Imperial College London

Enhancing Node Representations for Real-World Complex Networks with Topological Augmentation

Xiangyu Zhao*, Zehui Li*, Mingzhu Shen, Guy-Bart Stan, Pietro Liò, Yiren Zhao
{**x.zhao22**, zehui.li22}@imperial.ac.uk

ECAI 2024
23/10/2024

Background
Graph Augmentation

Purpose of Graph Augmentation:

Generate augmented graphs that can enhance the information from the original graphs

Existing Graph Augmentation Methods:

- Graph structure perturbation: DropNode, DropEdge, node feature masking, edge feature masking
- Generating synthetic data: Mixup, diffusion models

Problems:

- Cannot capture higher-order node relations beyond pairwise
- Cannot increase GNNs' expressiveness

Background

Representation Learning on Higher-Order Graphs

There is a lack of data that can be used to form higher-order edges

Library	Deep Hypergraph	PyTorch Geometric
Graph Type	Hypergraph	Simple Graph
#Real-World Datasets	17	93
Maximum #Nodes	240 094	59 249 719
Maximum #Edges	679 302	978 147 253

Background
Motivations

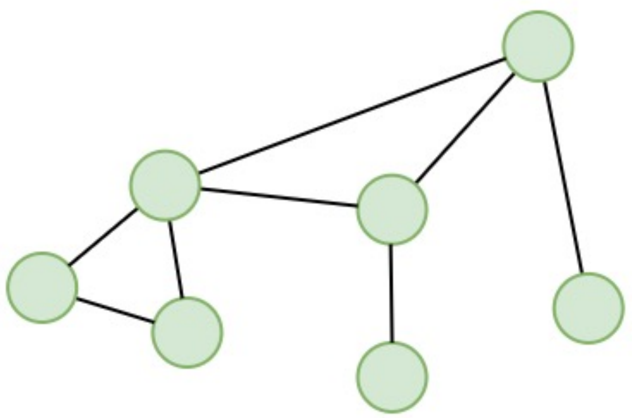
There is a need of:

- A graph augmentation method that integrates higher-order edge information into the original graphs
- A hyperedge construction strategy to deal with the scarcity of available hyperedge data
- A collection of real-world graph datasets containing both simple edges and hyperedges

Our solution: Topological Augmentation (TopoAug)

Method

Topological Deep Learning

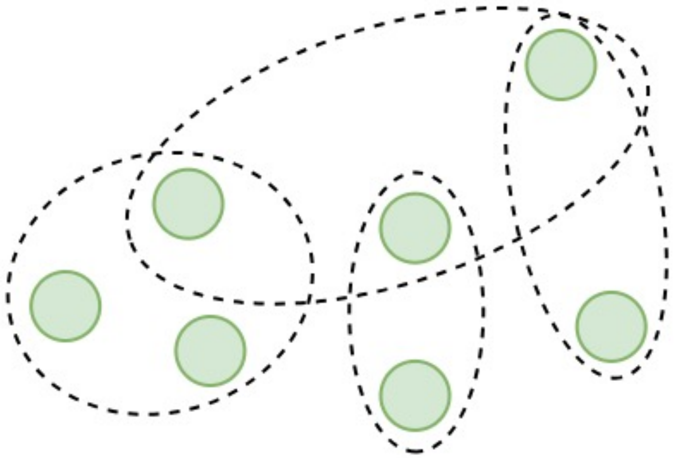


Simple Graphs

Contains only simple edges:

$$\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$$

Each simple edges connects two nodes

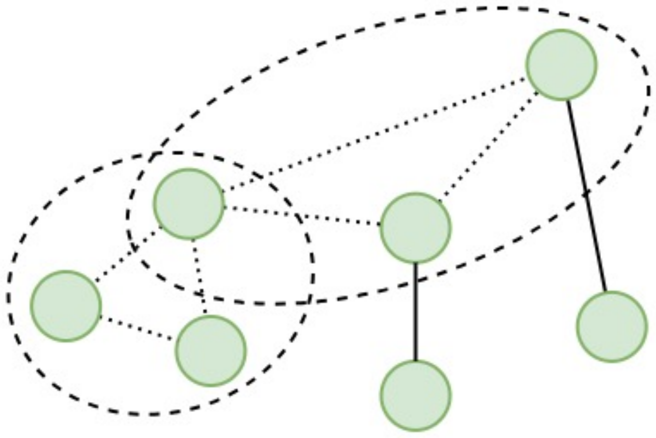


Hypergraphs

Contains only hyperedges:

$$\mathcal{E} \subseteq \mathcal{P}(\mathcal{V}) \setminus \emptyset$$

Each hyperedge connects two or more nodes



Combinatorial Complexes

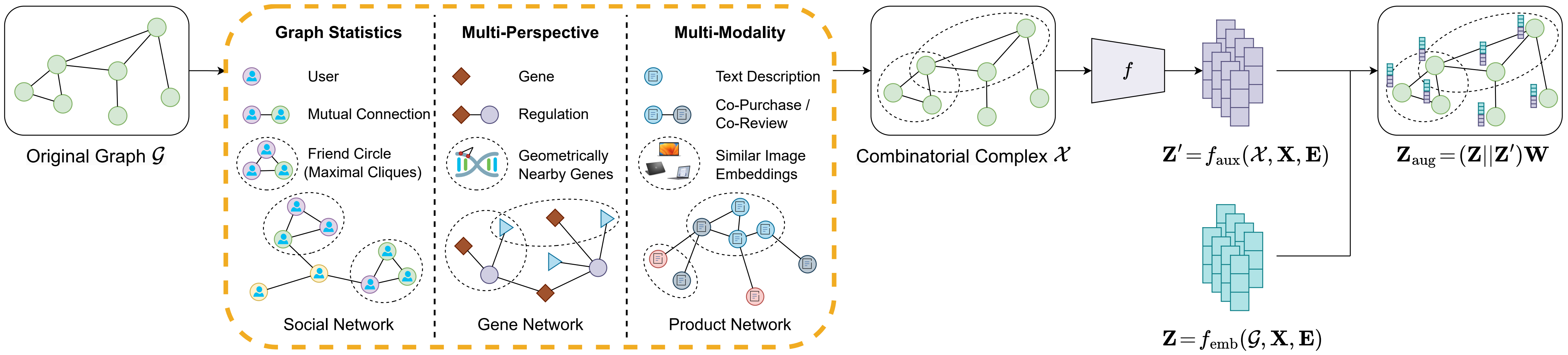
Contains both simple edges and hyperedges using

$$\mathcal{X} \subseteq \mathcal{P}(\mathcal{V}) \setminus \emptyset$$

and rank function rk

Method

Our Model: Topological Augmentation



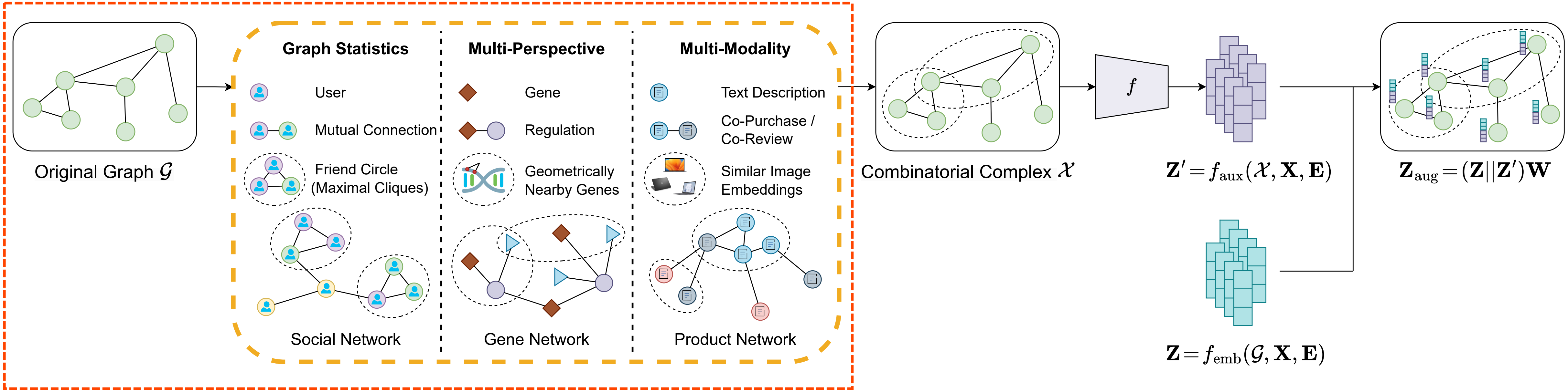
Hyperedge Extraction:

$$\mathcal{E}_h = h(\mathcal{G}, \mathbf{X}, \mathbf{E})$$

- From the graph statistics: social networks
- From a different data perspective: biological networks
- From a different data modality: commercial networks

Method

Our Model: Topological Augmentation



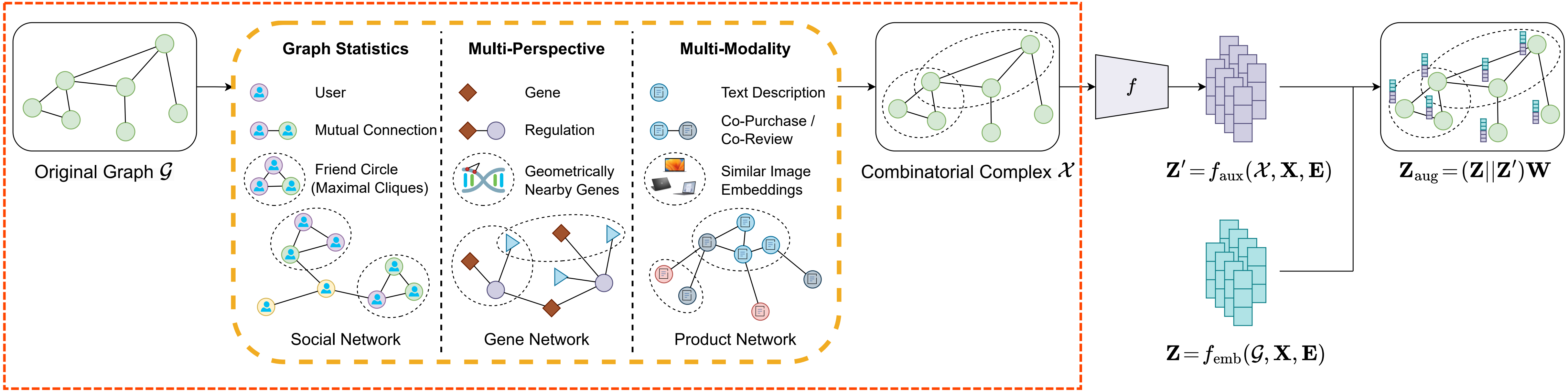
Hyperedge Extraction:

$$\mathcal{E}_h = h(\mathcal{G}, \mathbf{X}, \mathbf{E})$$

- From the graph statistics: social networks
- From a different data perspective: biological networks
- From a different data modality: commercial networks

Method

Our Model: Topological Augmentation



Combinatorial Complex Construction:

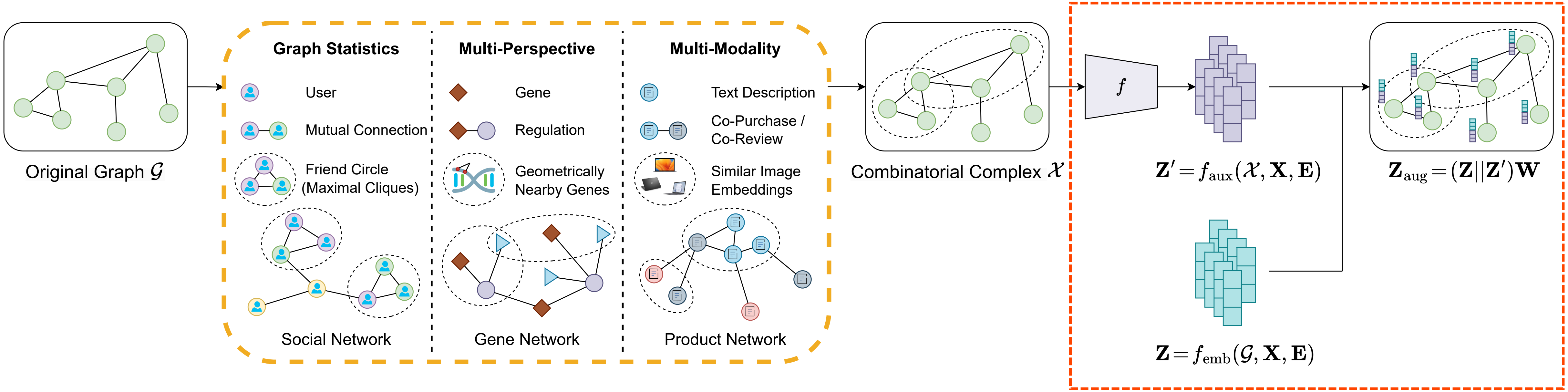
$\mathcal{V}$ remains unchanged from the original graph

$$\mathcal{X} = \{\{v\} | v \in \mathcal{V}\} \cup \mathcal{E} \cup \mathcal{E}_h$$

$$\forall x \in \mathcal{X}.\, \mathrm{rk}(x) = \begin{cases} 0 & \text{for } x = \{v\} \\ 1 & \text{for } x \in \mathcal{E} \\ 2 & \text{for } x \in \mathcal{E}_h \end{cases}$$

Method

Our Model: Topological Augmentation

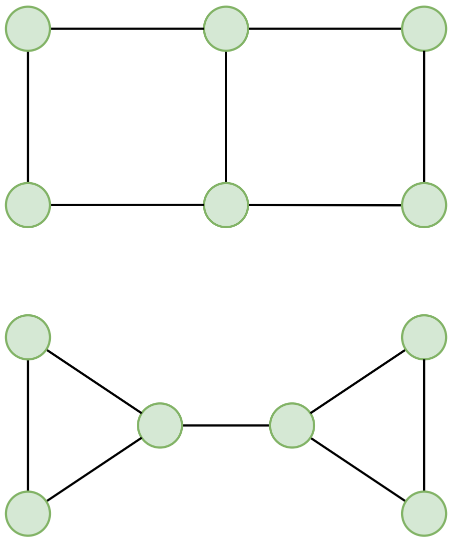


Graph Augmentation Pipeline:

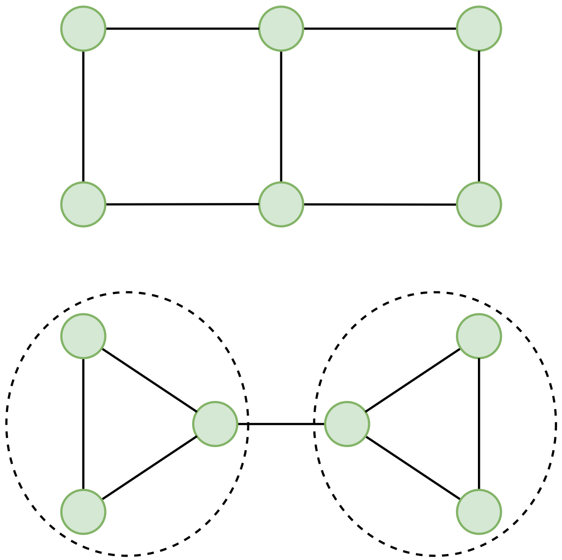
$$\mathrm{TopoAug}(\mathcal{G}, \mathbf{X}, \mathbf{E}) = \Big(\underbrace{f_{\text{emb}}(\mathcal{G}, \mathbf{X}, \mathbf{E})}_{\substack{\text{Original node embeddings} \\ \text{from original simple edges}}} \,\|\, \underbrace{f_{\text{aux}}\big(\overbrace{f_{CC}(\mathcal{G})}^{\substack{\text{Combinatorial complex} \\ \text{construction}}}, \mathbf{X}, \mathbf{E}\big)}_{\substack{\text{Auxiliary node embeddings} \\ \text{from extracted hyperedges}}} \Big) \mathbf{W}$$

Method

Beyond 1-WL Limits



Without TopoAug



With TopoAug

TopoAug helps GNNs to distinguish non-isomorphic graphs that have the same node degrees…

…and therefore surpass the limitations posed by the 1-WL test!

Data

Filling the Gap: 26 Novel Real-World Graph Datasets

Library	Deep Hypergraph	Our Work
Graph Type	Hypergraph	Combinatorial Complex
#Real-World Datasets	17	26
Average #Nodes	35 240	12 218
Average #Edges	N/A	149 262
Average #Hyperedges	51 158	73 490

Data

Datasets at a Glance

Name	Hyperedge Construction Mechanism	#Datasets	Average #Nodes	Average #Edges	Average #Hyperedges	#Classes
MUSAE-GitHub	Graph Statistics	1	37 700	578 006	223 672	4
MUSAE-Facebook		1	22 470	342 004	236 663	4
MUSAE-Twitch		6	5 686	143 038	110 142	2
MUSAE-Wiki		3	6 370	266 998	118 920	Regression
GRAND-Tissues	Multi-Perspective	6	5 931	5 926	11 472	3
GRAND-Diseases		4	4 596	6 252	7 743	3
Cora		2	2 708	5 429	1 326	7
PubMed		1	19 717	44 338	7 963	3
Amazon-Computers	Multi-Modality	1	10 226	55 324	10 226	10
Amazon-Photos		1	6 777	45 306	6 777	10

Data

Datasets at a Glance

Name	Hyperedge Construction Mechanism	#Datasets	Average #Nodes	Average #Edges	Average #Hyperedges	#Classes
MUSAE-GitHub	Graph Statistics	1	37 700	578 006	223 672	4
MUSAE-Facebook		1	22 470	342 004	236 663	4
MUSAE-Twitch		6	5 686	143 038	110 142	2
MUSAE-Wiki		3	6 370	266 998	118 920	Regression
GRAND-Tissues	Multi-Perspective	6	5 931	5 926	11 472	3
GRAND-Diseases		4	4 596	6 252	7 743	3
Cora		2	2 708	5 429	1 326	7
PubMed		1	19 717	44 338	7 963	3
Amazon-Computers	Multi-Modality	1	10 226	55 324	10 226	10
Amazon-Photos		1	6 777	45 306	6 777	10

Data

Datasets at a Glance

Name	Hyperedge Construction Mechanism	#Datasets	Average #Nodes	Average #Edges	Average #Hyperedges	#Classes
MUSAE-GitHub	Graph Statistics	1	37 700	578 006	223 672	4
MUSAE-Facebook		1	22 470	342 004	236 663	4
MUSAE-Twitch		6	5 686	143 038	110 142	2
MUSAE-Wiki		3	6 370	266 998	118 920	Regression
GRAND-Tissues	Multi-Perspective	6	5 931	5 926	11 472	3
GRAND-Diseases		4	4 596	6 252	7 743	3
Cora		2	2 708	5 429	1 326	7
PubMed		1	19 717	44 338	7 963	3
Amazon-Computers	Multi-Modality	1	10 226	55 324	10 226	10
Amazon-Photos		1	6 777	45 306	6 777	10

Data

Datasets at a Glance

Name	Hyperedge Construction Mechanism	#Datasets	Average #Nodes	Average #Edges	Average #Hyperedges	#Classes
MUSAE-GitHub	Graph Statistics	1	37 700	578 006	223 672	4
MUSAE-Facebook		1	22 470	342 004	236 663	4
MUSAE-Twitch		6	5 686	143 038	110 142	2
MUSAE-Wiki		3	6 370	266 998	118 920	Regression
GRAND-Tissues	Multi-Perspective	6	5 931	5 926	11 472	3
GRAND-Diseases		4	4 596	6 252	7 743	3
Cora		2	2 708	5 429	1 326	7
PubMed		1	19 717	44 338	7 963	3
Amazon-Computers	Multi-Modality	1	10 226	55 324	10 226	10
Amazon-Photos		1	6 777	45 306	6 777	10

Data

Datasets at a Glance

Name	Hyperedge Construction Mechanism	#Datasets	Average #Nodes	Average #Edges	Average #Hyperedges	#Classes
MUSAE-GitHub	Graph Statistics	1	37 700	578 006	223 672	4
MUSAE-Facebook		1	22 470	342 004	236 663	4
MUSAE-Twitch		6	5 686	143 038	110 142	2
MUSAE-Wiki		3	6 370	266 998	118 920	Regression
GRAND-Tissues	Multi-Perspective	6	5 931	5 926	11 472	3
GRAND-Diseases		4	4 596	6 252	7 743	3
Cora		2	2 708	5 429	1 326	7
PubMed		1	19 717	44 338	7 963	3
Amazon-Computers	Multi-Modality	1	10 226	55 324	10 226	10
Amazon-Photos		1	6 777	45 306	6 777	10

Results

Node Classification Accuracy (%)

Method	Graph Statistics		Multi-Perspective			Multi-Modality	
	GitHub	TwitchDE	CoraCoCite	Brain	LungCancer	Computers	Photos
GCN	87.2	65.5	81.4	62.5	59.6	75.6	29.5
GAT	86.4	64.5	83.0	62.5	59.6	74.2	43.4
GraphSAGE	87.1	65.7	83.2	61.8	61.5	75.0	36.6
HyperConv	80.8	65.4	79.1	62.5	59.3	84.2	33.7
ED-HNN	86.2	68.1	80.3	66.3	60.2	97.3	78.6
GCN+DropNode	86.2	**67.9**	85.5	63.3	58.2	91.8	71.1
GCN+DropEdge	86.6	67.7	86.3	63.2	60.7	92.2	78.6
GCN+Mixup	85.8	67.6	85.6	65.0	59.0	87.7	78.0
GCN+NodeFeatureMasking	85.9	67.8	85.2	64.0	59.2	91.7	80.7
GCN+TopoAug (Ours)	**87.4**	**67.9**	**86.6**	**66.7**	**63.7**	**98.1**	**80.9**

Results

Node Classification Accuracy (%)

Method	Graph Statistics		Multi-Perspective			Multi-Modality	
	GitHub	TwitchDE	CoraCoCite	Brain	LungCancer	Computers	Photos
GCN	87.2	65.5	81.4	62.5	59.6	75.6	29.5
GAT	86.4	64.5	83.0	62.5	59.6	74.2	43.4
GraphSAGE	87.1	65.7	83.2	61.8	61.5	75.0	36.6
HyperConv	80.8	65.4	79.1	62.5	59.3	84.2	33.7
ED-HNN	86.2	68.1	80.3	66.3	60.2	97.3	78.6
GCN+DropNode	86.2	**67.9**	85.5	63.3	58.2	91.8	71.1
GCN+DropEdge	86.6	67.7	86.3	63.2	60.7	92.2	78.6
GCN+Mixup	85.8	67.6	85.6	65.0	59.0	87.7	78.0
GCN+NodeFeatureMasking	85.9	67.8	85.2	64.0	59.2	91.7	80.7
GCN+TopoAug (Ours)	**87.4**	**67.9**	**86.6**	**66.7**	**63.7**	**98.1**	**80.9**

Results

Node Classification Accuracy (%)

Method	Graph Statistics		Multi-Perspective			Multi-Modality	
	GitHub	TwitchDE	CoraCoCite	Brain	LungCancer	Computers	Photos
GCN	87.2	65.5	81.4	62.5	59.6	75.6	29.5
GAT	86.4	64.5	83.0	62.5	59.6	74.2	43.4
GraphSAGE	87.1	65.7	83.2	61.8	61.5	75.0	36.6
HyperConv	80.8	65.4	79.1	62.5	59.3	84.2	33.7
ED-HNN	86.2	68.1	80.3	66.3	60.2	97.3	78.6
GCN+DropNode	86.2	**67.9**	85.5	63.3	58.2	91.8	71.1
GCN+DropEdge	86.6	67.7	86.3	63.2	60.7	92.2	78.6
GCN+Mixup	85.8	67.6	85.6	65.0	59.0	87.7	78.0
GCN+NodeFeatureMasking	85.9	67.8	85.2	64.0	59.2	91.7	80.7
GCN+TopoAug (Ours)	**87.4**	**67.9**	**86.6**	**66.7**	**63.7**	**98.1**	**80.9**

Results

Node Classification Accuracy (%)

Method	Graph Statistics		Multi-Perspective			Multi-Modality	
	GitHub	TwitchDE	CoraCoCite	Brain	LungCancer	Computers	Photos
GCN	87.2	65.5	81.4	62.5	59.6	75.6	29.5
GAT	86.4	64.5	83.0	62.5	59.6	74.2	43.4
GraphSAGE	87.1	65.7	83.2	61.8	61.5	75.0	36.6
HyperConv	80.8	65.4	79.1	62.5	59.3	84.2	33.7
ED-HNN	86.2	68.1	80.3	66.3	60.2	97.3	78.6
GCN+DropNode	86.2	**67.9**	85.5	63.3	58.2	91.8	71.1
GCN+DropEdge	86.6	67.7	86.3	63.2	60.7	92.2	78.6
GCN+Mixup	85.8	67.6	85.6	65.0	59.0	87.7	78.0
GCN+NodeFeatureMasking	85.9	67.8	85.2	64.0	59.2	91.7	80.7
GCN+TopoAug (Ours)	**87.4**	**67.9**	**86.6**	**66.7**	**63.7**	**98.1**	**80.9**

Results

Node Classification Accuracy (%)

Method	Graph Statistics		Multi-Perspective			Multi-Modality	
	GitHub	TwitchDE	CoraCoCite	Brain	LungCancer	Computers	Photos
GCN	87.2	65.5	81.4	62.5	59.6	75.6	29.5
GAT	86.4	64.5	83.0	62.5	59.6	74.2	43.4
GraphSAGE	87.1	65.7	83.2	61.8	61.5	75.0	36.6
HyperConv	80.8	65.4	79.1	62.5	59.3	84.2	33.7
ED-HNN	86.2	68.1	80.3	66.3	60.2	97.3	78.6
GraphSAGE+DropNode	86.2	**68.3**	86.4	63.3	64.5	95.9	69.8
GraphSAGE+DropEdge	86.8	67.8	87.1	64.2	64.7	95.9	80.5
GraphSAGE+Mixup	85.9	67.1	**87.2**	64.2	63.4	92.2	71.5
GraphSAGE+NodeFeatureMasking	86.1	68.1	87.1	64.4	64.9	95.7	79.3
GraphSAGE+TopoAug (Ours)	**87.3**	**68.3**	**87.2**	**66.6**	**66.4**	**98.2**	**80.9**

Results
Summary

TopoAug consistently outperforms vanilla GNNs and other graph augmentation methods

- No preference in backbone GNN: more expressive GNN + TopoAug → better results
- More effective on larger datasets
- More effective when hyperedges are constructed using different information than simple edges
  → Build hyperedges from a different data perspective or modality

Imperial College London

Paper

Code

Datasets

Thank you

Enhancing Node Representations for Real-World Complex Networks with Topological Augmentation

x.zhao22@imperial.ac.uk

23/10/2024