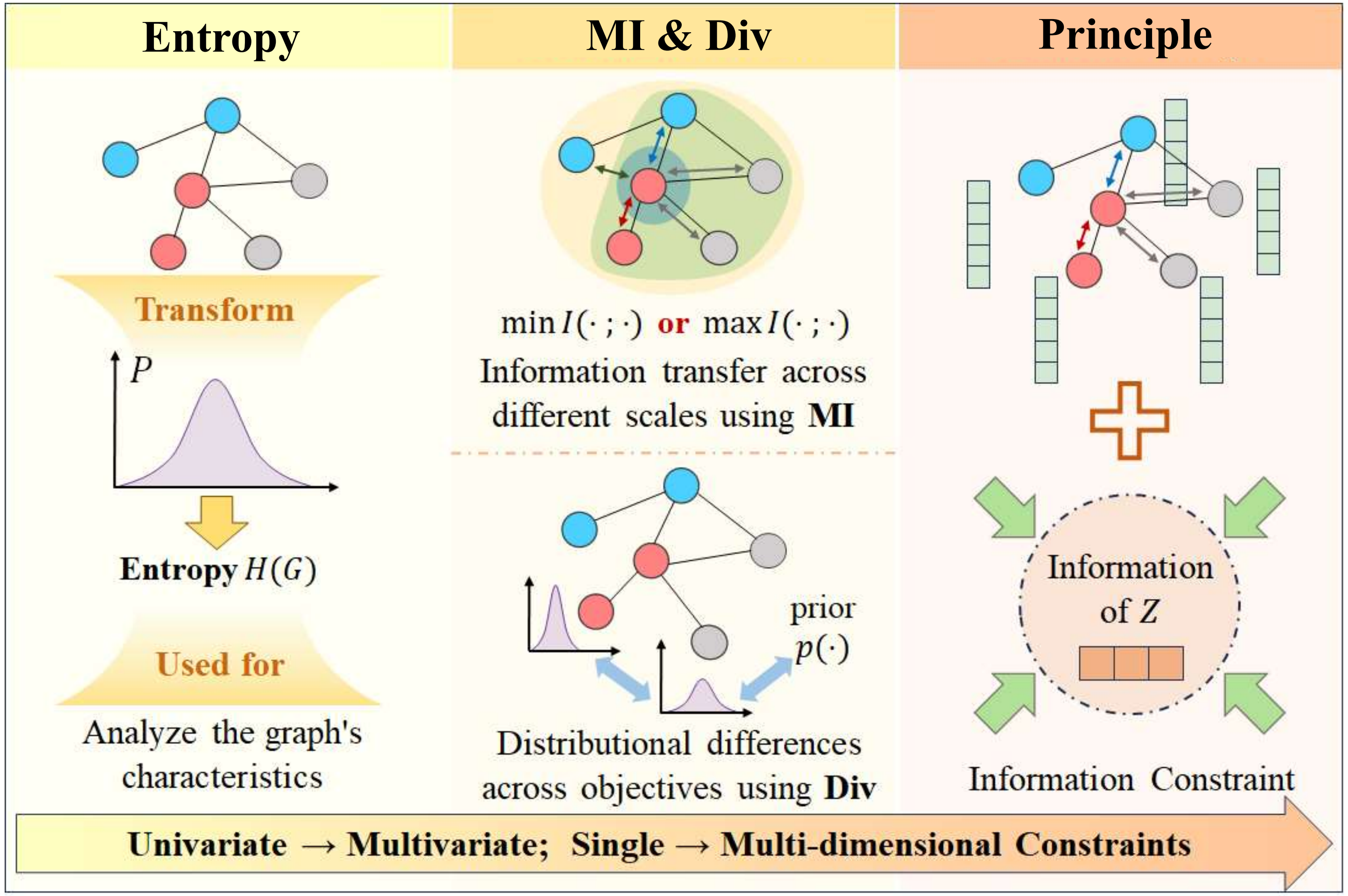
Assess the intrinsic uncertainty and complexity of graph.

- MI & Divergence

Capture both interdependencies and variations inherent in learning.

- Principle

Offer a unified and general objective for representation learning



How to Capture and Utilize Information?

- Entropy

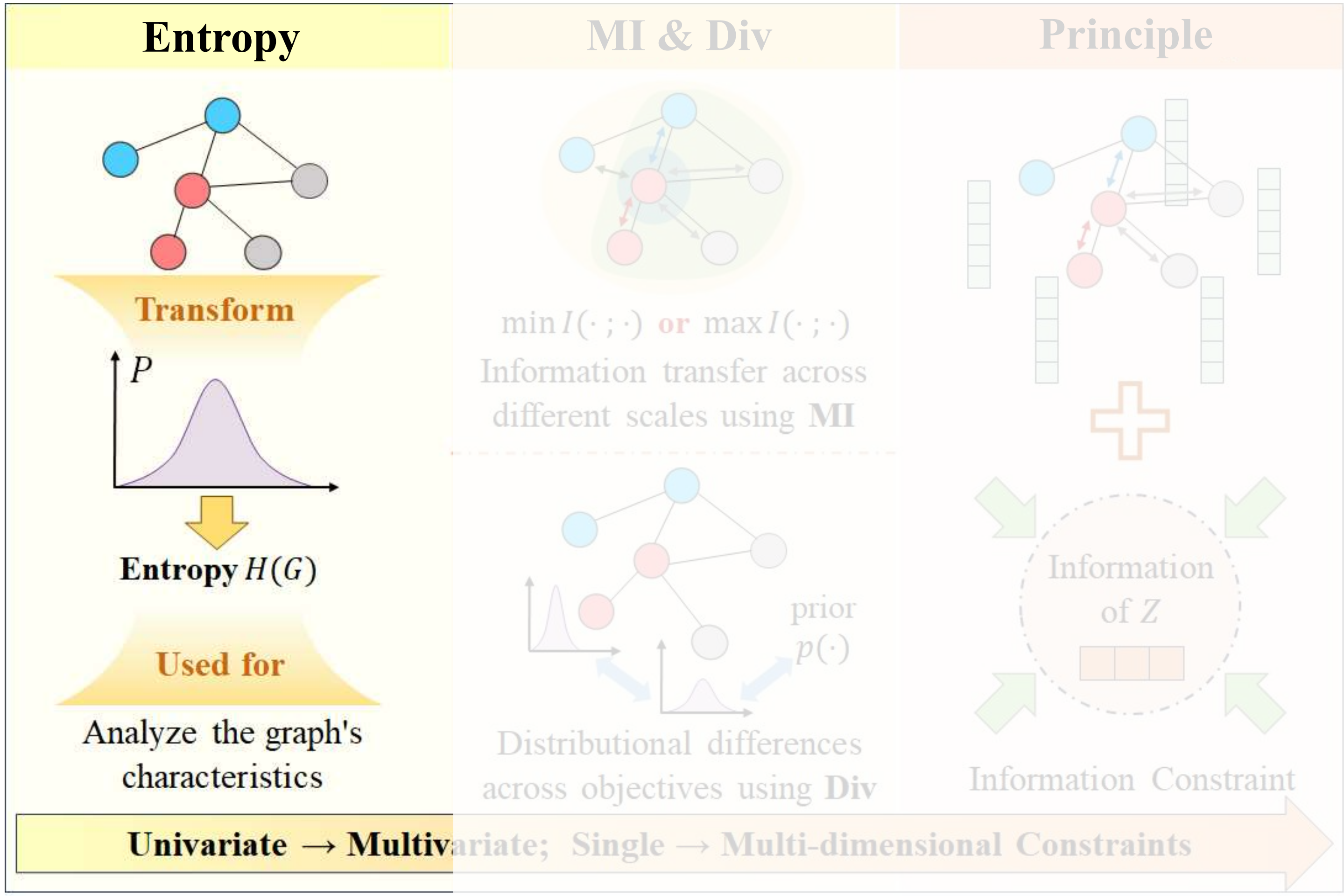
Assess the intrinsic uncertainty and complexity of graph.

- MI & Divergence

Capture both interdependencies and variations inherent in learning.

- Principle

Offer a unified and general objective for representation learning



Entropy: Assess intrinsic uncertainty and complexity

Classical Entropy

Shannon Entropy $H_S(P) = -\sum_i p_i \log p_i$

Rényi Entropy $H_R(P) = \frac{1}{1-q}\sum_i \log\left(\sum_i p_i^q\right)$

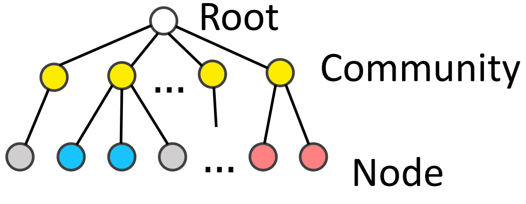
von Neumann Entropy $H_{vN}(\rho) = -\mathrm{Tr}(\rho \log \rho)$

- Probability-based
- Symmetry (i.i.d. data)
- Global information

Entropy

Graph-Specific Entropy

Structure Entropy (1-D, 2-D, ...)



$$H^1(G) = \sum_i \frac{d_i}{\mathrm{vol}(G)} \log \frac{d_i}{\mathrm{vol}(G)}$$

von Neumann Graph Entropy

$$H_{vN}(G) = \sum_i \frac{\lambda_i}{\mathrm{vol}(G)} \log \frac{\lambda_i}{\mathrm{vol}(G)}$$

- Structure-oriented
- Dependency (non i.i.d)
- Relation information

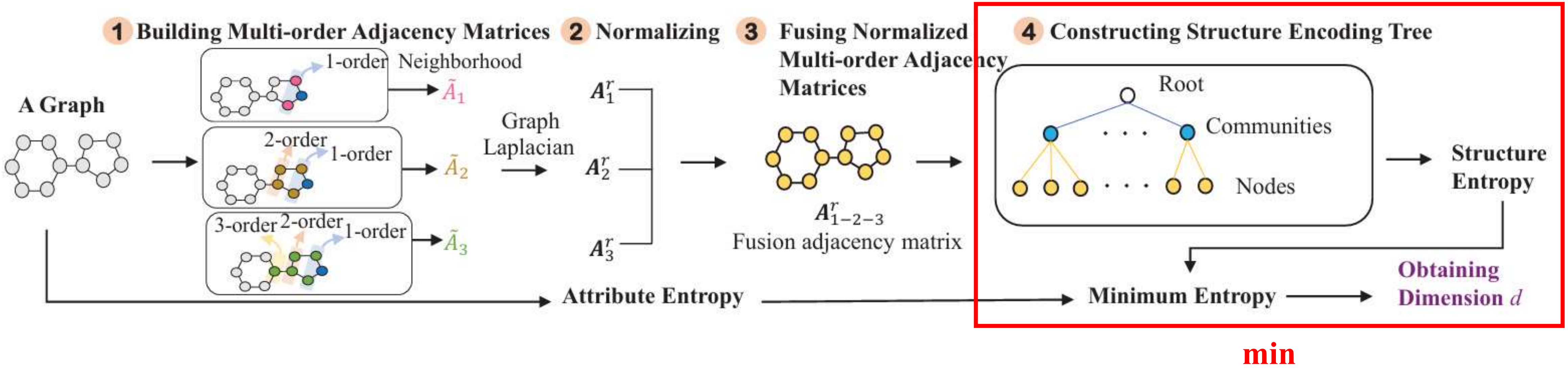
Entropy: Assess intrinsic uncertainty and complexity

	Entropy	Graph Structure Component
Classical Entropy	Shannon Entropy	——
	Rényi's q-order Entropy	——
	von Neumann Entropy	——
Graph-Specific Entropy	von Neumann Graph Entropy	Laplacian matrix eigenvalues
	1-D Structure Entropy	Node degree
	2-D Structure Entropy	Graph node partition
	Edge Entropy	Class of two nodes on the edge
	Körner Graph Entropy	Independent sets
	Residual Entropy	Graph node partition

Leveraging graph structure is key to designing graph-specific entropy

Entropy: Assess intrinsic uncertainty and complexity

- **Undertanding Graph Data: Node Embedding Dimension Selection**
  - One of the most fundamental setting method design: **Embedding Dimension**.
  - Optimal dimension? **Minimum Entropy!**
  - **MGEDE's Idea**[1]: min attribute/structure entropy → **min uncertainty** → **optimal dimension**

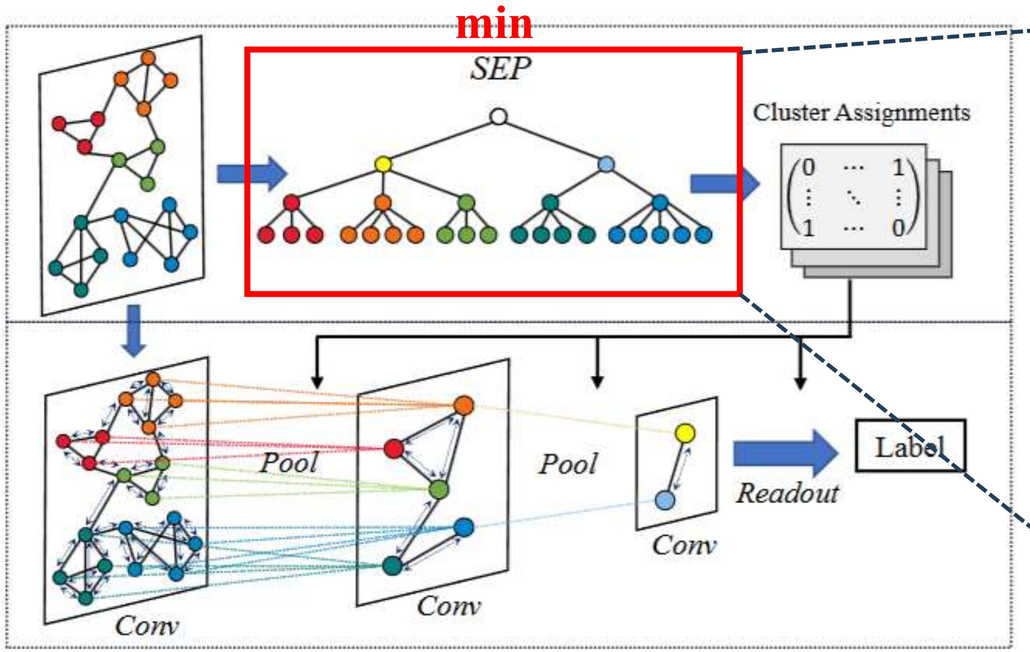


1. Yang Z, Zhang G, Wu J, et al. Minimum entropy principle guided graph neural networks, WSDM 2023.

Entropy: Assess intrinsic uncertainty and complexity

- **Undertanding Graph Data: Hierarchical Information Extraction**

  - MGEDE only uses global information, while SEP[1] takes **hierarchical** information into account
  - 2-D Structure Entropy has an inherent hierarchical structure in the calculation
  - **SEP's Idea:** min structure entropy → **optimal** coding tree → **hierarchical** relationships → pooling



How to optimize structure entropy?
3 steps, greedy algorithm:

1. **MERGE**: Bottom to top construction
2. **REMOVE**: Compress tree to the certain height
3. **FILL**: Avoid cross-layer links.

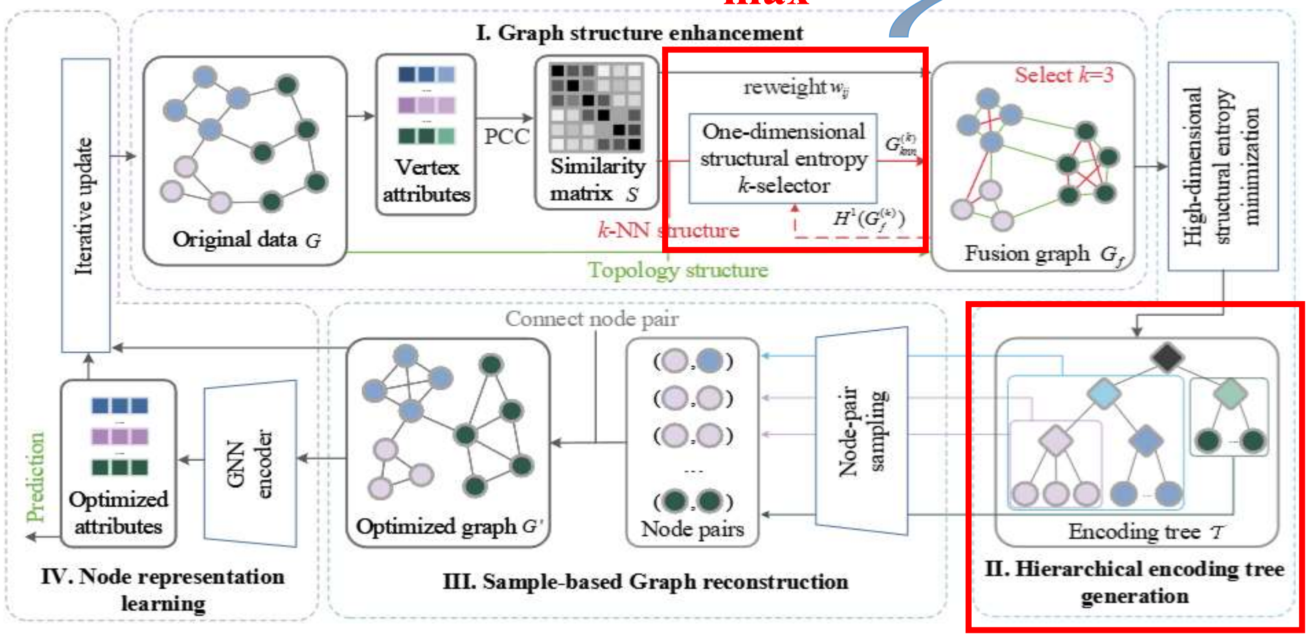
1. Wu J, Chen X, Xu K, et al. Structural entropy guided graph hierarchical pooling, ICML 2022

Entropy: Assess intrinsic uncertainty and complexity

- **Undertanding Graph Data: The entropy is not always minimized**
  - ➢ **SE-GSL's Idea**: **max** entropy for **richer** information while **min** entropy for **ordered** hierarchy

build a graph using KNN

max



min

generate coding tree

max 1-D structure entropy: guide the selection of k for **larger encoding information**

offer more information

extract meaningful information

min 2-D structure entropy: gain optimal **hierarchical** relationships

1. Zou D, Peng H, Huang X, et al. Se-gsl: A general and effective graph structure learning framework through structural entropy optimization, WWW 2023

How to Capture and Utilize Information?

- Entropy

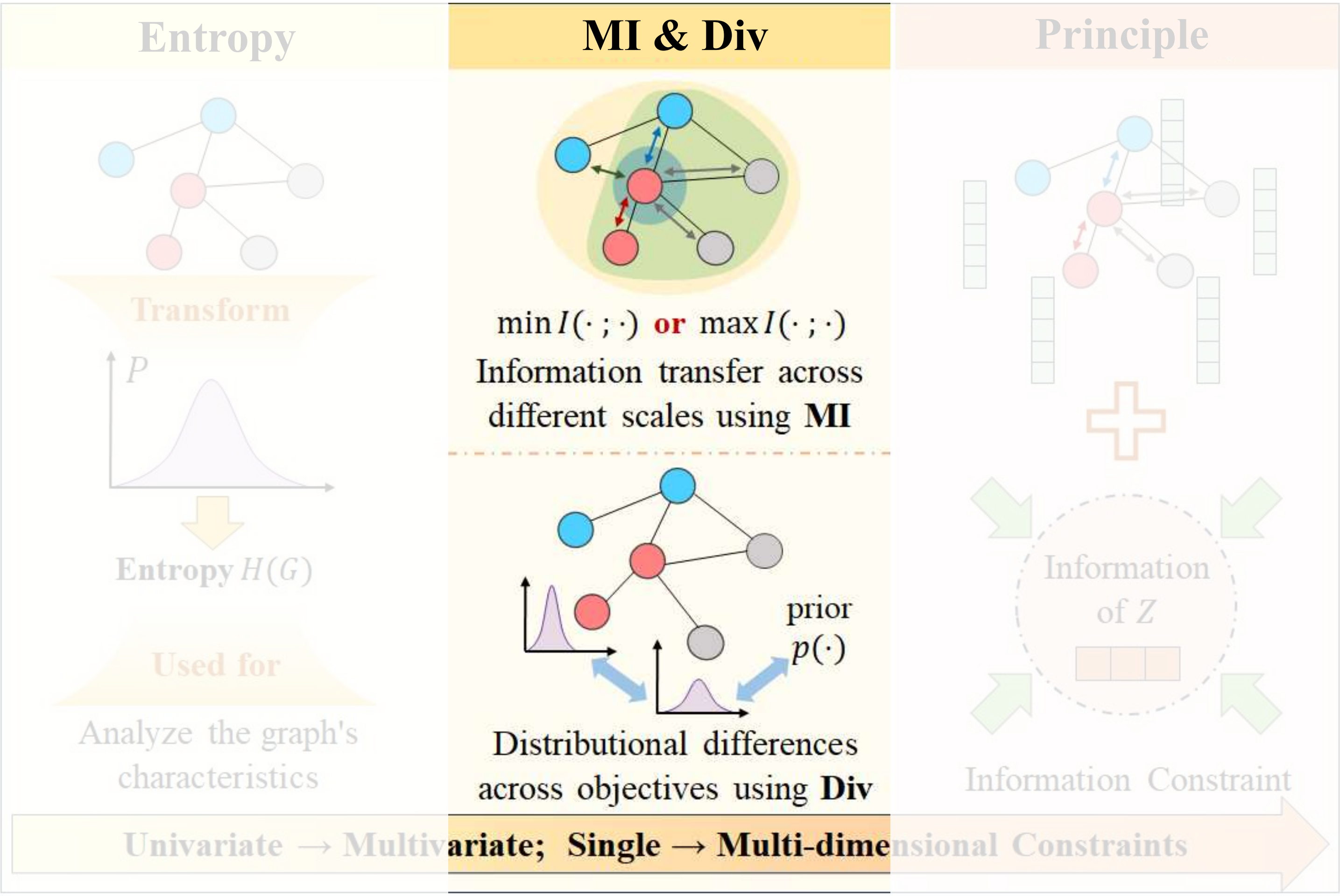
Assess the intrinsic uncertainty and complexity of graph.

- MI & Divergence

Capture both interdependencies and variations inherent in learning.

- Principle

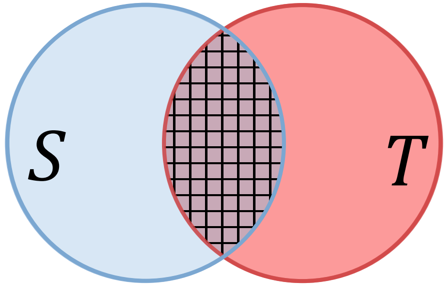
Offer a unified and general objective for representation learning



MI & Divergence

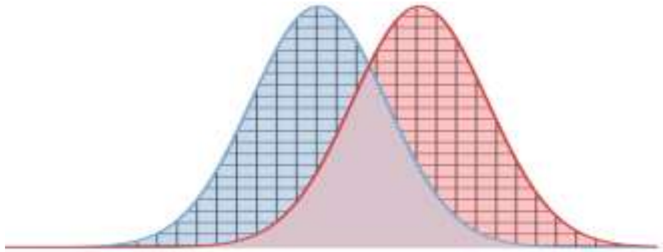
- **Mutual Information (MI): quantifying the amount of information transmitted**
- **Divergence: measuring distribution differences**

Mutual Information



$$I(S,T) = \int_S \int_T f(s,t) \log \frac{f(s,t)}{f(s)f(t)} dsdt$$

Divergence

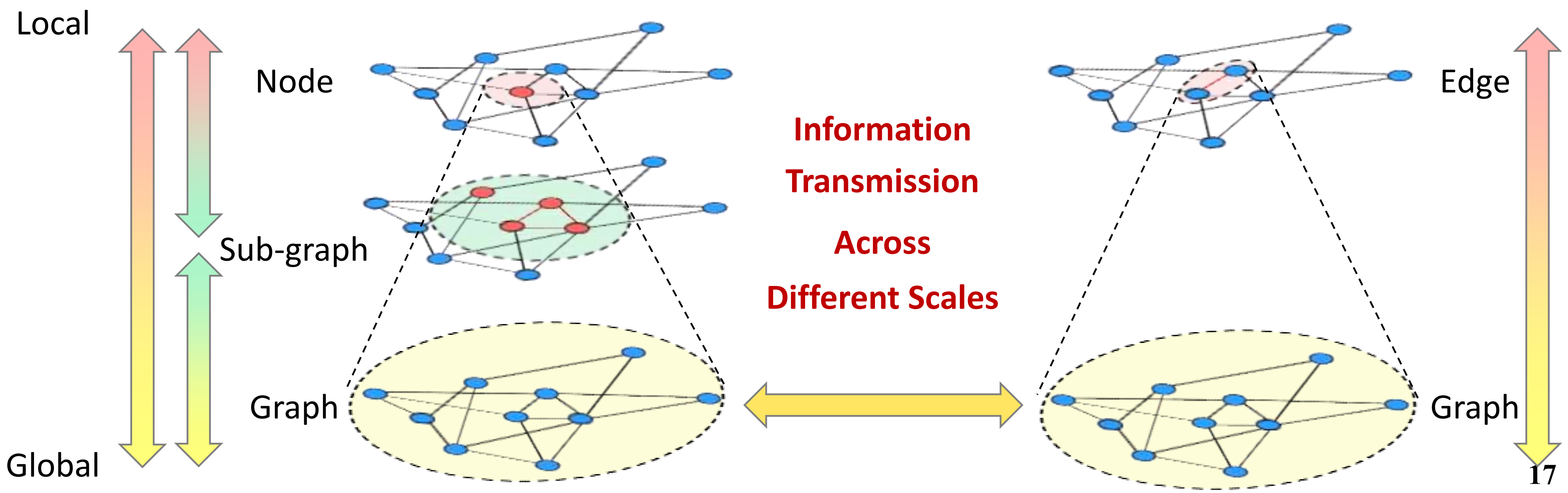


$$D_{KL}(f||g) = \int_T f(t) \log \frac{f(t)}{g(t)} dt$$

$$I(S,T) = D_{KL}\big(f(s,t)||f(s)f(t)\big)$$

MI for Information Transmission

- Message passing is a fundamental paradigm in graph learning, where **controlling information flow** is the key.
- Information can be filtered and compressed by capturing the information flow among different views (**local ↔ global**)



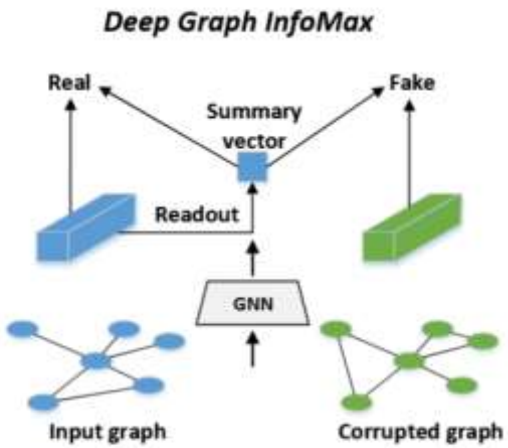
MI for Information Transmission

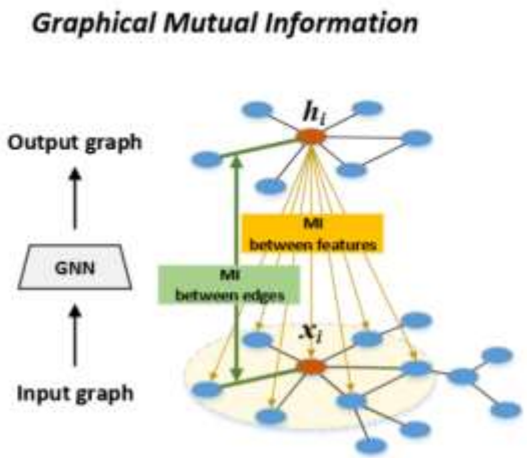
- **MI for Information Transmission: local ↔ global**

➢ **DGI**[1]: Maximize MI between local node representations and the global graph representation

➢ By **aligning local and global views**, DGI learns embeddings that preserve rich structural and feature information **without supervision.**

➢ **DMI**[2]: introducing **Feature** Mutual Information and **Topology-Aware** Mutual Information → **comprehensively capture the information in graph**





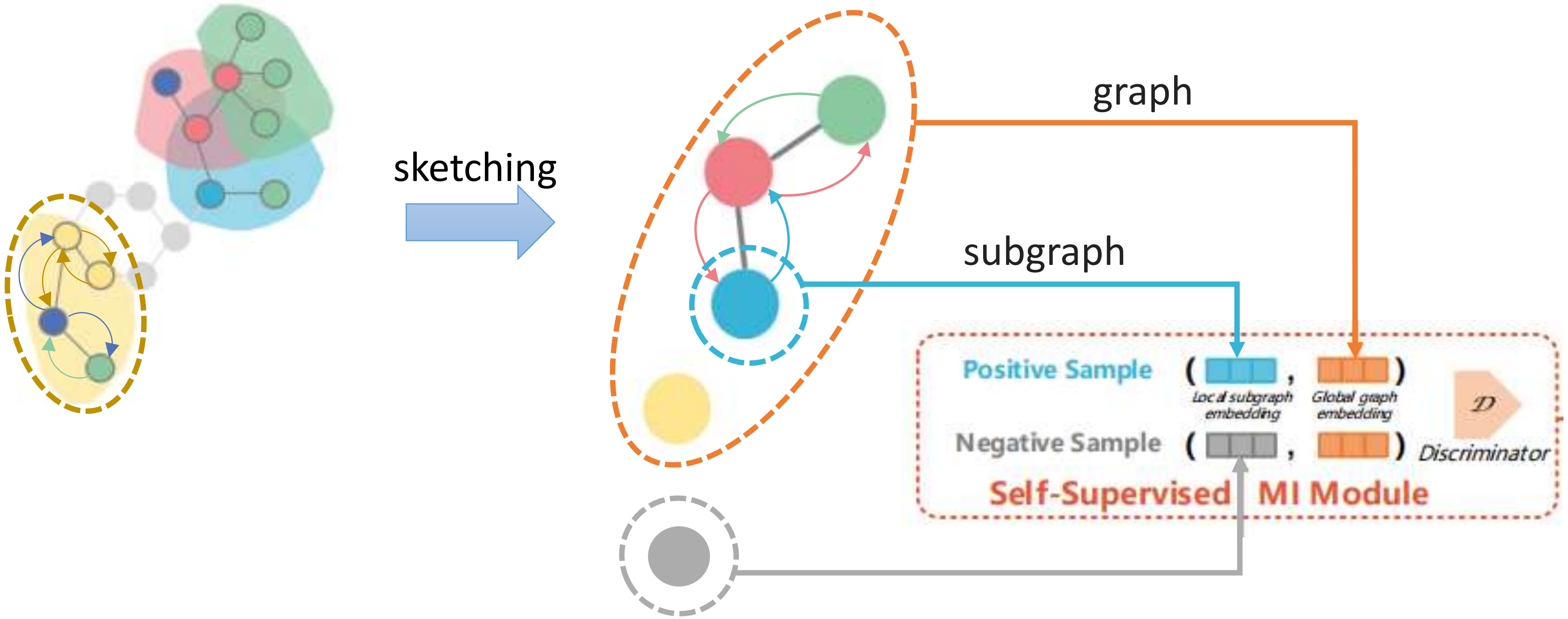
1. Veličković P, Fedus W, Hamilton W L, et al. Deep Graph Infomax, ICLR 2018.
2. Peng Z, Huang W, Luo M, et al. Graph representation learning via graphical mutual information maximization, WWW 2020.

MI for Information Transmission

- **MI for Information Transmission: local ↔ global**

➢ Considering **higher-scale** information transmission **(subgraph ↔ graph)**, SUGAR[1] provided the answer.

➢ SUGAR adaptively selects **critical subgraphs** and encourage **subgraph** representations to **preserve global properties** by maximizing MI between subgraphs and the global graph.



1. Sun Q, Li J, Peng H, et al. Sugar: Subgraph neural network with reinforcement pooling and self-supervised mutual information mechanism, WWW 2021.

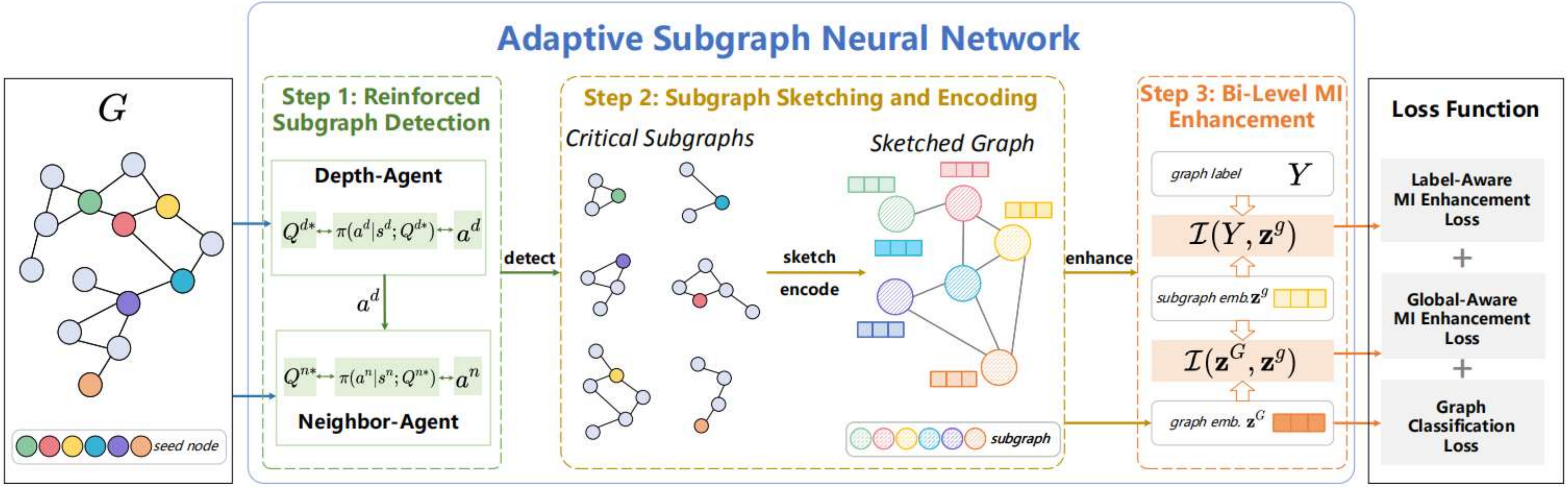
MI for Information Transmission

- **MI for Information Transmission: local ↔ global**

➢ AdaSNN[1] further extends SUGAR[1] into a **Bi-level MI Enhancement mechanism**.

➢ **Maximize MI between subgraphs and graph**: ensuring subgraph capture comprehensive structural context.

➢ **Maximize MI between subgraphs and labels**: injecting discriminative power into subgraph representations.



1. Li J, Sun Q, Peng H, et al. Adaptive subgraph neural network with reinforced critical structure mining, TPAMI 2023

MI for Information Transmission

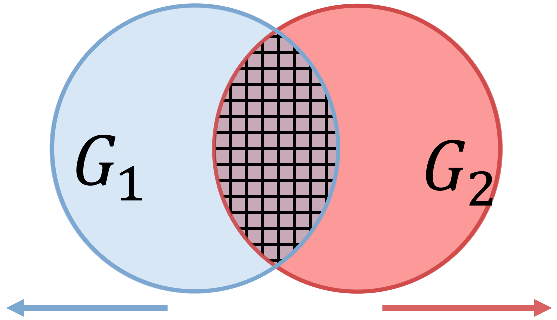
- **MI for Information Transmission: always max MI?**

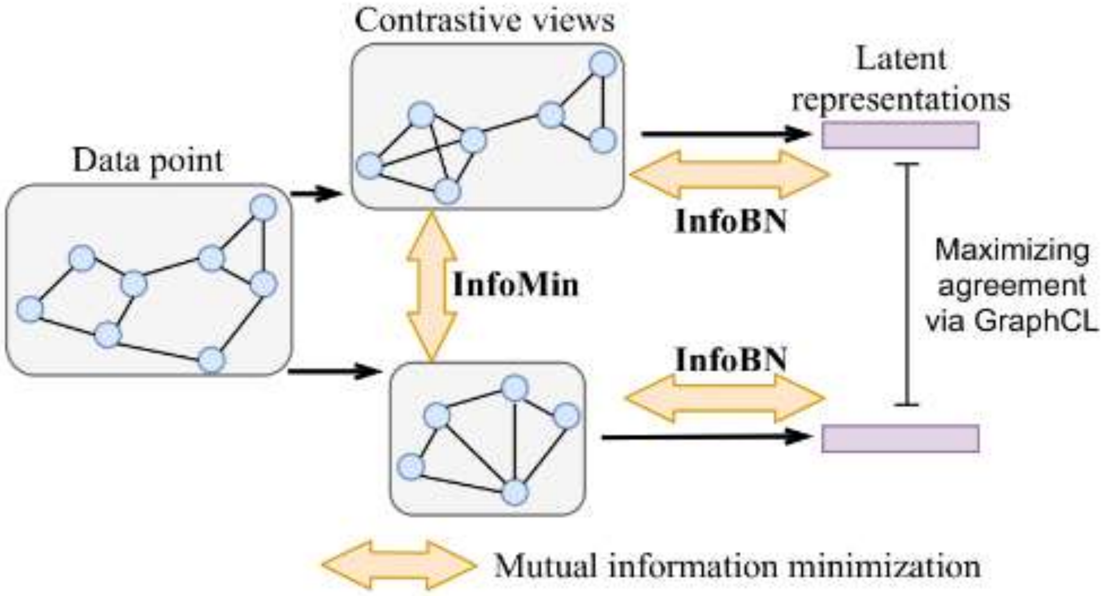
➢ While maximizing MI maintains consistency across views, doing so alone often leads to **homogeneous**, collapsed representations.

➢ **GraphCL's Idea**[1]: Introducing a **minimization MI** step between views helps **preserve diversity** by avoiding overly similar multi-view embeddings.

$\min I(G_1, G_2) \rightarrow$ low information

overlap between $G_1, G_2 \rightarrow$ diversity





1. You Y, Chen T, Wang Z, et al. Bringing your own view: Graph contrastive learning without prefabricated data augmentations, WSDM 2022

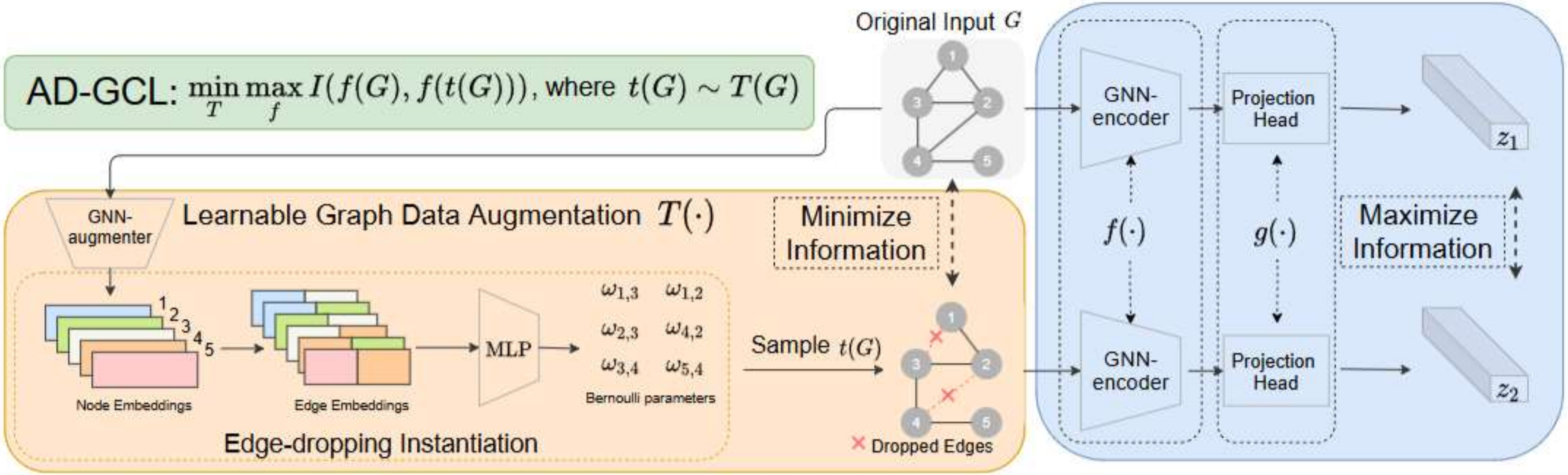
MI for Information Transmission

- **MI for Information Transmission: always max MI?**

➢ How to trade off between consistency and diversity?

➢ **AD-GCL's Idea**[1]: adopting an **adversarial min–max mutual information scheme**.

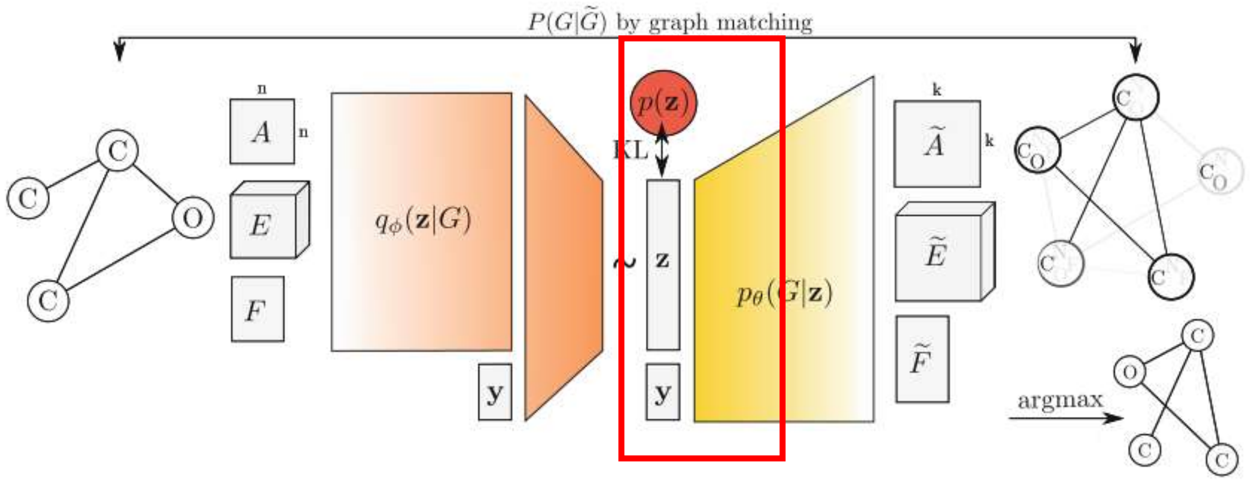
➢ It **maximizes MI** between the original graph and its augmented view (**to preserve relevant information**), while simultaneously **minimizing MI** between trivial or redundant augmentations (**to discourage collapse and redundancy**).



1. Suresh S, Li P, Hao C, et al. Adversarial graph augmentation to improve graph contrastive learning, NeurIPS 2021

Divergence for Distribution Learning

- Divergence measures the distance between two distributions.
- Divergence is always used to **enforce** the **latent distribution** to approximate **prior distribution in deep learning**.
- **GraphVAE**[1]: Divergence regularization shapes latent space, preserving information while enabling diverse, realistic graph generation.



GraphVAE[1]:

$$\mathcal{L}(\phi,\theta;G) = \\ = \mathbb{E}_{q_\phi(\mathbf{z}|G)}[-\log p_\theta(G|\mathbf{z})] + \mathrm{KL}[q_\phi(\mathbf{z}|G)||p(\mathbf{z})]$$

$q_\phi(\mathbf{z}|G)$: obtained through the model

$p(\mathbf{z})$: unknown, usually is *N(0,I)*

1. Simonovsky M, Komodakis N. Graphvae: Towards generation of small graphs using variational autoencoders, ICANN, 2018.

How to Capture and Utilize Information?

- Entropy

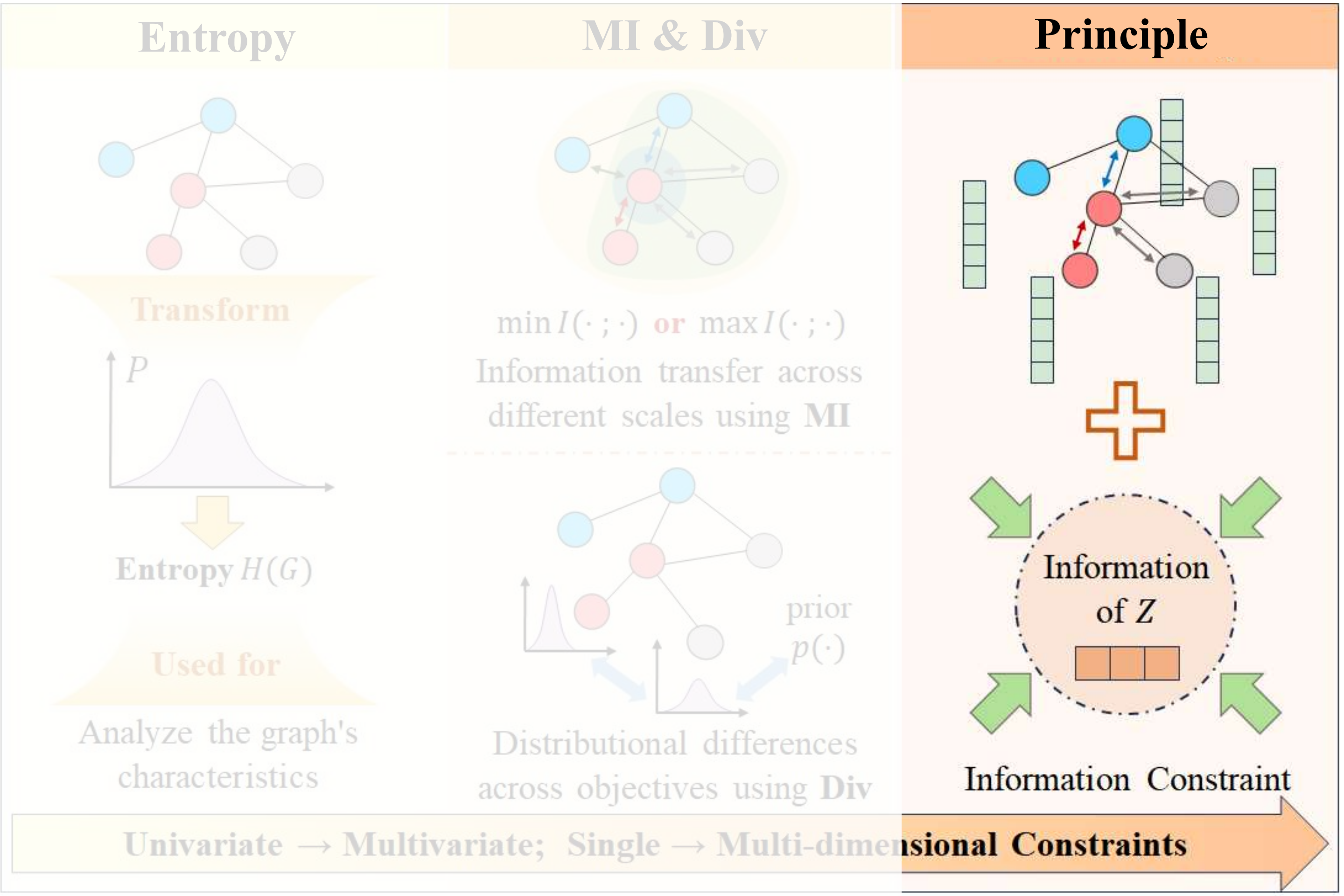
Assess the intrinsic uncertainty and complexity of graph.

- MI & Divergence

Capture both interdependencies and variations inherent in learning.

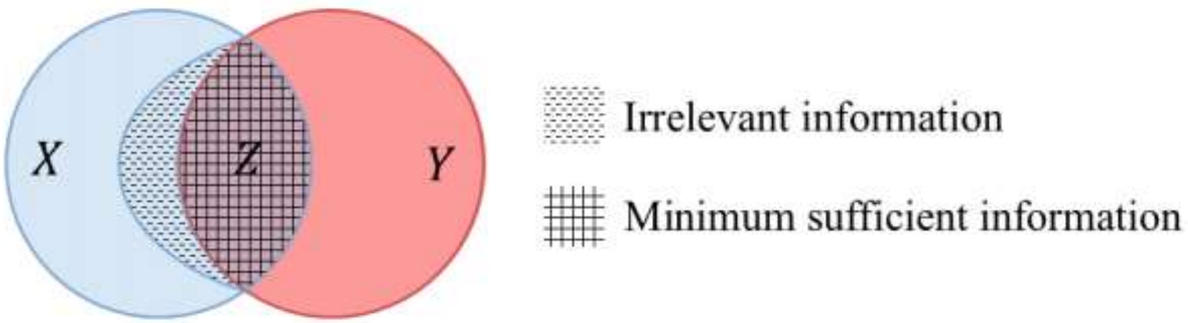
- Principle

Offer a unified and general objective for representation learning



Principled Graph Learning

- Combining Information-Theoretic principles offers a holistic strategy for developing advanced graph learning models.
- The most popular principle is **Information Bottleneck (IB)**
- IB explains representation learning as a **trade-off: retain task-relevant information while compressing irrelevant information**.



IB Objective:

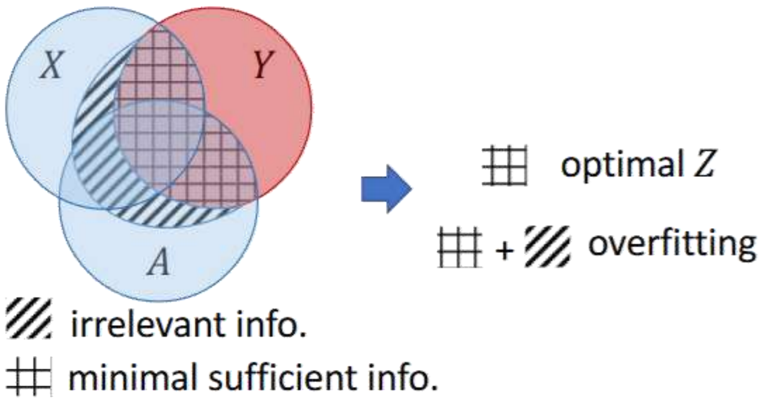
$$\min -I(Z;Y) + \beta I(Z;X)$$

Prediction term   Compression term

- $I(Z;Y)$: Efficient task-relevant information
- $I(Z;X)$: Minimal irrelevant information

IB Extentions to Graph Data

- **Graph Information Bottleneck (GIB)**[1]: the first work to introduce IB into graph learning.
- **Addressing Graph-Specific Challenges**: GIB assumes local dependency and formulates a tractable search space via a Markov chain to hierarchically extract information from structure and features.
- GIB significantly increase robustness against adversarial attacks on both graph structure and node features.



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

Graph Information Bottleneck:

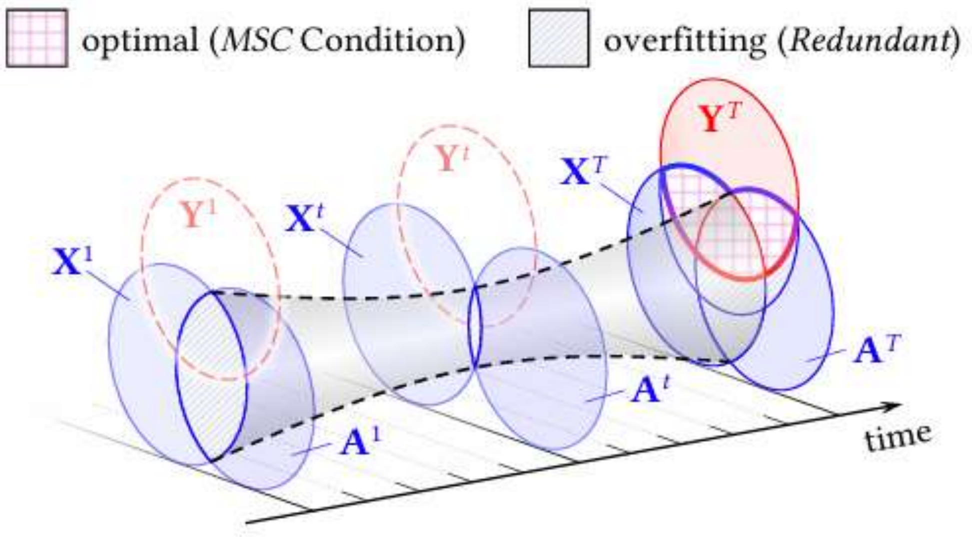
$$\min_{\mathbb{P}(Z|\mathcal{D})\in\Omega} \mathrm{GIB}_\beta(\mathcal{D}, Y; Z) \triangleq [-I(Y;Z) + \beta I(\mathcal{D};Z)]$$

1. Wu T, Ren H, Li P, et al. Graph information bottleneck, NeurIPS 2020

IB Extentions to Graph Data

More IB Extension: Data - Space - Method

- Dynamic GIB (DGIB)[1] extends IB to dynamic graphs, directs and refines the information flow passing through graph snapshots.
- DGIB aims to extract **Minimum** & **Sufficient** & **Consensual** representation.



(a) Illustration of the DGIB principle.

DGIB Objective:

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

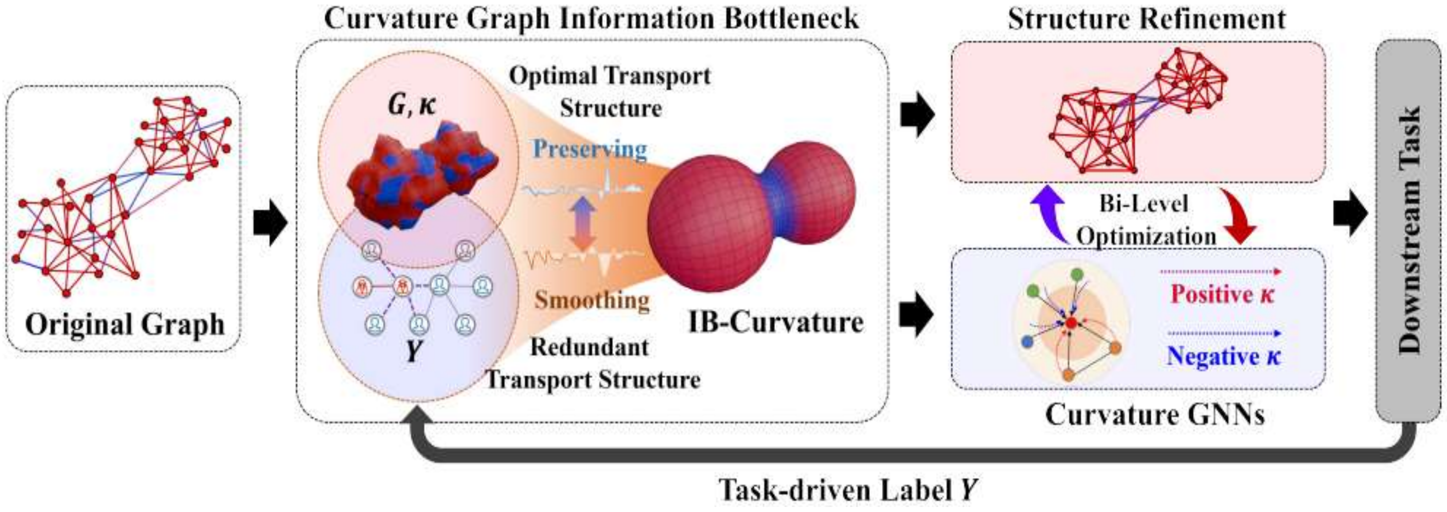
Consensual: constrains

1. Yuan H, Sun Q, Fu X, et al. Dynamic graph information bottleneck, WWW 2024.

IB Extentions to Graph Data

More IB Extension: Data - Space - Method

- CurvGIB[1]: From Information Bottleneck → to Geometry-Aware Bottleneck.
- CurvGIB introducing discrete curvature to guide the information compression along with the optimal transport structure, aiming to extract information from a more suitable embedding space.



CurvIB Objective:

$$\mathbf{Z}_{\mathrm{IB}}, \kappa_{\mathrm{IB}} = \arg\min_{\mathbf{Z},\kappa} \mathrm{CurvGIB}(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \kappa)$$
$$\triangleq \arg\min_{\mathbf{Z},\kappa} [-I(\mathbf{Z} \mid \kappa; \mathbf{Y}) + \beta I(\mathbf{Z} \mid \kappa; \mathbf{X})],$$

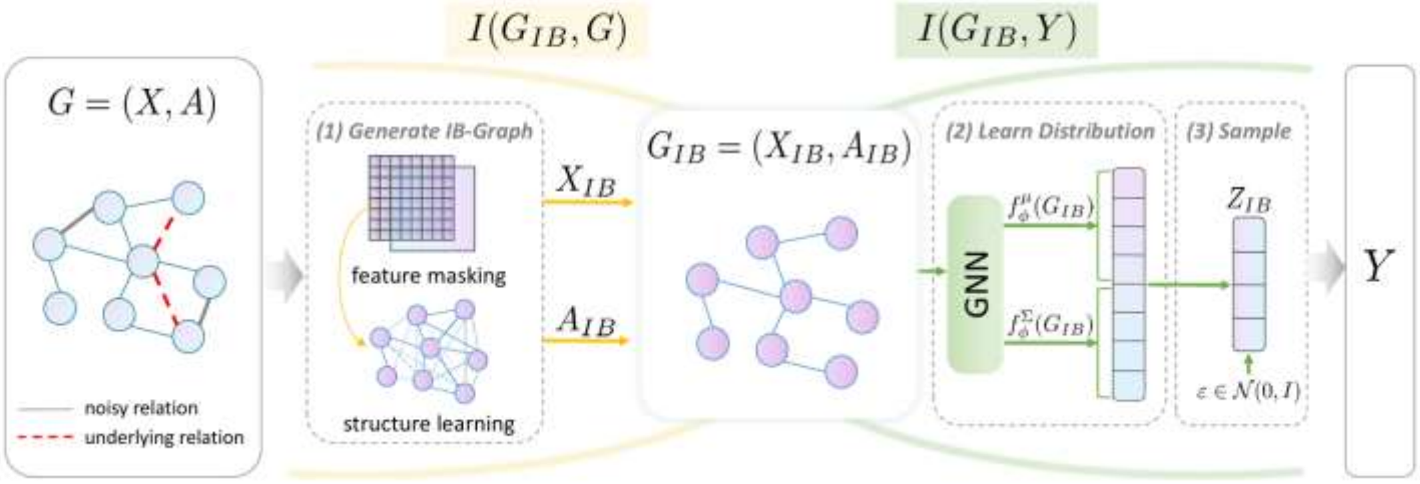
Preserving — Compression (Smoothing)

1. Fu X, Wang J, Gao Y, et al. Discrete curvature graph information bottleneck, AAAI 2025

IB Extentions to Graph Data

More IB Extension: Data - Space - Method

- VIB-GSL[1]: Apply Information Bottleneck (IB) to noisy, incomplete, or spurious graph structure.
- VIB-GSL uses variational approximation to provide a tractable bound to optimize adjacency matrix for task-relevant graph.



VIB-GSL Objective:

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

Preserve task-relevant structure

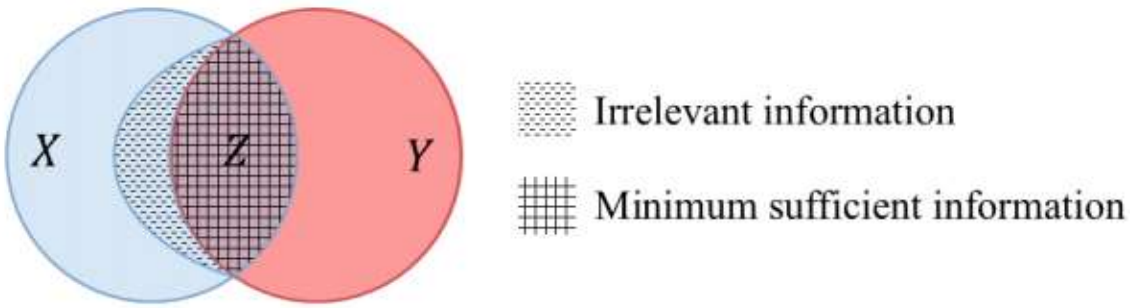
Remove irrelevant structure

1. Sun Q, Li J, Peng H, et al. Graph structure learning with variational information bottleneck, AAAI 2022

IB vs. PRI: Two Views of Information

- However, the IB strongly depends on Y. What if we want to extract and compress information solely from G?
- **PRI**: Does not require labels → takes the representation learning as a trade-off between information redundancy and preservation with respect to intrinsic data structure.
- **Viewpoint**: PRI = "label-free" extension of IB, focusing on relevance without supervision.

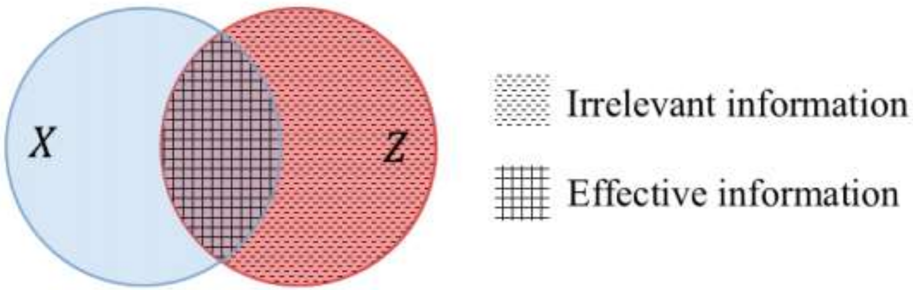
Information Bottleneck (IB):



$$\min -I(Z;Y) + \beta I(Z;X)$$

Sufficient Minimum

Principle of Relevant Information (PRI):



$$\min H(Z) + \beta D(Z||X)$$

Redundancy Discrepancy

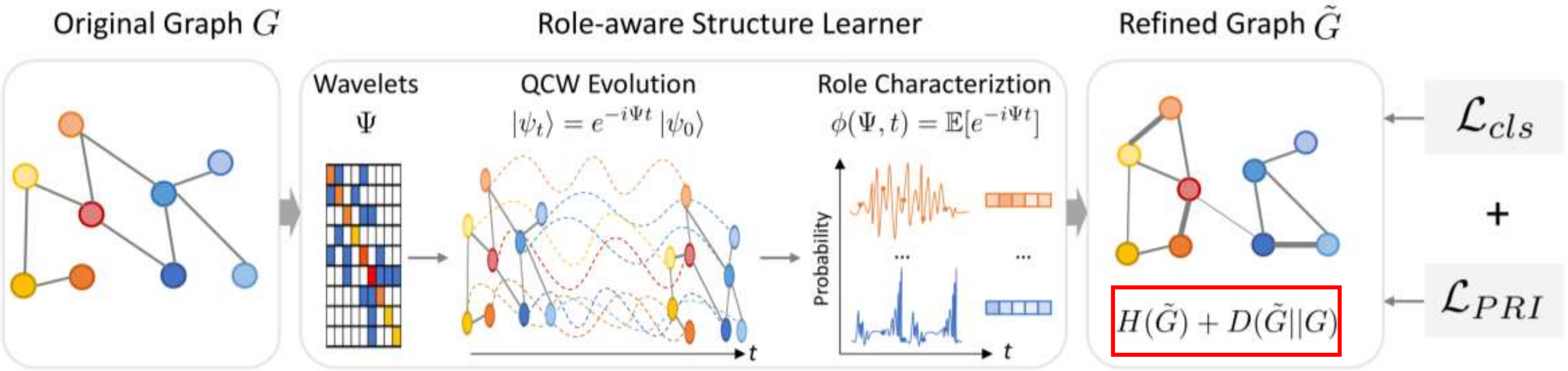
PRI for Graph Data

- PRI-GSL[1]: learn task-relevant graphs while preserving intrinsic self-organization patterns (clusters, communities).

PRI-GSL Objective: $\mathcal{L}_{\mathrm{PRI}} = H(\tilde{G}) + \beta D(\tilde{G}||G)$

Redundant Term: measure the disorder of graph

Distortion term: measure the discrepancy between two graphs



1. Sun Q, Li J, Yang B, et al. Self-organization preserved graph structure learning with principle of relevant information, AAAI 2023

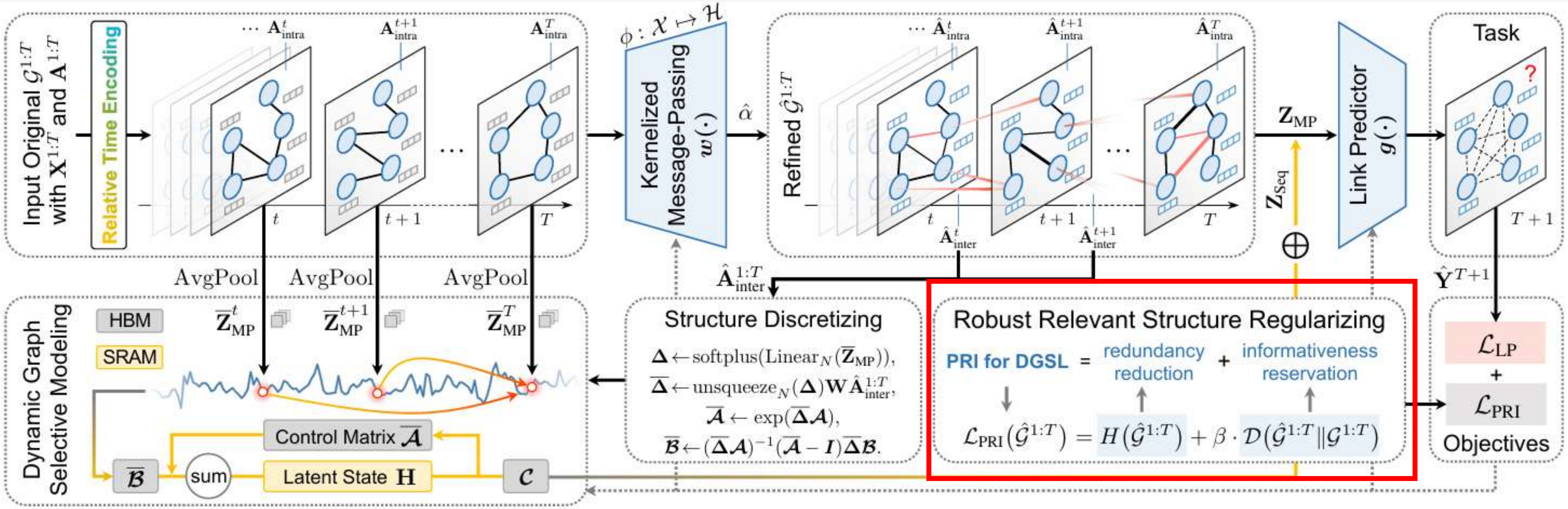
PRI for Graph Data

- DG-Mamba[1]: Even in more complex scenarios, effective information compression can still be achieved by imposing constraints on the dynamic graph formed from multiple graphs.

DG-Mamba Objective: $\mathcal{L}_{\mathrm{PRI}}(\hat{\mathcal{G}}^{1:T}) = H(\hat{\mathcal{G}}^{1:T}) + \beta \cdot \mathcal{D}(\hat{\mathcal{G}}^{1:T} \| \mathcal{G}^{1:T})$

Redundant Term: fltering out noise and redundant information

Distortion term: preserve discriminative and invariant temporal-spatial patterns



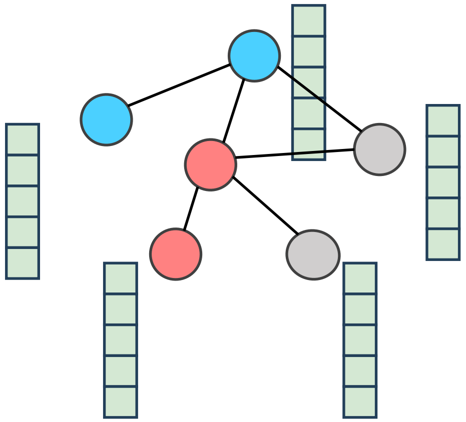
1. Yuan H, Sun Q, Wang Z, et al. Dg-mamba: Robust and efficient dynamic graph structure learning with selective state space models, AAAI 2025

Outline:

- Why Information Theory for Graph Learning?
- How to Capture and Leverage Information in Graph?
- **What's Next? Future Directions of Information-Theoretic GRL**

What's Next? Future Directions of Info-GRL

- How to Build a Graph-Specific Information-Theoretic Framework?



Graph Information

≠

Feature Information

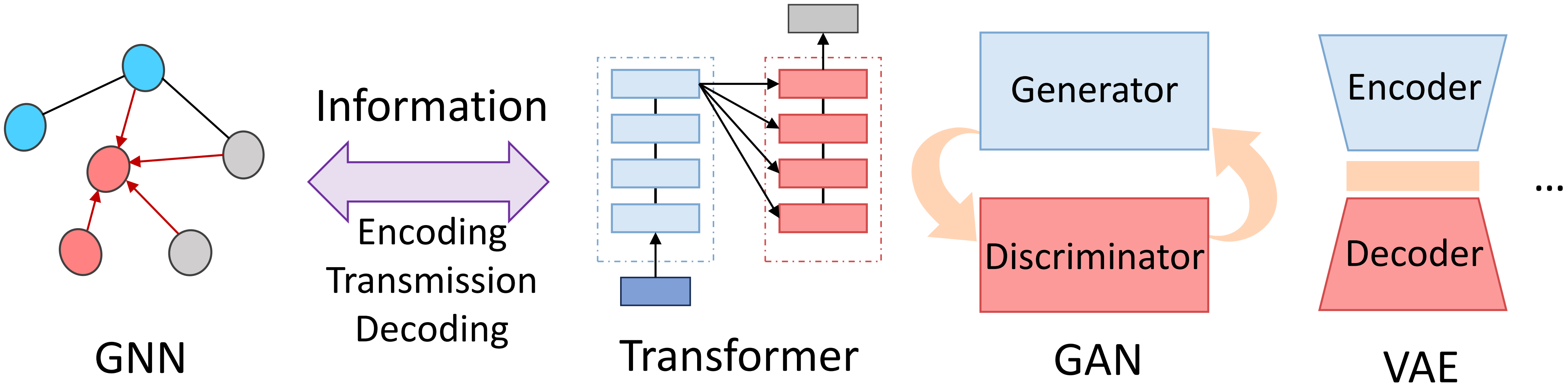
+

Structure Information

- **Graph-specific information Measures:** node/edge/subgraph entropy; context MI...
- **Graph Distortion Measures**: topology distortion (edges, motifs), feature distortion, spectral/cut distortion.
- **Low–Distortion Objectives**: minimal bits to encode a graph (or embeddings) under bounded structural + feature loss.
- **Evaluation Protocols**: predictiveness vs compression trade-offs; robustness under perturbations; OOD transfer.

What’s Next? Future Directions of Info-GRL

- How to bridge GNNs and alternative architectures from the Information-Theoretic perspective?

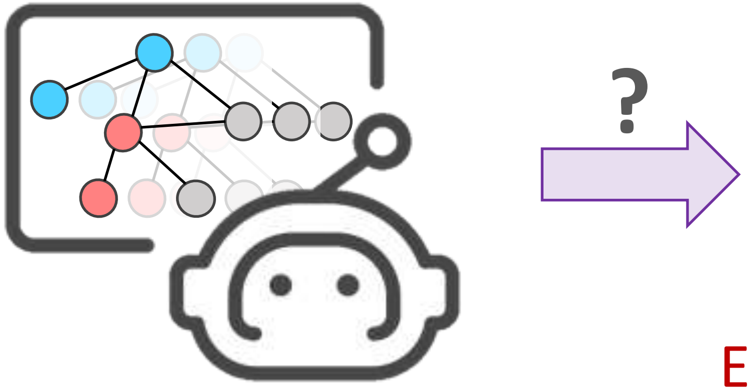


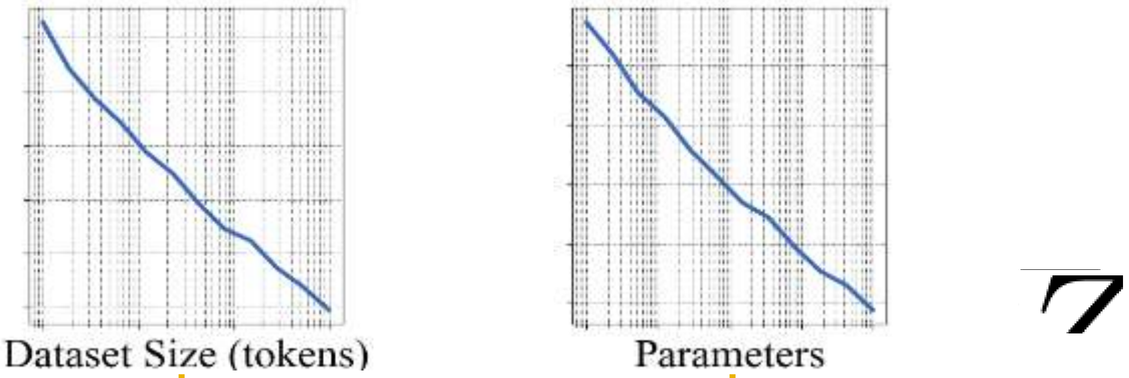
Massage passing: most suitable for graph data?

Other architectures: garnered attention due to their ability to model graph data distributions.

What's Next? Future Directions of Info-GRL

- How to Understand the Scaling Law of Graph Foundation Models (GFMs)?
- How to Design Low-Distortion Constraints in the Era of GFM?





Entropy convergence?
Information saturation?

Information capacity?
Expressive power?

Conclusion

- **Unifying Principle**: Information Theory provides a unified framework for low-distortion graph representation learning through compression, transmission, and preservation of information.
- **Emerging Interface**: Information-theoretic tools are already applied in graph learning for data modeling, capturing dependencies, and designing optimization objectives.
- **The Road Ahead**: Information Theory will inspire new directions in next-generation graph learning, including graph foundational models.

Thank you!

Qingyun Sun

sunqy@buaa.edu.cn

Beihang University, Beijing, China