CIKM-22 Full Paper Track

Relational Self-Supervised Learning on Graphs

Namkyeong Lee, Dongmin Hyun,
Junseok Lee, Chanyoung Park

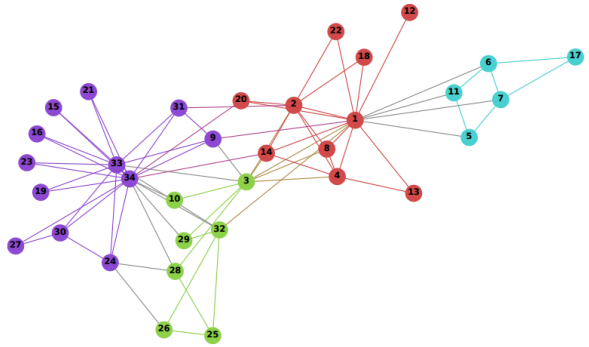
Korea Advanced Institute of Science and Technology (KAIST)
Pohang Institute of Science and Technology (POSTECH)

DSAIL Data Science & Artificial Intelligence

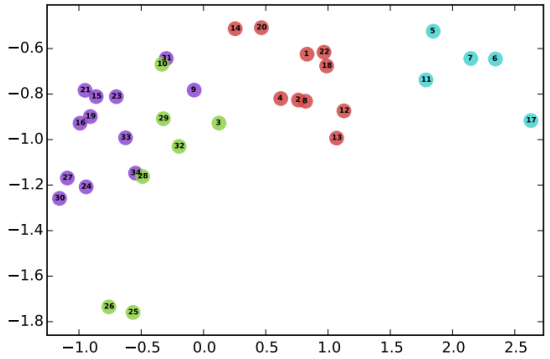
TABLE OF CONTENTS

BACKGROUND GRAPH REPRESENTATION LEARNING

Graph is ubiquitous data structure, employed extensively within computer science and related fields.



Graph Neural Network



Node Classification

Link Prediction

Community Detection

Network Similarity

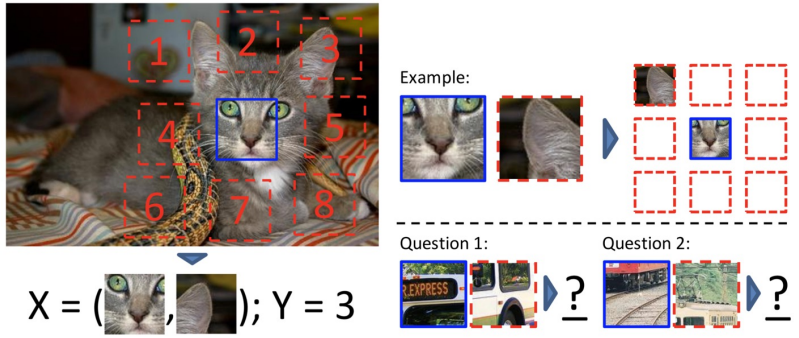
Graph representation learning means mapping the nodes or entire graphs, as points in a low-dimensional vector space.

Graph representation learning has been a powerful strategy for analyzing graph-structured data such as social network, especially by using Graph Neural Networks!

BACKGROUND SELF-SUPERVISED LEARNING ON IMAGES

Self-Supervised Learning automatically generates some kind of supervisory signal to solve some task.
(Typically, to learn representations of data or to automatically label a dataset.)

Key Idea
→ Define pretext training task that captures the information of the input data.
→ Use the dependencies among different dimensions of the input data!



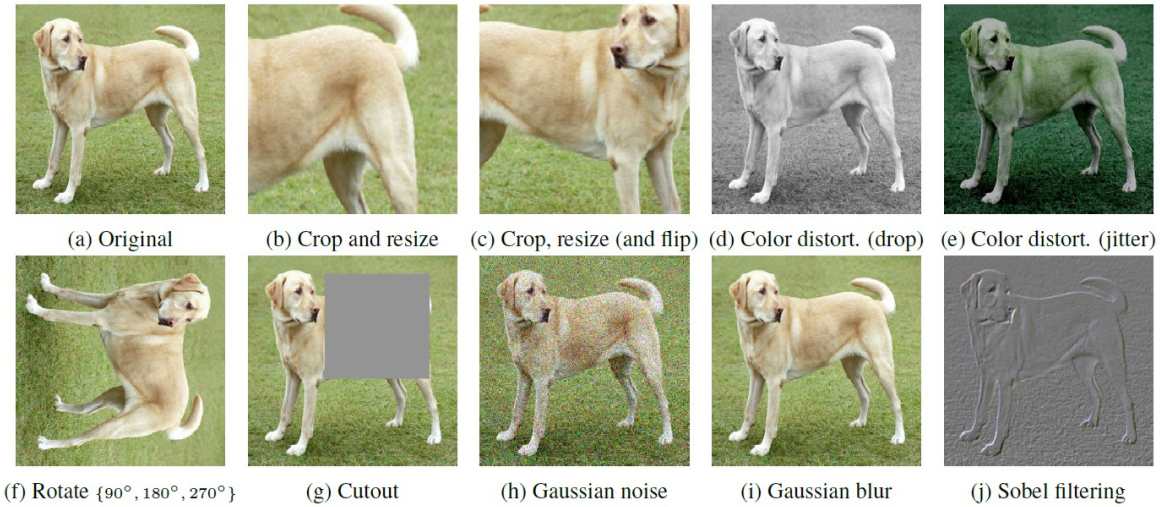
Self-Supervised Learning uses way more supervisory than supervised learning, and enormously more than reinforcement learning. That's why calling it "unsupervisory" is totally misleading.

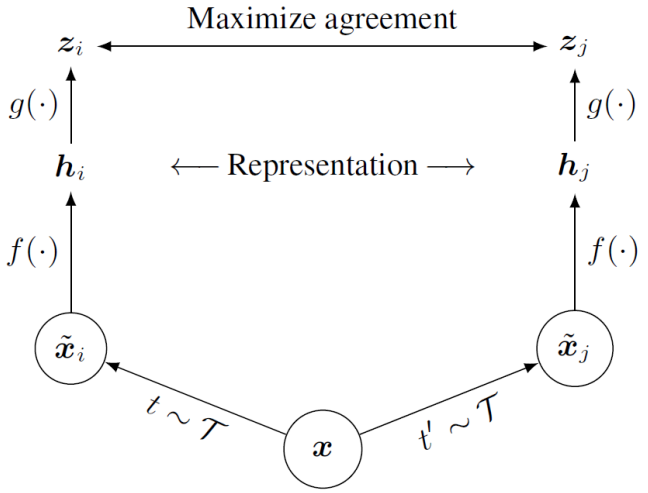
- Yann LeCun, 2019

DOERSCH, Carl; GUPTA, Abhinav; EFROS, Alexei A. Unsupervised visual representation learning by context prediction. In: *Proceedings of the IEEE international conference on computer vision*. 2015. p. 1422-1430.

BACKGROUND SELF-SUPERVISED LEARNING ON IMAGES _ SIMCLR

SimCLR is trained by reducing the distance between representations of augmented views of the same image (Positive), and increasing the distance between representations of augmented views from different images (Negative).



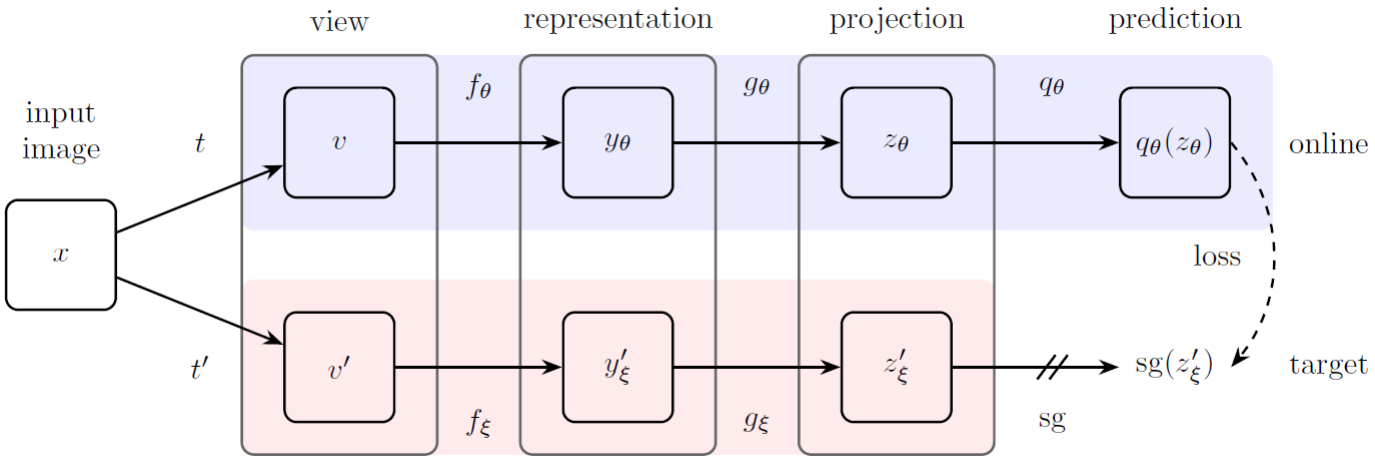


Sample mini batch of N examples.
→ Create 2N data points via Data Augmentation.
→ Given a positive pair, treat other 2(N-1) points as negative examples.
→ Instance Discrimination!

CHEN, Ting, et al. A simple framework for contrastive learning of visual representations. In: *International conference on machine learning*. PMLR, 2020. p. 1597-1607.

BACKGROUND SELF-SUPERVISED LEARNING ON IMAGES _ BYOL

BYOL learns representations of images without using negative samples
→ predicting the target representation with a given online representation



$$\theta \leftarrow \mathrm{optimizer}\left(\theta, \nabla_\theta \mathcal{L}^{\mathtt{BYOL}}_{\theta,\xi}, \eta\right)$$

Online network

$$\xi \leftarrow \tau\xi + (1-\tau)\theta$$

Target network Online network

At each training iteration, online network is trained to minimize the cosine similarity loss,
while target network's parameters are updated using the exponential moving average of online network's parameter.

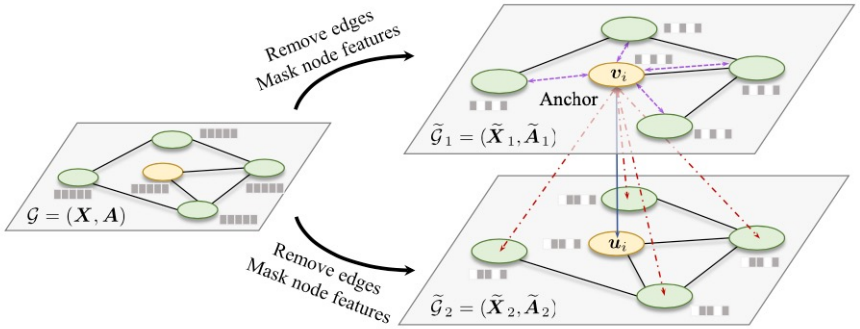
GRILL, Jean-Bastien, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

BACKGROUND SELF-SUPERVISED LEARNING ON GRAPHS

Inspired by the success of contrastive methods in computer vision applied on images, those methods have been recently adopted to graph-structured data.

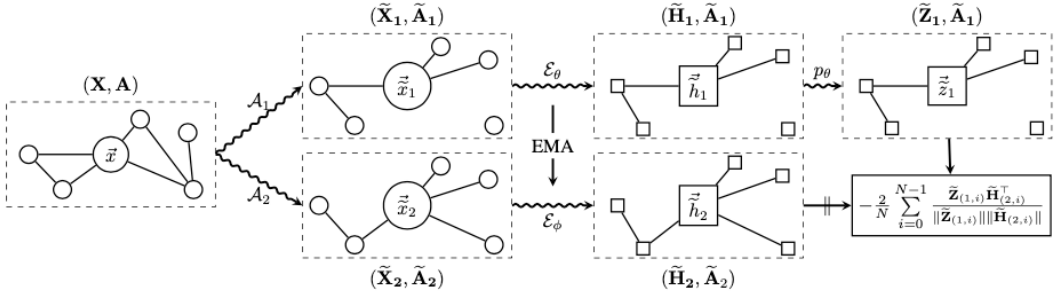
GRACE (Inspired by SimCLR)

Learns representations by pulling the representation of the same node in the two augmented views of graph while pushing apart representations of every other node.



BGRL (Inspired by BYOL)

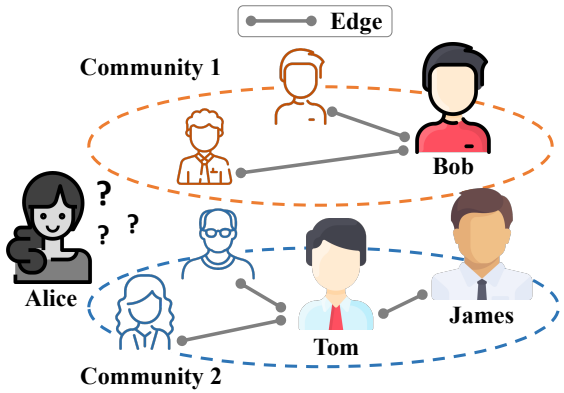
Learns representations by predicting the augmented view of node itself without using negative samples.



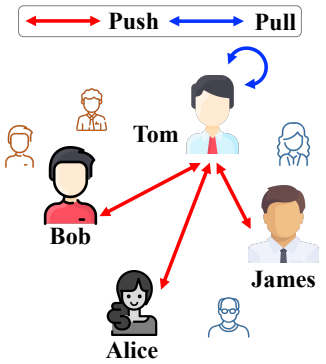
Learning features that are invariant under the augmentation!

MOTIVATION GRAPHS EXHIBIT RELATIONAL INFORMATION

Recent graph representation learning (GRL) methods do not reflect the nature of the graph
→ Recall that *Graphs exhibit relational information among nodes*



(a) Graph-Structured Data



(b) Recent GRL methods

Previous methods (contrastive & non-contrastive) cannot fully benefit from relational information of graph structured data

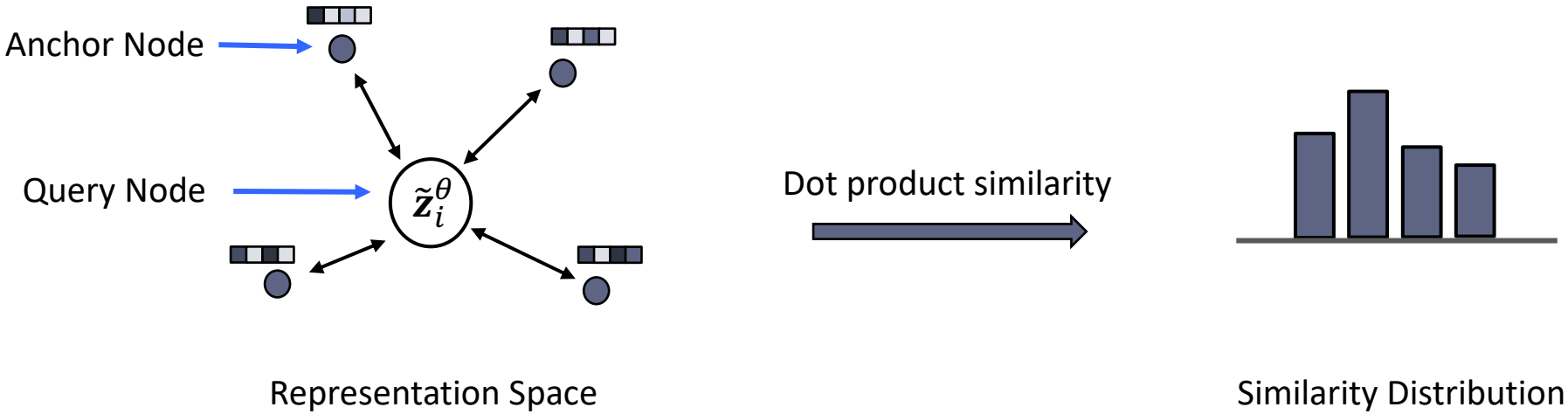
Moreover,
Contrastive methods (e.g. GRACE) are prone to sampling bias issue

Sampling Bias?
Some negative samples are in fact semantically similar to query nodes

How about learning augmentation-invariant relationships?

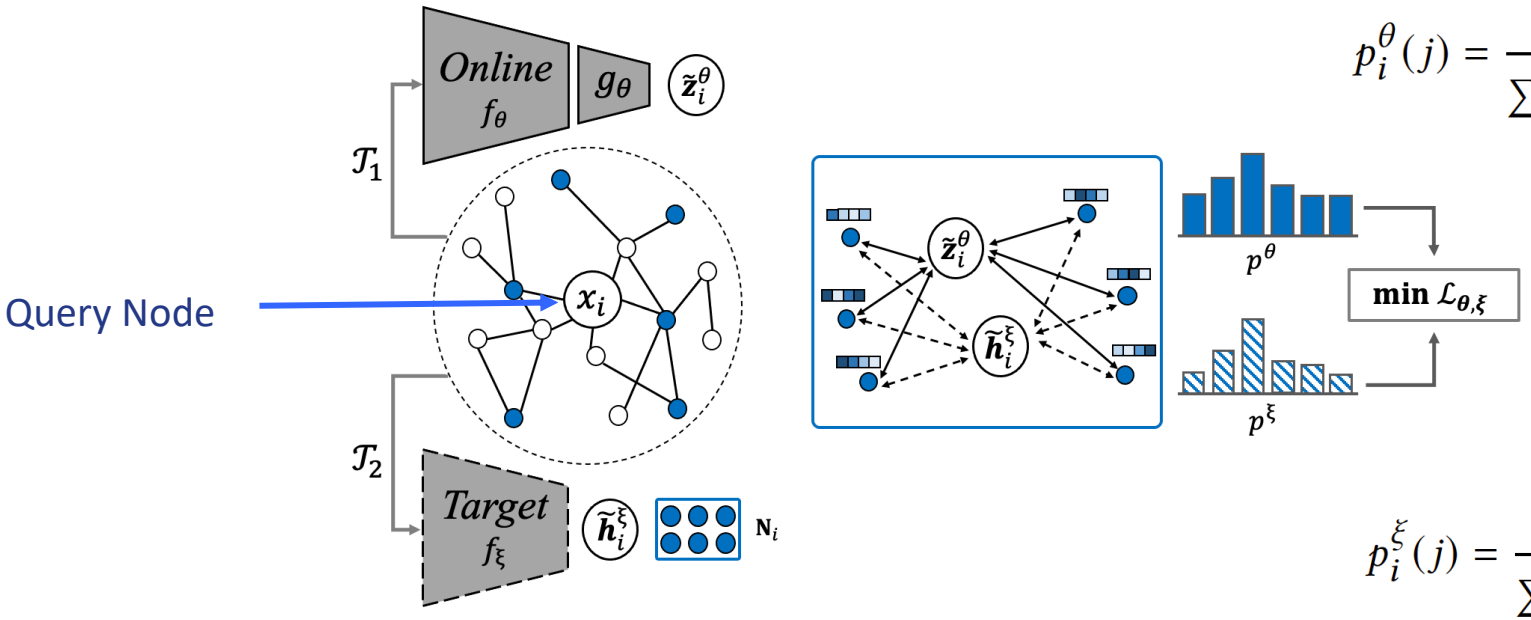
METHODOLOGY RELATIONAL SELF-SUPERVISED LEARNING ON GRAPHS

How can we define relationship between a query node and anchor nodes?



We define cosine similarity as relationship between a query node and anchor nodes

METHODOLOGY RELATIONAL SELF-SUPERVISED LEARNING ON GRAPHS



$$p_i^\theta(j) = \frac{\exp(\mathrm{sim}(\tilde{\mathbf{z}}_i^\theta, \tilde{\mathbf{h}}_j^\xi)/\tau_\theta)}{\sum_{k\in\mathbf{N}_i}\exp(\mathrm{sim}(\tilde{\mathbf{z}}_i^\theta, \tilde{\mathbf{h}}_k^\xi)/\tau_\theta)}, \forall v_j \in \mathbf{N}_i$$

$$\mathcal{L}_{\theta,\xi} = \sum_{v_i\in\mathcal{V}} KL(p_i^\theta \,||\, p_i^\xi)$$

$$p_i^\xi(j) = \frac{\exp(\mathrm{sim}(\tilde{\mathbf{h}}_i^\xi, \tilde{\mathbf{h}}_j^\xi)/\tau_\xi)}{\sum_{k\in\mathbf{N}_i}\exp(\mathrm{sim}(\tilde{\mathbf{h}}_i^\xi, \tilde{\mathbf{h}}_k^\xi)/\tau_\xi)}, \forall v_j \in \mathbf{N}_i$$

Online network is trained to mimic the relational information captured by target network

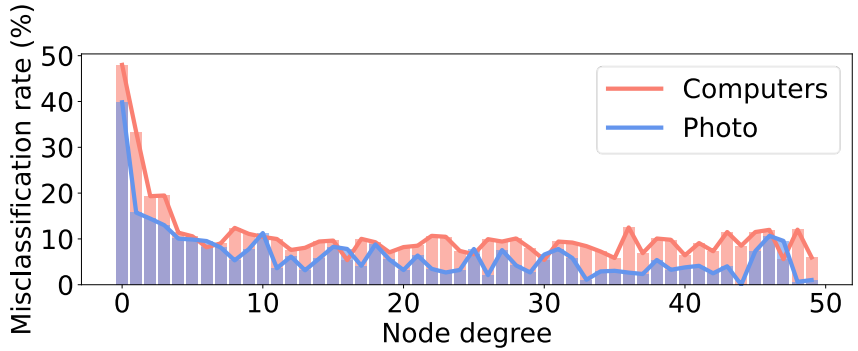
→ **Learning augmentation-invariant relationship!** (Instead of augmentation-invariant node representation)

Next research question: How to sample anchor nodes?

Diverse relational information regarding both global and local perspectives should be considered

METHODOLOGY GLOBAL ANCHOR NODE SAMPLING

Global anchor nodes: Structurally distant nodes



Misclassification rate for certain degree of nodes

Misclassification rate of low-degree nodes is significantly high
→ *Degree-bias* issue!

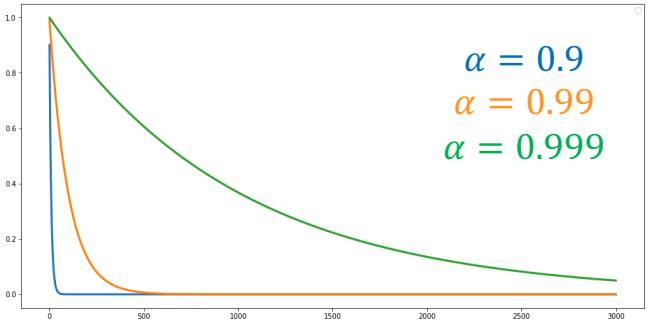
We should focus on low-degree nodes while training RGRL

Approach: Sample anchor nodes from ***inverse degree-weighted distribution***

→ Sample more from low-degree nodes

$$w_j = \alpha^{\log(\deg_j+1)} + \beta$$

$0 < \alpha < 1$



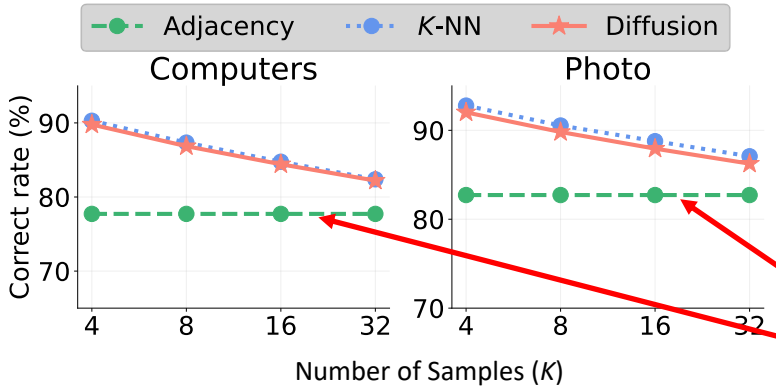
Setting $0 < \alpha < 1$ approximates the misclassification rate

$$p_{sample}(j) = \frac{w_j}{\sum_{v_k \in \mathcal{V}} w_k}, \forall v_j \in \mathcal{V}$$

Inverse degree-weighted distribution

METHODOLOGY LOCAL ANCHOR NODE SAMPLING

Local anchor nodes: Structurally close nodes



Ratio of its neighboring nodes being the same label

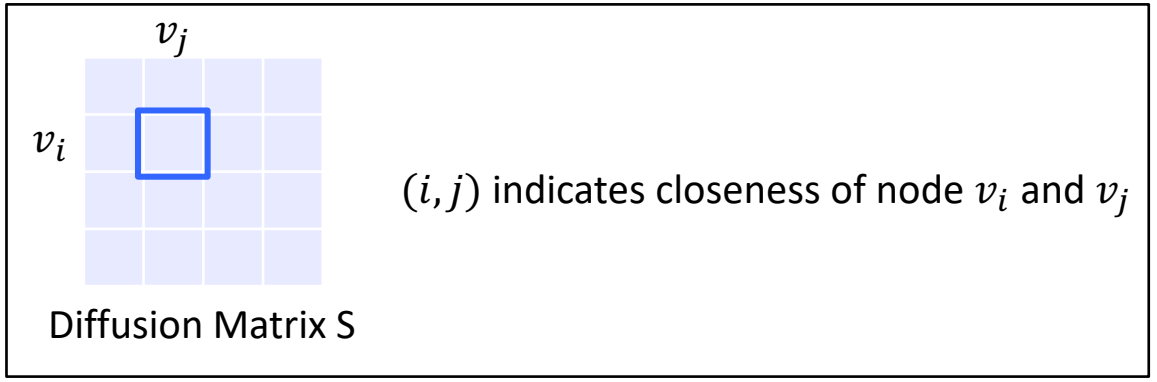
- Adjacency may fail to capture fine-grained relationship among nodes
  - ex) “Data Mining” vs. “Machine Learning” community
    - Structurally close but different class
- **We should sample anchor nodes that are**
  - 1) Structurally close with query node in the graph structure
  - 2) Share the same label with the query node

Approach: Sample anchor nodes based on ***diffusion score matrix (Personalized PageRank)***

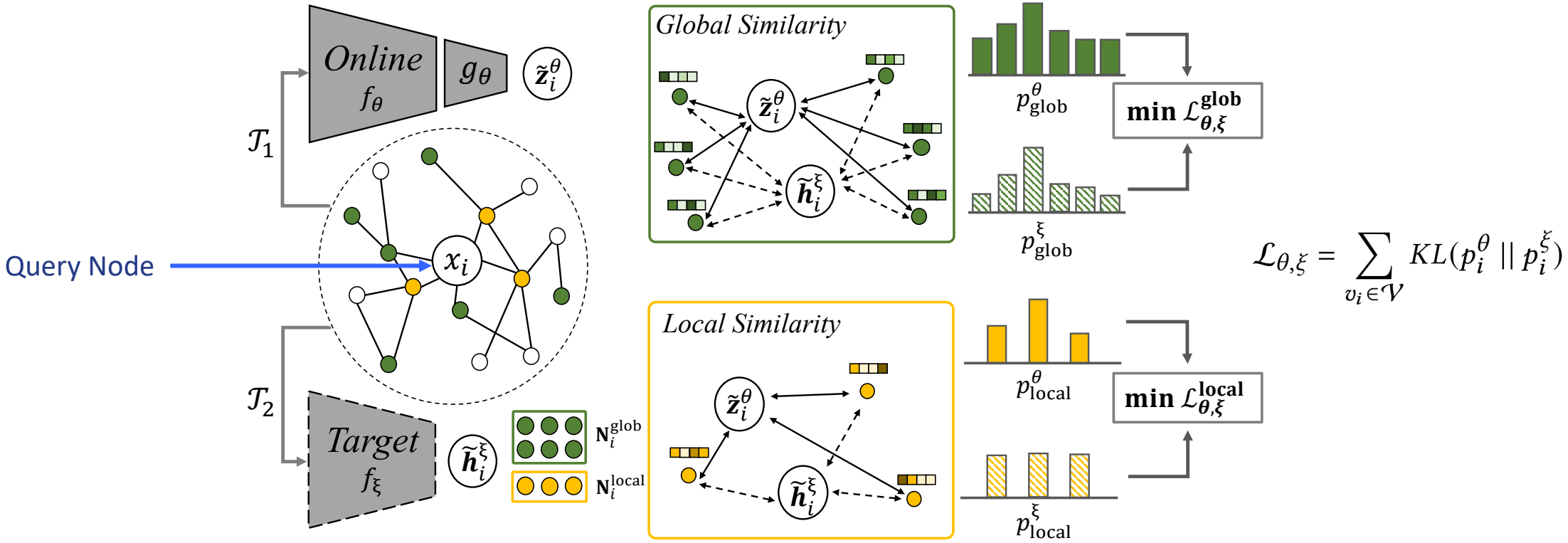
$$\mathbf{S} = \sum_{k=0}^{\infty} t(1-t)^k \mathbf{T}^k$$

t : Teleport probability ($t \in (0,1)$)

$\mathbf{T}$: Symmetric transition matrix



METHODOLOGY RELATIONAL SELF-SUPERVISED LEARNING ON GRAPHS



$$\mathcal{L}_{\theta,\xi} = \sum_{v_i \in \mathcal{V}} KL(p_i^\theta \,||\, p_i^\xi)$$

$$p_i^\theta(j) = \frac{\exp(\text{sim}(\tilde{\mathbf{z}}_i^\theta, \tilde{\mathbf{h}}_j^\xi)/\tau_\theta)}{\sum_{k \in \mathbf{N}_i} \exp(\text{sim}(\tilde{\mathbf{z}}_i^\theta, \tilde{\mathbf{h}}_k^\xi)/\tau_\theta)}, \forall v_j \in \mathbf{N}_i$$

(Relational information regarding **online network**)

$$p_i^\xi(j) = \frac{\exp(\text{sim}(\tilde{\mathbf{h}}_i^\xi, \tilde{\mathbf{h}}_j^\xi)/\tau_\xi)}{\sum_{k \in \mathbf{N}_i} \exp(\text{sim}(\tilde{\mathbf{h}}_i^\xi, \tilde{\mathbf{h}}_k^\xi)/\tau_\xi)}, \forall v_j \in \mathbf{N}_i$$

(Relational information regarding **target network**)

DISCUSSION HOW RGRL OVERCOMES THE LIMITATIONS OF PREVIOUS WORKS

Previous works: **1) Contrastive methods**, **2) Non-contrastive methods**

1) Limitation of **Contrastive methods**

- Sampling bias: Simply treating all other nodes as negatives incurs false negatives
- Another problem occurs when sampling bias is combined with the **contrastive loss** that is defined as follows [1]:

Positive pair

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

Negative pair

- As τ decreases, the model gives larger penalty to hard negative samples (push away)
  - Makes sense if we know true negatives (supervised setting)
  - But, harmful in self-supervised learning where false negatives exist

Contrastive loss is “Hardness-aware loss”

- Gives larger penalties to similar nodes → similar nodes that belong to negative samples become more dissimilar

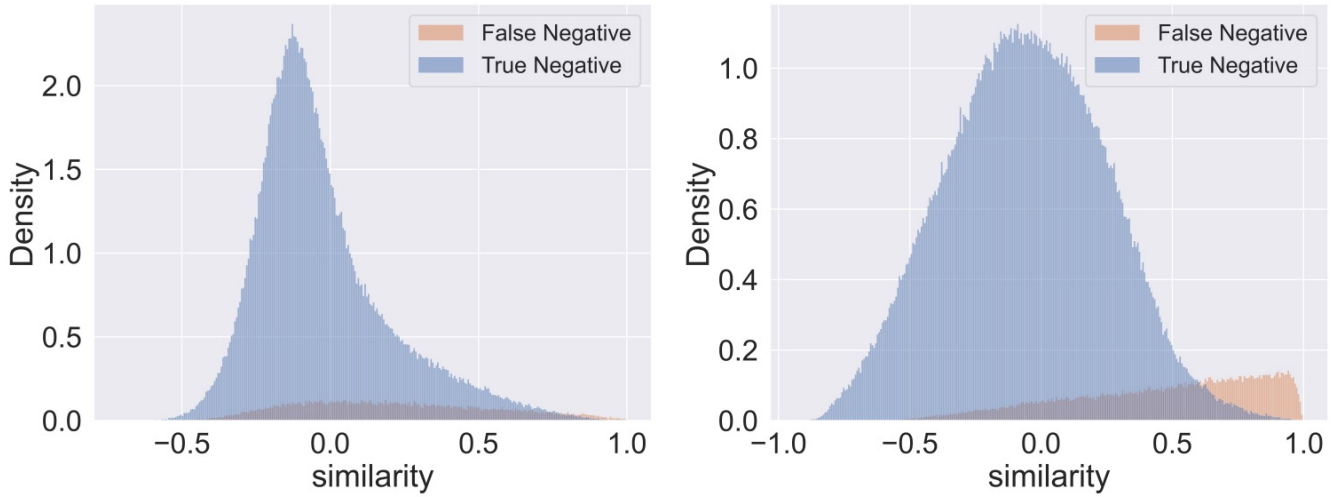
Thus, false negative is trained to be more dissimilar

[1] Understanding the behaviour of contrastive loss, CVPR 2021

DISCUSSION HOW RGRL OVERCOMES THE LIMITATIONS OF PREVIOUS WORKS

The problem gets even more severe in graph domain,

- In graphs, most “HARD” negatives are indeed “FALSE” negatives



(a) CIFAR-10 (Image) (b) Coauthor-CS (Graph)

RGRL relaxes the strict binary classification of contrastive methods with **soft labeling**

- RGRL can decide how much to push or pull other nodes based on the relational information among the nodes with out relying on the binary decisions of positives and negatives

DISCUSSION HOW RGRL OVERCOMES THE LIMITATIONS OF PREVIOUS WORKS

Previous works: **1) Contrastive methods**, **2) Non-contrastive methods**

2) Limitation of **Non-contrastive methods**

- Since we don't use any negative samples, **node features should be fully informative**
  - Performance actually degrades if features contain noise (as will be shown later)
  - Overfit to a few non-informative feature

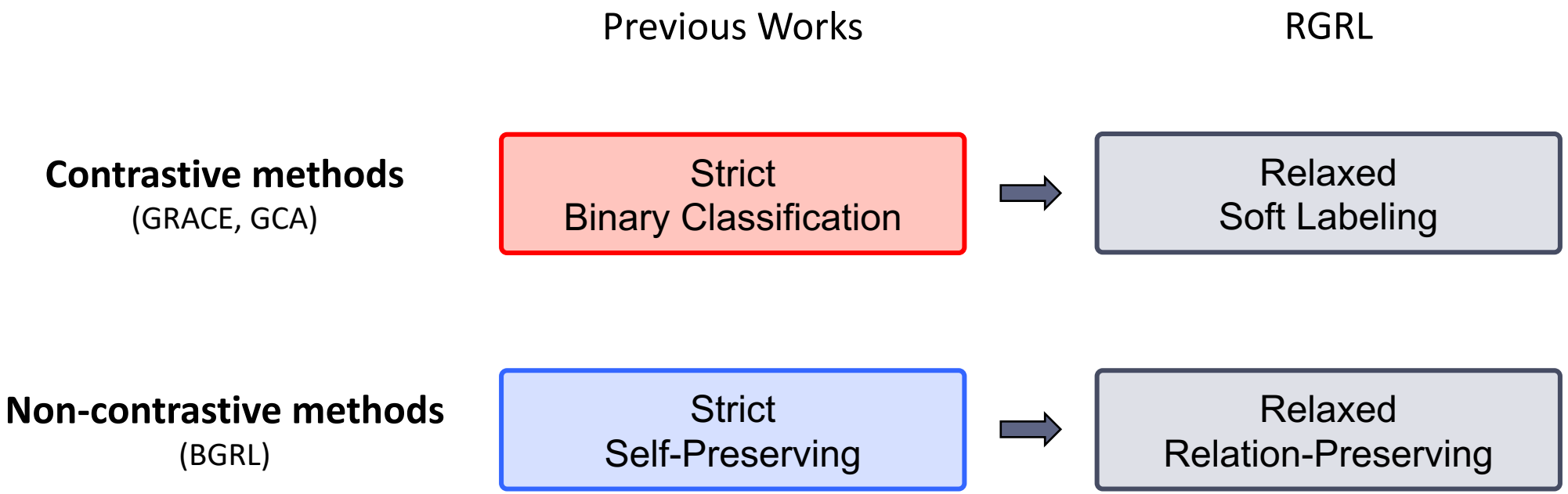
RGRL alleviates the overfitting problem with a little help from other nodes in the graph

- Learn from the relationship with other nodes

RGRL relaxes the strict self-preserving loss with **relation-preserving loss**

- Allows the representations to vary as long as the **relationship among the representations is preserved**

DISCUSSION HOW RGRL OVERCOMES THE LIMITATIONS OF PREVIOUS WORKS



RGRL achieves the best of both worlds by relaxing strict constraints of previous works

EXPERIMENTS NODE CLASSIFICATION

	WikiCS	Computers	Photo	Co.CS	Co.Physics
GCN	77.19 (0.12)	86.51 (0.54)	92.42 (0.22)	93.03 (0.31)	95.65 (0.16)
Feats.	71.98 (0.00)	73.81 (0.00)	78.53 (0.00)	90.37 (0.00)	93.58 (0.00)
n2v	71.79 (0.05)	84.39 (0.08)	89.67 (0.12)	85.08 (0.03)	91.19 (0.04)
DW	74.35 (0.06)	85.68 (0.06)	89.44 (0.11)	84.61 (0.22)	91.77 (0.15)
DW+Feats.	77.21 (0.03)	86.28 (0.07)	90.05 (0.08)	87.70 (0.04)	94.90 (0.09)
DGI	75.35 (0.14)	83.95 (0.47)	91.61 (0.22)	92.15 (0.63)	94.51 (0.52)
GMI	74.85 (0.08)	82.21 (0.31)	90.68 (0.17)	OOM	OOM
MVGRL	77.52 (0.08)	87.52 (0.11)	91.74 (0.07)	92.11 (0.12)	95.33 (0.03)
GRACE	78.25 (0.65)	88.15 (0.43)	92.52 (0.32)	92.60 (0.11)	OOM
GCA	78.30 (0.62)	88.49 (0.51)	92.99 (0.27)	92.76 (0.16)	OOM
CCA-SSG	77.88 (0.41)	87.01 (0.41)	92.59 (0.25)	92.77 (0.17)	95.16 (0.10)
BGRL	79.60 (0.60)	89.23 (0.34)	93.06 (0.30)	92.90 (0.15)	95.43 (0.09)
RGRL	**80.29 (0.72)**	**89.70 (0.44)**	**93.43 (0.31)**	**92.94 (0.13)**	**95.46 (0.10)**

Performance on node classification tasks

	Transductive						Inductive	
	Cora	Cite-seer	Pub-med	Cora Full	ogbn-arXiv Valid	ogbn-arXiv Test	Reddit	PPI
GRACE	83.38 (0.95)	70.79 (0.83)	83.96 (0.29)	64.19 (0.36)	OOM	OOM	94.84 (0.03)	67.12 (0.05)
GCA	82.79 (1.01)	70.70 (0.91)	84.19 (0.32)	64.34 (0.42)	OOM	OOM	94.85 (0.06)	66.72 (0.08)
CCA-SSG	83.01 (0.66)	70.35 (1.23)	84.81 (0.22)	64.09 (0.37)	59.43 (0.05)	58.50 (0.08)	94.89 (0.02)	66.09 (0.01)
BGRL	82.82 (0.86)	69.06 (0.80)	**86.16 (0.19)**	63.94 (0.39)	70.66 (0.06)	69.61 (0.09)	94.90 (0.04)	68.89 (0.08)
RGRL	**83.98 (0.78)**	**71.29 (0.87)**	85.33 (0.20)	**64.62 (0.39)**	**72.34 (0.09)**	**71.49 (0.08)**	**95.04 (0.03)**	**69.28 (0.06)**

Performance on various datasets (transductive/inductive)

RGRL outperforms previous methods that overlook the relationship among nodes

EXPERIMENTS NODE CLASSIFICATION

	WikiCS	Computers	Photo	Co.CS	Co.Physics
GCN	77.19 (0.12)	86.51 (0.54)	92.42 (0.22)	93.03 (0.31)	95.65 (0.16)
Feats.	71.98 (0.00)	73.81 (0.00)	78.53 (0.00)	90.37 (0.00)	93.58 (0.00)
n2v	71.79 (0.05)	84.39 (0.08)	89.67 (0.12)	85.08 (0.03)	91.19 (0.04)
DW	74.35 (0.06)	85.68 (0.06)	89.44 (0.11)	84.61 (0.22)	91.77 (0.15)
DW+Feats.	77.21 (0.03)	86.28 (0.07)	90.05 (0.08)	87.70 (0.04)	94.90 (0.09)
DGI	75.35 (0.14)	83.95 (0.47)	91.61 (0.22)	92.15 (0.63)	94.51 (0.52)
GMI	74.85 (0.08)	82.21 (0.31)	90.68 (0.17)	OOM	OOM
MVGRL	77.52 (0.08)	87.52 (0.11)	91.74 (0.07)	92.11 (0.12)	95.33 (0.03)
GRACE	78.25 (0.65)	88.15 (0.43)	92.52 (0.32)	92.60 (0.11)	OOM
GCA	78.30 (0.62)	88.49 (0.51)	92.99 (0.27)	92.76 (0.16)	OOM
CCA-SSG	77.88 (0.41)	87.01 (0.41)	92.59 (0.25)	92.77 (0.17)	95.16 (0.10)
BGRL	79.60 (0.60)	89.23 (0.34)	93.06 (0.30)	92.90 (0.15)	95.43 (0.09)
RGRL	**80.29 (0.72)**	**89.70 (0.44)**	**93.43 (0.31)**	**92.94 (0.13)**	**95.46 (0.10)**

Performance on node classification tasks

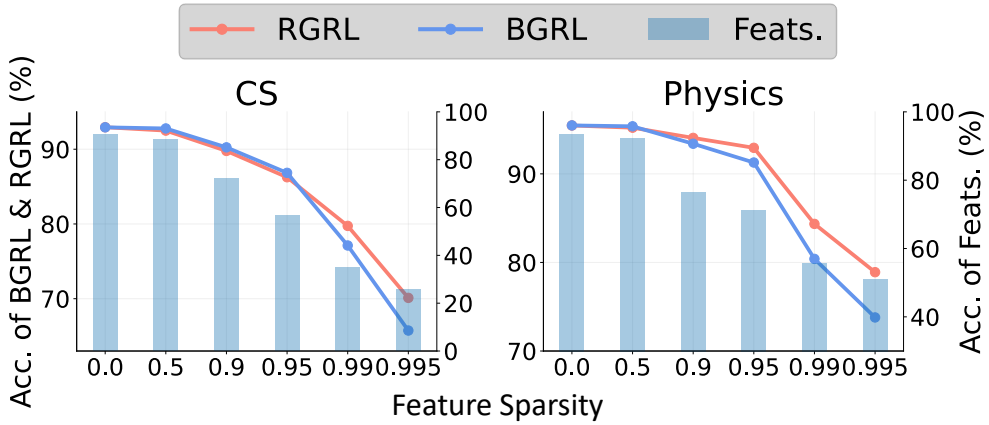
Less Informative Features

More Informative Features

EXPERIMENTS NODE CLASSIFICATION

	WikiCS	Computers	Photo	Co.CS	Co.Physics
GCN	77.19 (0.12)	86.51 (0.54)	92.42 (0.22)	93.03 (0.31)	95.65 (0.16)
Feats.	71.98 (0.00)	73.81 (0.00)	78.53 (0.00)	90.37 (0.00)	93.58 (0.00)
n2v	71.79 (0.05)	84.39 (0.08)	89.67 (0.12)	85.08 (0.03)	91.19 (0.04)
DW	74.35 (0.06)	85.68 (0.06)	89.44 (0.11)	84.61 (0.22)	91.77 (0.15)
DW+Feats.	77.21 (0.03)	86.28 (0.07)	90.05 (0.08)	87.70 (0.04)	94.90 (0.09)
DGI	75.35 (0.14)	83.95 (0.47)	91.61 (0.22)	92.15 (0.63)	94.51 (0.52)
GMI	74.85 (0.08)	82.21 (0.31)	90.68 (0.17)	OOM	OOM
MVGRL	77.52 (0.08)	87.52 (0.11)	91.74 (0.07)	92.11 (0.12)	95.33 (0.03)
GRACE	78.25 (0.65)	88.15 (0.43)	92.52 (0.32)	92.60 (0.11)	OOM
GCA	78.30 (0.62)	88.49 (0.51)	92.99 (0.27)	92.76 (0.16)	OOM
CCA-SSG	77.88 (0.41)	87.01 (0.41)	92.59 (0.25)	92.77 (0.17)	95.16 (0.10)
BGRL	79.60 (0.60)	89.23 (0.34)	93.06 (0.30)	92.90 (0.15)	95.43 (0.09)
RGRL	**80.29 (0.72)**	**89.70 (0.44)**	**93.43 (0.31)**	**92.94 (0.13)**	**95.46 (0.10)**

Performance on node classification tasks



Classification accuracy over feature sparsity

Dataset with less informative features

→ Large improvements in performance

→ External self-supervisory signals from other nodes help RGRL to learn from less informative features

Dataset with more informative features

→ RGRL is more robust than BGRL as the quality of input features gets worse

APPENDIX EXPERIMENTS: LINK PREDICTION & MULTIPLEX NETWORK

		Computers		Photo		Co. CS		Co. Physics	
		AUC	AP	AUC	AP	AUC	AP	AUC	AP
Random Neg.	GRACE	0.939	0.939	0.962	0.960	0.970	0.970	OOM	OOM
	GCA	0.954	0.954	0.965	0.960	**0.971**	**0.970**	OOM	OOM
	CCA-SSG	0.961	0.959	0.973	0.970	0.949	0.950	0.943	0.936
	BGRL	0.964	0.961	0.978	0.976	0.952	0.948	0.952	0.947
	RGRL	**0.974**	**0.972**	**0.983**	**0.981**	0.967	0.968	**0.964**	**0.963**
Hard Neg.	GRACE	0.933	0.933	0.939	0.929	0.870	0.868	OOM	OOM
	GCA	0.938	0.929	0.948	0.939	0.874	0.869	OOM	OOM
	CCA-SSG	0.954	0.952	0.947	0.943	0.847	0.835	0.871	0.856
	BGRL	0.959	0.956	0.959	0.956	0.845	0.832	0.903	0.892
	RGRL	**0.969**	**0.968**	**0.967**	**0.964**	**0.878**	**0.881**	**0.923**	**0.919**

Performance on link prediction

Random Negative

→ Randomly select pair of nodes that are not connected

Hard Negative

→ Select pair of nodes that are within 3-hop distances

→ Harder than random negatives!

Improvements on hard negative edges is more significant than random negatives

→ RGRL detects more fine-grained relational information

EXPERIMENTS QUALITATIVE ANALYSIS – CASE STUDY

Query Author	Model	Top-1 Similar Author	# Co-authored Papers	Student?
Jiawei Han	BGRL	Ke Wang	14	✗
	RGRL	Xifeng Yan	87	✓
Christos Faloutsos	BGRL	Tina Eliassi-Rad	27	✗
	RGRL	Hanghang Tong	47	✓

Case 1) Which author is the most similar?

- Discovers author who have more co-authored papers
- Discovers former Ph.D. students of the query authors

Advisor-advisee relationship
→ Core relationship in the academia network!

Query Author	Model	Top-1 Similar Author	# Co-authored Papers	Research Keywords
Jiawei Han	BGRL	Zhou Aoying	0	Query Processing
	RGRL	Ee-Peng Lim	0	Data & Text Mining
Christos Faloutsos	BGRL	Michael J. Pazzani	0	Machine Learning
	RGRL	David Jensen	2	Machine Learning

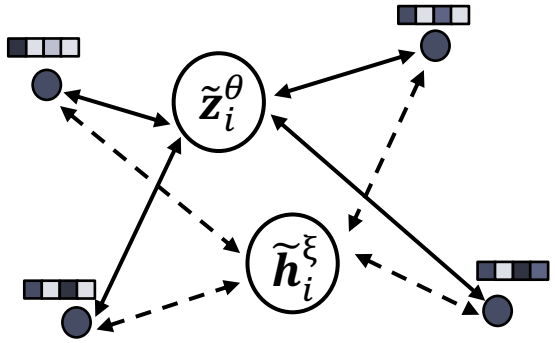
Case 2) Which author will co-work in the future?

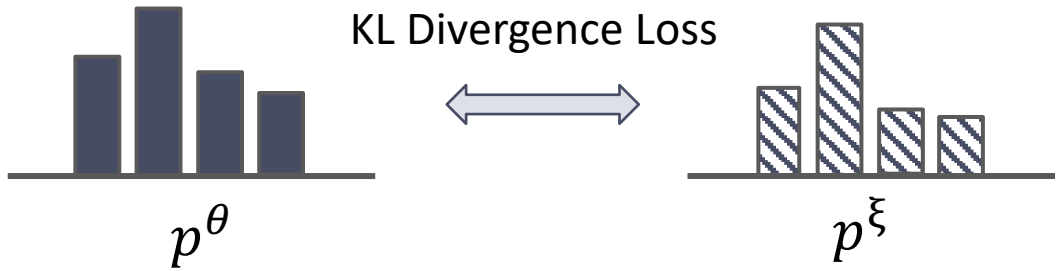
- Discovers author of more relevant research area
- Discovers author of actually co-authored in the past

RGRL discovers core relationship and meaningful knowledge that is not revealed in the given graph

CONCLUSION

Proposed to learn node representations such that the relationship among nodes is invariant to augmentations
→ **"Augmentation-Invariant" Relationship**





By learning augmentation invariant relationship,
RGRL relaxes several strict constraints of previous works thereby achieving the best of both worlds

Extensive experiments on 14 datasets demonstrate that RGRL
1) is robust to less informative or noisy features
2) improves performance on low degree nodes
3) discovers core relationship and meaningful knowledge that is not revealed in the given graph

THANK YOU!

[Full Paper] https://arxiv.org/abs/2208.10493

[Source Code] https://github.com/Namkyeong/RGRL

[Lab Homepage] http://dsail.kaist.ac.kr/

[Contact] namkyeong96@kaist.ac.kr

APPENDIX EXPERIMENTS: LINK PREDICTION & MULTIPLEX NETWORK

		Computers		Photo		Co. CS		Co. Physics	
		AUC	AP	AUC	AP	AUC	AP	AUC	AP
Random Neg.	GRACE	0.939	0.939	0.962	0.960	0.970	0.970	OOM	OOM
	GCA	0.954	0.954	0.965	0.960	**0.971**	**0.970**	OOM	OOM
	CCA-SSG	0.961	0.959	0.973	0.970	0.949	0.950	0.943	0.936
	BGRL	0.964	0.961	0.978	0.976	0.952	0.948	0.952	0.947
	RGRL	**0.974**	**0.972**	**0.983**	**0.981**	0.967	0.968	**0.964**	**0.963**
Hard Neg.	GRACE	0.933	0.933	0.939	0.929	0.870	0.868	OOM	OOM
	GCA	0.938	0.929	0.948	0.939	0.874	0.869	OOM	OOM
	CCA-SSG	0.954	0.952	0.947	0.943	0.847	0.835	0.871	0.856
	BGRL	0.959	0.956	0.959	0.956	0.845	0.832	0.903	0.892
	RGRL	**0.969**	**0.968**	**0.967**	**0.964**	**0.878**	**0.881**	**0.923**	**0.919**

Performance on link prediction

Dataset	IMDB		DBLP	
Metric	Macro-F1	Micro-F1	Macro-F1	Micro-F1
HAN	0.599	0.607	0.716	0.708
DMGI	0.648	0.648	0.771	0.766
$DMGI_{attn}$	0.602	0.606	0.778	0.770
HDMI	0.650	**0.658**	0.820	0.811
BGRL	0.631	0.634	0.819	0.807
RGRL	**0.653**	**0.658**	**0.830**	**0.818**

Performance on multiplex network

Link Prediction

→ Improvements on hard negative edges (within 3-hop distances) is more significant than random negatives

→ RGRL detects more fine-grained relational information

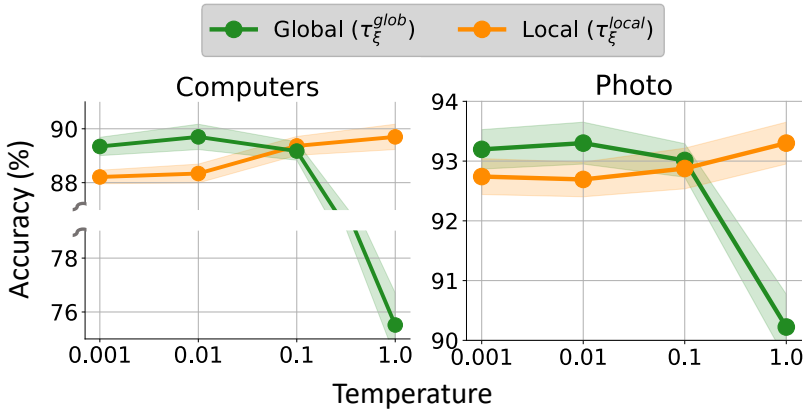
Multiplex Network

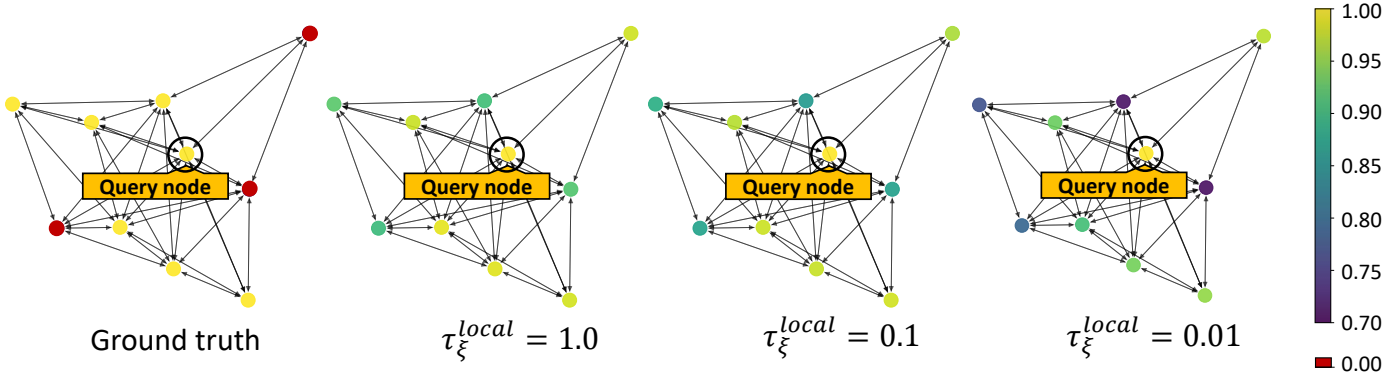
→ RGRL can learn from diverse relationship inherent in multiplex network due to its flexibility

APPENDIX EXPERIMENTS: MODEL ANALYSIS

$$p_i^{\xi}(j) = \frac{\exp(\mathrm{sim}(\tilde{\mathbf{h}}_i^{\xi}, \tilde{\mathbf{h}}_j^{\xi})/\tau_{\xi})}{\sum_{k \in \mathbf{N}_i} \exp(\mathrm{sim}(\tilde{\mathbf{h}}_i^{\xi}, \tilde{\mathbf{h}}_k^{\xi})/\tau_{\xi})}, \forall v_j \in \mathbf{N}_i$$

Recall that temperature controls sharpness of similarity distribution
→ Learns discriminative features as temperature decreases





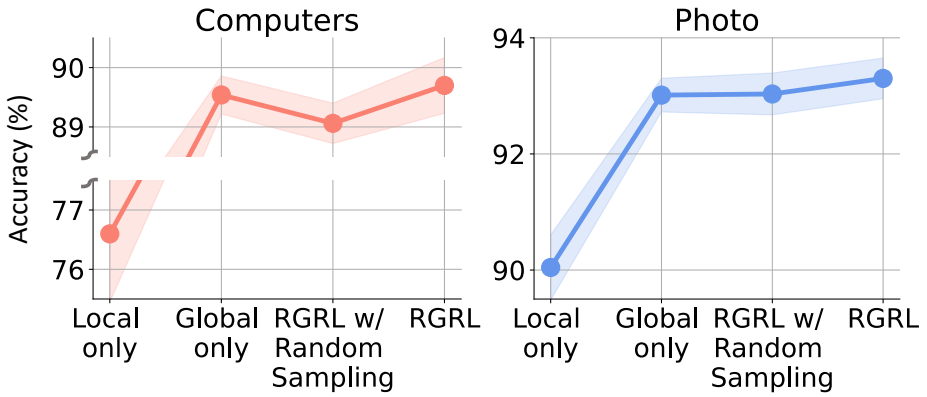
Global Temperature

→ Target distribution should be sharpened to provide strong supervisory signal for the model training

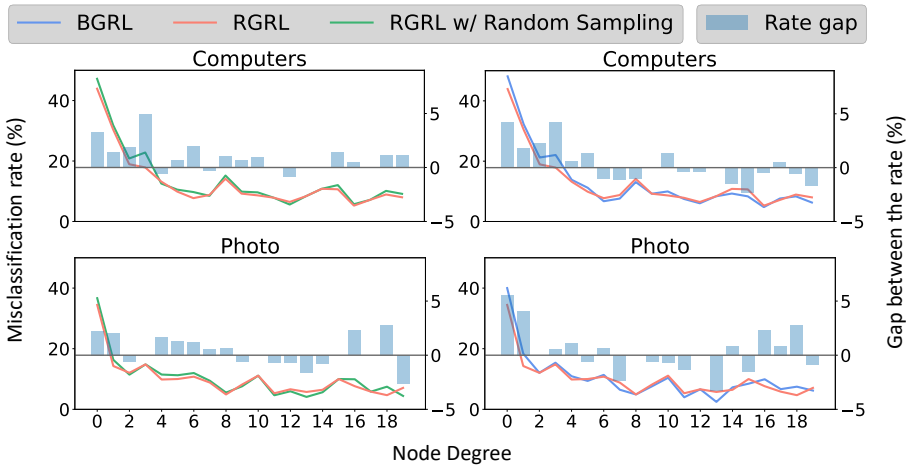
Local Temperature

→ Less discriminative features are required (high temperature)
→ Structurally close and semantically identical nodes should be close in representation space

APPENDIX EXPERIMENTS: MODEL ANALYSIS



Ablation Studies



Misclassification rate comparison

Considering the global similarity (i.e., Global Only) is more beneficial than considering the local similarity
→ However, considering the both perspective (i.e., RGRL) shows the best performance

RGRL's inverse degree sampling strategy successfully alleviates the *degree-bias* issue