Publish at: 37th Conference on Neural Information Processing Systems (NeurIPS 2023)
Present at: Singapore ACM SIGKDD Symposium 2024

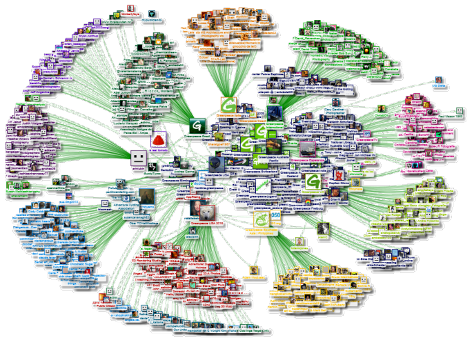
LD^2: Scalable Heterophilous GNN with Decoupled Embeddings

Ningyi Liao **Siqiang Luo** **Xiang Li** **Jieming Shi**

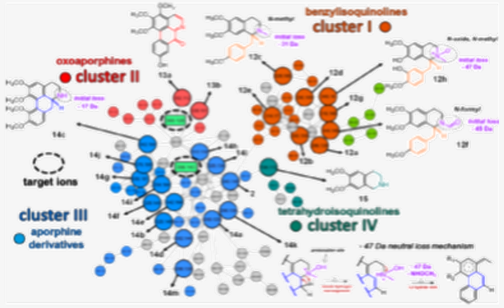
PRESENT BY: Ningyi Liao

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE
EAST CHINA NORMAL UNIVERSITY 華東師範大學
THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學

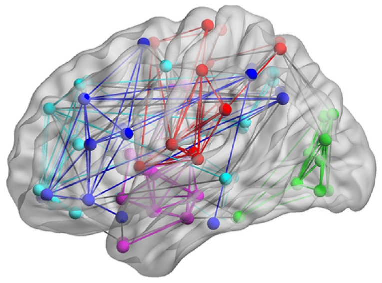
Background: Learning on Graphs



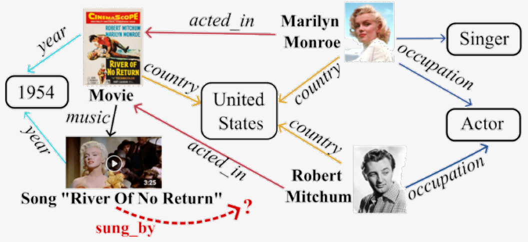
Social Networks



Molecular Networks

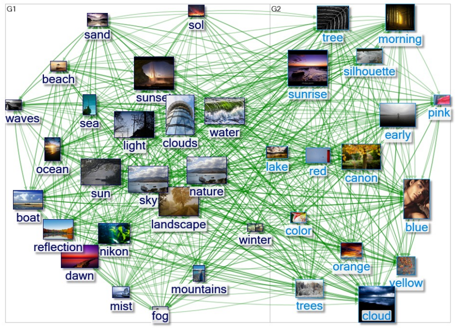


Human Brain Networks

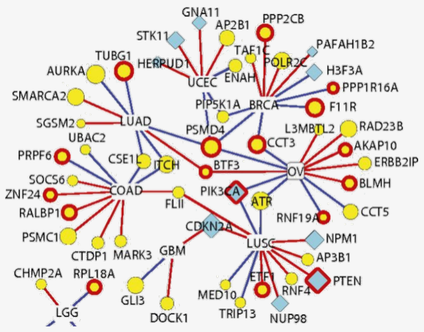


Knowledge Graphs

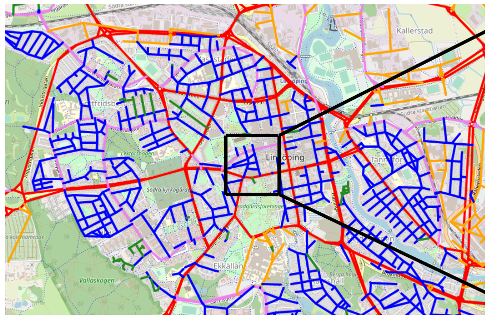
Transportation Networks



Tag Networks



Biological Networks

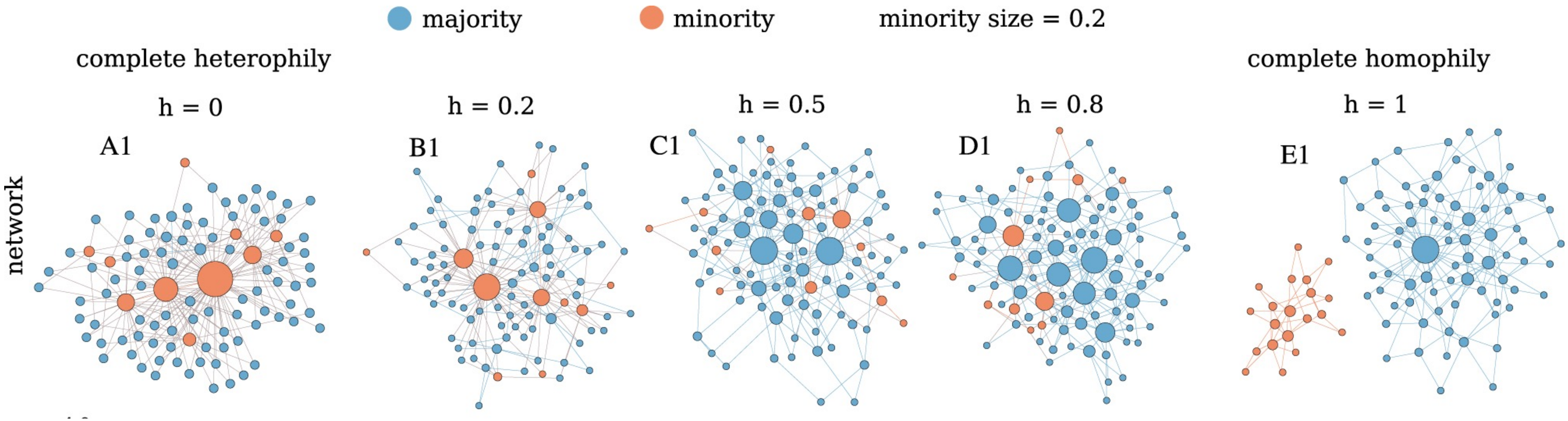


Road Networks

Images downloaded from the Internet.

Background: Graph Heterophily

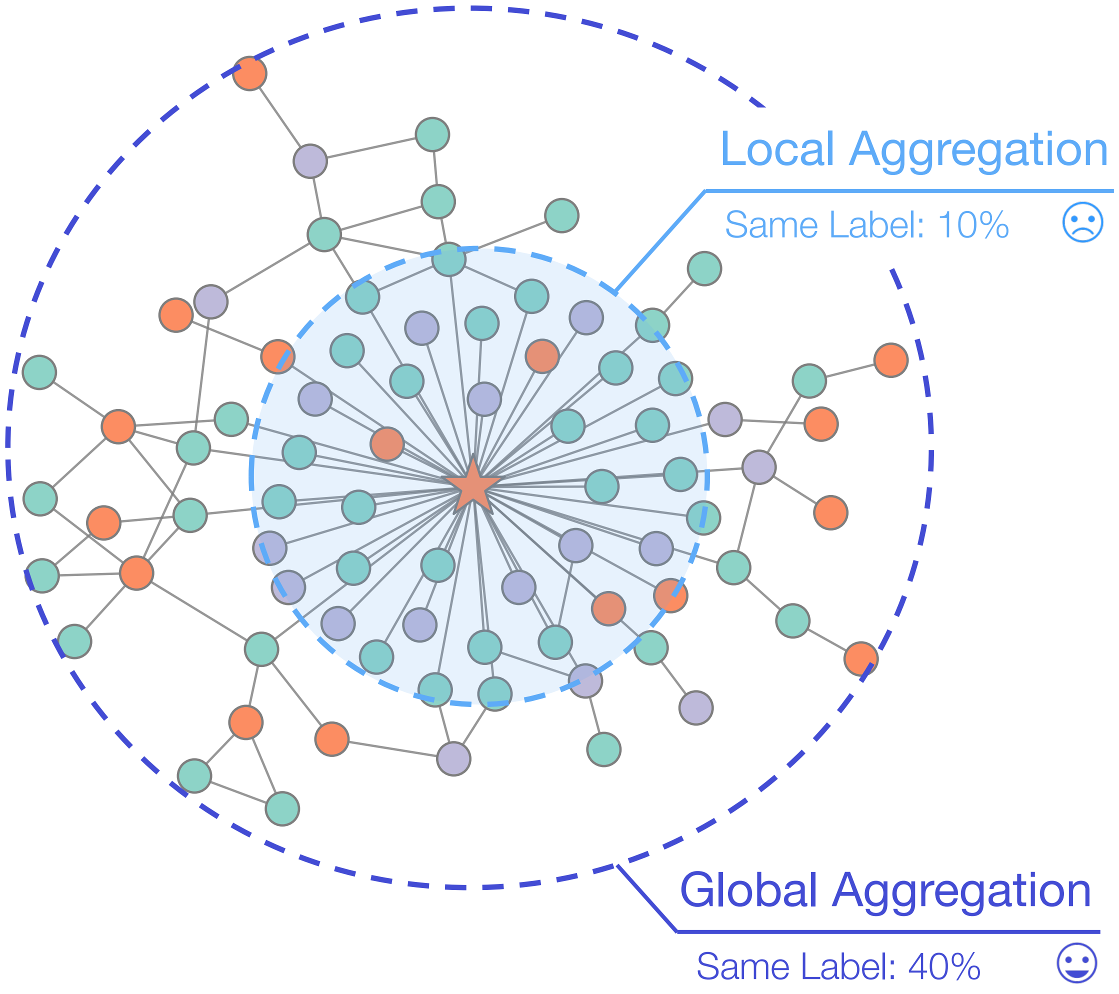
- Homophily: connected nodes tend to be of *similar* classes
- Heterophily: connected nodes tend to be of *dissimilar* classes



Karimi, F., et al. Homophily influences ranking of minorities in social networks. Sci Rep (2018).

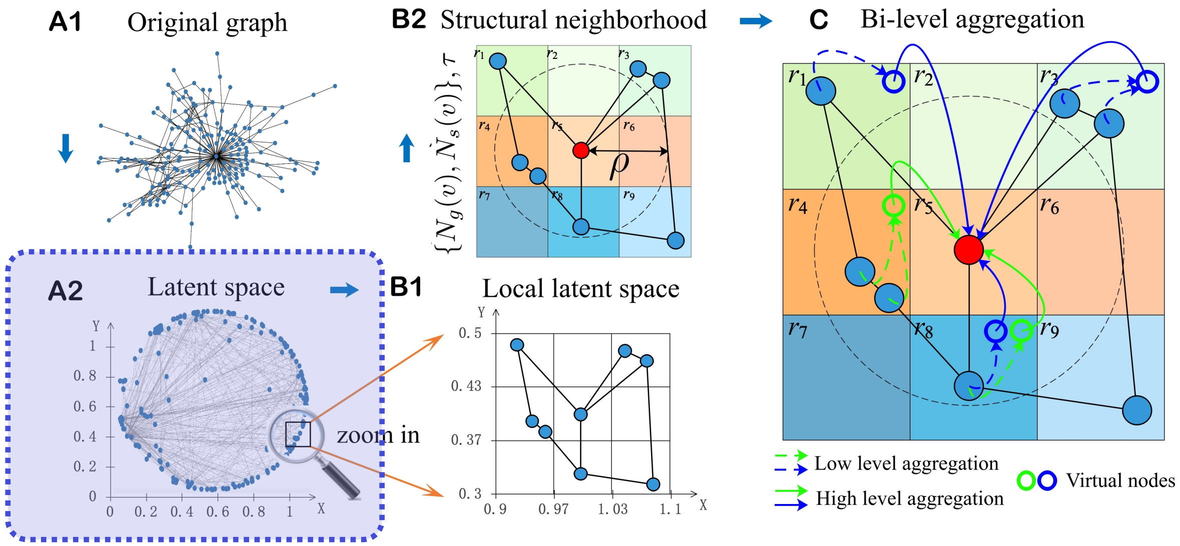
Background: Heterophilous GNN

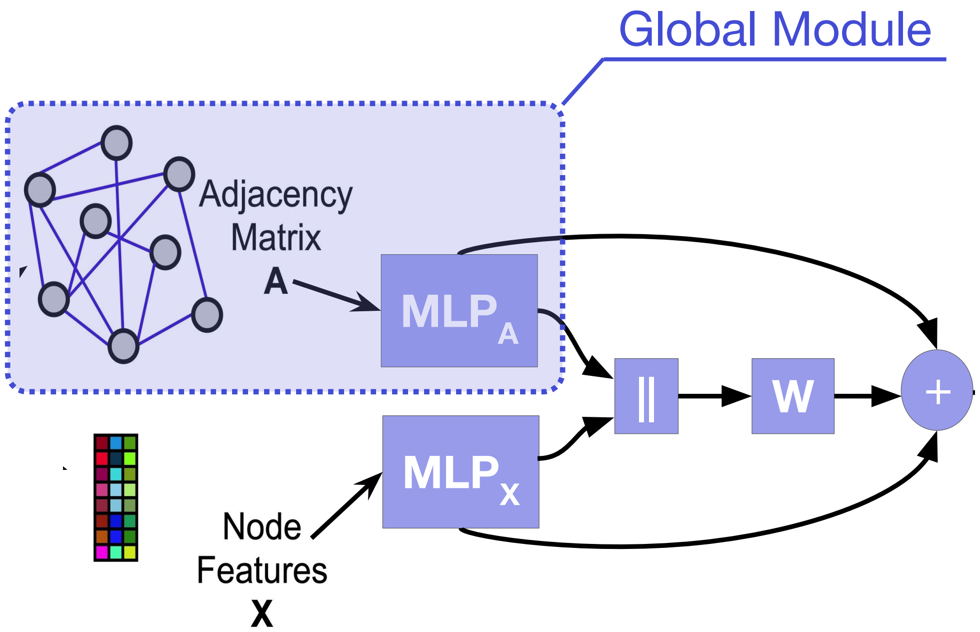
- Example of heterophily: fraudster – normal user in transaction networks
- Conventional locality-based GNNs not suitable
- Existing heterophilous GNNs rely on global computation



Background: Global Hetero GNN

- Geom-GCN: global embedding
- LINKX: full-graph adjacency MLP

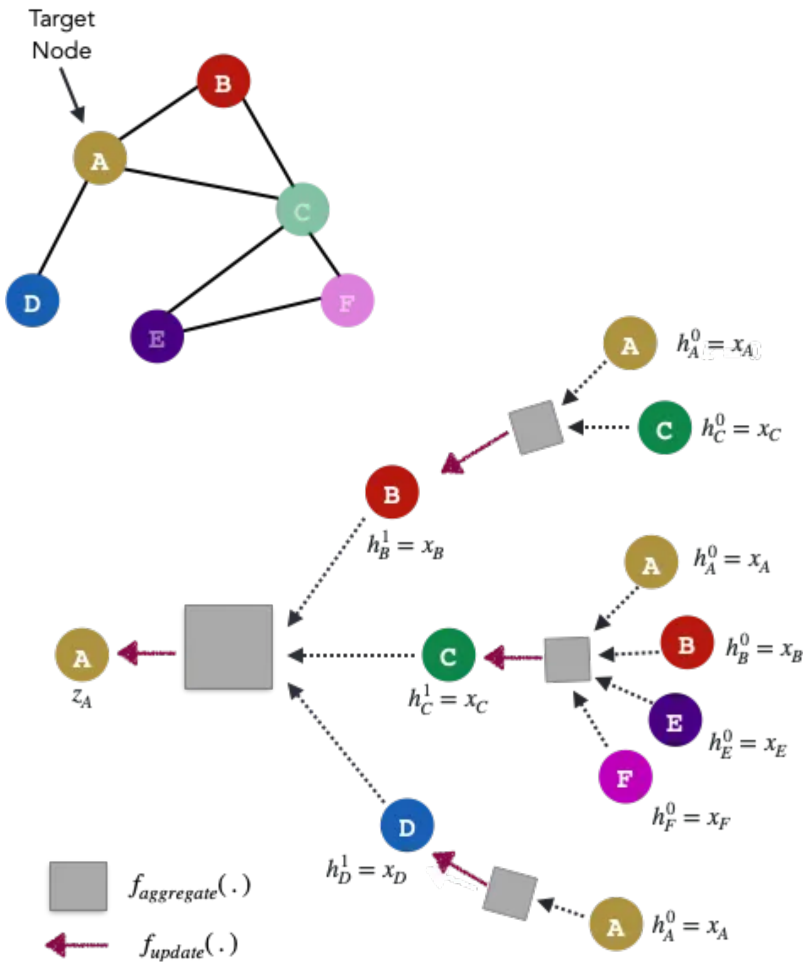




[1] H Pei et al. "Geom-GCN: Geometric Graph Convolutional Networks". ICLR 2020.
[2] D Lim et al. "Large scale learning on non-homophilous graphs: New benchmarks and strong simple methods". NeurIPS 2021.

Motivation: GNN Scalability

- <u>Neighborhood explosion</u>: overhead increases for long-range computation
- Real-world graphs are on the scale of millions or billions
- Global computation not scalable to large graphs



Xiao Li, et al., "A survey of graph neural network based recommendation in social networks", Neurocomputing, 2023.

Motivation: Hetero GNNs not Scalable Enough

Natural conflict:

Global Computation *vs* Scalability & Minibatch

Model	Time - Precomp	Time - Train	Time - Test	GPU Memory
GPRGNN	$O(m)$	$O(IL_PmF + ILnF^2)$	$O(L_PmF + LnF^2)$	$O(LnF + m + LF^2)$
GCNJK	–	$O(ILmF + ILnF^2)$	$O(LmF + LnF^2)$	$O(L_CnF + L_CF^2)$
MixHop	–	$O(IL_PLmF + ILnF^2)$	$O(L_PLmF + LnF^2)$	$O(CLnF + CLF^2)$
LINKX	–	$O(ImF + ILnF^2)$	$O(mF + LnF^2)$	$O(L_Cn_bF + L_CF^2 + nF)$
LD² (ours)	$O(L_PmF)$	$O(ILnF^2)$	$O(LnF^2)$	$O(L_Cn_bF + L_CF^2)$

Terms that not scalable to large m and n

Terms that not suitable for minibatch

Method: LD^2 Framework

- Precomputation:

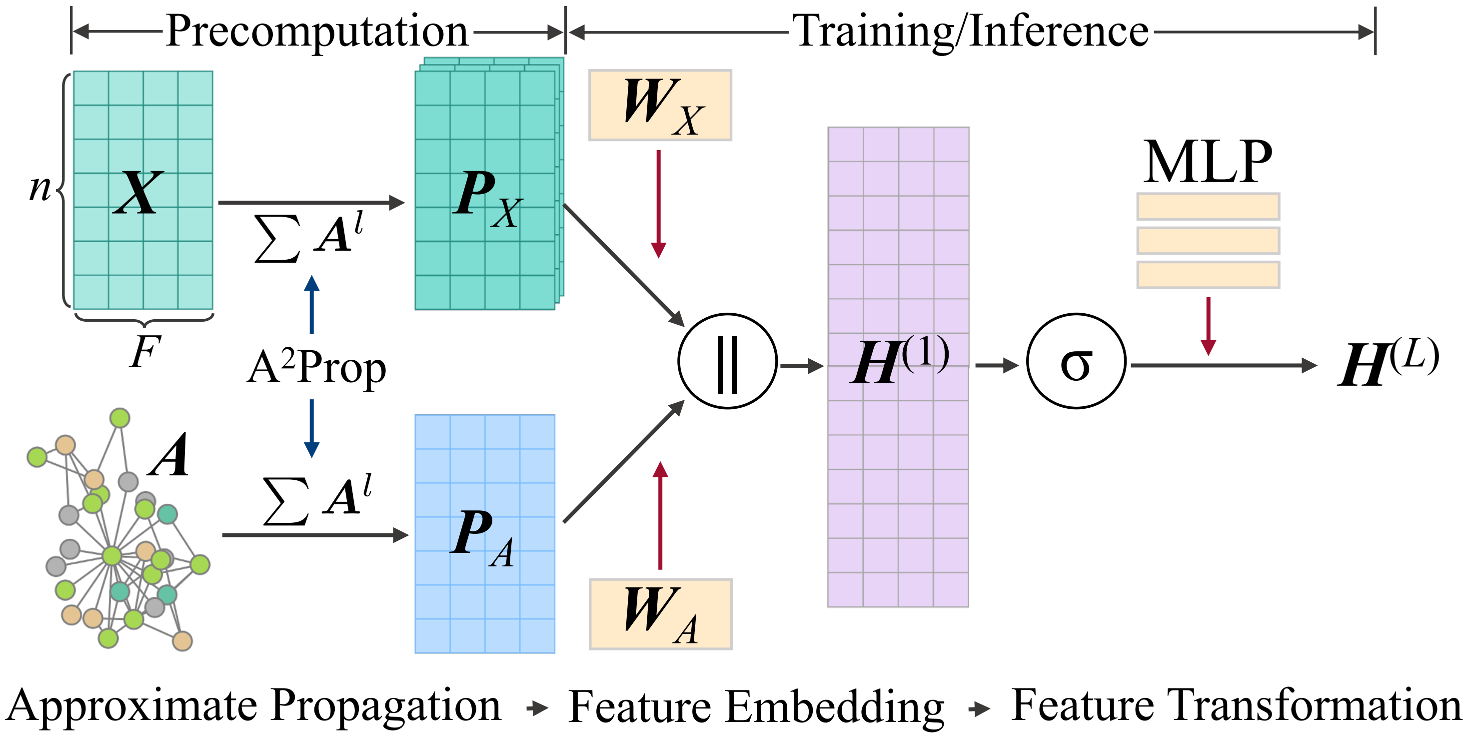
$$\boldsymbol{P}_A, \boldsymbol{P}_X = \mathrm{A}^2\mathrm{Prop}(\boldsymbol{A}, \boldsymbol{X})$$

- Feature Transformation:

$$\boldsymbol{H}^{(L)} = \mathrm{MLP}(\boldsymbol{P}_A \boldsymbol{W}_A \| \boldsymbol{P}_X \boldsymbol{W}_X)$$

Long-Distance Feature Embedding

Low-Dimensional Adj Embedding



Linear to m and n

Decoupled simple minibatching

Method: Adjacency Embedding

- <u>L</u>ow-<u>D</u>imensional 2-hop adjacency decomposition

Adj Embedding

Dense $n \times F$

$$\boldsymbol{P}_A \cdot \boldsymbol{P}_A^\top \approx \boldsymbol{A}^2$$

2-hop Adj Matrix

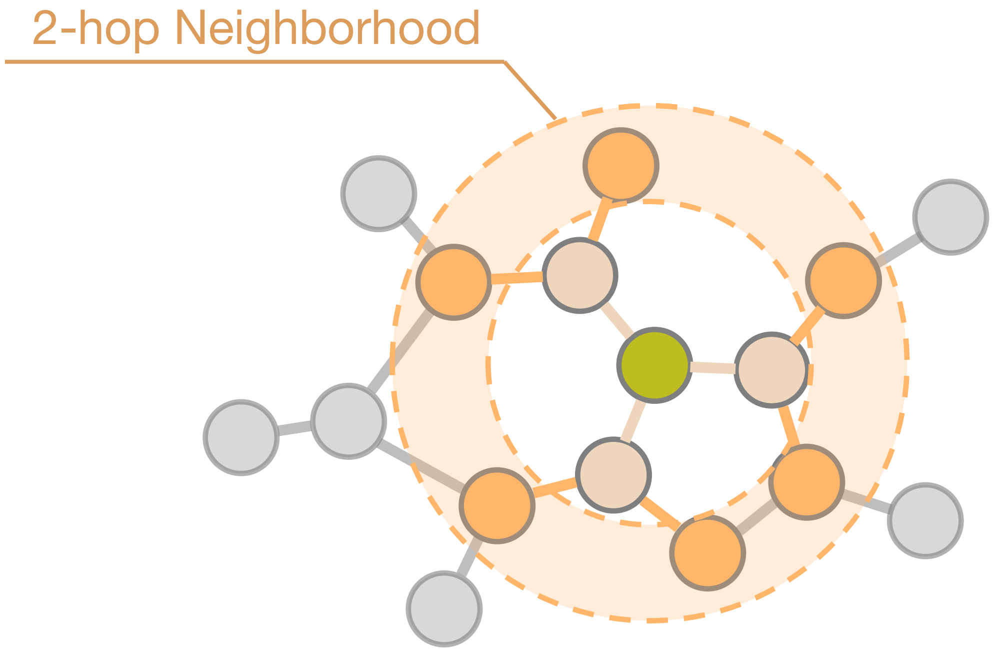
Sparse $n \times n$

- Calculation:

Orthonormalize

$$\boldsymbol{P}_A = \phi\left(\boldsymbol{A}^2 \cdots \phi\left(\boldsymbol{A}^2 \mathcal{N}\right)\right)$$

Gaussian Matrix

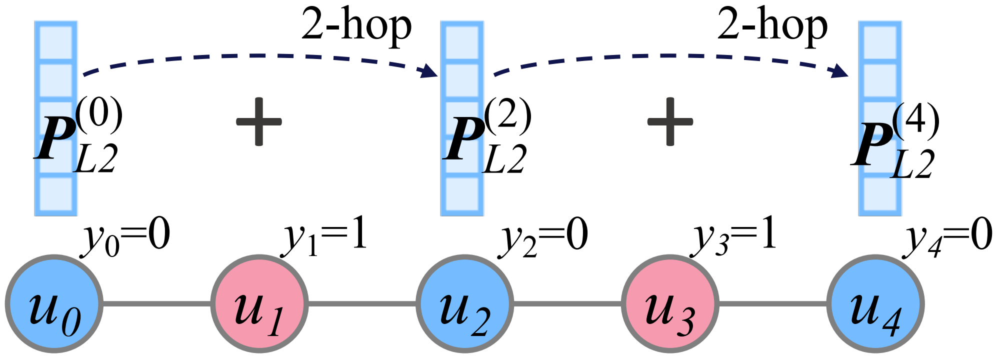


Method: Feature Embedding

- Long-Distance generalized graph propagation

CHANNEL①: *Constant* 2-hop Adjacency Propagation

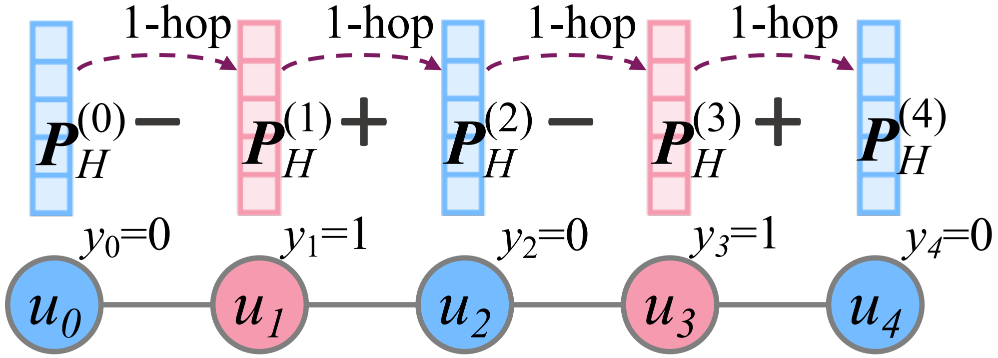
$$\boldsymbol{P}_{X,L2} = \sum_{l=1}^{L} \bar{\boldsymbol{A}}^{2l} \cdot \boldsymbol{X}$$



CHANNEL②: *Inverse* 1-hop Laplacian Propagation

Laplacian $\tilde{L} = I - \tilde{A}$

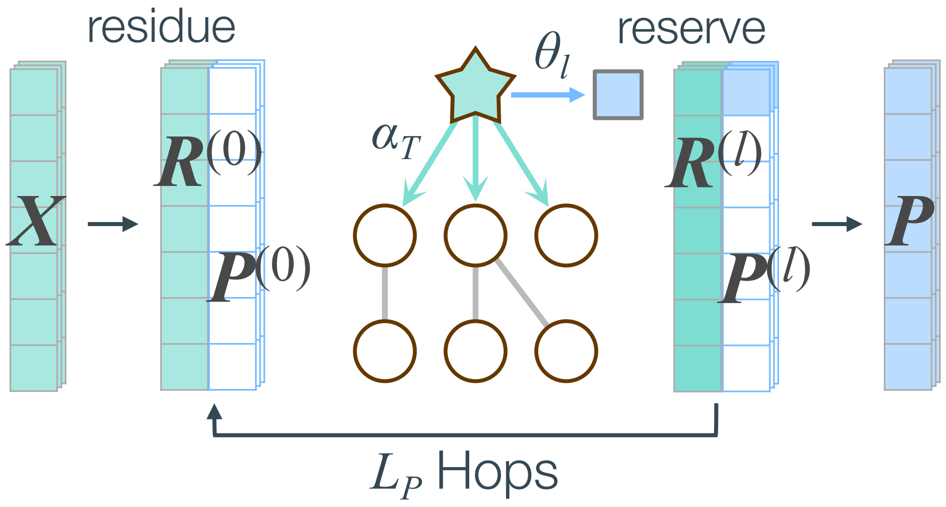
$$\boldsymbol{P}_{X,H} = \sum_{l=1}^{L} (\tilde{\boldsymbol{L}} + \boldsymbol{I})^{l} \cdot \boldsymbol{X}$$



Method: Embedding Precomputation

- Multi-channel embedding, one-time computation:

CHANNEL	
Adjacency	$\boldsymbol{P}_A = \phi\left(\boldsymbol{A}^{2L} \cdot \boldsymbol{\mathcal{N}}\right)$
Feature – 2-hop Constant	$\boldsymbol{P}_{X,L2} = \sum_{l=1}^{L} \bar{\boldsymbol{A}}^{2l} \cdot \boldsymbol{X}$
Feature – 1-hop Inverse	$\boldsymbol{P}_{X,H} = \sum_{l=1}^{L} \tilde{\boldsymbol{L}}^{l} \cdot \boldsymbol{X}$



Precompute Time

$O(L_P m F)$

Evaluation: Effectiveness

- Top 1 accuracy on 6/8 large-scale heterophilous datasets
- No accuracy drop for minibatch

Dataset	genius	tolokers	arxiv-year	penn94	twitch-gamers	pokec	snap-patents	wiki
Nodes n	421,858	11,758	169,343	41,536	168,114	1,632,803	**2,738,035**	1,770,981
Edges m	922,864	1,038,000	1,157,799	1,362,220	6,797,557	22,301,964	13,967,949	**242,507,069**
F / N_c	12 / 2	10 / 2	128 / 5	4,814 / 2	7 / 2	65 / 2	269 / 5	600 / 5
MLP	82.47 ±0.06	73.38 ±0.25	37.23 ±0.31	74.41 ±0.48	61.26 ±0.19	61.81 ±0.07	23.03 ±1.48	35.64 ±0.10
PPRGo	79.81 ±0.00	78.16 ±0.00	39.35 ±0.12	58.75 ±0.31	47.19 ±2.26	50.61 ±0.04	(>12h)	(>12h)
SGC	79.85 ±0.01	71.16 ±0.06	43.40 ±0.16	68.31 ±0.27	57.05 ±0.21	56.58 ±0.06	37.70 ±0.06	28.12 ±0.08
GCNJK-GS	80.65 ±0.07	74.41 ±0.73	48.26 ±0.64	65.91 ±0.16	59.91 ±0.42	59.38 ±0.21	33.64 ±0.05	42.95 ±0.39
MixHop-GS	80.63 ±0.04	77.47 ±0.40	49.26 ±0.16	75.00 ±0.37	61.80 ±0.00	64.02 ±0.02	34.73 ±0.15	45.52 ±0.11
LINKX	82.51 ±0.10	77.74 ±0.13	**50.44** ±0.30	**78.63** ±0.25	64.15 ±0.18	68.64 ±0.65	52.69 ±0.05	50.59 ±0.12
LD^2 (ours)	**85.31** ±0.06	**79.76** ±0.26	50.29 ±0.11	75.52 ±0.10	**64.33** ±0.19	**74.93** ±0.10	**58.58** ±0.34	**52.91** ±0.16

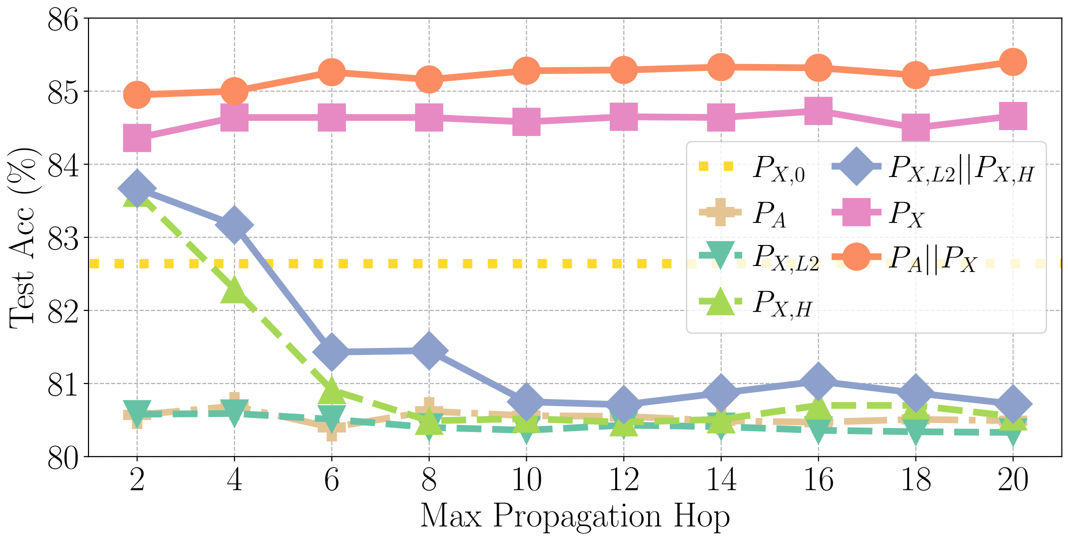
Evaluation: Efficiency

- 3-15× faster minibatch training, significantly fast inference
- Up to 5× lower GPU memory for large graphs

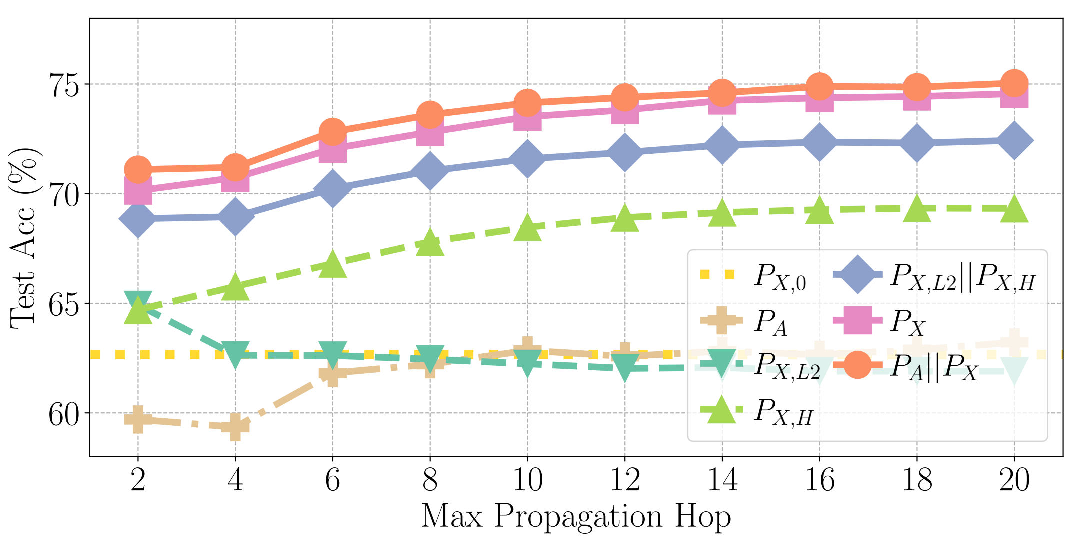
Dataset	twitch-gamers			pokec			snap-patents			wiki		
	Learn	Infer	Mem.	Learn	Infer	Mem.	Learn	Infer	Mem.	Learn	Infer	Mem.
MLP	6.36	0.02	0.61	47.86	0.11	13.77	27.39	0.28	9.33	133.55	0.62	18.15
PPRGo	10.46+15.88	0.41	9.64	121.95+56.11	2.69	3.82	(>12h)			(>12h)		
SGC	0.09+0.74	0.01	0.28	1.05+8.08	0.01	0.28	4.94+23.54	0.01	0.42	12.66+7.98	0.01	0.52
GCNJK-GS	71.48	0.02*	7.33	27.33	0.09*	9.03	19.02	0.23*	9.21	95.52	0.69*	16.36
MixHop-GS	52.12	0.01*	1.49	71.35	0.03*	12.91	45.24	0.16*	19.58	84.22	0.23*	16.28
LINKX	10.99	0.19	2.35	28.77	0.33	9.03	39.80	0.22	21.53	180.71	1.14	14.53
LD^2 (ours)	0.85+**1.96**	**0.01**	**1.44**	17.95+**6.18**	**0.01**	**3.82**	31.32+**6.96**	**0.02**	**3.96**	28.12+**6.50**	**0.01**	**4.47**

Evaluation: Efficiency

- Higher propagation hops are important for distant information
- Combination of channels is effective for different graphs



genius



pokec

Summary

- **LD^2 Framework**: Pre-Propagation Decoupled Heterophilous GNN, optimized minibatch training
- **Low Dimension Adjacency Embedding**: embed full graph topology in low dimensional matrix representation
- **Long Distance Feature Embeddings**: multi-channel node features by different graph propagation schemes
- **Performance Improvement**: 3–15× faster minibatch training and inference, up to 5× smaller memory footprint

Following work ...

Benchmarking Spectral Graph Neural Networks:
A Comprehensive Study on Effectiveness and Efficiency

arXiv: 2406.09675

GitHub: gdmnl/Spectral-GNN-Benchmark

- Unified framework for over 30 spectral GNN filters
- Scalable pipeline for training on million-scale graphs
- Observation and guideline on homophilous & heterophilous data

THANK YOU

Acknowledgments

This research is supported by Singapore MOE AcRF Tier-2 funding (MOE-T2EP20122-0003), NTU startup grant (020948-00001), and the Joint NTU-WeBank Research Centre on FinTech. Xiang Li is supported by Shanghai Science and Technology Committee General Program No. 22ZR1419900 and National Natural Science Foundation of China No. 62202172. Jieming Shi is supported by Hong Kong RGC ECS No. 25201221, National Natural Science Foundation of China No. 62202404.