Multi-label Node Classification Task On Graph-structured Data

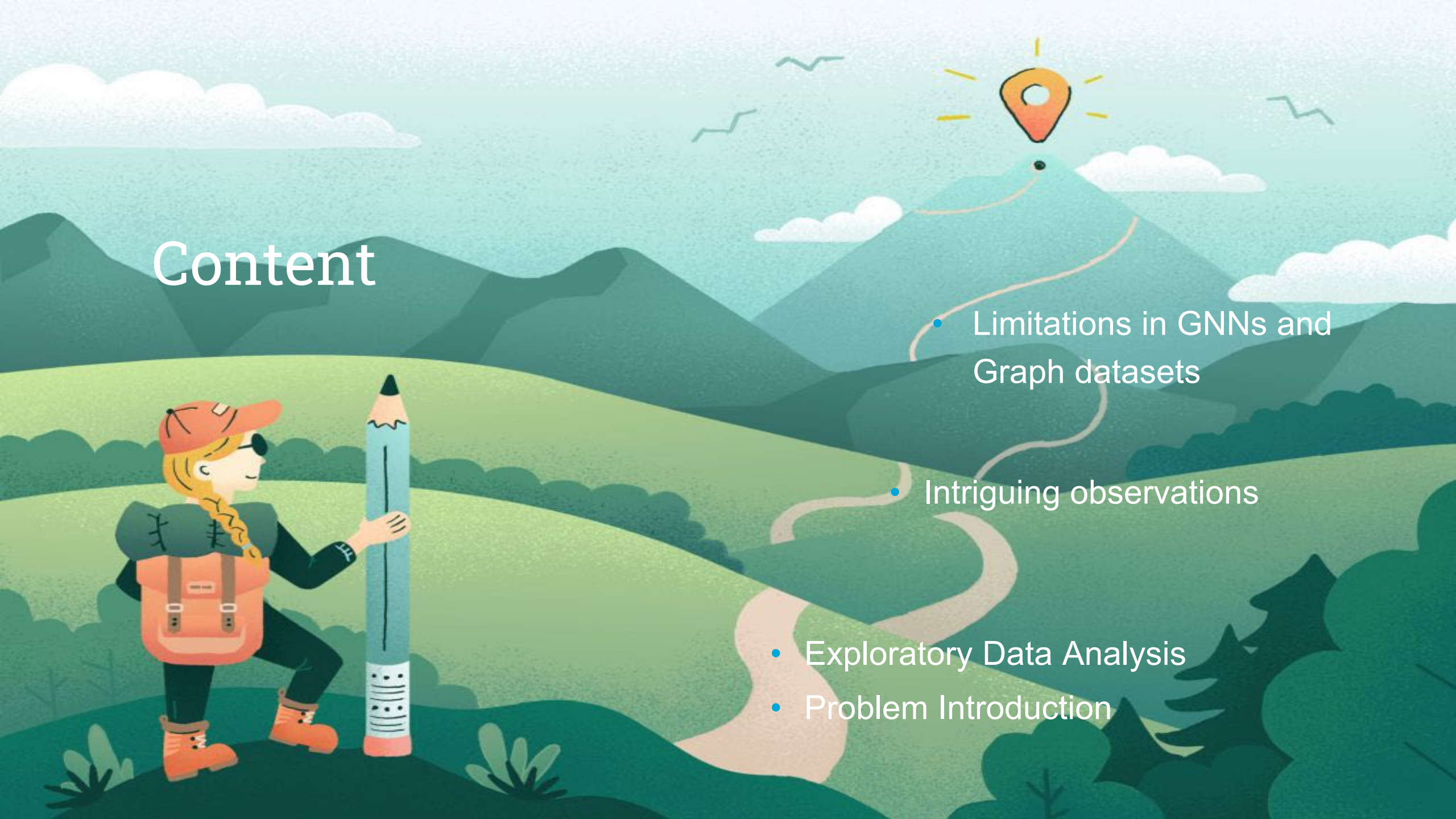
Tianqi Zhao

09.11.2023

https://openreview.net/forum?id=EZhkV2BjDP

TU Delft

Content

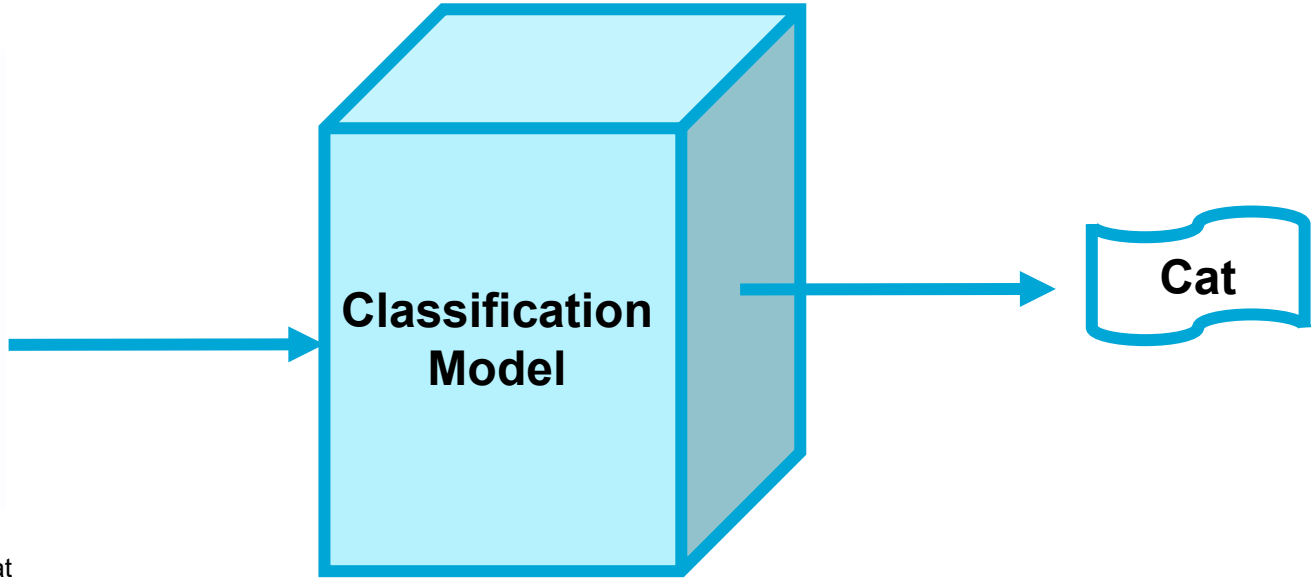


- Limitations in GNNs and Graph datasets
- Intriguing observations
- Exploratory Data Analysis
- Problem Introduction

Classification Task

- Input: description of the data
- Output: label

Image source: https://stock.adobe.com/nl/search?k=cat



TU Delft

A More Realistic Scenario: Multi-label

- Cat
- Dog

Illustration of multi-label classification. One input Image contains multiple labels (cat and dog).

Image source: https://www.pixelstalk.net/cute-dog-and-cat-wallpaper/=0

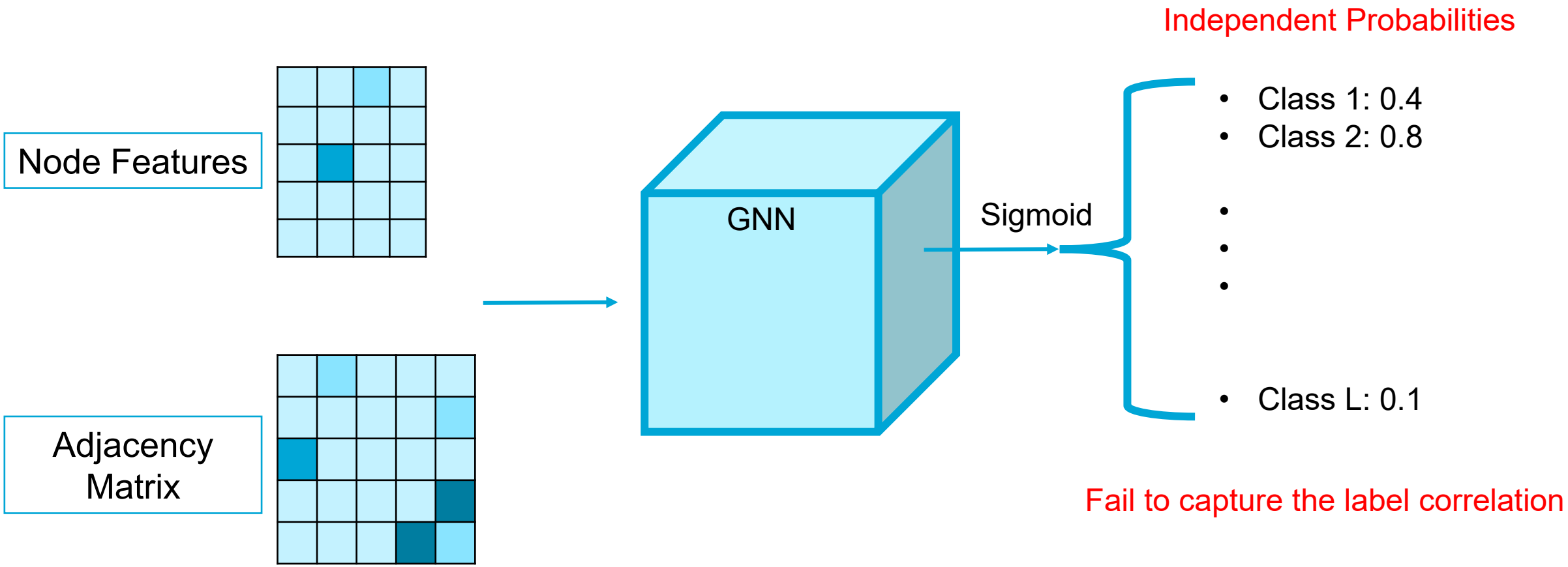
Multi-label Node Classification On Graphs

- **Graph:** Datapoints that are connected by the topological structure
- **Node classification task:**
  - **Given:** a graph, incomplete labelling
  - **Task:** predict the classes of the unlabeled nodes

Illustration of multi-label nodes in graphs

Image Source: https://minutes.co/5-ways-to-improve-networking-skills-from-professionals-who-mastered-it/

GNNs Deployed For Multi-label Node Classification Task



Performance of GNNs

- **Homophily in Single-label graphs:**
  - High homophily, GNNs show high performance
  - Most of the connected nodes share the same label
- **Homophily in Multi-label graphs:**
  - **Fraction of shared labels**
  - **Would GNNs work well on these datasets?**

Dataset	Label Homophily
Blogcatalog	0.10
Yelp	0.22
Ogbn-proteins	0.15

TU Delft

Would GNNs designed for single-label low homophily graphs work on multi-label graphs?

- **Low homophily: Opposites attract**

GNN Designed For Single Label Low Homophily Graphs: H2GCN

- Assumption: useful information in higher neighbourhoods

Method	BlogCat	Yelp	OGB-Proteins	DBLP
MLP	0.043	0.096	0.026	0.350
DeepWalk	**0.190**	0.096	0.044	0.585
LANC	<u>0.050</u>	OOM	<u>0.045</u>	0.836
Gcn	0.037	0.131	**0.054**	**0.893**
Gat	0.041	0.150	0.021	0.829
GraphSage	0.045	**0.251**	0.027	<u>0.868</u>
H2Gcn	0.039	<u>0.226</u>	0.036	0.858
GCN-LPA	0.043	0.116	0.023	0.801

TU Delft

Different Semantics Of Homophily

- Low homophily: each neighbour only shares a small fraction of labels
- But all labels can be found among the neighbouring nodes

Real-life Scenario

- Read
- Travel
- Computer Science

- Each friend shares part of one's interests: low homophily
- All the interests are in the direct neighbourhood

TU Delft

Characteristics Of Ogbn-proteins

OGB Get Started Updates Large-Scale Challenge Datasets Leaderboards Papers Team Github

Leaderboard for ogbn-proteins

The ROC-AUC score on the test and validation sets. The higher, the better.

Package: >=1.1.1

Rank	Method	Ext. data	Test ROC-AUC	Validation ROC-AUC	Contact	References	#Params	Hardware	Date
1	**LD+GAT**	Yes	0.8942 ± 0.0007	0.9527 ± 0.0007	Zhihao Shi (MIRA Lab, USTC & CityBrain Lab, Alibaba Cloud)	Paper, Code	664,233,700	GeForce RTX 3090 (24GB GPU)	Sep 27, 2023
2	**GIPA(Wide&Deep)**	No	0.8917 ± 0.0007	0.9472 ± 0.0020	Houyi Li	Paper, Code	17,438,716	Tesla V100-SXM2(32G)	Jan 19, 2023
3	**AGDN**	No	0.8865 ± 0.0013	0.9418 ± 0.0005	Chuxiong Sun	Paper, Code	8,605,486	Tesla V100 (16GB GPU)	Sep 2, 2022
4	**RevGNN-Wide**	No	0.8824 ± 0.0015	0.9450 ± 0.0008	Guohao Li - DeepGCNs.org	Paper, Code	68,471,608	NVIDIA RTX 6000 (48G)	Jun 16, 2021
5	**GAT+BOT+NGNN**	No	0.8809 ± 0.0016	0.9375 ± 0.0019	Xiang song (DGL team)	Paper, Code	11,740,552	Tesla V100 (32GB)	Jan 23, 2022
6	**RevGNN-Deep**	No	0.8774 ± 0.0013	0.9326 ± 0.0006	Guohao Li - DeepGCNs.org	Paper, Code	20,031,384	NVIDIA RTX 6000 (48G)	Jun 16, 2021
7	**GAT+BoT**	No	0.8765 ± 0.0008	0.9280 ± 0.0008	Yangkun Wang (DGL Team)	Paper, Code	2,484,192	Tesla A100 (40GB GPU)	Jun 16,

Image source: https://ogb.stanford.edu/docs/leader_nodeprop/

Characteristics Of Ogbn-proteins

- Huge number of **unlabelled nodes**
  - in Ogbn-proteins, over **40% nodes without labels**
  - **89% test nodes without labels**
- Evaluation
  - Sparse label matrix, Long training epochs
  - encourages model to assign negative class and achieve high AUROC

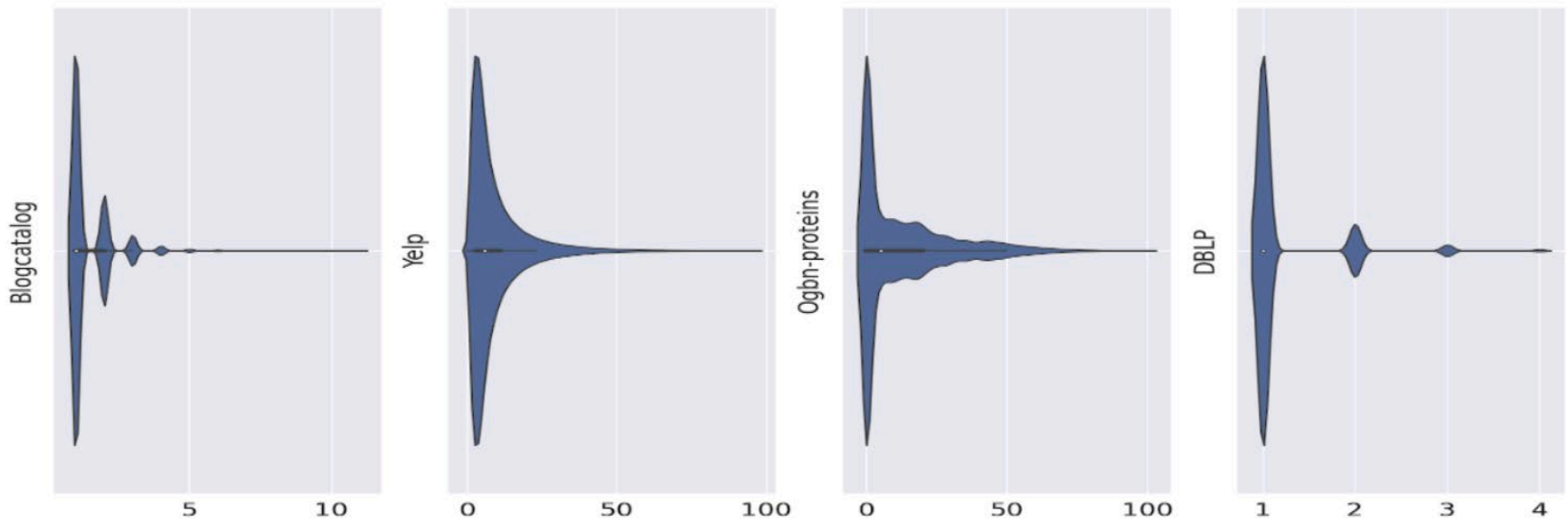
	Model	MicroF1	MacroF1	AUROC	AP
OGB-PROTEINS	MLP	2.55	2.40	54.05	2.59
	DEEPWALK	**2.88**	**2.75**	<u>68.75</u>	4.41
	LANC	2.35	2.21	68.03	<u>4.48</u>
	GCN	<u>2.77</u>	<u>2.63</u>	**71.48**	**5.36**
	GAT	2.55	2.40	50.64	2.14
	GRAPHSAGE	2.59	2.43	55.83	2.68
	H2GCN	2.55	2.39	62.75	3.61
	GCN-LPA	2.56	2.41	53.22	2.33

TU Delft

Characteristics Of Available Multi-label Graphs

- Label assignment
  - Most of the nodes have one label
    - in Blogcatalog, 72% nodes have one label
    - Nodes: bloggers
    - Labels: social groups
    - Edges: friendships

Is this true?



Multi-label graph generator model

- Varying feature quality and homophily

Method	r_{ori_feat} → high					r_{homo} → high				
	0.0	0.2	0.5	0.8	1.0	0.2	0.4	0.6	0.8	1.0
MLP	0.172	0.187	0.220	0.277	0.343	**0.343**	0.343	0.343	0.343	0.343
DEEPWALK	**0.487**	**0.487**	**0.487**	**0.487**	**0.487**	0.181	**0.522**	**0.813**	**0.869**	0.552
LANC	0.337	0.342	0.365	0.353	0.391	0.190	0.380	0.434	0.481	0.629
GCN	0.313	0.316	0.311	0.301	0.337	0.261	0.343	0.388	0.450	0.493
GAT	0.311	0.339	0.329	0.338	0.360	0.172	0.359	0.390	0.428	0.439
GRAPHSAGE	0.300	0.328	0.377	0.393	0.430	0.289	0.426	0.458	0.533	0.553
H2GCN	0.376	0.401	0.427	0.442	0.467	0.297	0.484	0.512	0.572	**0.652**
GCN-LPA	0.337	0.333	0.368	0.363	0.391	0.170	0.408	0.495	0.604	0.583

TU Delft

Conclusions

- Multi-label Node Classification Task is still an open research field
- Check the datasets before the experiments
- Choose the right model on the datasets
- Use the right evaluation metric

Paper: Zhao et al. "Multi-label Node Classification On Graph-structured Data", TMLR 2023. https://openreview.net/forum?id=EZhkV2BjDP

TU Delft