ICML | 2021

Graph Contrastive Learning Automated

Yuning You[1], Tianlong Chen[2], Yang Shen[1], Zhangyang Wang[2]

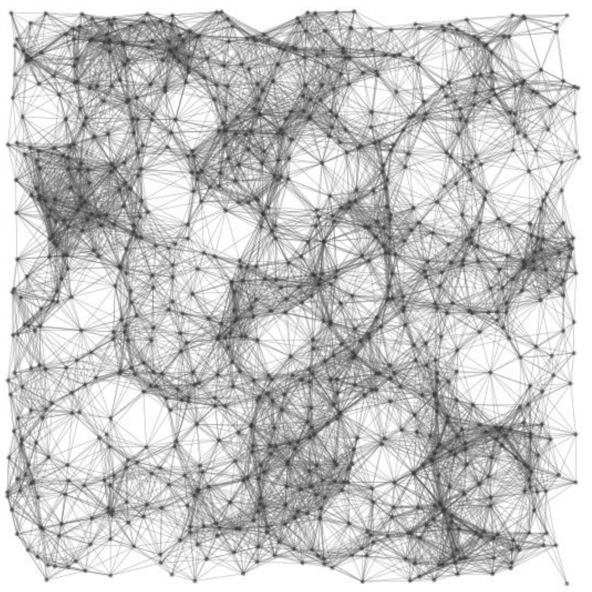
[1]Texas A&M University, [2]University of Texas at Austin

The Learning on Graphs and Geometry Reading Group (LoGaG) Presentation

August 17, 2021

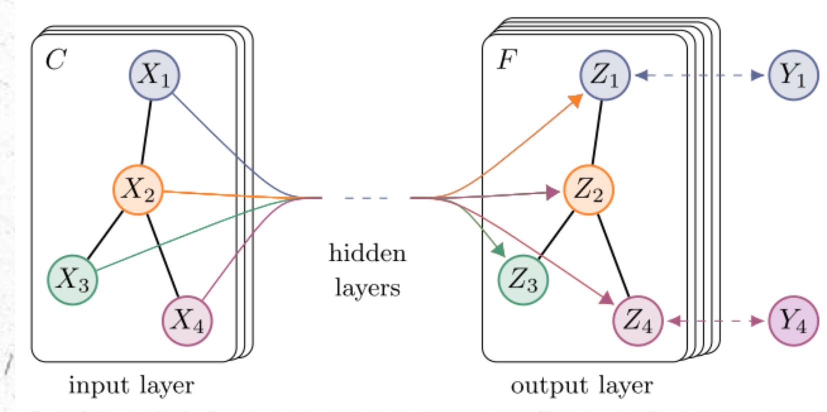
Background

- Data-structure of graphs

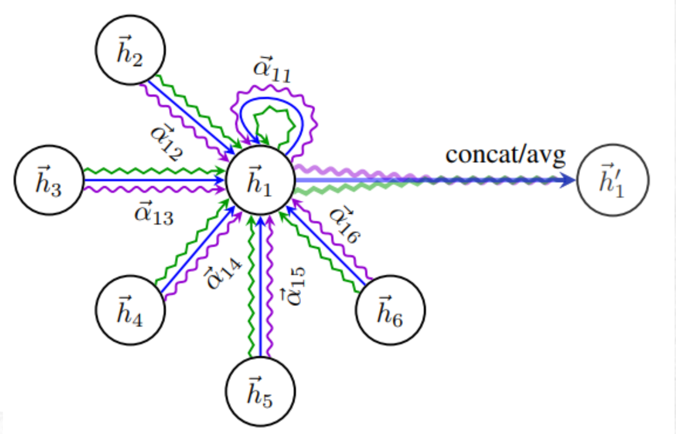


LoGaG cover

- Graph networks



GCN, ICLR'17

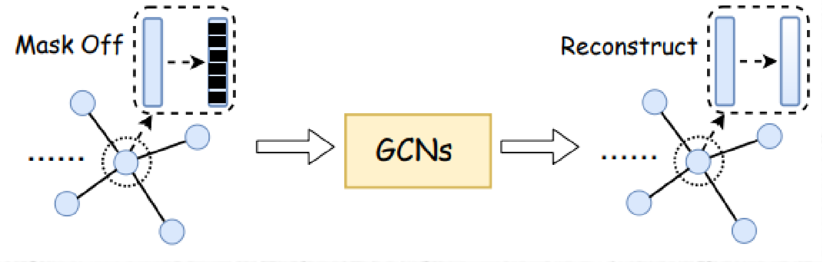


GAT, ICLR'18

- Self-supervised learning: Warming-up models with unlabeled data

Background

- Self-supervision on graphs
  - Predictive task



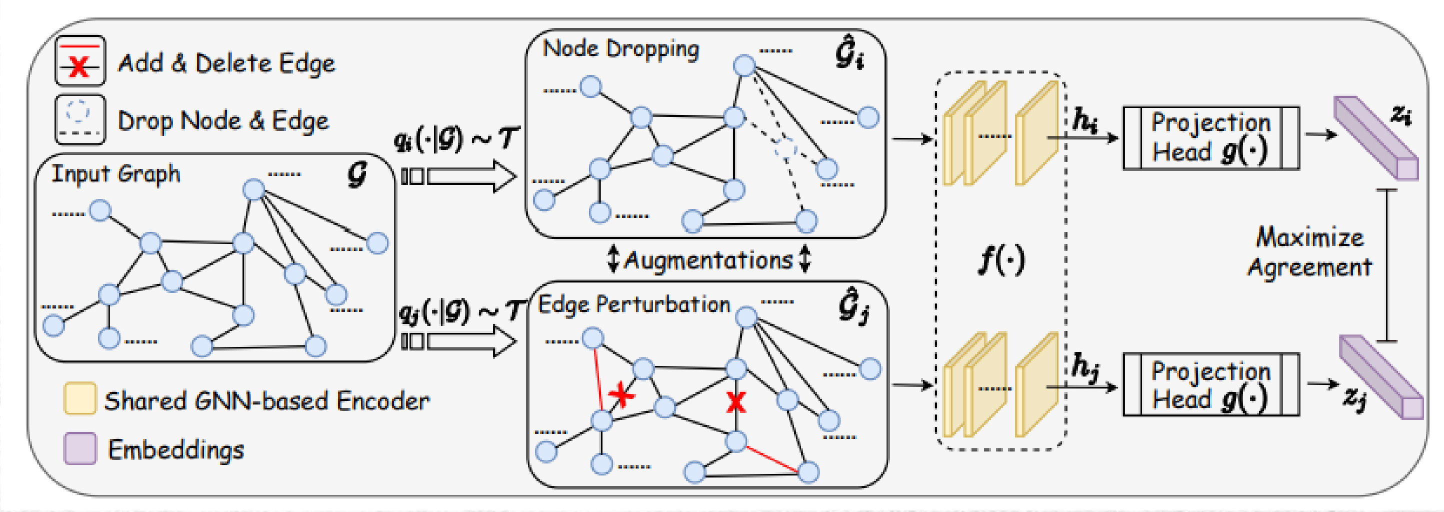
E.g. Graph completion, ICML'20

AutoSSL, arXiv'21

Self-Supervised Learning of Graph Neural Networks: A Unified Review

Yaochen Xie, Zhao Xu, Jingtun Zhang, Zhengyang Wang, Shuiwang Ji, *Senior Member, IEEE*

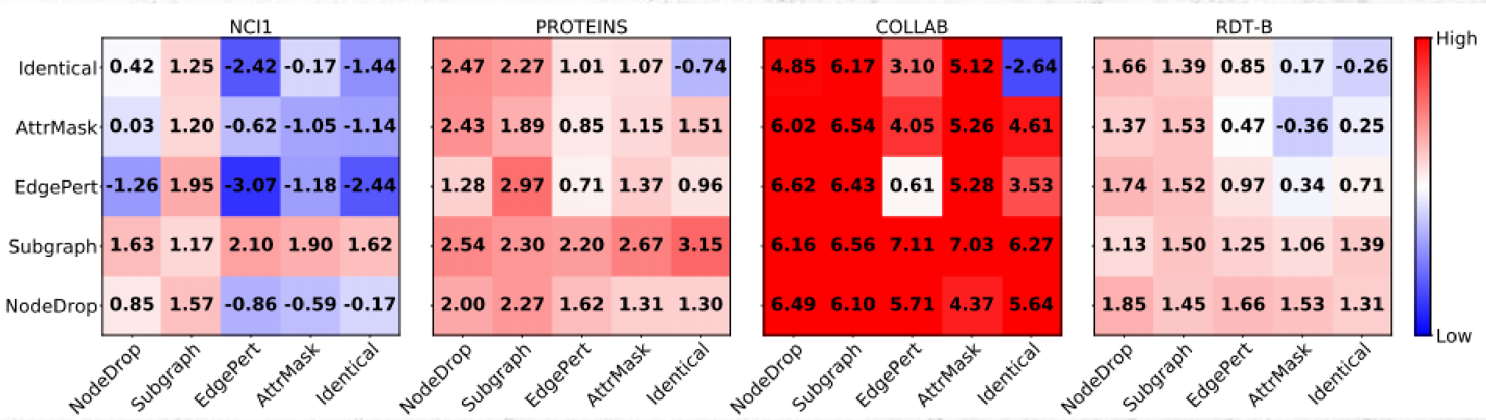
  - Contrastive learning
    - GraphCL, NeurIPS'20
    - Augmentations + perturbation invariance



Background

- Augmentation is crucial in GraphCL

Data augmentation	Type	Underlying Prior
Node dropping	Nodes, edges	Vertex missing does not alter semantics.
Edge perturbation	Edges	Semantic robustness against connectivity variations.
Attribute masking	Nodes	Semantic robustness against losing partial attributes.
Subgraph	Nodes, edges	Local structure can hint the full semantics.



- Reason: challenge of heterogeneous nature of graph data


Fig 1. Social networks

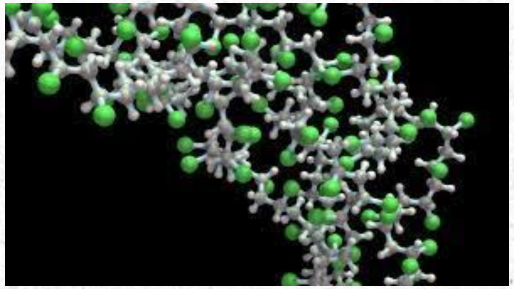


Fig 2. Polymers

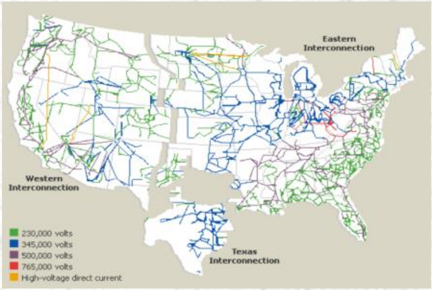
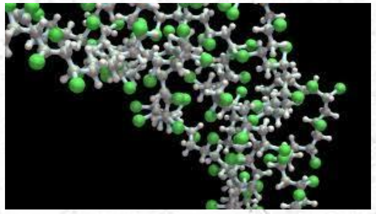
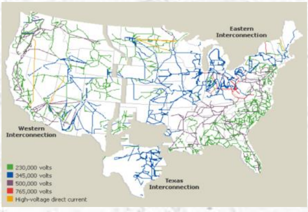


Fig 3. Power grids

Background

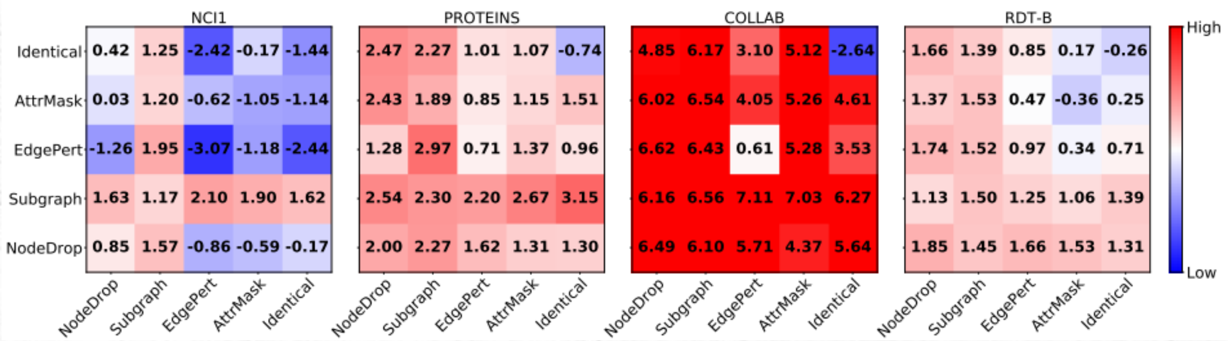
- Data heterogeneity

- Ad-hoc choices of augmentations in GraphCL

Data augmentation	Type	Underlying Prior
Node dropping	Nodes, edges	Vertex missing does not alter semantics.
Edge perturbation	Edges	Semantic robustness against connectivity variations.
Attribute masking	Nodes	Semantic robustness against losing partial attributes.
Subgraph	Nodes, edges	Local structure can hint the full semantics.

- Rules derived from tedious tuning



- Question: Can we be more principled and automated?

Contributions

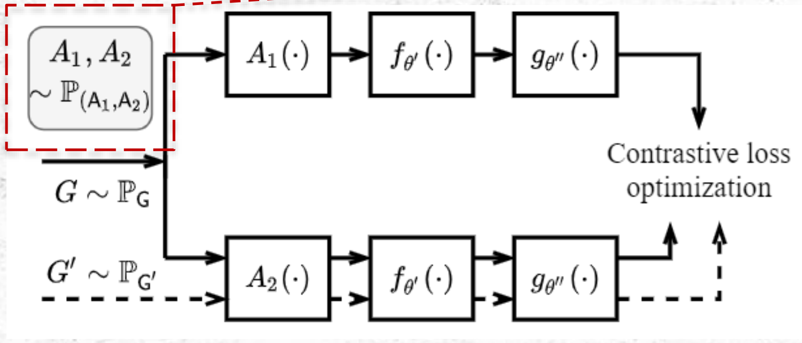
- Given a new and unseen graph dataset, can GraphCL automatically select augmentations, avoiding ad-hoc choices or tedious tuning?

- Joint augmentation optimization (JOAO)
  - A principled bi-level optimization framework
  - Automatic, free of human labor of trial-and-error
  - Adaptive, generalizing smoothly to handling diverse graph data
  - Dynamic, allowing for augmentation types varying at different steps

Method. JOAO

- GraphCL
  - Enforcing perturbation invariance

Data augmentation	Type	Underlying Prior
Node dropping	Nodes, edges	Vertex missing does not alter semantics.
Edge perturbation	Edges	Semantic robustness against connectivity variations.
Attribute masking	Nodes	Semantic robustness against losing partial attributes.
Subgraph	Nodes, edges	Local structure can hint the full semantics.



$$\begin{aligned} &\min_\theta \ \mathcal{L}(\mathsf{G}, \mathsf{A}_1, \mathsf{A}_2, \theta) \\ &= \min_\theta \ \Big\{ \big(-\mathbb{E}_{\mathbb{P}_\mathsf{G} \times \mathbb{P}_{(\mathsf{A}_1, \mathsf{A}_2)}} \mathrm{sim}(\overbrace{\mathsf{T}_{\theta,1}(\mathsf{G}), \mathsf{T}_{\theta,2}(\mathsf{G})}^{\text{Positive pairs}}) \\ &\quad + \mathbb{E}_{\mathbb{P}_\mathsf{G} \times \mathbb{P}_{\mathsf{A}_1}} \log \big(\mathbb{E}_{\mathbb{P}_{\mathsf{G}'} \times \mathbb{P}_{\mathsf{A}_2}} \exp (\mathrm{sim}(\underbrace{\mathsf{T}_{\theta,1}(\mathsf{G}), \mathsf{T}_{\theta,2}(\mathsf{G}')}_{\text{Negative pairs}}))\big)\big) \Big\}, \end{aligned} \quad (1)$$

- The unified framework,
  joint augmentation optimization (JOAO)
  as a bi-level optimization

$$\begin{aligned} &\min_\theta \quad \mathcal{L}(\mathsf{G}, \mathsf{A}_1, \mathsf{A}_2, \theta), \\ &\text{s.t.} \ \boxed{\mathbb{P}_{(\mathsf{A}_1, \mathsf{A}_2)}} \in \arg\min_{\mathbb{P}_{(\mathsf{A}'_1, \mathsf{A}'_2)}} \mathcal{D}(\mathsf{G}, \mathsf{A}'_1, \mathsf{A}'_2, \theta), \end{aligned} \quad (2)$$

Method. Instantiation of JOAO

- A min-max optimization instantiation

$$\min_\theta \quad \mathcal{L}(\mathsf{G}, \mathsf{A}_1, \mathsf{A}_2, \theta),$$
$$\text{s.t.} \quad \mathbb{P}_{(\mathsf{A}_1,\mathsf{A}_2)} \in \arg\max_{\mathbb{P}_{(\mathsf{A}'_1,\mathsf{A}'_2)}} \Big\{ \boxed{\mathcal{L}(\mathsf{G}, \mathsf{A}'_1, \mathsf{A}'_2, \theta)}$$
$$\boxed{- \frac{\gamma}{2} \mathrm{dist}(\mathbb{P}_{(\mathsf{A}'_1,\mathsf{A}'_2)}, \mathbb{P}_{\mathrm{prior}})} \Big\}, \qquad (3)$$

- Principles
  - Exploiting challenging augmentations: model-based adversarial training MBRDL, arXiv'20
  - Regularization with prior
    - Uniform distribution avoiding collapse
    - Squared Euclidean distance Ref 5. Wang et al., arXiv'19
  - Trade-off by γ

$$\mathrm{dist}(\mathbb{P}_{(\mathsf{A}_1,\mathsf{A}_2)}, \mathbb{P}_{\mathrm{prior}}) = \sum_{i=1}^{|\mathcal{A}|}\sum_{j=1}^{|\mathcal{A}|}(p_{ij} - \frac{1}{|\mathcal{A}|^2})^2,$$
$$p_{ij} = \mathrm{Prob}(\mathsf{A}_1 = A^i, \mathsf{A}_2 = A^j)$$

Method. AGD for JOAO

- Alternating gradient descent (AGD) Wang et al., arXiv'19
  - Upper-level minimization
  - Lower-lever maximization

$$\min_{\theta} \quad \mathcal{L}(\mathsf{G}, \mathsf{A}_1, \mathsf{A}_2, \theta),$$
$$\text{s.t.} \quad \mathbb{P}_{(\mathsf{A}_1,\mathsf{A}_2)} \in \arg\max_{\mathbb{P}_{(\mathsf{A}'_1,\mathsf{A}'_2)}} \Big\{ \mathcal{L}(\mathsf{G}, \mathsf{A}'_1, \mathsf{A}'_2, \theta) - \frac{\gamma}{2} \mathrm{dist}(\mathbb{P}_{(\mathsf{A}'_1,\mathsf{A}'_2)}, \mathbb{P}_{\mathrm{prior}}) \Big\}, \quad (3)$$

- Upper-level minimization
  - GraphCL optimization given sampling distribution

$$\theta^{(n)} = \theta^{(n-1)} - \alpha' \nabla_{\theta} \mathcal{L}(\mathsf{G}, \mathsf{A}_1, \mathsf{A}_2, \theta), \quad (4)$$

where $\alpha' \in \mathcal{R}_{>0}$ is the learning rate.

Method. AGD for JOAO

- Lower-level maximization
  - Gradient is not intuitive
  - Analytical rewrite

$$\min_\theta \quad \mathcal{L}(\mathsf{G}, \mathsf{A}_1, \mathsf{A}_2, \theta),$$
$$\text{s.t.} \quad \mathbb{P}_{(\mathsf{A}_1,\mathsf{A}_2)} \in \arg\max_{\mathbb{P}_{(\mathsf{A}'_1,\mathsf{A}'_2)}} \Big\{ \mathcal{L}(\mathsf{G}, \mathsf{A}'_1, \mathsf{A}'_2, \theta) - \frac{\gamma}{2}\,\mathrm{dist}(\mathbb{P}_{(\mathsf{A}'_1,\mathsf{A}'_2)}, \mathbb{P}_{\mathrm{prior}}) \Big\}, \qquad (3)$$

$$\mathcal{L}(\mathsf{G}, \mathsf{A}_1, \mathsf{A}_2, \theta) = \sum_{i=1}^{|\mathcal{A}|}\sum_{j=1}^{|\mathcal{A}|} \overbrace{p_{ij}}^{\text{Targeted}} \Big\{ -\mathbb{E}_{\mathbb{P}_\mathsf{G}} \mathrm{sim}(T_\theta^i(\mathsf{G}), T_\theta^j(\mathsf{G})) + \mathbb{E}_{\mathbb{P}_\mathsf{G}} \log\Big(\sum_{j'=1}^{|\mathcal{A}|} \underbrace{p_{j'}}_{\text{Undesired}} \mathbb{E}_{\mathbb{P}_{\mathsf{G}'}} \exp(\mathrm{sim}(T_\theta^i(\mathsf{G}), T_\theta^{j'}(\mathsf{G}')))\Big) \Big\}, \qquad (5)$$

- Undesired marginal probability $p_{j'}$ entangled in negative term

Method. AGD for JOAO

- A lower-bound approximation to decouple $p_{j'}$

$$
\begin{aligned}
&\mathbb{E}_{\mathbb{P}_{\mathsf{G}} \times \mathbb{P}_{\mathsf{A}_1}} \log(\mathbb{E}_{\mathbb{P}_{\mathsf{G}'} \times \mathbb{P}_{\mathsf{A}_2}} \exp(\mathrm{sim}(\mathsf{T}_{\theta,1}(\mathsf{G}), \mathsf{T}_{\theta,2}(\mathsf{G}')))) \\
\geq\, &\mathbb{E}_{\mathbb{P}_{\mathsf{G}} \times \mathbb{P}_{\mathsf{A}_1} \times \mathbb{P}_{\mathsf{A}_2}} \log(\mathbb{E}_{\mathbb{P}_{\mathsf{G}'}} \exp(\mathrm{sim}(\mathsf{T}_{\theta,1}(\mathsf{G}), \mathsf{T}_{\theta,2}(\mathsf{G}')))) \\
\approx\, &\mathbb{E}_{\mathbb{P}_{\mathsf{G}} \times \mathbb{P}_{(\mathsf{A}_1, \mathsf{A}_1)}} \log(\mathbb{E}_{\mathbb{P}_{\mathsf{G}'}} \exp(\mathrm{sim}(\mathsf{T}_{\theta,1}(\mathsf{G}), \mathsf{T}_{\theta,2}(\mathsf{G}')))),
\end{aligned} \tag{6}
$$

$$
\begin{aligned}
\min_\theta \quad & \mathcal{L}(\mathsf{G}, \mathsf{A}_1, \mathsf{A}_2, \theta), \\
\text{s.t.} \quad & \mathbb{P}_{(\mathsf{A}_1, \mathsf{A}_2)} \in \arg\max_{\mathbb{P}_{(\mathsf{A}'_1, \mathsf{A}'_2)}} \Big\{ \mathcal{L}(\mathsf{G}, \mathsf{A}'_1, \mathsf{A}'_2, \theta) \\
& \qquad - \frac{\gamma}{2} \mathrm{dist}(\mathbb{P}_{(\mathsf{A}'_1, \mathsf{A}'_2)}, \mathbb{P}_{\text{prior}}) \Big\},
\end{aligned} \tag{3}
$$

- Approximated contrastive loss:

$$
\begin{aligned}
\mathcal{L}(\mathsf{G}, \mathsf{A}_1, \mathsf{A}_2, \theta) &\approx \sum_{i=1}^{|\mathcal{A}|} \sum_{j=1}^{|\mathcal{A}|} \overbrace{p_{ij}}^{\text{Targeted}} \ell(\mathsf{G}, A^i, A^j, \theta) \\
= \sum_{i=1}^{|\mathcal{A}|} \sum_{j=1}^{|\mathcal{A}|} p_{ij} \Big\{ &- \mathbb{E}_{\mathbb{P}_{\mathsf{G}}} \mathrm{sim}(T_\theta^i(\mathsf{G}), T_\theta^j(\mathsf{G})) \\
&+ \mathbb{E}_{\mathbb{P}_{\mathsf{G}}} \log(\mathbb{E}_{\mathbb{P}_{\mathsf{G}'}} \exp(\mathrm{sim}(T_\theta^i(\mathsf{G}), T_\theta^j(\mathsf{G}')))) \Big\}.
\end{aligned} \tag{7}
$$

Method. AGD for JOAO

- Rewrote lower-level optimization

$$\min_\theta \quad \mathcal{L}(\mathsf{G}, \mathsf{A}_1, \mathsf{A}_2, \theta),$$
$$\text{s.t.} \quad \mathbb{P}_{(\mathsf{A}_1,\mathsf{A}_2)} \in \arg\max_{\mathbb{P}_{(\mathsf{A}'_1,\mathsf{A}'_2)}} \Big\{ \mathcal{L}(\mathsf{G}, \mathsf{A}'_1, \mathsf{A}'_2, \theta) - \frac{\gamma}{2} \operatorname{dist}(\mathbb{P}_{(\mathsf{A}'_1,\mathsf{A}'_2)}, \mathbb{P}_{\text{prior}}) \Big\}, \quad (3)$$

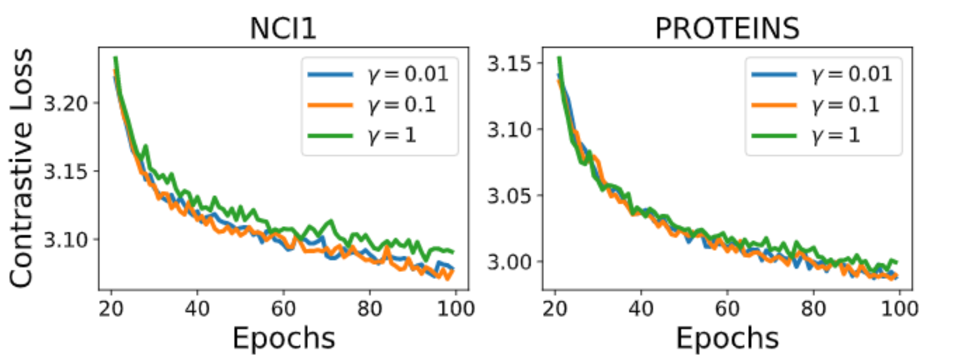
$$\mathbb{P}_{(\mathsf{A}_1,\mathsf{A}_2)} \in \arg\max_{\boldsymbol{p}\in\mathcal{P},\boldsymbol{p}=[p_{ij}],i,j=1,\dots,|\mathcal{A}|} \{\psi(\boldsymbol{p})\},$$
$$\psi(\boldsymbol{p}) = \sum_{i=1}^{|\mathcal{A}|}\sum_{j=1}^{|\mathcal{A}|} p_{ij}\ell(\mathsf{G}, A^i, A^j, \theta) - \frac{\gamma}{2}\sum_{i=1}^{|\mathcal{A}|}\sum_{j=1}^{|\mathcal{A}|}(p_{ij} - \frac{1}{|\mathcal{A}|^2})^2, \quad (8)$$

- Projected gradient descent Boyd et al., 2004

$$\boldsymbol{b} = \boldsymbol{p}^{(n-1)} + \alpha''\nabla_{\boldsymbol{p}}\psi(\boldsymbol{p}^{(n-1)}), \boldsymbol{p}^{(n)} = (\boldsymbol{b} - \mu\mathbf{1})_+, \quad (9)$$

where $\alpha'' \in \mathcal{R}_{>0}$ is the learning rate, μ is the root of the equation $\mathbf{1}^\mathsf{T}(\boldsymbol{b} - \mu\mathbf{1}) = 1$, and $(\cdot)_+$ is the element-wise non-negative operator. μ can be efficiently found via the bi-jection method.

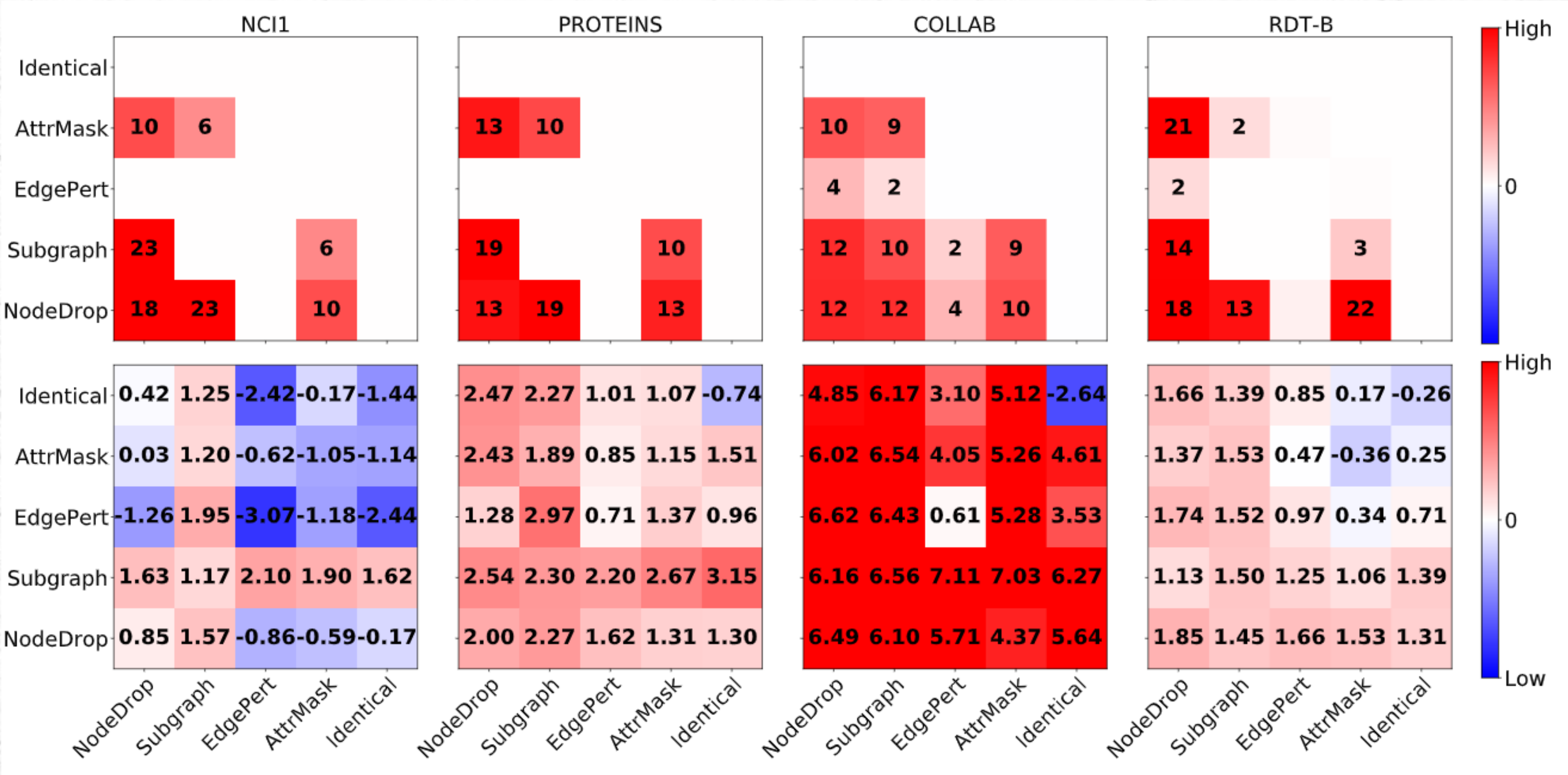
❖ Empirical convergence



Method. JOAO Sanity Check

- Are JOAO selected augmentation reasonable?

- Selections align with “best practices”



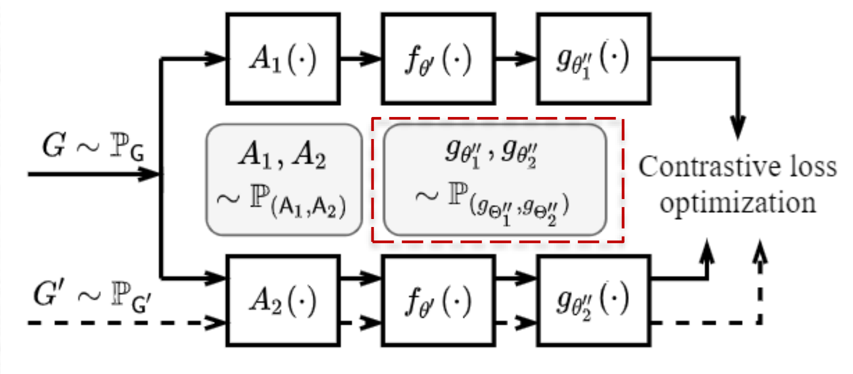
Method. JOAOv2 Addressing Distortion

- JOAO selects automatic, adaptive and dynamic augmentations
- However, more diverse, aggressive and challenging
- Potentially distorting training distribution

SLA+AG, ICML'20
DistAug, ICML'20

Datasets	A.S.	JOAO	JOAOv2
NCI1	0.2	61.77±1.61	62.52±1.16
	0.25	60.95±0.55	61.67±0.72
PROTEINS	0.2	71.45±0.89	71.66±1.10
	0.25	71.61±1.65	73.01±1.02

- JOAOv2 = JOAO + augmentation-aware multi-projection heads



$$\begin{aligned}
&\min_\theta \quad \mathcal{L}_{\text{v2}}(\mathsf{G}, \mathsf{A}_1, \mathsf{A}_2, \theta', \Theta''_1, \Theta''_2), \\
&\text{s.t.} \quad \mathbb{P}_{(\mathsf{A}_1,\mathsf{A}_2)} \in \arg\max_{\mathbb{P}_{(\mathsf{A}'_1,\mathsf{A}'_2)}} \Big\{ \mathcal{L}_{\text{v2}}(\mathsf{G}, \mathsf{A}_1, \mathsf{A}_2, \theta', \Theta''_1, \Theta''_2) \\
&\qquad\qquad - \frac{\gamma}{2} \operatorname{dist}(\mathbb{P}_{(\mathsf{A}'_1,\mathsf{A}'_2)}, \mathbb{P}_{\text{prior}}) \Big\}, \\
&\qquad \boxed{\mathbb{P}_{(g_{\Theta''_1}, g_{\Theta''_2})}} = \mathbb{P}_{(\mathsf{A}_1,\mathsf{A}_2)}. \qquad (10)
\end{aligned}$$

Experiments & Discussions

- Settings
  - Semi-supervised
  - Unsupervised
  - Transfer
- Competitors
  - Heuristic designed pretexts
  - GraphCL with rules
- Datasets
  - Across diverse fields
  - On bioinformatics domains
- Summary of JOAO performance

	v.s. GraphCL	v.s. Heuristic methods
Across diverse fields	Comparable	Better
On specific domains	Better	Worse

Experiments & Discussions. Across Diverse Datasets

- JOAO performs on par with ad-hoc rules

- Augmentation-aware projection heads strengths JOAO

Semi-supervised learning

L.R.	Methods	NCI1	PROTEINS	DD	COLLAB	RDT-B	RDT-M5K	GITHUB	A.R.↓
1%	No pre-train.	60.72±0.45	-	-	57.46±0.25	-	-	54.25±0.22	7.6
	Augmentations	60.49±0.46	-	-	58.40±0.97	-	-	56.36±0.42	6.6
	GAE	61.63±0.84	-	-	63.20±0.67	-	-	59.44±0.44	4.0
	Infomax	**62.72**±0.65	-	-	61.70±0.77	-	-	58.99±0.50	3.3
	ContextPred	61.21±0.77	-	-	57.60±2.07	-	-	56.20±0.49	6.6
	GraphCL	**62.55**±0.86	-	-	**64.57**±1.15	-	-	58.56±0.59	**2.6**
	JOAO	61.97±0.72	-	-	**63.71**±0.84	-	-	60.35±0.24	3.0
	JOAOv2	**62.52**±1.16	-	-	**64.51**±2.21	-	-	**61.05**±0.31	**2.0**
10%	No pre-train.	73.72±0.24	70.40±1.54	73.56±0.41	73.71±0.27	86.63±0.27	51.33±0.44	60.87±0.17	7.0
	Augmentations	73.59±0.32	70.29±0.64	74.30±0.81	74.19±0.13	87.74±0.39	52.01±0.20	60.91±0.32	6.2
	GAE	74.36±0.24	70.51±0.17	74.54±0.68	75.09±0.19	87.69±0.40	**53.58**±0.13	63.89±0.52	4.5
	Infomax	**74.86**±0.26	72.27±0.40	**75.78**±0.34	73.76±0.29	88.66±0.95	**53.61**±0.31	65.21±0.88	3.0
	ContextPred	73.00±0.30	70.23±0.63	74.66±0.51	73.69±0.37	84.76±0.52	51.23±0.84	62.35±0.73	7.2
	GraphCL	**74.63**±0.25	**74.17**±0.34	**76.17**±1.37	74.23±0.21	**89.11**±0.19	52.55±0.45	65.81±0.79	**2.4**
	JOAO	74.48±0.27	72.13±0.92	**75.69**±0.67	**75.30**±0.32	88.14±0.25	52.83±0.54	65.00±0.30	3.5
	JOAOv2	**74.86**±0.39	73.31±0.48	**75.81**±0.73	**75.53**±0.18	88.79±0.65	52.71±0.28	**66.66**±0.60	**1.8**

Unsupervised learning

Methods	NCI1	PROTEINS	DD	MUTAG	COLLAB	RDT-B	RDT-M5K	IMDB-B	A.R.↓
GL	-	-	-	81.66±2.11	-	77.34±0.18	41.01±0.17	65.87±0.98	7.4
WL	**80.01**±0.50	72.92±0.56	-	80.72±3.00	-	68.82±0.41	46.06±0.21	**72.30**±3.44	5.7
DGK	**80.31**±0.46	73.30±0.82	-	87.44±2.72	-	78.04±0.39	41.27±0.18	66.96±0.56	4.9
node2vec	54.89±1.61	57.49±3.57	-	72.63±10.20	-	-	-	-	8.6
sub2vec	52.84±1.47	53.03±5.55	-	61.05±15.80	-	71.48±0.41	36.68±0.42	55.26±1.54	9.5
graph2vec	73.22±1.81	73.30±2.05	-	83.15±9.25	-	75.78±1.03	47.86±0.26	71.10±0.54	5.7
MVGRL	-	-	-	75.40±7.80	-	82.00±1.10	-	63.60±4.20	7.2
InfoGraph	76.20±1.06	**74.44**±0.31	72.85±1.78	**89.01**±1.13	**70.65**±1.13	82.50±1.42	53.46±1.03	**73.03**±0.87	3.0
GraphCL	77.87±0.41	**74.39**±0.45	**78.62**±0.40	86.80±1.34	**71.36**±1.15	**89.53**±0.84	**55.99**±0.28	**71.14**±0.44	**2.6**
JOAO	**78.07**±0.47	**74.55**±0.41	**77.32**±0.54	**87.35**±1.02	**69.50**±0.36	**85.29**±1.35	**55.74**±0.63	70.21±3.08	3.3
JOAOv2	78.36±0.53	74.07±1.10	**77.40**±1.15	**87.67**±0.79	69.33±0.34	**86.42**±1.45	**56.03**±0.27	70.83±0.25	**2.8**

Experiments & Discussions. Across Diverse Datasets

- ➢ JOAOv2 generally outperforms heuristic self-supervised pretext tasks

Semi-supervised learning

L.R.	Methods	NCI1	PROTEINS	DD	COLLAB	RDT-B	RDT-M5K	GITHUB	A.R.↓
1%	No pre-train.	60.72±0.45	-	-	57.46±0.25	-	-	54.25±0.22	7.6
	Augmentations	60.49±0.46	-	-	58.40±0.97	-	-	56.36±0.42	6.6
	GAE	61.63±0.84	-	-	63.20±0.67	-	-	59.44±0.44	4.0
	Infomax	**62.72**±0.65	-	-	61.70±0.77	-	-	58.99±0.50	3.3
	ContextPred	61.21±0.77	-	-	57.60±2.07	-	-	56.20±0.49	6.6
	GraphCL	**62.55**±0.86	-	-	**64.57**±1.15	-	-	58.56±0.59	**2.6**
	JOAO	61.97±0.72	-	-	**63.71**±0.84	-	-	60.35±0.24	3.0
	JOAOv2	**62.52**±1.16	-	-	**64.51**±2.21	-	-	**61.05**±0.31	**2.0**
10%	No pre-train.	73.72±0.24	70.40±1.54	73.56±0.41	73.71±0.27	86.63±0.27	51.33±0.44	60.87±0.17	7.0
	Augmentations	73.59±0.32	70.29±0.64	74.30±0.81	74.19±0.13	87.74±0.39	52.01±0.20	60.91±0.32	6.2
	GAE	74.36±0.24	70.51±0.17	74.54±0.68	75.09±0.19	87.69±0.40	**53.58**±0.13	63.89±0.52	4.5
	Infomax	**74.86**±0.26	72.27±0.40	**75.78**±0.34	73.76±0.29	88.66±0.95	**53.61**±0.31	65.21±0.88	3.0
	ContextPred	73.00±0.30	70.23±0.63	74.66±0.51	73.69±0.37	84.76±0.52	51.23±0.84	62.35±0.73	7.2
	GraphCL	**74.63**±0.25	**74.17**±0.34	**76.17**±1.37	74.23±0.21	**89.11**±0.19	52.55±0.45	65.81±0.79	**2.4**
	JOAO	74.48±0.27	72.13±0.92	**75.69**±0.67	**75.30**±0.32	88.14±0.25	52.83±0.54	65.00±0.30	3.5
	JOAOv2	**74.86**±0.39	73.31±0.48	**75.81**±0.73	**75.53**±0.18	88.79±0.65	52.71±0.28	**66.66**±0.60	**1.8**

Unsupervised learning

Methods	NCI1	PROTEINS	DD	MUTAG	COLLAB	RDT-B	RDT-M5K	IMDB-B	A.R.↓
GL	-	-	-	81.66±2.11	-	77.34±0.18	41.01±0.17	65.87±0.98	7.4
WL	**80.01**±0.50	72.92±0.56	-	80.72±3.00	-	68.82±0.41	46.06±0.21	**72.30**±3.44	5.7
DGK	**80.31**±0.46	73.30±0.82	-	87.44±2.72	-	78.04±0.39	41.27±0.18	66.96±0.56	4.9
node2vec	54.89±1.61	57.49±3.57	-	72.63±10.20	-	-	-	-	8.6
sub2vec	52.84±1.47	53.03±5.55	-	61.05±15.80	-	71.48±0.41	36.68±0.42	55.26±1.54	9.5
graph2vec	73.22±1.81	73.30±2.05	-	83.15±9.25	-	75.78±1.03	47.86±0.26	71.10±0.54	5.7
MVGRL	-	-	-	75.40±7.80	-	82.00±1.10	-	63.60±4.20	7.2
InfoGraph	76.20±1.06	**74.44**±0.31	72.85±1.78	**89.01**±1.13	**70.65**±1.13	82.50±1.42	53.46±1.03	**73.03**±0.87	3.0
GraphCL	77.87±0.41	**74.39**±0.45	**78.62**±0.40	86.80±1.34	**71.36**±1.15	**89.53**±0.84	**55.99**±0.28	**71.14**±0.44	**2.6**
JOAO	**78.07**±0.47	**74.55**±0.41	**77.32**±0.54	**87.35**±1.02	**69.50**±0.36	**85.29**±1.35	**55.74**±0.63	70.21±3.08	3.3
JOAOv2	78.36±0.53	74.07±1.10	**77.40**±1.15	**87.67**±0.79	69.33±0.34	**86.42**±1.45	**56.03**±0.27	70.83±0.25	**2.8**

Experiments & Discussions.
On Bioinformatics Datasets

- JOAOv2 underperforms heuristic self-supervised pretext tasks, without incorporating domain expertise
- JOAOv2 generalizes better than GraphCL on unseen / domain specific datasets

Transfer learning

Methods	BBBP	Tox21	ToxCast	SIDER	ClinTox	MUV	HIV	BACE	PPI	A.R.↓
No pre-train.	65.8±4.5	74.0±0.8	63.4±0.6	57.3±1.6	58.0±4.4	71.8±2.5	75.3±1.9	70.1±5.4	64.8±1.0	6.6
Infomax	68.8±0.8	75.3±0.5	62.7±0.4	58.4±0.8	69.9±3.0	**75.3**±2.5	76.0±0.7	75.9±1.6	64.1±1.5	5.3
EdgePred	67.3±2.4	**76.0**±0.6	**64.1**±0.6	60.4±0.7	64.1±3.7	74.1±2.1	76.3±1.0	**79.9**±0.9	**65.7**±1.3	3.8
AttrMasking	64.3±2.8	**76.7**±0.4	**64.2**±0.5	**61.0**±0.7	71.8±4.1	**74.7**±1.4	77.2±1.1	**79.3**±1.6	**65.2**±1.6	3.1
ContextPred	68.0±2.0	**75.7**±0.7	**63.9**±0.6	**60.9**±0.6	65.9±3.8	**75.8**±1.7	**77.3**±1.0	**79.6**±1.2	64.4±1.3	3.4
GraphCL	**69.68**±0.67	73.87±0.66	62.40±0.57	**60.53**±0.88	**75.99**±2.65	69.80±2.66	**78.47**±1.22	75.38±1.44	**67.88**±0.85	4.6
JOAO	**70.22**±0.98	74.98±0.29	62.94±0.48	59.97±0.79	**81.32**±2.49	71.66±1.43	76.73±1.23	77.34±0.48	64.43±1.38	4.5
JOAOv2	**71.39**±0.92	74.27±0.62	63.16±0.45	60.49±0.74	**80.97**±1.64	73.67±1.00	**77.51**±1.17	75.49±1.27	63.94±1.59	4.3

Experiments & Discussions. On Large-Scale Datasets

- JOAOv2 achieves a better generalizability and scalability, outperforms on large-scale datasets

Semi-supervised learning on large-scale datasets

L.R.	Methods	ogbg-ppa	ogbg-code
1%	No pre-train.	16.04±0.74	6.06±0.01
	GraphCL	40.81±1.33	**7.66**±0.25
	JOAO	**47.19**±1.30	6.84±0.31
	JOAOv2	**44.30**±1.67	**7.74**±0.24
10%	No pre-train.	56.01±1.05	17.85±0.60
	GraphCL	57.77±1.25	**22.45**±0.17
	JOAO	**60.91**±0.83	22.06±0.30
	JOAOv2	**59.32**±1.11	**22.65**±0.22

Conclusions

- Problem: Handling heterogenous graph data with less manual efforts

- Contributions:
  - JOAO, a unified automatic framework
  - An instantiation as min-max optimization, with AGD for solution
  - JOAOv2, addressing distortion with multi-projection heads
  - Thorough experiments verifying the rationale and performance advantage

Further Discussions

- Limitation:
  - Automating augmentation selection, while requiring human to construct & config augmentation pool: “full” automation is still desired
- Potential:
  - In parallel to the principled formulation of bi-level optimization, a meta-learning formulation can also be pursued

References & Figures

- References
  - GCN: Semi-Supervised Classification with Graph Convolutional Networks
  - GAT: Graph Attention Networks
  - Graph completion: When Does Self-Supervision Help Graph Convolutional Networks?
  - AutoSSL: Automated Self-Supervised Learning for Graphs
  - GraphCL: Graph Contrastive Learning with Augmentation
  - MBRDL: Model Based Robust Deep Learning
  - Wang et al: Towards A Unified Min-Max Framework for Adversarial Exploration and Robustness
  - Boyd et al: Convex Optimization
  - DLA+AG: Self-Supervised Label Augmentation via Input Transformations
  - DistAug: Distribution Augmentation for Generative Modeling

- Figures
  - 1. https://www.euroscientist.com/imagine-a-social-network-like-facebook-with-no-facebook/
  - 2. https://www.thoughtco.com/what-is-a-polymer-820536
  - 3. https://www.e-education.psu.edu/ebf483/node/643

Thank you for listening!

Paper: https://arxiv.org/abs/2106.07594

Code: https://github.com/Shen-Lab/GraphCL_Automated