ILLINOIS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Transfer Learning of Graph Neural Networks with Ego-graph Information Maximization

Qi Zhu*[1], Carl Yang*[2], Yidan Xu[3], Haonan Wang[1], Chao Zhang[4], Jiawei Han[1]

*Equal Contribution

[1]University of Illinois at Urbana-Champaign

[2]Emory University

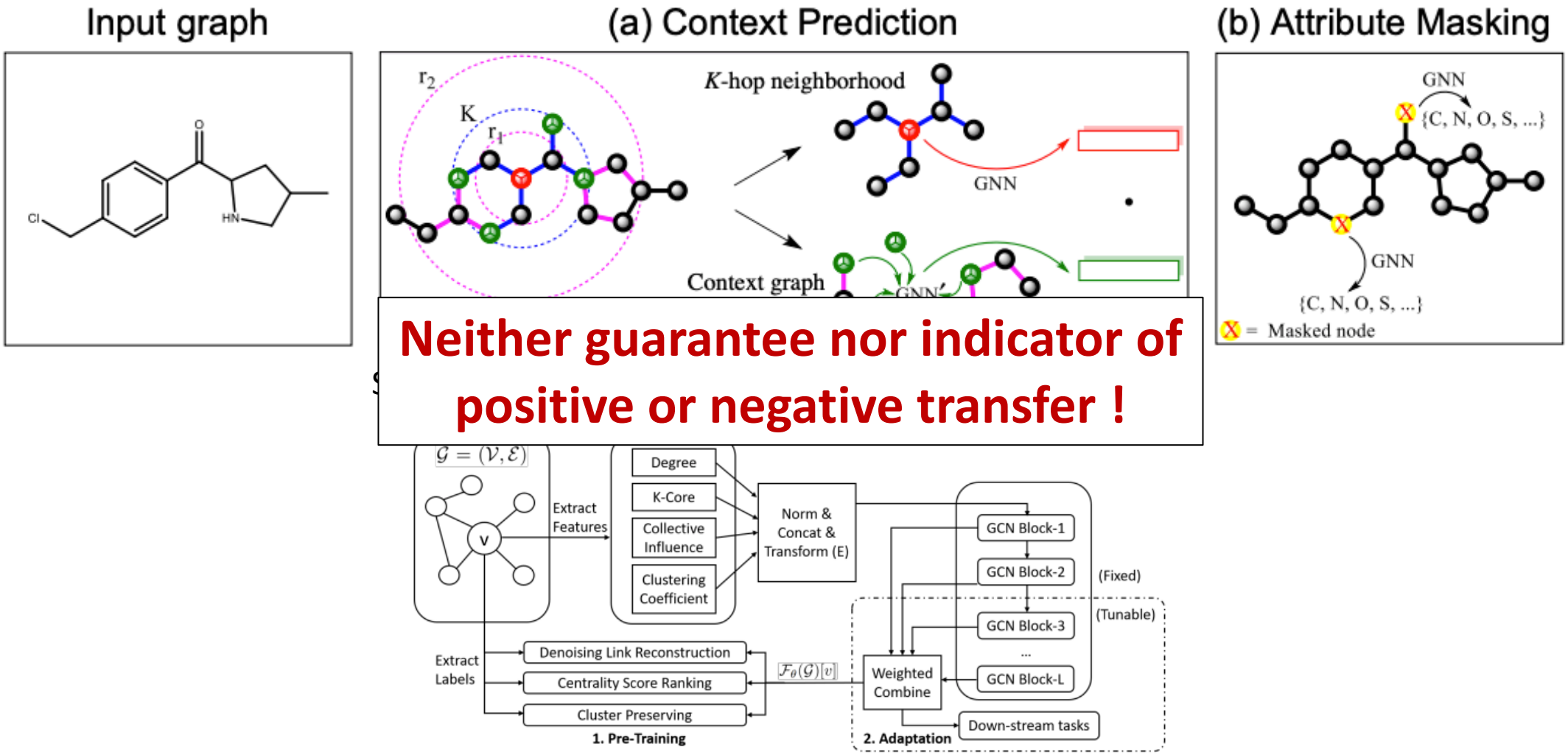
[3]University of Washington

[4]Georgia Institute of Technology

Overview

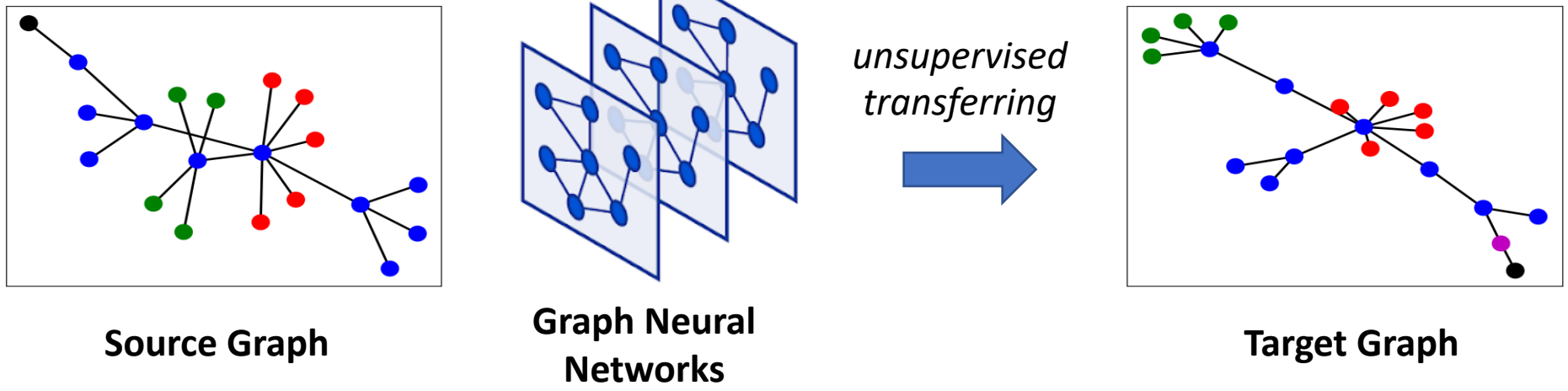
- Existing Study on GNN Pre-training
- Conditions on transferable GNNs
- Proposed transferable framework
  - Input space of GNN
  - Ego-graph Information Maximization objective
- Experiments
- Model analysis

Existing Study on GNN Pre-training



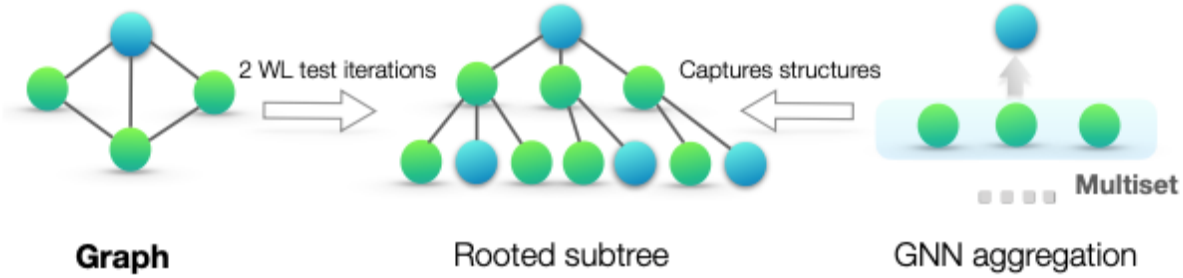
Pre-Training Graph Neural Networks for Generic Structural Feature Extraction

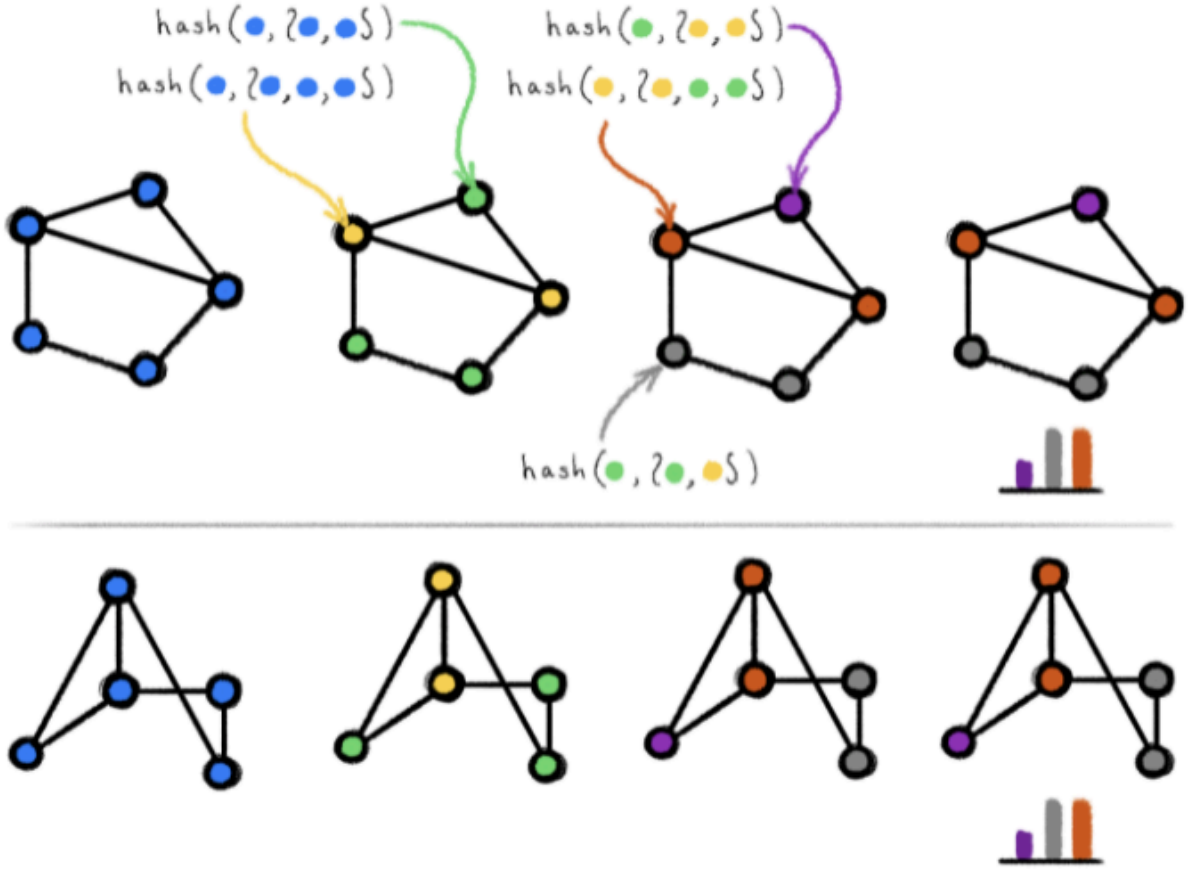
A transfer learning perspective on GNNs



Graph Similarity as an indicator

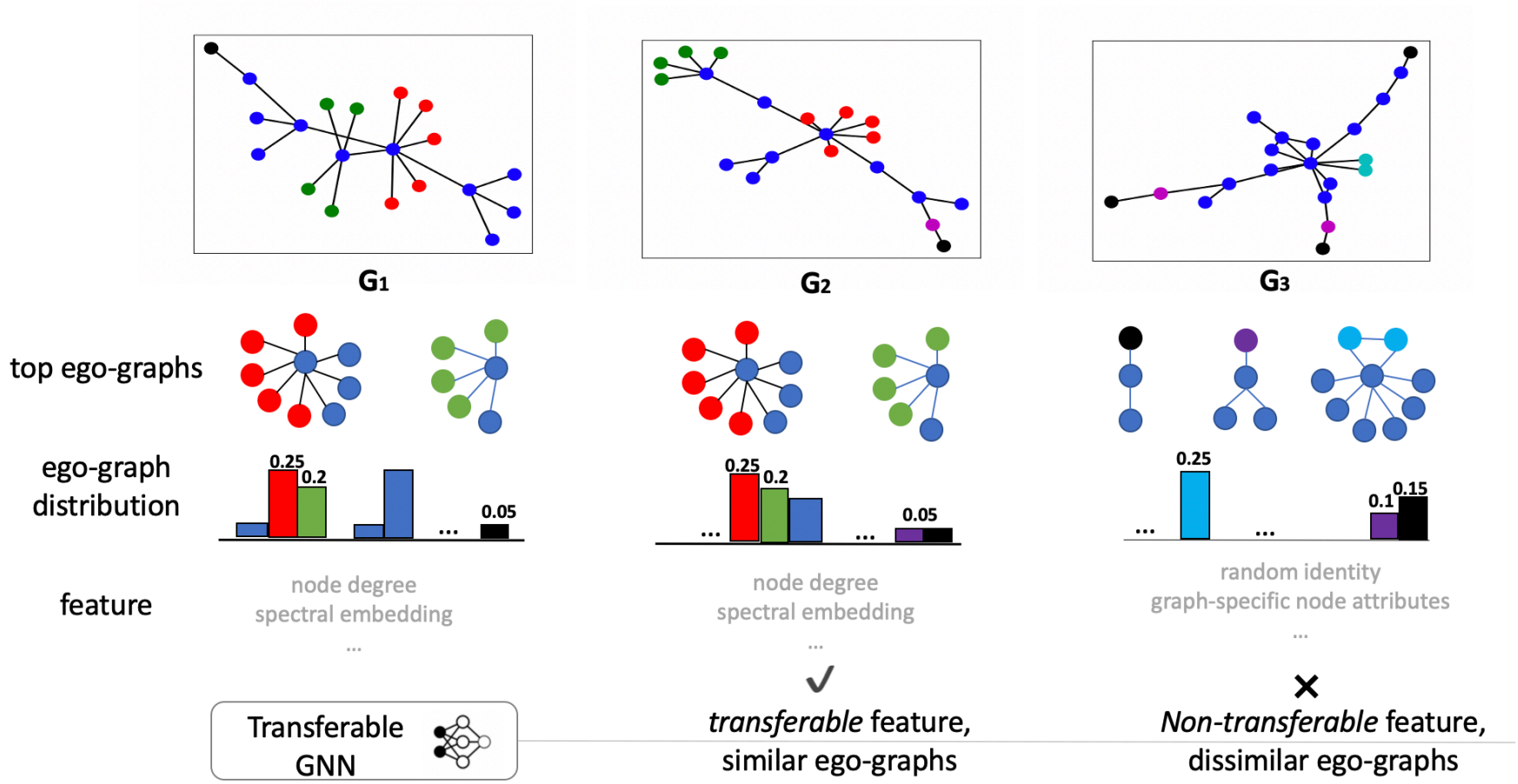
- WL-test use rooted subtree to distinguish different graphs.





Can we use rooted subtree (ego-graph) to measure the similarity between graphs ?

Ego-graph distribution difference as indicator



A natural view of graph neural network is a function F over graph(ego-graph) and node features.
Hence, transferability is measured upon domain (feature) discrepancy.

Definition of structural information

Definition 3.1 (K-hop ego-graph). *We call a graph $g_i = \{V(g_i), E(g_i)\}$ a k-hop ego-graph centered at node v_i if it has a k-layer centroid expansion [4] such that the greatest distance between v_i and any other nodes in the ego-graph is k,* i.e. *$\forall v_j \in V(g_i), |d(v_i, v_j)| \leq k$, where $d(v_i, v_j)$ is the graph distance between v_i and v_j.*

Definition 3.2 (Structural information). *Let $\mathcal{G}$ be a topological space of sub-graphs, we view a graph G as samples of k-hop ego-graphs $\{g_i\}_{i=1}^n$ drawn* i.i.d. *from $\mathcal{G}$ with probability μ,* i.e., *$g_i \overset{\text{i.i.d.}}{\sim} \mu \ \forall i = 1, \cdots, n$. The structural information of G is then defined to be the set of k-hop ego-graph of $\{g_i\}_{i=1}^n$ and their empirical distribution.*

Design of transferable learning objective

- Motivation, if self-supervised model approximates the ego-graph distribution of the source graph. The inference error on target graph ε_t therefore, captures the structural difference if ε_s is small.
- We further use empirical loss different Δl between source and target graph to evaluate the potential of such transfer.

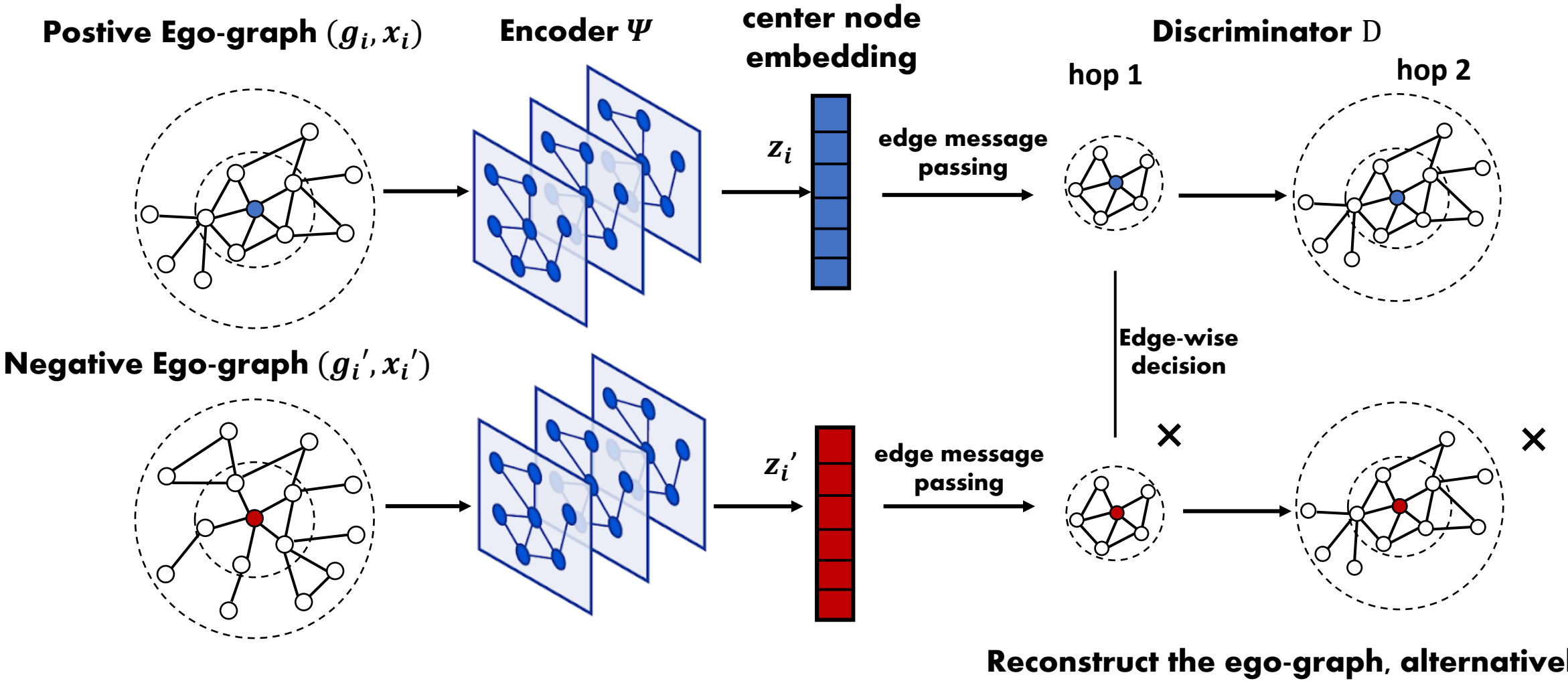
Ego-graph Information Maximization (EGI)

- To capture the joint distribution of structural information and node features, an idea GNN maximize the mutual information between structural information $\{g_i, x_i\}$ and its output Ψ. Such that,

$$\mathcal{I}^{(\mathrm{JSD})}(\mathcal{G}, \Psi) = \mathbb{E}_{\mathbb{P}}\left[-\mathrm{sp}\left(-T_{\mathcal{D},\Psi}(g_i, \Psi(g_i, x_i))\right)\right] - \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{U}}}\left[\mathrm{sp}\left(T_{\mathcal{D},\Psi}(g_i, \Psi(g_i', x_i'))\right)\right]$$

- Discriminator D is asked to distinguish the samples from joint distribution and product of two marginal distributions.

EGI Model Optimization



Transferability of EGI

Theorem A.2. *Let $G_a = \{(g_i, x_i)\}_{i=1}^n$ and $G_b = \{(g_{i'}, x_{i'})\}_{i'=1}^m$ be two graphs and node features are structure-respecting with $x_i = f(L_{g_i}), x_{i'} = f(L_{g_{i'}})$ for some function $f : \mathbb{R}^{|V(g_i)| \times |V(g_i)|} \to \mathbb{R}^d$. Consider GCN Ψ_θ with k layers and a 1-hop polynomial filter ϕ,the empirical performance difference of Ψ_θ with $\mathcal{L}_{\text{EGI}}$ satisfies*

$$|\mathcal{L}_{\text{EGI}}(G_a) - \mathcal{L}_{\text{EGI}}(G_b)| \leq \mathcal{O}\left(\frac{1}{nm}\sum_{i=1}^{n}\sum_{i'=1}^{m}[M + C\lambda_{\max}(L_{g_i} - L_{g_{i'}}) + \tilde{C}\lambda_{\max}(\tilde{L}_{g_i} - \tilde{L}_{g_{i'}}))]\right), \tag{1}$$

where M is dependant on Ψ, $\mathcal{D}$, node features, and the largest eigenvalue of L_{g_i} and $\tilde{L}_{g_i}$. C is a constant dependant on the encoder, while $\tilde{C}$ is a constant dependant on the decoder. With a slight abuse of notation, we denote $\lambda_{\max}(A) := \lambda_{\max}(A^T A)^{1/2}$. Note that, in the main paper, we have $C := M + C\lambda_{\max}(L_{g_i} - L_{g_{i'}})$, and $\Delta_{\mathcal{D}}(G_a, G_b) := \tilde{C}\lambda_{\max}(\tilde{L}_{g_i} - \tilde{L}_{g_{i'}})$.

- The above theorem states the empirical risk difference on source and target graph are bounded by the Laplacian difference on in-degree and out-degree adjacency matrices.
- Specifically, the EGI bound term $\Delta_D(G_a, G_b)$ describes the transferability of the EGI objective.

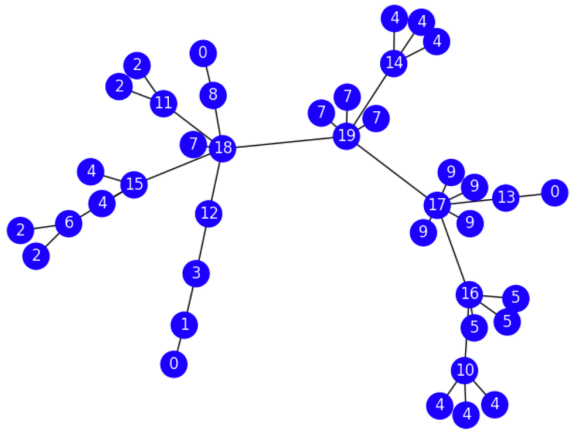
Application of EGI

- Usage of EGI
  - Have a series of similar large graph on different task, train EGI embedding on any of the graph and get transferable embedding easily.
- Usage of EGI gap term $\Delta_D(G_a, G_b)$
  - *point-wise pre-judge: compute the term between source and target graph to assess the potential of positive transfer (< 1.0 in practice)*
  - *pair-wise pre-selection: when multiple source graphs are available G_a^1, G_a^2, G_a^n select most suitable source graph G_a^* with the smallest EGI gap Δ_D*

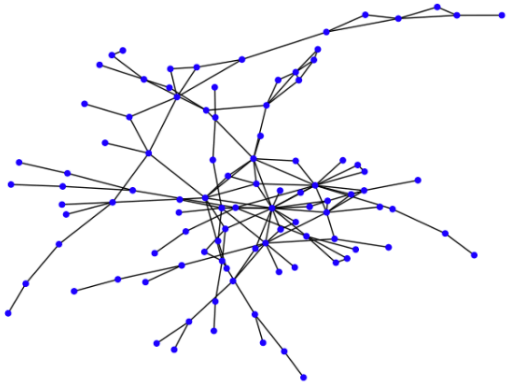
Experiments

- Synthetic Experiment
  - Limit the power of rooted subtree by number of hop and still try to find structural equivalent nodes
- Unsupervised Transfer on node classification
  - Train self-supervised encoder on source graph. Obtain node embeddings on target graph without fine-tuning.
- Few-shot fine-tuning on relation classification
  - Jointly train the encoder and task-specific loss

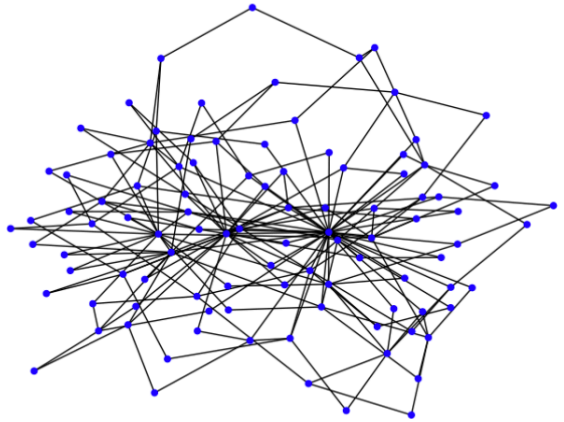
Synthetic experiments



Synthetic task: finding structural equivalent nodes



(a) Forest-fire graph example



(b) Barabasi-albert graph example

Method	transferable features			non-transferable feature			structural difference	
	F-F	B-F	δ(acc.)	F-F	B-F	δ(acc.)	$\Delta_{\mathcal{D}}$(F,F)	$\Delta_{\mathcal{D}}$(B,F)
GIN (untrained)	0.572	0.572	/	0.358	0.358	/		
VGAE (GIN)	0.498	0.432	+0.066	0.240	0.239	0.001		
DGI (GIN)	0.578	0.591	-0.013	0.394	0.213	+0.181	0.752	0.883
EGI (GIN)	**0.710**	0.616	+0.094	0.376	0.346	+0.03		

Real Data Experiments

Task: Unsupervised transferring on role identification

Dataset: Airport (USA, Europe, Brazil), role – level of popularity

Table 2: Results of role identification with direct-transfering on the Airport dataset. The performance reported (%) are the average over 100 runs. The scores marked with ** passed t-test with $p < 0.01$ over the second best results.

Method	Europe (source)		USA (target)		Brazil (target)	
	node degree	uniform	node degree	uniform	node degree	uniform
features	52.81	20.59	55.67	20.22	67.11	19.63
GIN (untrained)	55.75	53.88	61.56	58.32	70.04	70.37
GVAE (Kipf & Welling, 2016)	53.90	21.12	55.51	22.39	66.33	17.70
DGI (Velickovic et al., 2019)	57.75	22.13	54.90	21.76	67.93	18.78
MaskGNN (Hu et al., 2019a)	56.37	55.53	60.82	54.64	66.71	74.54
ContextPredGNN (Hu et al., 2019a)	52.69	49.95	50.38	54.75	62.11	70.66
Structural Pre-train (Hu et al., 2019b)	56.00	53.83	62.17	57.49	68.78	72.41
EGI	**59.15****	54.98	**64.55****	57.40	**73.15****	70.00

Common self-supervised algorithms such as DGI and GVAE fails to positive transfer.

Real Data Experiments

Task: Unsupervised transferring + fine-tuning on Link Prediction

Dataset: knowledge graph (YAGO)

Post-fine-tuning: use transferred encoder Ψ

Joint-fine-tuning: jointly optimize the EGI and task objective on target

Method	post-fine-tuning		joint-fine-tuning	
	AUROC	MRR	AUROC	MRR
No pre-train	0.6866	0.5962	N.A.	N.A
GVAE [24]	0.7009	0.6009	0.6786	0.5676
DGI [45]	0.6885	0.5861	0.6880	0.5366
Mask-GIN [19]	0.7041	0.6242	0.6720	0.5603
ContextPred-GIN [19]	0.6882	0.6589	0.5293	0.3367
EGI	**0.7389****	**0.6695**	**0.7870****	**0.7289****

Model Analysis

- Efficient Computation of term Δ_D
  - Enumerating every single pair of ego-graph between source and target graph can easily blow up the memory (N by M pairs – N,M is the number of nodes).
  - In practice, we can estimate it by uniformly down sample such pairs

Sampling frequency	Europe-USA	Europe-Brazil
100 pairs	0.872±0.039	0.854±0.042
1000 pairs	0.859±0.012	0.848±0.007
Full	0.869	0.851

- Relation to the depth of rooted subtree (ego-graph)

	Europe (source)	USA (target)	Brazil (target)
Method	acc	acc, Δ_D	acc, Δ_D
EGI (k=1)	58.25	60.08, 0.385	60.74, 0.335
EGI (k=2)	59.15	64.55, 0.869	73.15, 0.851
EGI (k=3)	57.63	64.12, 0.912	72.22, 0.909

Thanks and Q&A

- More results are available: https://arxiv.org/abs/2009.05204
- Questions and discussions: qiz3@Illinois.edu