Cluster-Enhanced Contrastive Learning on Graphs

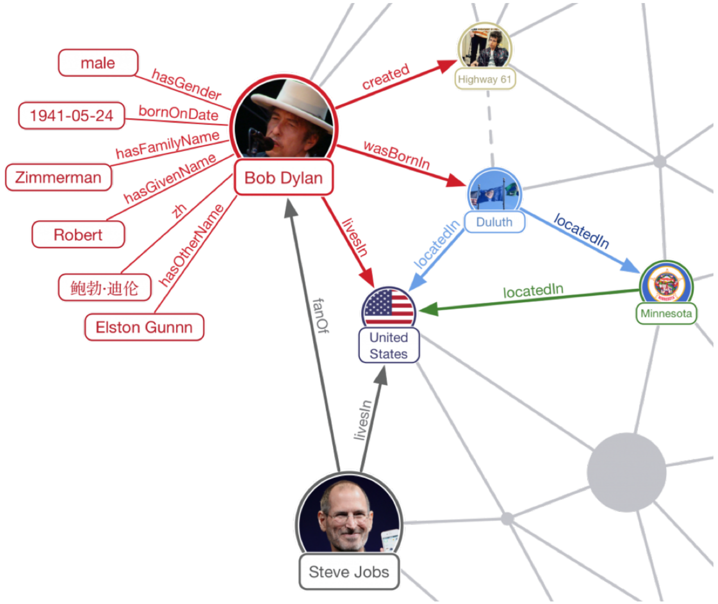
Presented by Jing Zhang (RUC)

Collaborated with Yanling Wang (RUC), Hongzhi Yin (UQ), Yuxiao Dong (THU)

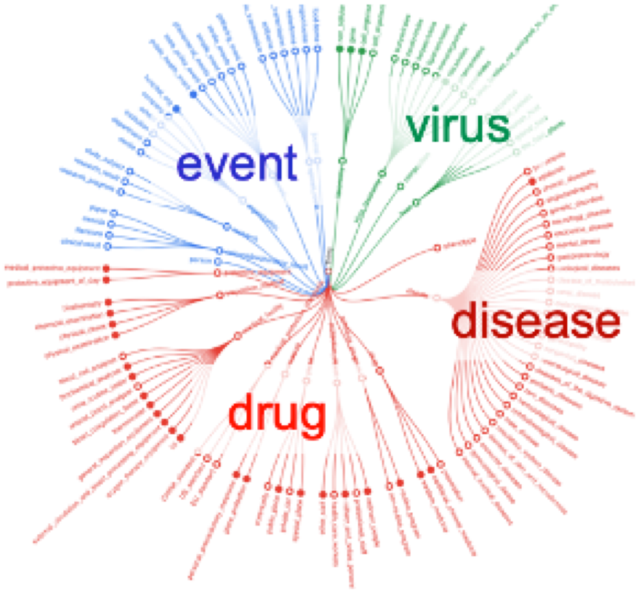
Shasha Guo (RUC), Haoyang Li (RUC), Cuiping Li (RUC), and Hong Chen (RUC)

Network Data

Social Network



Knowledge Graph



Medical Graph

Supervised Learning for Graph Neural Networks

- End-to-End Training

Data ⇨ Encoder ⇨ Classifier ⇨ Cross-entropy Loss

- Two-Stage Training

Data ⇨ Encoder ⇨ Pre-train Loss

(1) Pre-Training Stage
Reduce the difficulty of learning the classifier.

Encoder ⇩ Classifier ⇨ Cross-entropy Loss

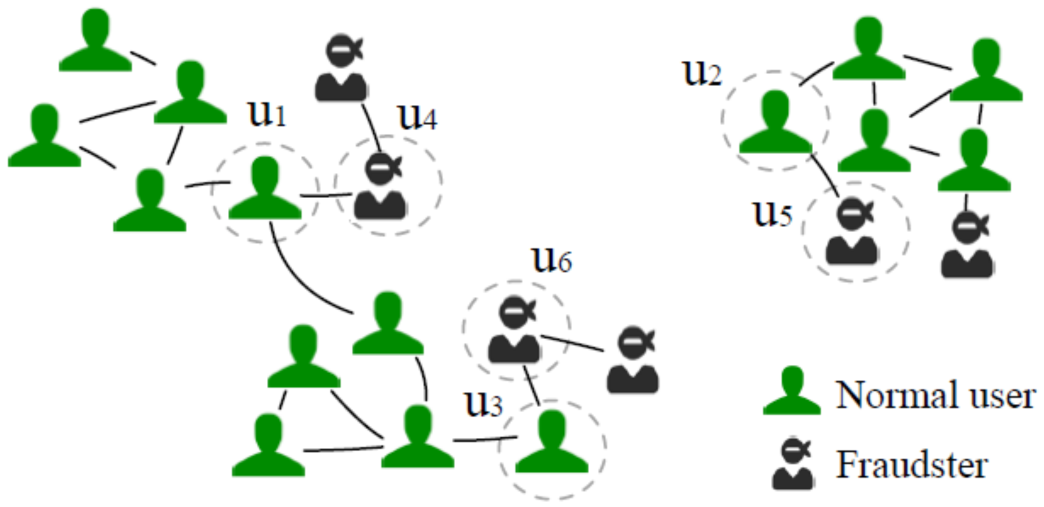
(2) Tuning Stage
Learn a task-specific classifier.

Pre-training by Contrastive Learning

- Basic Idea of Contrastive Learning
  - Contrastive learning (CL) aims to learn such an embedding space in which samples of the same class stay closer to each other while samples of different classes are far apart.
  - CL can be applied to both supervised and unsupervised settings.
- Challenges
  - Data inconsistence may impact the performance of contrastive learning.
  - Two kinds of data inconsistence
    - Intra-class variance: samples of the same class do not always share similar patterns
    - Inter-class similarity: samples of different classes may share similar patterns

Example: Anomaly Detection

- Fraudsters impersonate normal users to disguise themselves.
- Fraudsters have diverse fraud strategies.
- Normal users have diverse interests and behavior patterns.



Pre-training by Contrastive Learning

- Our Solution

Make the cluster information of the data and perform cluster-aware contrastive learning.

The clustering information can reduce the interference of the data inconsistence.

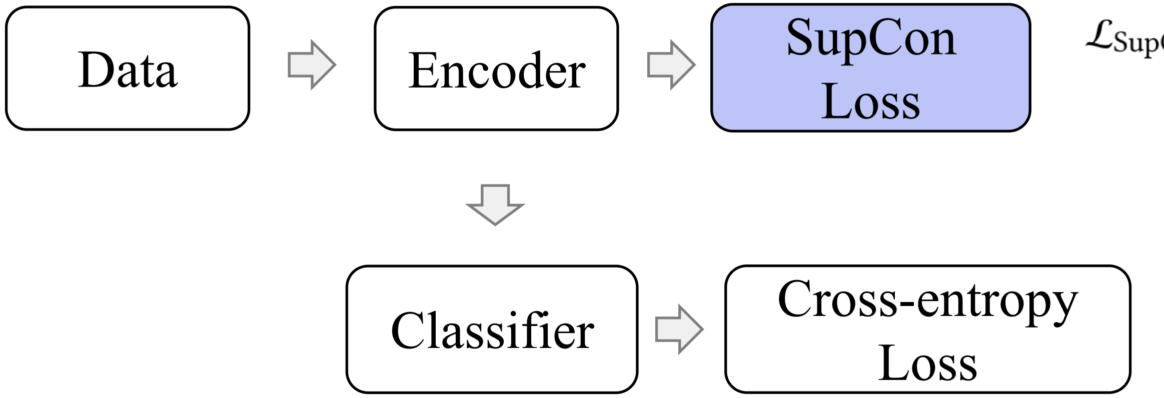
ClusterSCL: Cluster-Aware Supervised Contrastive Learning on Graphs

Yanling Wang[1, 2], Jing Zhang[1, 2], Haoyang Li[1, 2], Yuxiao Dong[3], Hongzhi Yin[4], Cuiping Li[1, 2], and Hong Chen[1, 2]

[1] School of Information, Renmin University of China
[2] Key Laboratory of Data Engineering and Knowledge Engineering of Ministry of Education, RUC
[3] Department of Computer Science and Technology, Tsinghua University
[4] School of Information Technology and Electrical Engineering, The University of Queensland

SupCon: Supervised Contrastive Learning

- SupCon pulls representations of the same class closer than those of different classes.
- SupCon shows advantages over cross-entropy on the ImageNet classification tasks.



$$\mathcal{L}_{\text{SupCon}} = -\sum_{v_i \in B} \frac{1}{|S_i|} \sum_{s_i \in S_i} \log \frac{\exp\left(\mathbf{h}_i^\top \mathbf{h}_{s_i}/\tau\right)}{\sum_{v_j \in B \setminus \{v_i\}} \exp\left(\mathbf{h}_i^\top \mathbf{h}_j/\tau\right)}$$

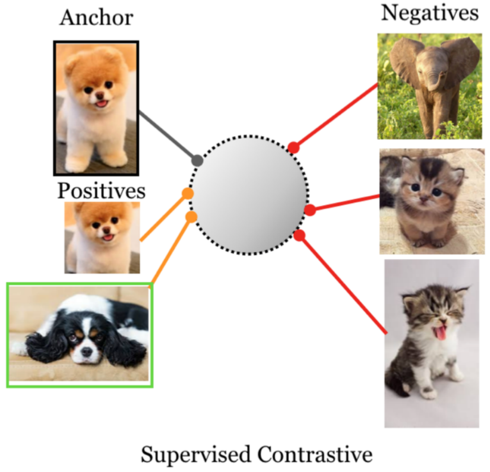
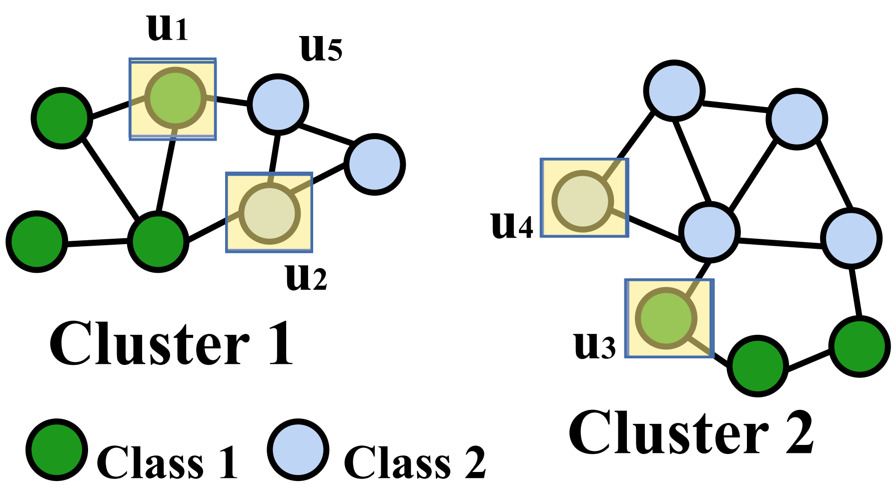


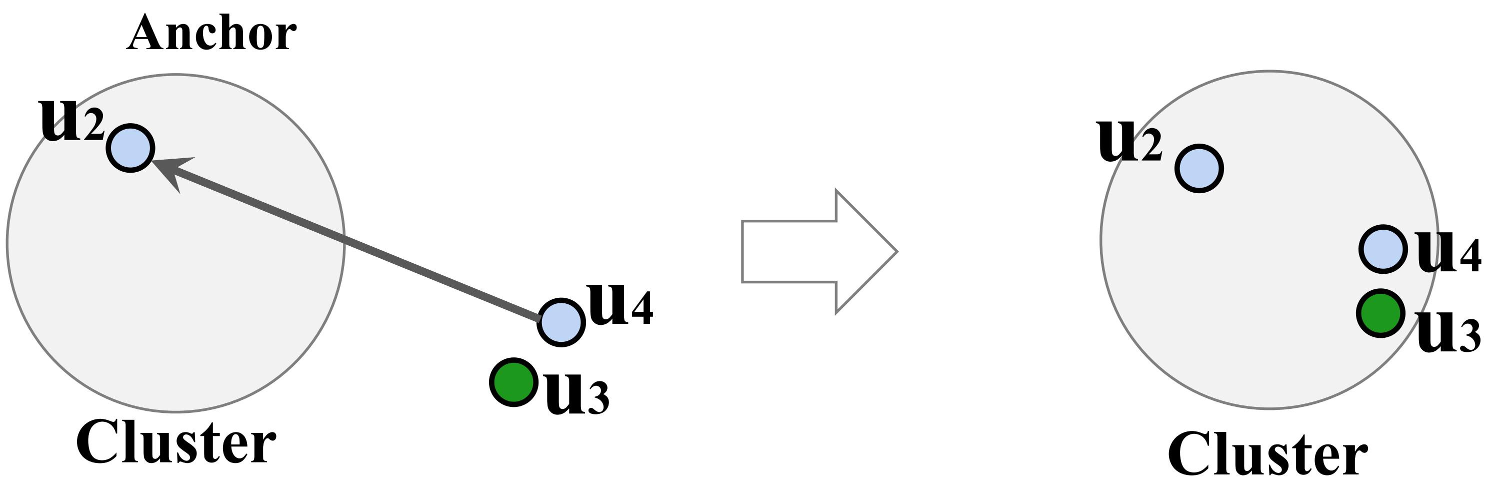
Image from Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. In NeurIPS. 18661–18673.

Limitation of SupCon Loss



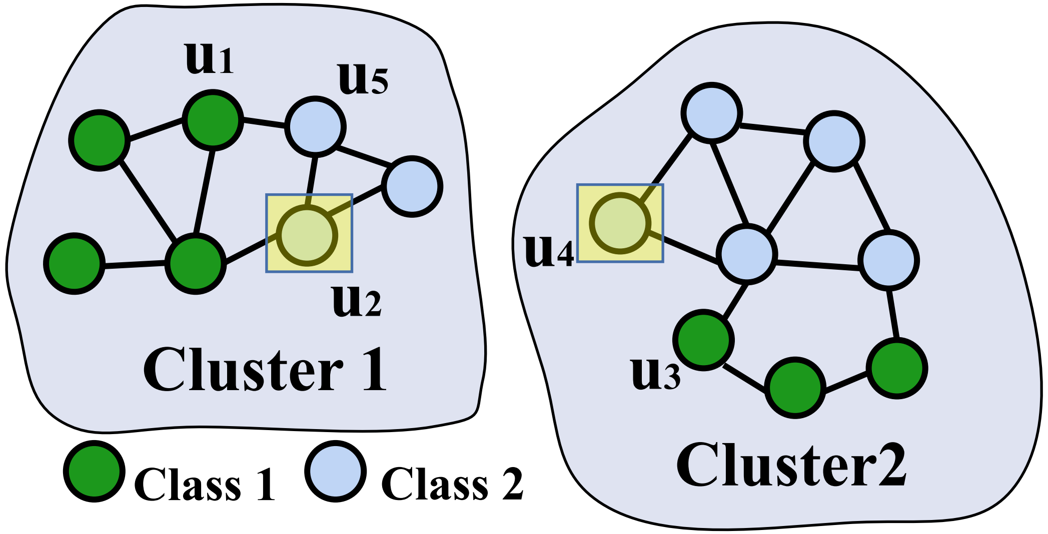
- Intra-class variances and inter-class similarities.
- Misinterpret the intrinsic data property.

The difficulty of learning the classifier is increased.



GS-SupCon: A Straightforward Solution

- Basic Idea: Express the intrinsic data property by the nodes' cluster distributions, and retain the distributions during supervised contrastive learinng.
- Solution: GS-SupCon conducts supervised contrastive learning within each cluster.

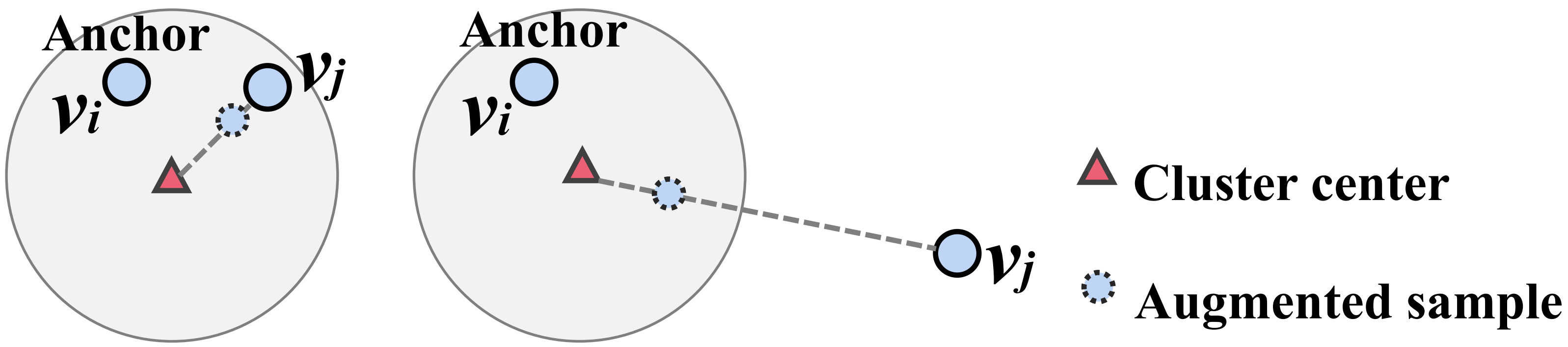


Overlook some potentially useful positve sample pairs.

CDA: Cluster-Aware Data Augmentation

- Basic Idea: Softly narrow the embedding space for supervised contrastive learning.

- Solution: CDA performs interpolation between a positive/negative sample and the anchor's cluster prototype in embedding space. The interpolation weight is adjusted.

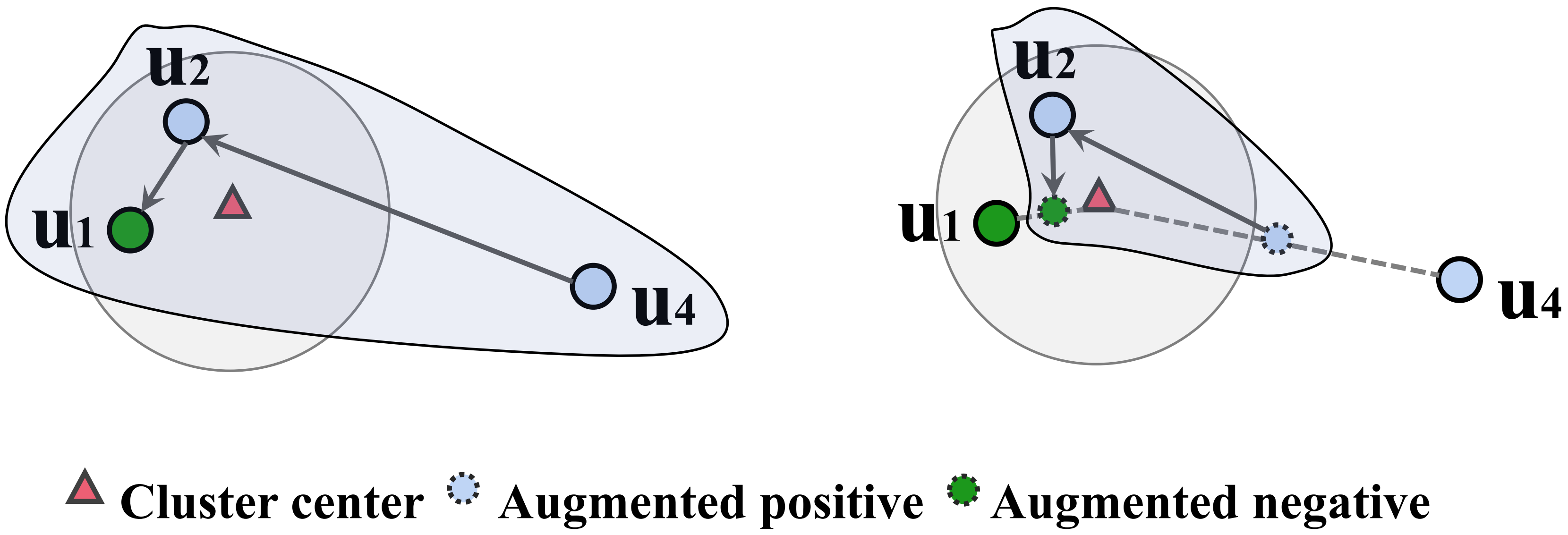


$$\tilde{\mathbf{h}}_j = \alpha \mathbf{h}_j + (1-\alpha)\, \mathbf{w}_{c_i}$$

$$\alpha = \frac{\exp\left(\mathbf{h}_i^\top \mathbf{h}_j\right)}{\exp\left(\mathbf{h}_i^\top \mathbf{h}_j\right) + \exp\left(\mathbf{h}_i^\top \mathbf{w}_{c_i}\right)}$$

Why can CDA Work?

The augmentations softly narrow the embedding space for supervised contrastive learning, so that the pulling strength and pushing strength between original sample pairs can be indirectly weakened to help retain the nodes' cluster distributions.



ClusterSCL

$$p(s_i|v_i) = \int p(c_i|v_i)\, p(s_i|v_i, c_i)\, dc_i$$

Soft Clustering

Cluster-Aware Discriminator

- Cluster-aware discriminator predicts the CDA-based positive sample for an anchor.

$$p(s_i|v_i,c_i) = \frac{\exp\left(\mathbf{h}_i^\top \tilde{\mathbf{h}}_{s_i}/\tau\right)}{\sum_{v_j \in V\setminus\{v_i\}} \exp\left(\mathbf{h}_i^\top \tilde{\mathbf{h}}_j/\tau\right)} = \frac{\exp\left(\mathbf{h}_i^\top \left(\alpha \mathbf{h}_{s_i} + (1-\alpha)\,\mathbf{w}_{c_i}\right)/\tau\right)}{\sum_{v_j \in V\setminus\{v_i\}} \exp\left(\mathbf{h}_i^\top \left(\alpha \mathbf{h}_j + (1-\alpha)\,\mathbf{w}_{c_i}\right)/\tau\right)}$$

- Soft clustering module calculates the cluster distribution for each anchor node.

$$p(c_i|v_i) = \frac{\exp\left(\mathbf{h}_i^\top \mathbf{w}_{c_i}/\kappa\right)}{\sum_{m=1}^{M} \exp\left(\mathbf{h}_i^\top \mathbf{w}_m/\kappa\right)}$$

- Variational EM algorithm for inference and learning.

$$\begin{aligned}\log p(s_i|v_i) &\geq \mathcal{L}_{\mathrm{ELBO}}(\boldsymbol{\theta}, \mathbf{w}; v_i, s_i) \\ &:= \mathbb{E}_{q(c_i|v_i,s_i)}\left[\log p(s_i|v_i,c_i)\right] \\ &\quad - \mathrm{KL}\left(q(c_i|v_i,s_i)\,||\,p(c_i|v_i)\right)\end{aligned}$$

Experiments

Table 3: Comparison with the Group Sensitive SupCon (GS-SupCon).

	GCN-encoder			GAT-encoder		
	SupCon	GS-SupCon	ClusterSCL	SupCon	GS-SupCon	ClusterSCL
Cora	0.793	0.788	**0.818**	0.816	0.822	**0.826**
Pubmed	0.788	0.797	**0.805**	0.797	0.801	**0.811**
Citeseer	0.687	0.684	**0.692**	0.693	0.695	**0.706**
LastFM Asia	0.756	**0.769**	0.752	0.776	0.775	**0.779**
Amazon Computers	0.831	0.833	**0.834**	0.842	0.848	**0.849**

- GS-SupCon derives comparable or better performance compared with SupCon.
- ClusterSCL outperforms GS-SupCon on most of the datasets.

Decoupling Representation Learning and Classification for GNN-based Anomaly Detection

Yanling Wang[1, 2], Jing Zhang[1, 2], Shasha Guo[1, 2], Hongzhi Yin[3],

Cuping Li[1, 2], and Hong Chen[1, 2]

[1] School of Information, Renmin University of China
[2] Key Laboratory of Data Engineering and Knowledge Engineering of Ministry of Education, RUC
[3] School of Information Technology and Electrical Engineering, The University of Queensland

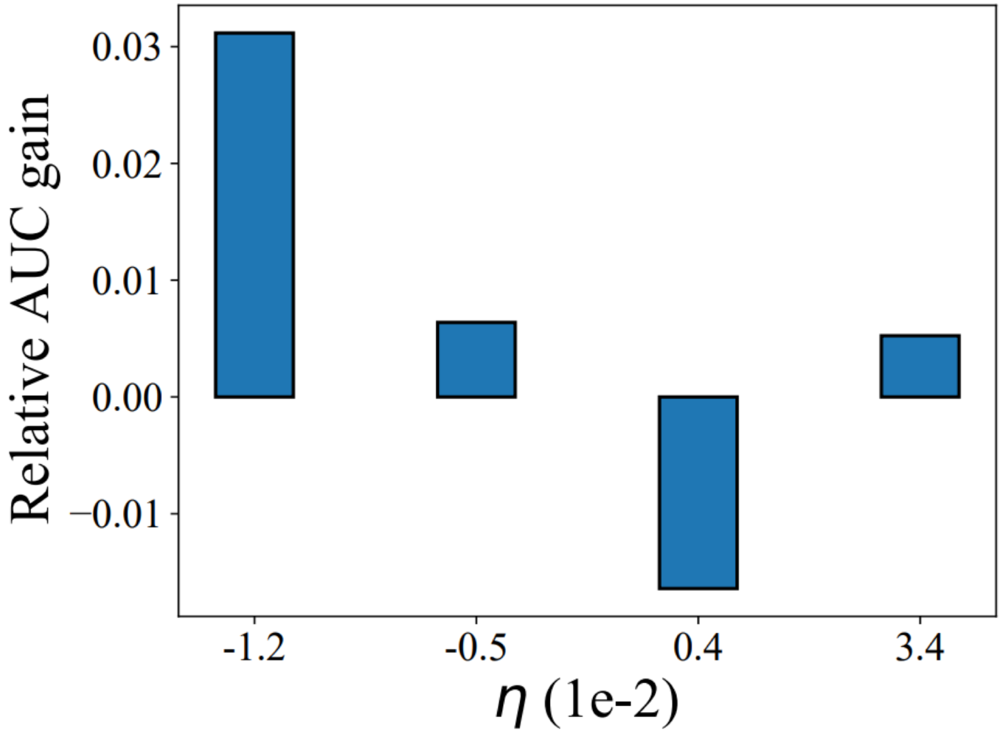
Deep Graph InfoMax -- DGI

- DGI encodes the global information into each node representation via a contrastive loss.

$$\mathcal{L}_{DGI} = -\frac{1}{2n} \sum_{i=1}^{n} \left(\mathbb{E}_{G} \log \mathcal{D}(\boldsymbol{h}_i^{(L)}, \boldsymbol{s}) + \mathbb{E}_{\tilde{G}} \log(1 - \mathcal{D}(\tilde{\boldsymbol{h}}_i^{(L)}, \boldsymbol{s})) \right)$$

Petar Velickovic, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R. Devon Hjelm. 2019. Deep Graph Infomax. In ICLR.

Does Decoupled Training with DGI Stably Perform Well?



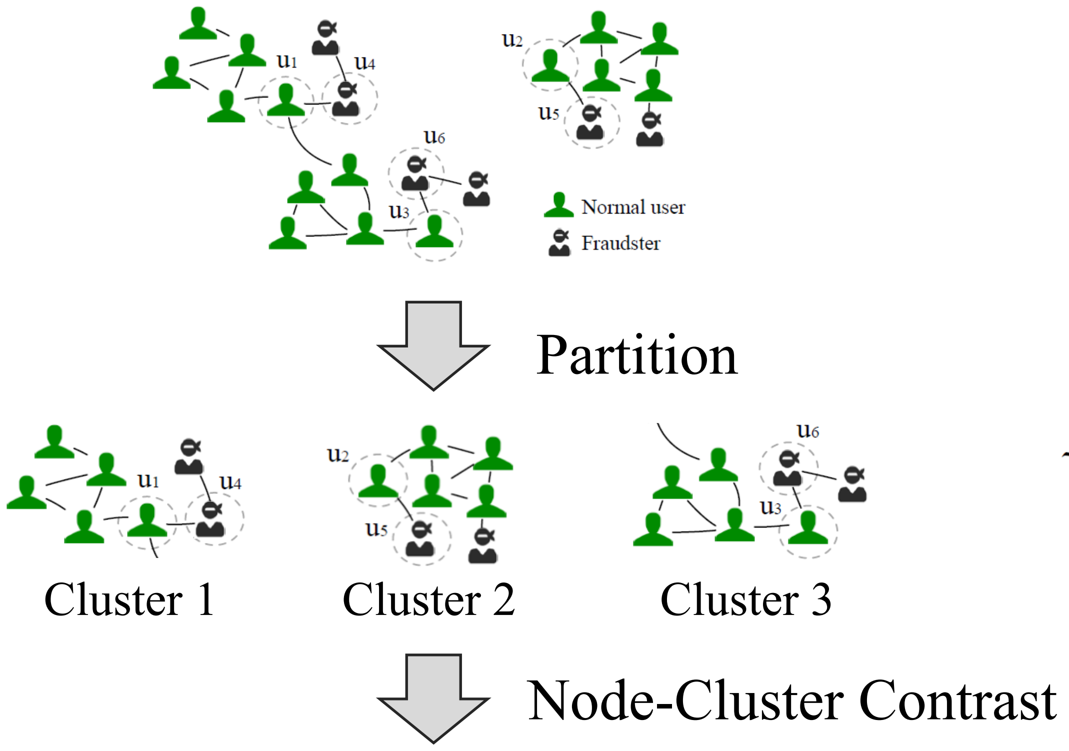
Settings:

- Inconsistency is the core factor impacting graph anomaly detection.
- Additive inverse of silhouette coefficient is used to quantify the inconsistency η.

Observation:

- Decoupled training may not always improve, and even brings negative influence when the data gets highly inconsistent.

Deep Cluster InfoMax -- DCI



DCI loss encodes the semi-global context into the node representations.

Cluster representation:

$$s_k = \sigma\left(\frac{1}{n_k}\sum_{v_i \in V_k} \boldsymbol{h}_i\right)$$

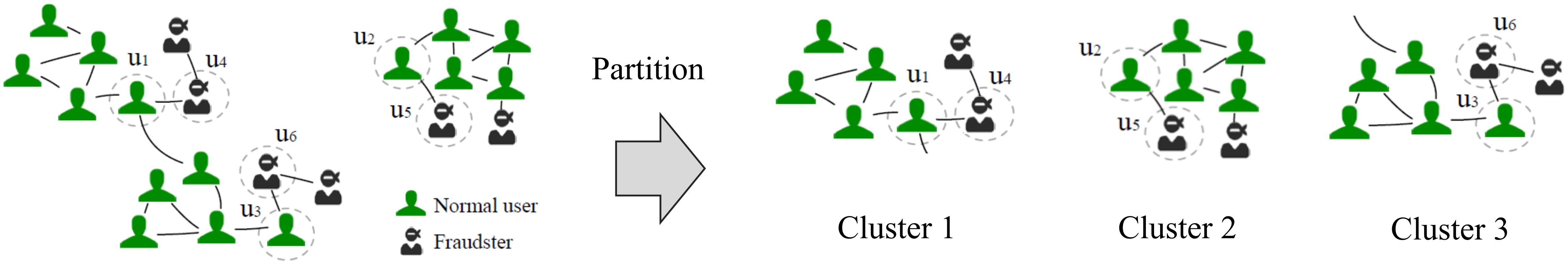
Maximize the local-semi-global affinity scores:

$$\mathcal{L}_{DCI}^{k} = -\frac{1}{2n_k}\sum_{v_i \in V_k}\left(\mathbb{E}_{C_k}\log \mathcal{D}(\boldsymbol{h}_i, \boldsymbol{s}_k) + \mathbb{E}_{\tilde{C}_k}\log(1-\mathcal{D}(\tilde{\boldsymbol{h}}_i, \boldsymbol{s}_k))\right)$$

$$\mathcal{L}_{DCI} = \frac{1}{K}\sum_{k=1}^{K}\mathcal{L}_{DCI}^{k}$$

In practice, we re-cluster nodes based on the node representations after every certain number of training epochs.

Why Can DCI Work?



Behaviors within the same cluster are often more concentrated than those in the whole graph.

The differences between normal users and fraudsters can be amplified in a concentrated space.

Experiments

Datasets:

Table 1: Statistics of the datasets.

Graph	#Users(% normal, abnormal)	#Objects	#Edges
Reddit	10,000 (96.34%, 3.66%)	984	78,516
Wiki	8,227 (97.36%, 2.64%)	1,000	18,257
Alpha	3,286 (61.21%, 38.79%)	3,754	24,186
Amazon	27,197 (91.73%, 8.27%)	5,830	52,156

Baselines:

- Joint learning algorithms:
CARE-GNN, GAT, GeniePath, and GIN

- SSL losses for decoupled training:
GAE, RW, GCC, and DGI.

Overall Evaluation

Table 2: Overall evaluation on four real-world datasets.

		Reddit	Wiki	Alpha	Amazon
Joint	CARE-GNN	0.700	0.702	0.802	0.729
	GAT	0.738	0.681	0.848	0.696
	GeniePath	0.720	0.689	0.849	0.738
	GIN	0.720	0.727	0.884	0.761
Decoupled	GAE	0.730	0.714	0.884	0.806
	RW	0.728	0.740	**0.908**	0.782
	GCC	0.669	0.695	0.865	0.733
	DGI	0.743	0.737	0.884	0.771
	DCI (ours)	**0.746**	**0.762**	0.907	**0.810**
Inconsistency η (1e-2)		-0.676	0.841	-	-

Note: All the decoupled models use GIN's encoder as the backbone.

Table 4: Evaluation of the multi-task learning.

		Reddit	Wiki	Alpha	Amazon
Joint	GIN	0.720	0.727	0.884	0.761
Multi-task	GAE	0.726	0.705	0.904	0.766
	DGI	0.647	0.664	0.891	0.806
	DCI	0.675	0.670	0.893	0.803

Note: All the multi-task models use GIN's encoder as the backbone.

- Decoupled training contributes to the anomaly detection.
- DCI is an effective self-supervised loss for decoupled training.
- Multi-task learning can outperform the joint training, but not always outperforms.
- Decoupled training shows advantages over multi-task learning.

Comparison between ClusterSCL and DCI

Table 2: Overall evaluation.The bold numbers are the best performance among all two-stage models.

		Cora	Pubmed	Citeseer	LastFM	Amazon
GCN-encoder	CE (E2E)	0.804	0.789	0.696	0.731	0.831
	DGI (Two-stage, Unsup)	0.801	0.796	**0.695**	0.749	0.838
	DCI (Two-stage, Unsup)	0.811	0.793	0.694	**0.757**	**0.844**
	SupCon (Two-stage, Sup)	0.793	0.788	0.687	0.756	0.831
	ClusterSCL (Two-stage, Sup)	**0.818**	**0.805**	0.692	0.752	0.834
GAT-encoder	CE (E2E)	0.799	0.786	0.691	0.772	0.828
	DGI (Two-stage, Unsup)	0.808	0.794	0.684	0.781	0.836
	DCI (Two-stage, Unsup)	0.821	0.790	0.695	**0.784**	0.836
	SupCon (Two-stage, Sup)	0.816	0.797	0.693	0.776	0.842
	ClusterSCL (Two-stage, Sup)	**0.826**	**0.811**	**0.706**	0.779	**0.849**

- The two-stage training generally performs better than the end-to-end training.
- Unsupervised DGI and DCI also obtain good performance.

Conclusions

- Contributions:
  - We emphasize the effectiveness of two-stage training for supervised graph learning tasks.
  - We study contrastive learning for the two-stage training.
  - We incorporate clustering techniques to reduce the influence of data inconsistence on contrastive learning.

- To discuss:
  - Is it possible to use the soft clustering technique under the unsupervised setting?
  - The inconsistency is difficult to be measured and controlled in real data.
  - Can we handle tough non-homophily graphs using the idea of cluster-enhanced contrastive learning?

中国人民大學 RENMIN UNIVERSITY OF CHINA
清華大学 Tsinghua University
THE UNIVERSITY OF QUEENSLAND AUSTRALIA

Thank you!

Q&A