How powerful are graph neural networks?

Authors: Keyulu Xu*, Weihua Hu*, Jure Leskovec, Stefanie Jegelka

Speaker: Ziyuan Ye (叶梓元)

Monday, May 23, 2022

Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2018, September). How Powerful are Graph Neural Networks?.
In *International Conference on Learning Representations (ICLR)*.

About the authors

Name	Organization	Research Interests	Other representative publications
Keyulu Xu	MIT	Graph Neural Networks, Deep Learning	1. Representation learning on graphs with jumping knowledge networks 2. What Can Neural Networks Reason About?
Weihua Hu	Standford	Machine Learning, Deep Learning	1. Open Graph Benchmark: Datasets for Machine Learning on Graphs
Jure Leskovec	Standford	Data mining, Machine Learning, Graph Neural Networks	1. node2vec: Scalable feature learning for networks 2. Inductive representation learning on large graphs 3. SNAP Datasets: Stanford large network dataset collection
Stefanie Jegelka	MIT	Machine Learning, Optimization, Submodularity	1. Max-value entropy search for efficient Bayesian optimization 2. Deep metric learning via lifted structured feature embedding

Content

1. Take-home message
2. Background
3. Research content
4. Experimental results
5. Future work

Content

1. Take-home message
2. Background
3. Research content
4. Experimental results
5. Future work

Take-home Message

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

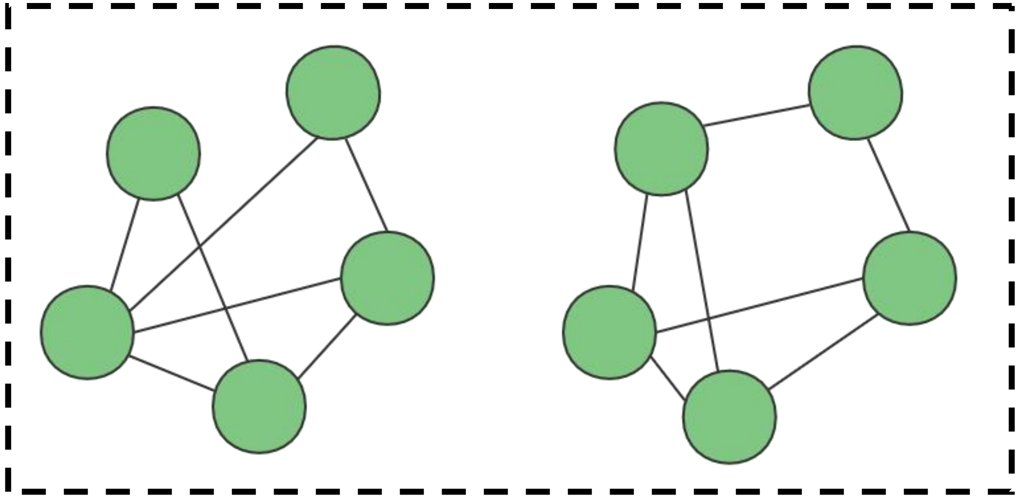
- **Motivation:**
  - Despite GNNs revolutionizing graph representation learning, there is limited understanding of their representational properties and limitations.
  - Can GNNs have as large discriminative power as the Weisfeiler-Lehman (WL) test if the GNN's aggregation scheme is highly expressive and can model injective functions?
- **Main contributions:**
  - They show that GNNs are at most as powerful as the WL test in distinguishing graph structures.
  - They develop Graph Isomorphism Network (GIN), and show that its discriminative power is equal to the power of the WL test.
- **Future work:**
  - Go beyond neighborhood aggregation (message passing) to pursue more powerful message passing ways.

GNN: Graph neural network

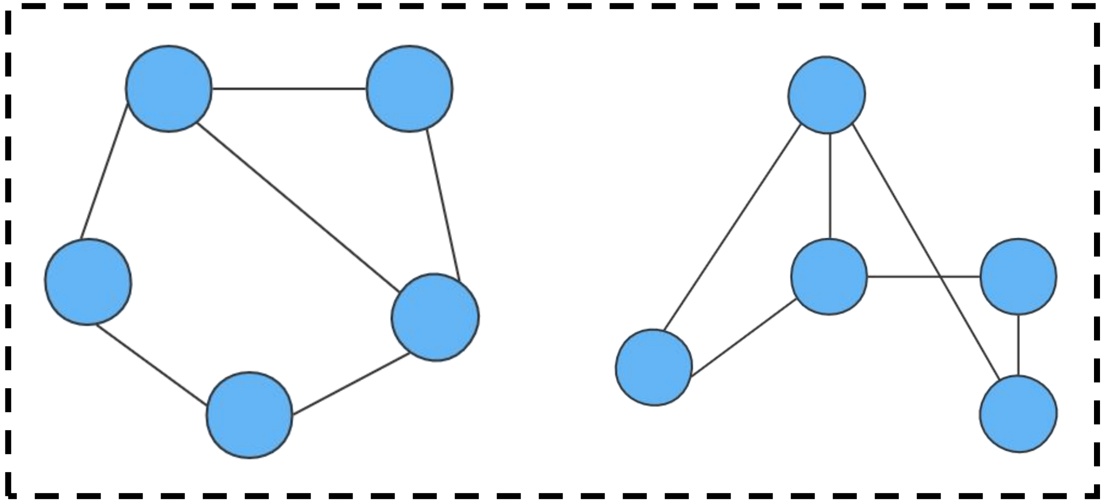
Content

1. Take-home message
2. Background
3. Research content
4. Experimental results
5. Future work

How to define a powerful GNN?



Non-isomorphic graphs



Isomorphic graphs

A powerful graph neural network model

Different representations

Same representation

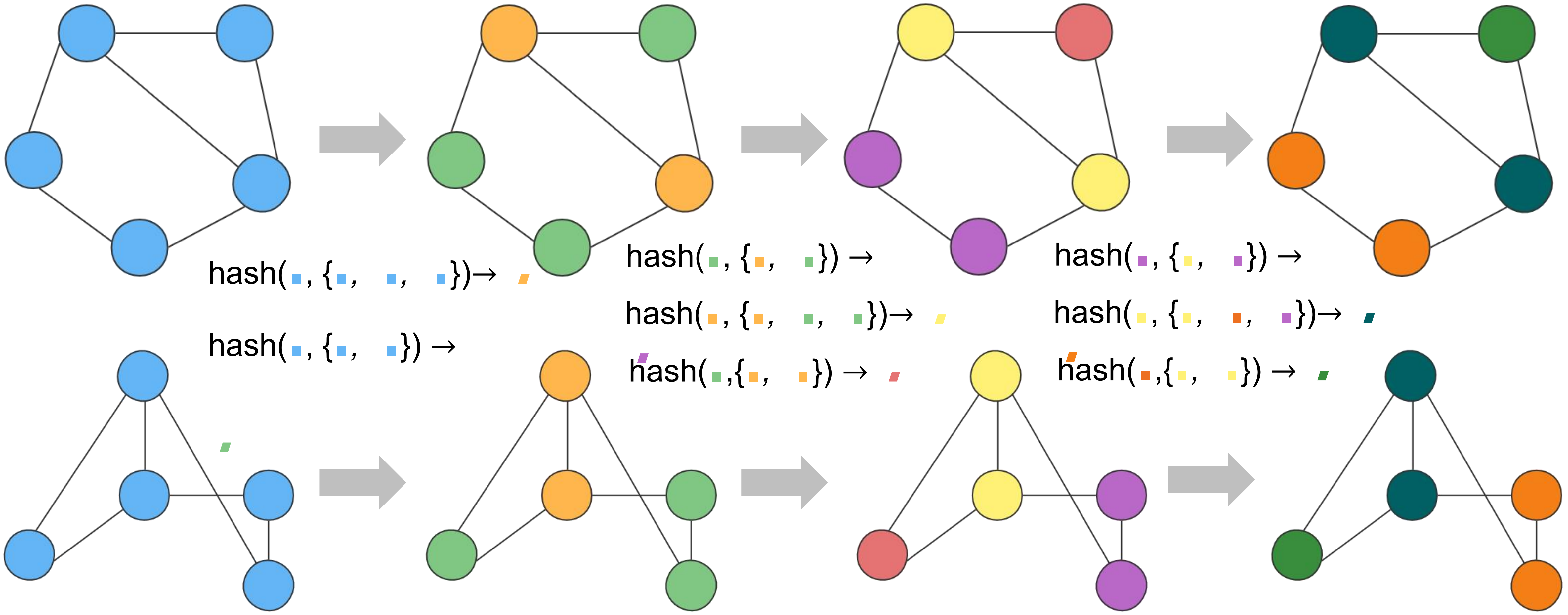
Traditional method to distinguish non-isomorphic graphs:

Weisfeiler-Lehman (WL) test

What is Weisfeiler-Lehman (WL) test?

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

The algorithm stops upon reaching a stable coloring

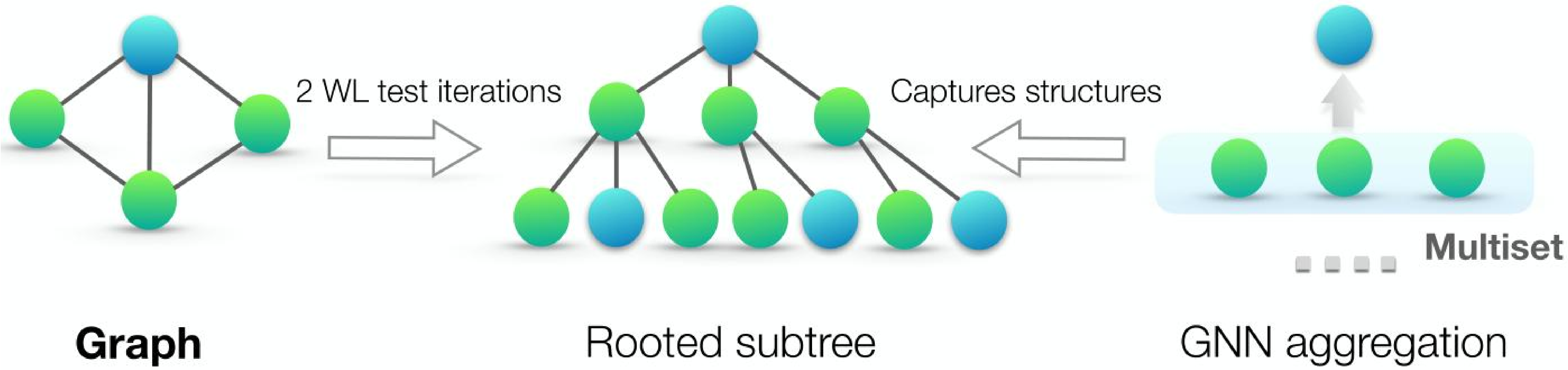


Adapted from *Michael Bronstein* blog: https://resources.experfy.com/ai-ml/expressive-power-graph-neural-networks-weisfeiler-lehman/

Content

1. Take-home message
2. Background
3. Research content
4. Experimental results
5. Future work

An overview of the framework



If GNN aggregation can capture the *full multiset* of node neighbors, whether there exist GNNs that are as powerful as the WL test?

If the neighbor aggregation and graph-level readout functions are injective, then the resulting GNN is as powerful as the WL test.

Building powerful GNN

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

A powerful GNN should hold the following two condition:

(a)

$$h_v^{(k)} = \phi\left(h_v^{(k-1)}, f\left(\left\{h_u^{(k-1)} : u \in \mathcal{N}(v)\right\}\right)\right)$$

$f(\cdot)$ which operates on multiset, and ϕ are injective.

(b)

GNN's graph-level readout, which operates on the multiset of node features $\{h_v^{(k)}\}$, is injective.

An important corollary:

Assume $\mathcal{X}$ is countable. There exists a function $f: \mathcal{X} \rightarrow \mathbb{R}^n$ so that for infinitely many choices of ϵ, including all irrational numbers, $h(c, X) = (1+\epsilon) \cdot f(c) + \sum_{x \in X} f(x)$ is unique for each pair (c, X), where $c \in \mathcal{X}$ and $X \subset \mathcal{X}$ is a multiset of bounded size. Any multiset function g can be decomposed as $g(c, X) = \phi((1+\epsilon) \cdot f(c) + \sum_{x \in X} f(x))$ for some function ϕ.

Building powerful GNN

An important corollary:

Assume $\mathcal{X}$ is countable. There exists a function $f: \mathcal{X} \rightarrow \mathbb{R}^n$ so that for infinitely many choices of ϵ, including all irrational numbers, $h(c, X) = (1+\epsilon) \cdot f(c) + \sum_{x \in X} f(x)$ is unique for each pair (c, X), where $c \in \mathcal{X}$ and $X \subset \mathcal{X}$ is a multiset of bounded size. Any multiset function g can be decomposed as $g(c, X) = \phi((1+\epsilon) \cdot f(c) + \sum_{x \in X} f(x))$ for some function ϕ.

+

Universal approximation theorem

Universal approximation theorem imply that neural networks (e.g. multi-layer perceptron, MLP) can *represent* a wide variety of interesting functions when given appropriate weights.

=

Graph Isomorphic Network (GIN):

Sum aggregators + MLP to model $f^{(k+1)} \circ \phi^{(k)}$

Comparison of different models

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Model	Aggregate functions	Update functions
GCN	$h_v^{(k)} = \text{ReLU}\left(W \cdot \text{MEAN}\left\{h_u^{(k-1)}, \forall u \in \mathcal{N}(v) \cup \{v\}\right\}\right)$	
GraphSAGE	$a_v^{(k)} = \text{MAX}\left(\left\{\text{ReLU}\left(W \cdot h_u^{(k-1)}\right), \forall u \in \mathcal{N}(v)\right\}\right)$	$h_v^{(k)} = \text{COMBINE}^{(k)}\left(h_v^{(k-1)}, a_v^{(k)}\right)$
GIN	$h_v^{(k)} = \text{MLP}^{(k)}\left(\left(1+\epsilon^{(k)}\right) \cdot h_v^{(k-1)} + \sum_{u \in \mathcal{N}(v)} h_u^{(k-1)}\right)$	

		Traditional representations	Representations by GIN
Node Classification	Node representation	$h_v^{(K)}$	$h_v^{(K)}$
Graph Classification	Graph representation	$h_G = \text{READOUT}\left(\left\{h_v^{(K)} \mid v \in G\right\}\right)$	$h_G = \text{CONCAT}\left(\text{READOUT}\left(\left\{h_v^{(k)} \middle| v \in G\right\}\right) \middle| k = 0,1,\dots,K\right)$

For GraphSAGE, these slides only provide *pooling* aggregator, the *mean* and *LSTM* aggregators are ignored for simplicity.

Content

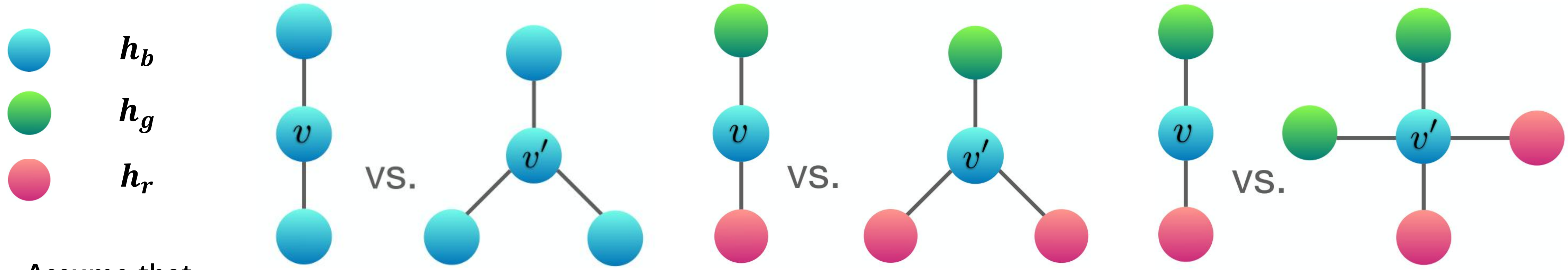
南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

1. Take-home message
2. Background
3. Research content
4. Experimental results
5. Future work

Aggregation: Mean or Max or Sum?

SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Does these aggregation functions work for distinguishing the non-isomorphic graph in the following examples?



Assume that $h_r > h_g > h_b$

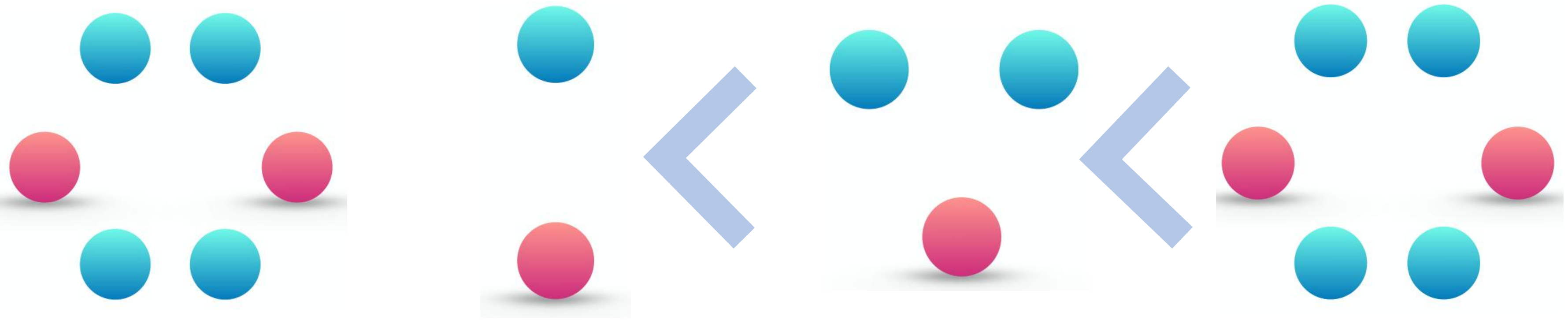
Example 1 **Example 2** **Example 3**

Aggregation	Example 1 Results	Example 1 Work?	Example 2 Results	Example 2 Work?	Example 3 Results	Example 3 Work?
Mean	h_b	*No*	$\frac{(h_r+h_g)}{2} / \frac{(2h_r+h_g)}{3}$	*Yes*	$\frac{(h_r+h_g)}{2}$	*No*
Max	h_b	*No*	h_r	*No*	h_r	*No*
Sum	$2h_b/3h_b$	*Yes*	$(h_r+h_g)/(2h_r+h_g)$	*Yes*	$(h_r+h_g)/(2h_r+2h_g)$	*Yes*

Aggregation: Mean or Max or Sum?

Aggregation	Max	Mean	Sum

Input



Information	Set	Distribution	Multiset
Perform well in which situations?	When representative elements or the "skeleton" are important	When statistical and distributional information are important	Suitable for all common situations

1-layer perceptron is sufficient?

SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

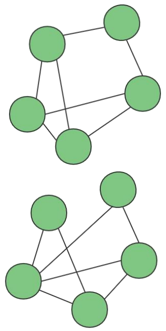
An important Lemma:

There exist finite multisets $X_1 \neq X_2$ so that for any linear mapping W, $\sum_{x \in X_1} ReLU(Wx) = \sum_{x \in X_2} ReLU(WX)$

Unlike models using MLPs, the 1-layer perceptron (even with the bias term) is *not a universal approximator* of multiset functions.

Answer to the question:

Not sufficient enough. Even if GNNs with 1-layer perceptron can embed different graphs to different locations to some degree, such embeddings may not adequately capture structural similarity, and can be difficult for simple classifiers, e.g., linear classifiers, to fit.



GIN-1-layer — Relatively large

GIN-multi-layer — Relatively small

Non-isomorphic graphs **Models** **Embeddings' Distribution Overlap**

Benefit of GIN beyond WL-test

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- **Capturing similarity of graph structures.**

Importance of structural similarity

Helpful for generalization of GNNs, especially the co-occurrence of subtrees is sparse / there are noisy edges and node features.

Limitation of WL-test

Node feature vectors in the WL test are essentially one-hot encodings and thus cannot capture the similarity between subtrees.

→

Solution by GIN

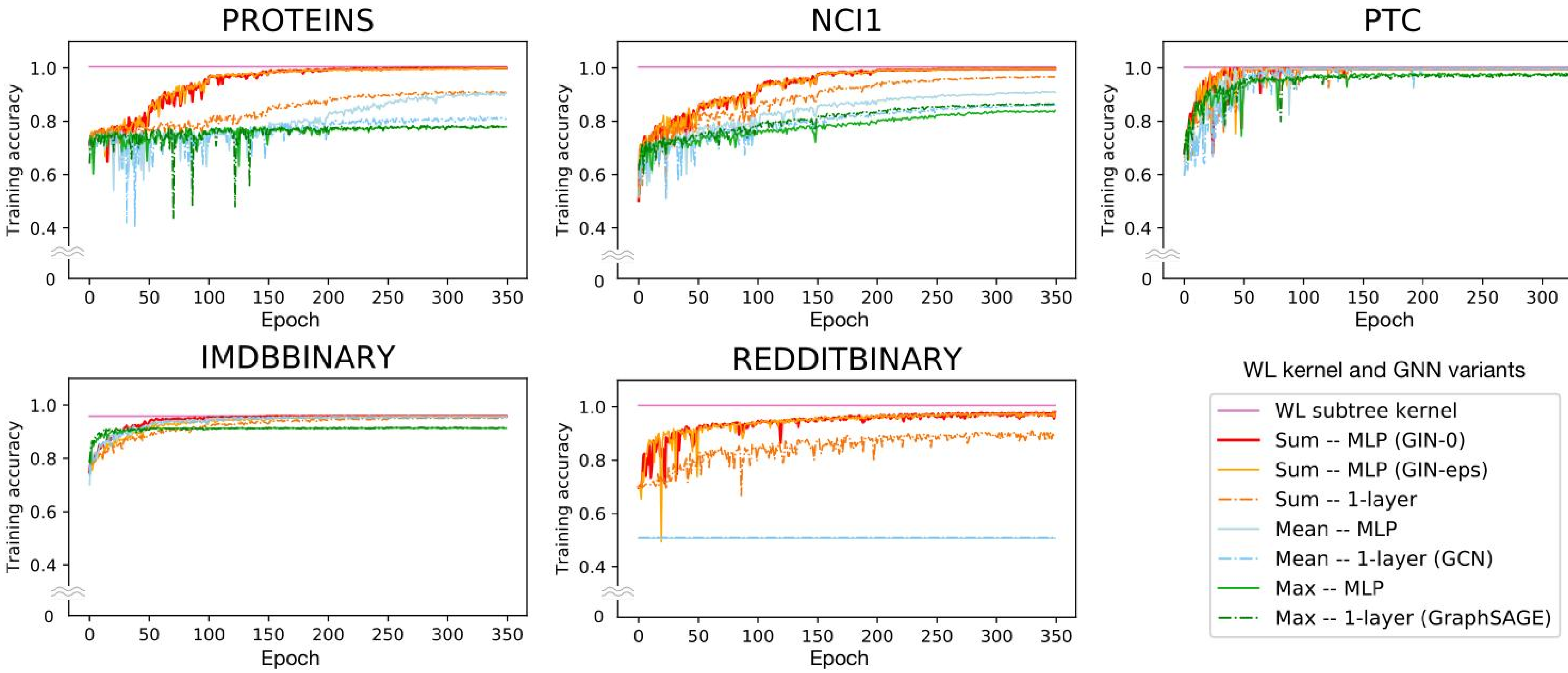
GIN satisfying the above criteria generalizes the WL test by learning to embed the subtrees to low-dimensional space.

Test set classification accuracies

	Datasets	IMDB-B	IMDB-M	RDT-B	RDT-M5K	COLLAB	MUTAG	PROTEINS	PTC	NCI1
Datasets	# graphs	1000	1500	2000	5000	5000	188	1113	344	4110
	# classes	2	3	2	5	3	2	2	2	2
	Avg # nodes	19.8	13.0	429.6	508.5	74.5	17.9	39.1	25.5	29.8
Baselines	WL subtree	73.8 ± 3.9	50.9 ± 3.8	81.0 ± 3.1	52.5 ± 2.1	78.9 ± 1.9	90.4 ± 5.7	75.0 ± 3.1	59.9 ± 4.3	**86.0 ± 1.8** *
	DCNN	49.1	33.5	–	–	52.1	67.0	61.3	56.6	62.6
	PATCHYSAN	71.0 ± 2.2	45.2 ± 2.8	86.3 ± 1.6	49.1 ± 0.7	72.6 ± 2.2	**92.6 ± 4.2** *	75.9 ± 2.8	60.0 ± 4.8	78.6 ± 1.9
	DGCNN	70.0	47.8	–	–	73.7	85.8	75.5	58.6	74.4
	AWL	74.5 ± 5.9	51.5 ± 3.6	87.9 ± 2.5	54.7 ± 2.9	73.9 ± 1.9	87.9 ± 9.8	–	–	–
GNN variants	SUM–MLP (**GIN-0**)	**75.1 ± 5.1**	**52.3 ± 2.8**	**92.4 ± 2.5**	**57.5 ± 1.5**	**80.2 ± 1.9**	**89.4 ± 5.6**	**76.2 ± 2.8**	**64.6 ± 7.0**	**82.7 ± 1.7**
	SUM–MLP (**GIN-ϵ**)	**74.3 ± 5.1**	**52.1 ± 3.6**	**92.2 ± 2.3**	**57.0 ± 1.7**	**80.1 ± 1.9**	**89.0 ± 6.0**	**75.9 ± 3.8**	63.7 ± 8.2	**82.7 ± 1.6**
	SUM–1-LAYER	74.1 ± 5.0	**52.2 ± 2.4**	90.0 ± 2.7	55.1 ± 1.6	**80.6 ± 1.9**	**90.0 ± 8.8**	**76.2 ± 2.6**	63.1 ± 5.7	82.0 ± 1.5
	MEAN–MLP	73.7 ± 3.7	**52.3 ± 3.1**	50.0 ± 0.0	20.0 ± 0.0	79.2 ± 2.3	83.5 ± 6.3	75.5 ± 3.4	**66.6 ± 6.9**	80.9 ± 1.8
	MEAN–1-LAYER (GCN)	74.0 ± 3.4	51.9 ± 3.8	50.0 ± 0.0	20.0 ± 0.0	79.0 ± 1.8	85.6 ± 5.8	76.0 ± 3.2	64.2 ± 4.3	80.2 ± 2.0
	MAX–MLP	73.2 ± 5.8	51.1 ± 3.6	–	–	–	84.0 ± 6.1	76.0 ± 3.2	64.6 ± 10.2	77.8 ± 1.3
	MAX–1-LAYER (GraphSAGE)	72.3 ± 5.3	50.9 ± 2.2	–	–	–	85.1 ± 7.6	75.9 ± 3.2	63.9 ± 7.7	77.7 ± 1.5

$$h_v^{(k)} = \mathrm{MLP}^{(k)}\left(\left(1+\epsilon^{(k)}\right)\cdot h_v^{(k-1)} + \sum\nolimits_{u\in\mathcal{N}(v)} h_u^{(k-1)}\right)$$

Training set performance

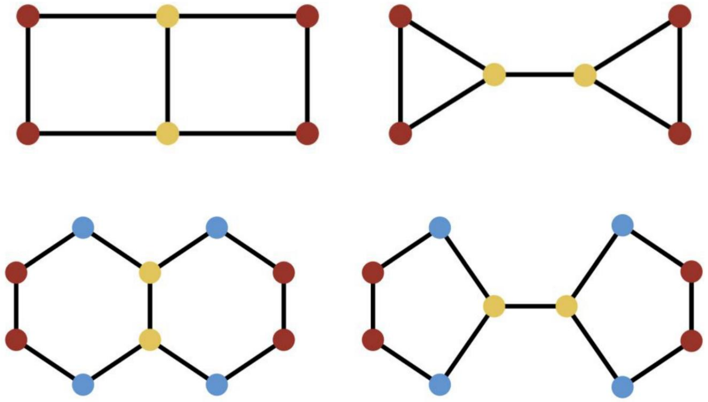


Content

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

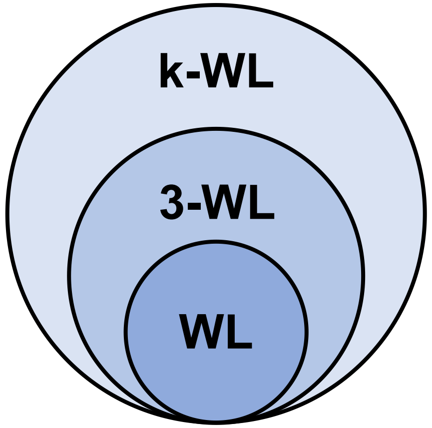
1. Take-home message
2. Background
3. Research content
4. Experimental results
5. Future work

Limitation of GIN



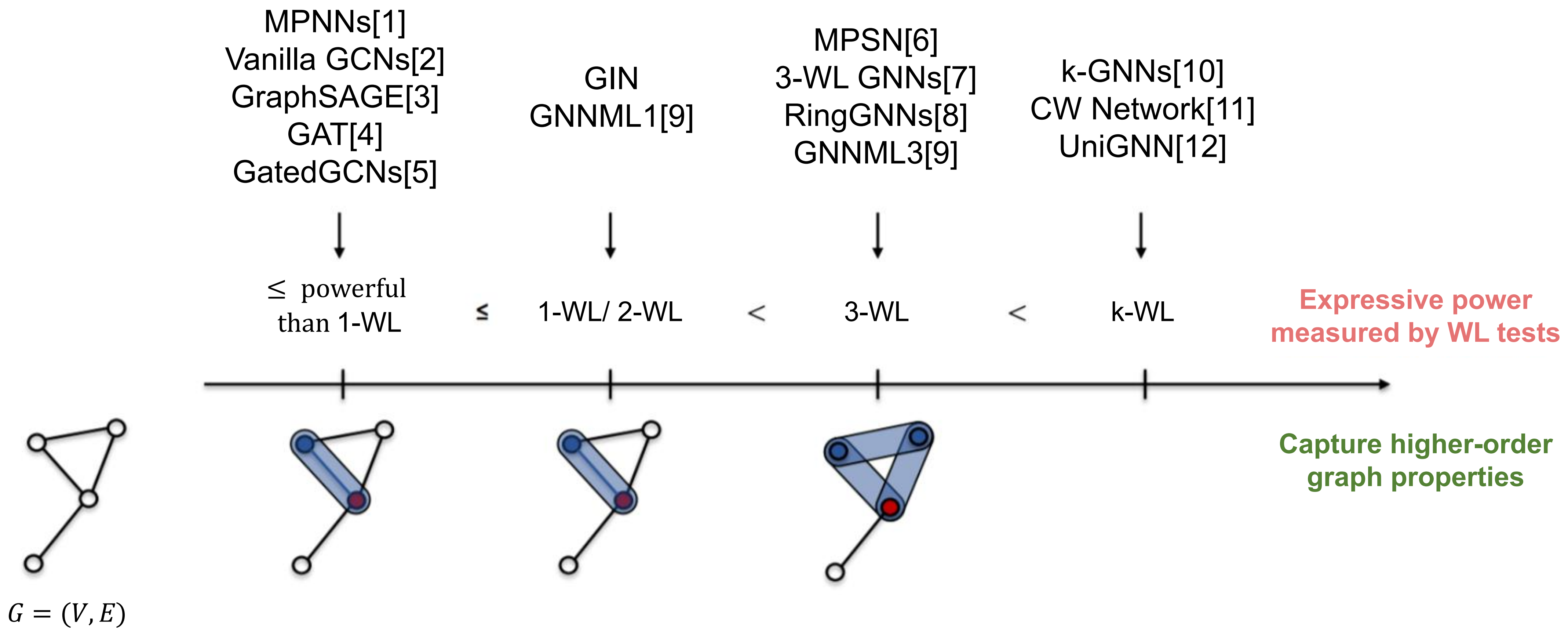
GIN fails to distinguish the higher-order structures

GIN fails to capture long-range interactions



GIN fails to break through WL

Future work



Adapted from *Xavier Bresson’s* slides

Future work

当科学家登上一座高山后，却发现神学家早就坐在那里了。

——爱因斯坦

当计算机科学家登上一座高山后，却发现数学家早就坐在那里了。

Random Walks on Graphs: a Survey

László Lovász

YALEU/DCS/TR-1029
May 1994

The PageRank Citation Ranking: Bringing Order to the Web

January 29, 1998

DeepWalk: Online Learning of Social Representations

Bryan Perozzi
Stony Brook University
Department of Computer Science

Rami Al-Rfou
Stony Brook University
Department of Computer Science

Steven Skiena
Stony Brook University
Department of Computer Science

2014 ACM SIGKDD

“It is a natural and powerful method to study discrete structures by ‘embedding’ them in the continuous world”

——Laszlo Lovasz

Reference

1. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017, July). Neural message passing for quantum chemistry. In International conference on machine learning (pp. 1263-1272). PMLR.
2. Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.
3. Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. Advances in neural information processing systems, 30.
4. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903*.
5. Li, Y., Tarlow, D., Brockschmidt, M., & Zemel, R. (2015). Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*.
6. Bodnar, C., Frasca, F., Wang, Y., Otter, N., Montufar, G. F., Lio, P., & Bronstein, M. (2021, July). Weisfeiler and lehman go topological: Message passing simplicial networks. In *International Conference on Machine Learning* (pp. 1026-1037). PMLR.

Reference

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

7. Maron, H., Ben-Hamu, H., Serviansky, H., & Lipman, Y. (2019). Provably powerful graph networks. *Advances in neural information processing systems*, *32*.
8. Chen, Z., Villar, S., Chen, L., & Bruna, J. (2019). On the equivalence between graph isomorphism testing and function approximation with gnns. *Advances in neural information processing systems*, *32*.
9. Balcilar, M., Héroux, P., Gauzere, B., Vasseur, P., Adam, S., & Honeine, P. (2021, July). Breaking the limits of message passing graph neural networks. In *International Conference on Machine Learning* (pp. 599-608). PMLR.
10. Morris, C., Ritzert, M., Fey, M., Hamilton, W. L., Lenssen, J. E., Rattan, G., & Grohe, M. (2019, July). Weisfeiler and leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 4602-4609).
11. Bodnar, C., Frasca, F., Otter, N., Wang, Y. G., Liò, P., Montufar, G. F., & Bronstein, M. (2021). Weisfeiler and lehman go cellular: Cw networks. *Advances in Neural Information Processing Systems*, *34*.
12. Huang, J., & Yang, J. (2021). UniGNN: a Unified Framework for Graph and Hypergraph Neural Networks. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence

Thanks for your attention!

Appendix

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Definition 1 (Multiset). A multiset is a generalized concept of a set that allows multiple instances for its elements. More formally, a multiset is a 2-tuple $X = (S, m)$ where S is the *underlying set* of X that is formed from its *distinct elements*, and $m : S \to \mathbb{N}_{>1}$ gives the *multiplicity* of the elements.

Lemma 2. *Let G_1 and G_2 be any two non-isomorphic graphs. If a graph neural network $\mathcal{A} : \mathcal{G} \to \mathbb{R}^d$ maps G_1 and G_2 to different embeddings, the Weisfeiler-Lehman graph isomorphism test also decides G_1 and G_2 are not isomorphic.*

Theorem 3. *Let $\mathcal{A} : \mathcal{G} \to \mathbb{R}^d$ be a GNN. With a sufficient number of GNN layers, $\mathcal{A}$ maps any graphs G_1 and G_2 that the Weisfeiler-Lehman test of isomorphism decides as non-isomorphic, to different embeddings if the following conditions hold:*

a) $\mathcal{A}$ aggregates and updates node features iteratively with

$$h_v^{(k)} = \phi\left(h_v^{(k-1)}, f\left(\left\{h_u^{(k-1)} : u \in \mathcal{N}(v)\right\}\right)\right),$$

where the functions f, which operates on multisets, and ϕ are injective.

b) $\mathcal{A}$'s graph-level readout, which operates on the multiset of node features $\left\{h_v^{(k)}\right\}$, is injective.

Appendix

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Lemma 4. *Assume the input feature space* $\mathcal{X}$ *is countable. Let* $g^{(k)}$ *be the function parameterized by a GNN's* k*-th layer for* $k = 1, ..., L$*, where* $g^{(1)}$ *is defined on multisets* $X \subset \mathcal{X}$ *of bounded size. The range of* $g^{(k)}$*, i.e., the space of node hidden features* $h_v^{(k)}$*, is also countable for all* $k = 1, ..., L$*.*

Lemma 5. *Assume* $\mathcal{X}$ *is countable. There exists a function* $f : \mathcal{X} \rightarrow \mathbb{R}^n$ *so that* $h(X) = \sum_{x \in X} f(x)$ *is unique for each multiset* $X \subset \mathcal{X}$ *of bounded size. Moreover, any multiset function* g *can be decomposed as* $g(X) = \phi\left(\sum_{x \in X} f(x)\right)$ *for some function* ϕ*.*

Corollary 6. *Assume* $\mathcal{X}$ *is countable. There exists a function* $f : \mathcal{X} \rightarrow \mathbb{R}^n$ *so that for infinitely many choices of* ϵ*, including all irrational numbers,* $h(c, X) = (1 + \epsilon) \cdot f(c) + \sum_{x \in X} f(x)$ *is unique for each pair* (c, X)*, where* $c \in \mathcal{X}$ *and* $X \subset \mathcal{X}$ *is a multiset of bounded size. Moreover, any function* g *over such pairs can be decomposed as* $g(c, X) = \varphi\left((1 + \epsilon) \cdot f(c) + \sum_{x \in X} f(x)\right)$ *for some function* φ*.*

Appendix

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Lemma 7. *There exist finite multisets* $X_1 \neq X_2$ *so that for any linear mapping* W, $\sum_{x \in X_1} \mathrm{ReLU}(Wx) = \sum_{x \in X_2} \mathrm{ReLU}(Wx)$.

Corollary 8. *Assume* $\mathcal{X}$ *is countable. There exists a function* $f : \mathcal{X} \to \mathbb{R}^n$ *so that for* $h(X) = \frac{1}{|X|} \sum_{x \in X} f(x)$, $h(X_1) = h(X_2)$ *if and only if multisets* X_1 *and* X_2 *have the same distribution. That is, assuming* $|X_2| \geq |X_1|$, *we have* $X_1 = (S, m)$ *and* $X_2 = (S, k \cdot m)$ *for some* $k \in \mathbb{N}_{\geq 1}$.

Corollary 9. *Assume* $\mathcal{X}$ *is countable. Then there exists a function* $f : \mathcal{X} \to \mathbb{R}^\infty$ *so that for* $h(X) = \max_{x \in X} f(x)$, $h(X_1) = h(X_2)$ *if and only if* X_1 *and* X_2 *have the same underlying set.*