DegUIL: Degree-aware Graph Neural Networks for Long-tailed User Identity Linkage

Meixiu Long, Siyuan Chen, Xin Du, and Jiahai Wang

Sun Yat-sen University, Guangzhou, China

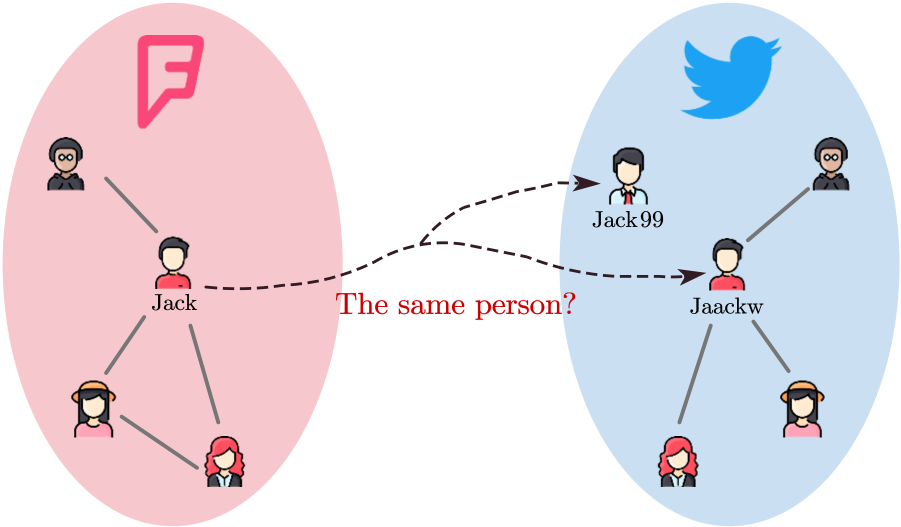
Outline

- **Background**
- **Problem & related work**
- **Challenge & insight**
- **Proposed model: DegUIL**
- **Experiments**
- **Conclusions**

Background

SUN YAT-SEN UNIVERSITY

- User identity linkage (UIL)
  - Link identities belonging to the same natural person across distinct social networks

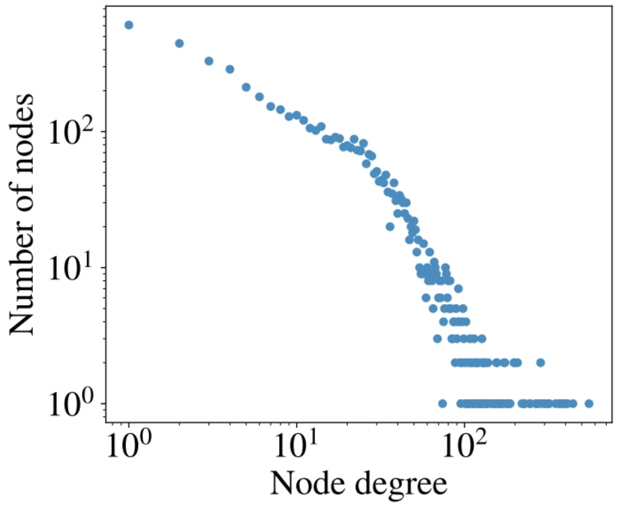


- Application
  - A data fusion and mining task
  - Cross-platform recommendation, etc

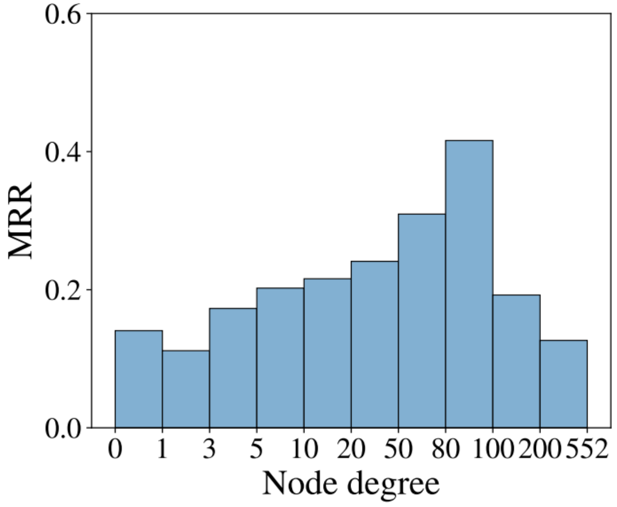
- UIL Methods
  - Mainly **structure-based** methods, encoded by graph neural networks (GNNs)

Problem: long-tailed distribution

whether social networks provide reliable and adequate information?



(a) Long-tailed node distribution

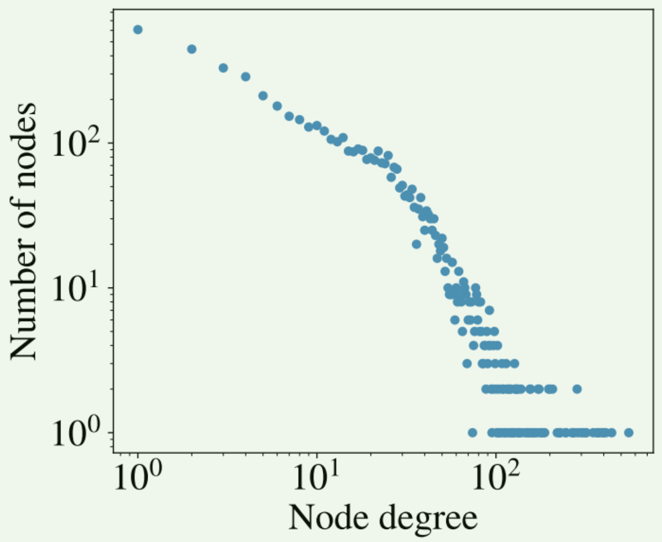


(b) MRR w.r.t degrees of test nodes

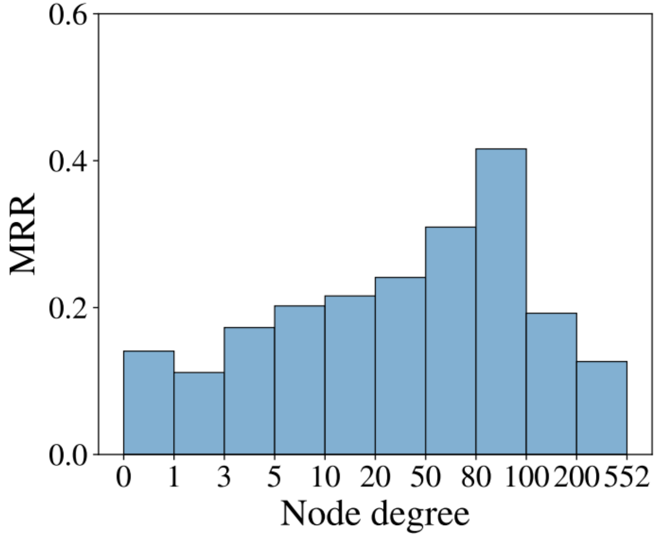
Problems

- An inherent structural gap exists among nodes
- The limited neighborhoods of tail nodes hinder the linkage performance
- Noise hidden in super head nodes exacerbates the quality of representation

Problem: long-tailed distribution



(a) Long-tailed node distribution

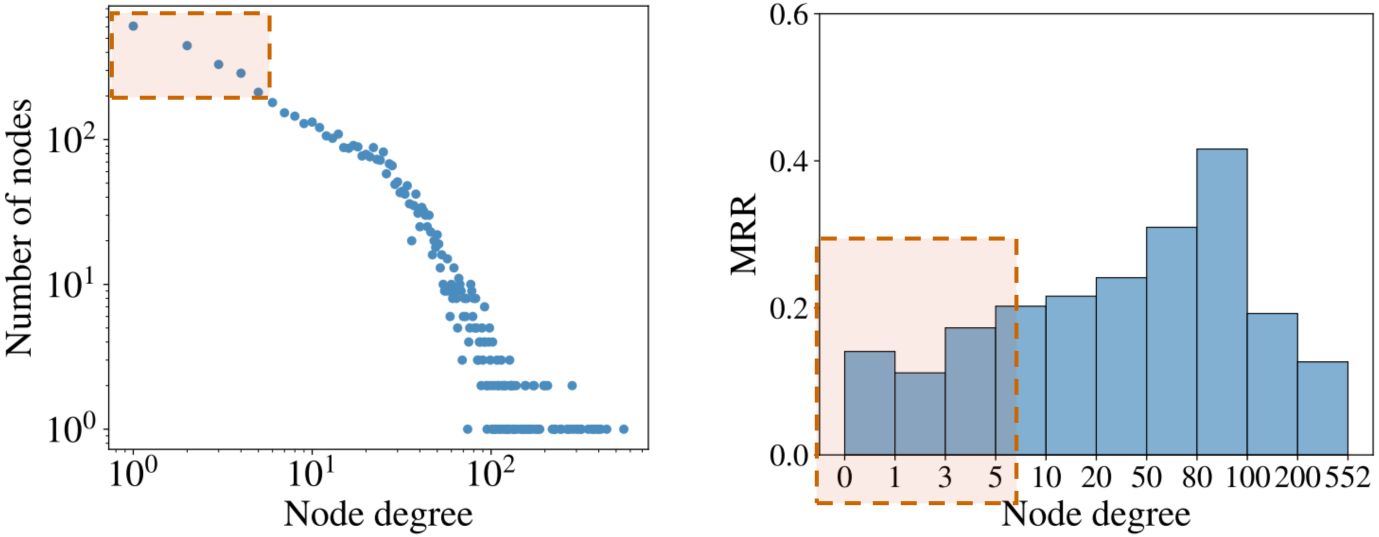


(b) MRR w.r.t degrees of test nodes

Problems

- An inherent structural gap exists among nodes
- The limited neighborhoods of tail nodes hinder the linkage performance
- Noise hidden in super head nodes exacerbates the quality of representation

Problem: long-tailed distribution



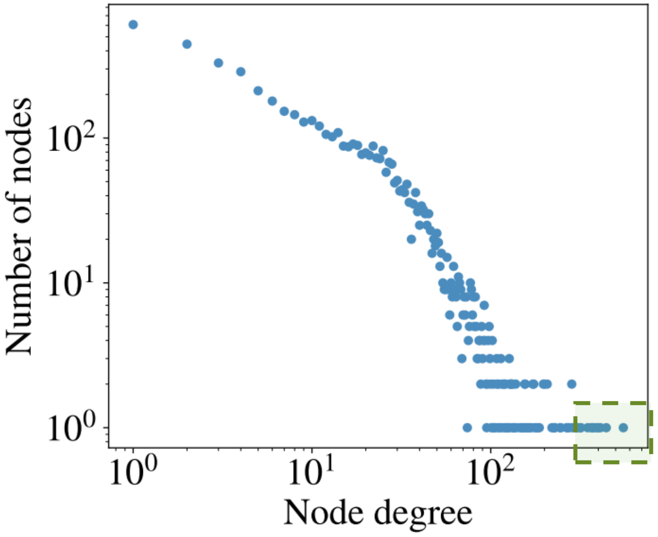
(a) Long-tailed node distribution (b) MRR w.r.t degrees of test nodes

Problems

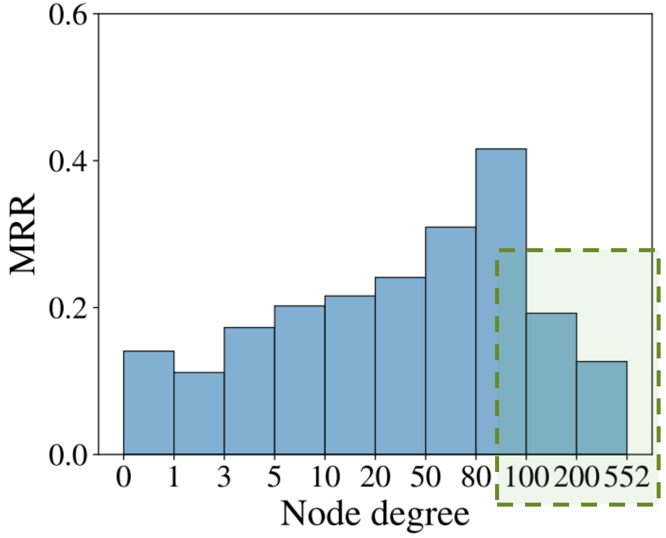
- An inherent structural gap exists among nodes
- The limited neighborhoods of **tail nodes** hinder the linkage performance
- Noise hidden in super head nodes exacerbates the quality of representation

Problem: long-tailed distribution

SUN YAT-SEN UNIVERSITY



(a) Long-tailed node distribution



(b) MRR w.r.t degrees of test nodes

Problems

- An inherent structural gap exists among nodes
- The limited neighborhoods of tail nodes hinder the linkage performance
- Noise hidden in **super head nodes** exacerbates the quality of representation

Related work

SUN YAT-SEN UNIVERSITY

- Degree-related UIL methods
  - SEA [1], learning embeddings
  - DAT [2], additional entity names

- Other long-tailed problems
  - Node degree long-tailed graphs [3][4]
  - Recommendation……

[1] 2019, WWW. Semi-supervised entity alignment via knowledge graph embedding with awareness of degree difference.
[2] 2020, SIGIR. Degree-aware alignment for entities in tail.
[3] 2020, CIKM. Towards locality-aware meta-learning of tail node embeddings on networks.
[4] 2021, KDD. Tail-GNN: Tail-node graph neural networks.

Challenge

- Problems

  Structural gap, limited neighborhoods, noise-filled graphs

- Goal

  How can we effectively **link identities for socially-inactive users in a noisy graph?**

- Challenges

  - C1: Tail nodes have no additional information but few neighbors
  - C2: How can noise be eliminated while preserving the intrinsic graph structure
  - C3: Each node owns both a unique locality and a generality

Insights: DegUIL

- Key idea

  Exploit the ideal neighborhood knowledge of head nodes to **correct structural bias** for meaningful aggregation in GNNs

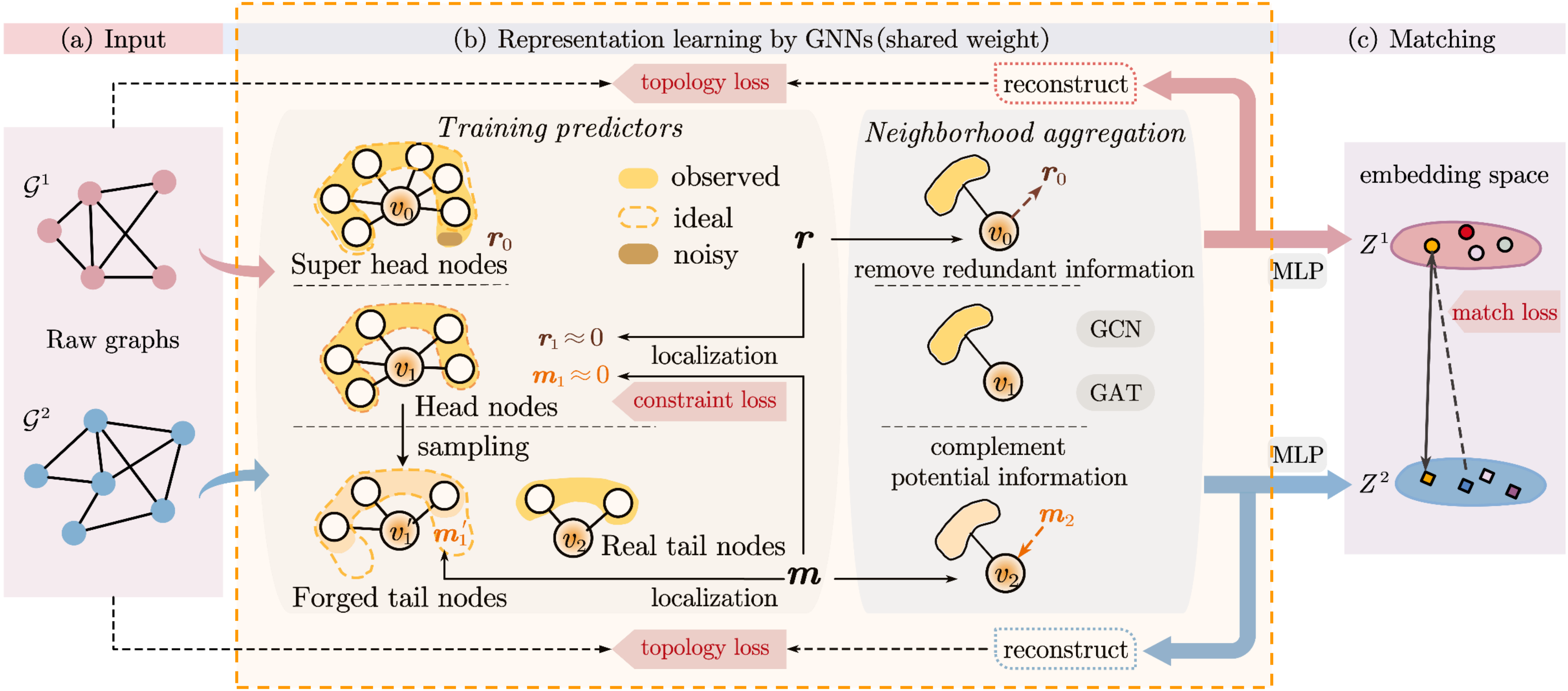
- First & Second Challenges

  - Train two modules to **enrich tail nodes and refine super head nodes** in embeddings

- Third Challenge

  - **Shared vectors** across the graph, adapt to the local context of each node

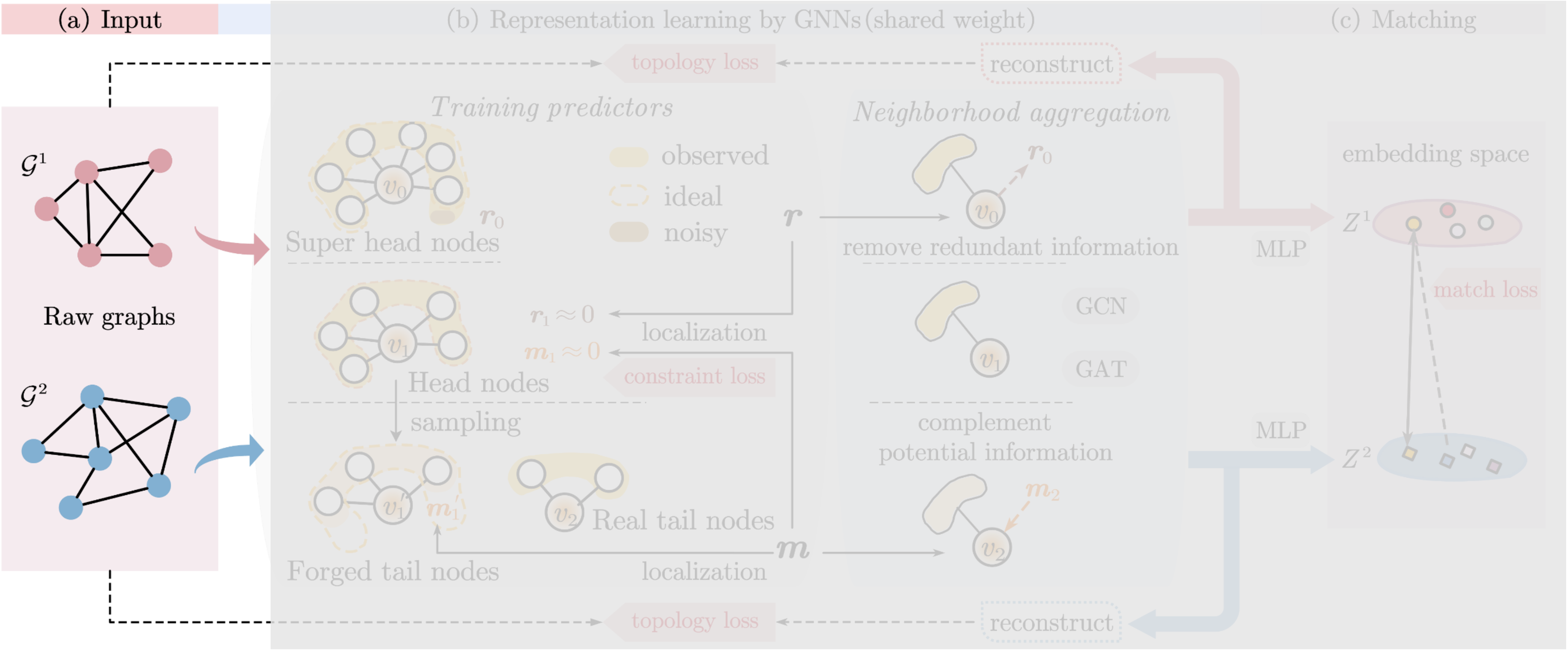
DegUIL: overall framework



Our work

SUN YAT-SEN UNIVERSITY

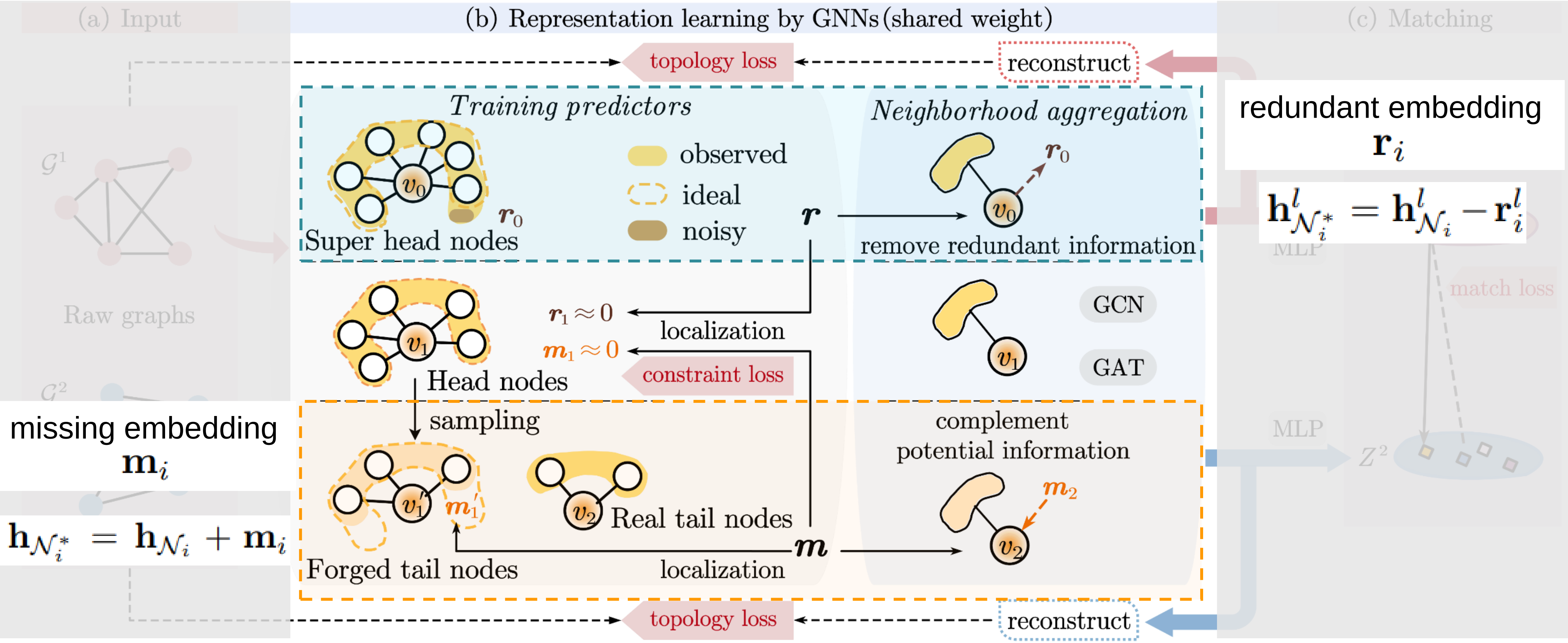
(a) Generate initial features



DegUIL: overall framework

SUN YAT-SEN UNIVERSITY

(b) correct observed neighborhood to be ideal



redundant embedding $\mathbf{r}_i$

$$\mathbf{h}^l_{\mathcal{N}^*_i} = \mathbf{h}^l_{\mathcal{N}_i} - \mathbf{r}^l_i$$

missing embedding $\mathbf{m}_i$

$$\mathbf{h}_{\mathcal{N}^*_i} = \mathbf{h}_{\mathcal{N}_i} + \mathbf{m}_i$$

Aggregation $\mathbf{h}^{l+1}_i = \mathrm{Agg}\left(\mathbf{h}^l_i, \{\mathbf{h}^l_k : v_k \in \mathcal{N}_i\} \cup \{I(v_i \in \mathcal{V}_{\mathrm{tail}})\, \mathbf{m}^l_i - I(v_i \in \mathcal{V}_{\mathrm{super}})\, \mathbf{r}^l_i\} ; \theta^{l+1}\right)$

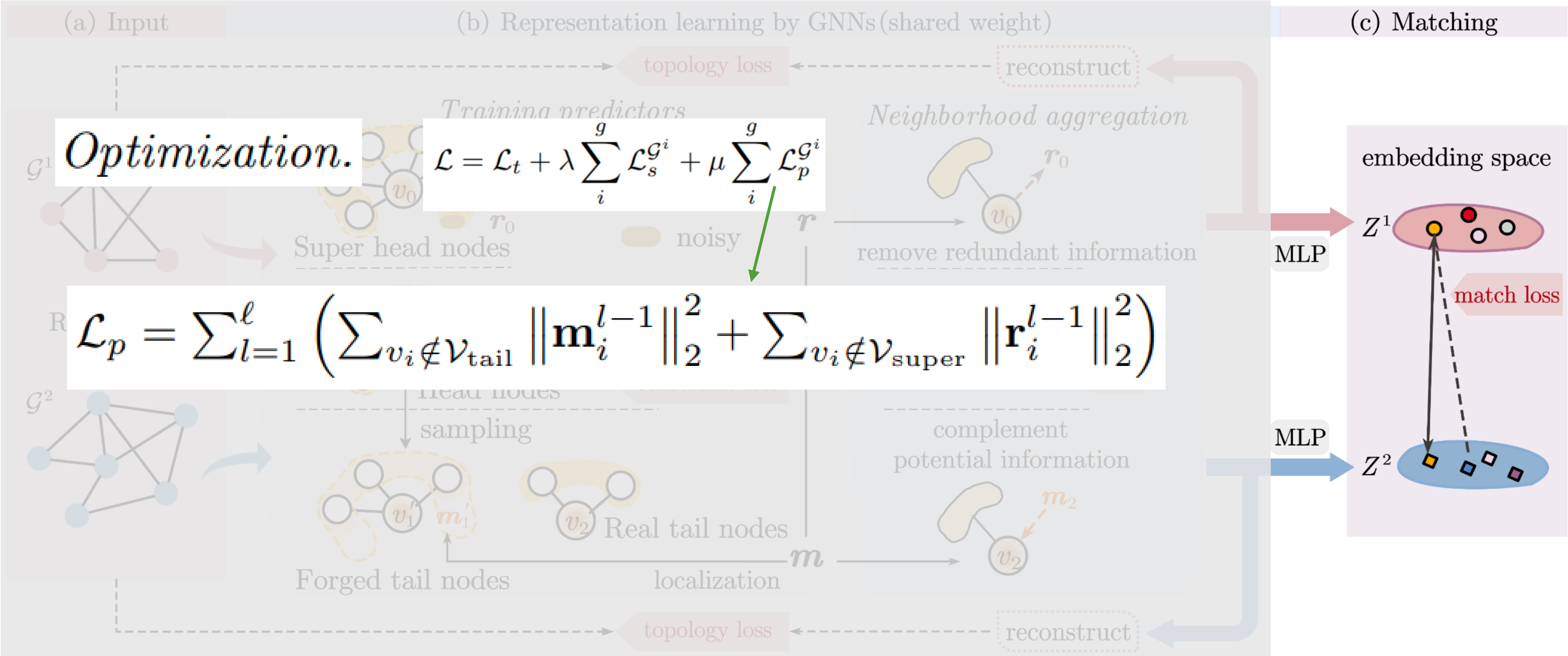
DegUIL: overall framework

(c) Matching identities

Optimization.

$$\mathcal{L} = \mathcal{L}_t + \lambda \sum_i^g \mathcal{L}_s^{\mathcal{G}^i} + \mu \sum_i^g \mathcal{L}_p^{\mathcal{G}^i}$$

$$\mathcal{L}_p = \sum_{l=1}^{\ell} \left(\sum_{v_i \notin \mathcal{V}_{\text{tail}}} \left\| \mathbf{m}_i^{l-1} \right\|_2^2 + \sum_{v_i \notin \mathcal{V}_{\text{super}}} \left\| \mathbf{r}_i^{l-1} \right\|_2^2 \right)$$



Experimental setting

Datasets

Table 1: Dataset statistics.

Networks	#Nodes	#Edges	#Anchor links	#Tail links
Foursquare	5313	76972	1609	443
Twitter	5120	164919		
DBLP17	9086	51700	2832	975
DBLP19	9325	47775		

Setup

- Tail nodes: degree ≤ 5
- Super head nodes: top-10% highest degree

Baselines

- ➢ conventional representation learning
  - node2vec [1]
- ➢ UIL approaches
  - PALE [2], NeXtAlign [3], SEA [4]
- ➢ Degree-related embedding approaches
  - Tail-GNN [5]

[1] 2016, KDD. node2vec: Scalable feature learning for networks.
[2] 2016, IJCAI. Predict anchor links across social networks via an embedding approach.
[3] 2019, WWW. Semi-supervised entity alignment via knowledge graph embedding with awareness of degree difference.
[4] 2021, KDD. Balancing consistency and disparity in network alignment.
[5] 2021, KDD. Tail-GNN: Tail-node graph neural networks.

Experimental results

Table 2: Overall performance. Best result appears in bold and the second best model is underlined except for ablation variants.

Dataset	Foursquare-Twitter				DBLP17-DBLP19			
Metric	Hits@1	Hits@10	Hits@30	MRR	Hits@1	Hits@10	Hits@30	MRR
node2vec	5.43	15.08	25.49	10.93	33.18	55.10	66.52	44.17
PALE	6.00	15.77	26.48	11.51	21.28	39.78	52.04	30.94
SEA	<u>6.93</u>	15.89	23.94	11.80	**38.62**	<u>60.13</u>	<u>71.01</u>	**49.27**
NeXtAlign	6.47	12.23	16.62	9.63	36.82	59.58	70.46	48.06
Tail-GNN	6.70	<u>17.67</u>	<u>28.39</u>	<u>12.66</u>	36.36	56.58	67.21	46.44
DegUIL	**9.33**	**21.70**	**32.81**	**16.00**	<u>37.59</u>	**60.73**	**71.51**	<u>48.96</u>
DegUIL_{w/o_AP}	8.11	19.39	30.39	14.30	36.26	59.29	70.32	47.67
DegUIL_{w/o_NR}	8.94	20.53	31.79	15.21	37.13	59.61	70.02	48.26

Observations

- DegUIL consistently outperforms other baselines.
- Degree-aware models perform better than traditional methods.
- DegUIL has a greater advantage in complex long-tailed datasets.

Experimental results

Table 2: Overall performance. Best result appears in bold and the second best model is underlined except for ablation variants.

Dataset	Foursquare-Twitter				DBLP17-DBLP19			
Metric	Hits@1	Hits@10	Hits@30	MRR	Hits@1	Hits@10	Hits@30	MRR
node2vec	5.43	15.08	25.49	10.93	33.18	55.10	66.52	44.17
PALE	6.00	15.77	26.48	11.51	21.28	39.78	52.04	30.94
SEA	6.93	15.89	23.94	11.80	**38.62**	60.13	71.01	**49.27**
NeXtAlign	6.47	12.23	16.62	9.63	36.82	59.58	70.46	48.06
Tail-GNN	6.70	17.67	28.39	12.66	36.36	56.58	67.21	46.44
DegUIL	**9.33**	**21.70**	**32.81**	**16.00**	37.59	**60.73**	**71.51**	48.96
DegUIL$_{w/o_AP}$	8.11	19.39	30.39	14.30	36.26	59.29	70.32	47.67
DegUIL$_{w/o_NR}$	8.94	20.53	31.79	15.21	37.13	59.61	70.02	48.26

Observations

- DegUIL consistently outperforms other baselines.
- Degree-aware models perform better than traditional methods.
- DegUIL has a greater advantage in complex long-tailed datasets.

Experimental results

Table 2: Overall performance. Best result appears in bold and the second best model is underlined except for ablation variants.

Dataset	Foursquare-Twitter				DBLP17-DBLP19			
Metric	Hits@1	Hits@10	Hits@30	MRR	Hits@1	Hits@10	Hits@30	MRR
node2vec	5.43	15.08	25.49	10.93	33.18	55.10	66.52	44.17
PALE	6.00	15.77	26.48	11.51	21.28	39.78	52.04	30.94
SEA	<u>6.93</u>	15.89	23.94	11.80	**38.62**	<u>60.13</u>	<u>71.01</u>	**49.27**
NeXtAlign	6.47	12.23	16.62	9.63	36.82	59.58	70.46	48.06
Tail-GNN	6.70	<u>17.67</u>	<u>28.39</u>	<u>12.66</u>	36.36	56.58	67.21	46.44
DegUIL	**9.33**	**21.70**	**32.81**	**16.00**	<u>37.59</u>	**60.73**	**71.51**	<u>48.96</u>
DegUIL$_{w/o_AP}$	8.11	19.39	30.39	14.30	36.26	59.29	70.32	47.67
DegUIL$_{w/o_NR}$	8.94	20.53	31.79	15.21	37.13	59.61	70.02	48.26

Observations

- DegUIL consistently outperforms other baselines.
- Degree-aware models perform better than traditional methods.
- DegUIL has a greater advantage in complex long-tailed datasets.

Experimental results

SUN YAT-SEN UNIVERSITY

Table 2: Overall performance. Best result appears in bold and the second best model is underlined except for ablation variants.

Dataset	Foursquare-Twitter				DBLP17-DBLP19			
Metric	Hits@1	Hits@10	Hits@30	MRR	Hits@1	Hits@10	Hits@30	MRR
node2vec	5.43	15.08	25.49	10.93	33.18	55.10	66.52	44.17
PALE	6.00	15.77	26.48	11.51	21.28	39.78	52.04	30.94
SEA	<u>6.93</u>	15.89	23.94	11.80	**38.62**	<u>60.13</u>	<u>71.01</u>	**49.27**
NeXtAlign	6.47	12.23	16.62	9.63	36.82	59.58	70.46	48.06
Tail-GNN	6.70	<u>17.67</u>	<u>28.39</u>	<u>12.66</u>	36.36	56.58	67.21	46.44
DegUIL	**9.33**	**21.70**	**32.81**	**16.00**	<u>37.59</u>	**60.73**	**71.51**	<u>48.96</u>
DegUIL$_{w/o_AP}$	8.11	19.39	30.39	14.30	36.26	59.29	70.32	47.67
DegUIL$_{w/o_NR}$	8.94	20.53	31.79	15.21	37.13	59.61	70.02	48.26

Observations

- DegUIL consistently outperforms other baselines.
- Degree-aware models perform better than traditional methods.
- DegUIL has a greater advantage in complex long-tailed datasets.

Ablation study

Table 2: Overall performance. Best result appears in bold and the second best model is underlined except for ablation variants.

Dataset	Foursquare-Twitter				DBLP17-DBLP19			
Metric	Hits@1	Hits@10	Hits@30	MRR	Hits@1	Hits@10	Hits@30	MRR
node2vec	5.43	15.08	25.49	10.93	33.18	55.10	66.52	44.17
PALE	6.00	15.77	26.48	11.51	21.28	39.78	52.04	30.94
SEA	<u>6.93</u>	15.89	23.94	11.80	**38.62**	<u>60.13</u>	<u>71.01</u>	**49.27**
NeXtAlign	6.47	12.23	16.62	9.63	36.82	59.58	70.46	48.06
Tail-GNN	6.70	<u>17.67</u>	<u>28.39</u>	<u>12.66</u>	36.36	56.58	67.21	46.44
DegUIL	**9.33**	**21.70**	**32.81**	**16.00**	<u>37.59</u>	**60.73**	**71.51**	<u>48.96</u>
DegUIL$_{w/o_AP}$	8.11	19.39	30.39	14.30	36.26	59.29	70.32	47.67
DegUIL$_{w/o_NR}$	8.94	20.53	31.79	15.21	37.13	59.61	70.02	48.26

- No absent neighborhood predictor (w/o AP): impairs the performance
- No noisy neighborhood remover (w/o NR): hurts the performance
- the gain of AP is more significant than that of NR.

Evaluation by degree

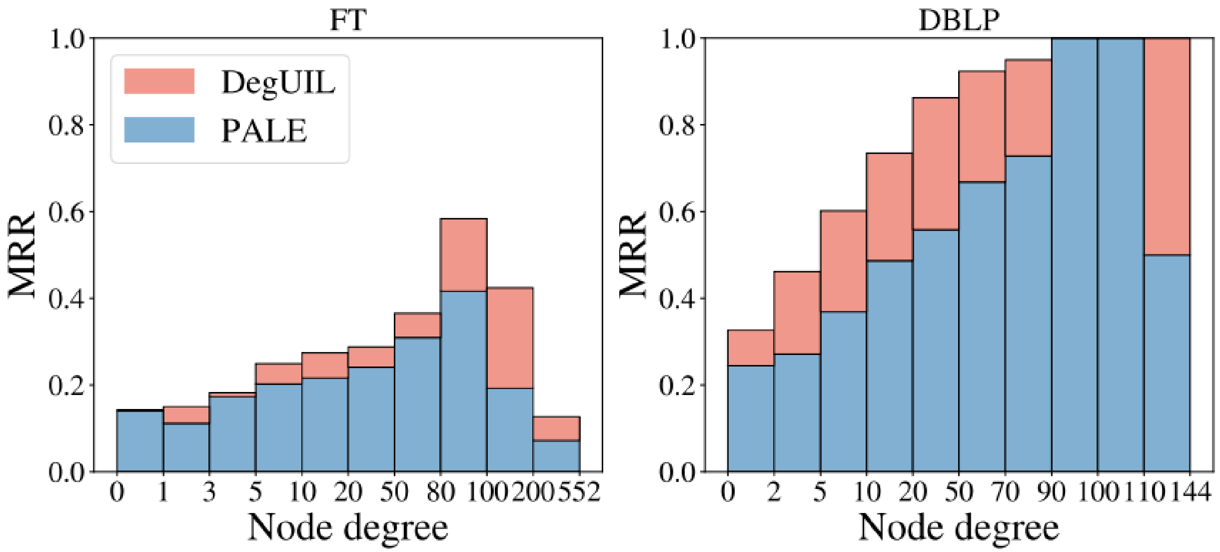


Fig. 4: MRR results by degrees.

- Low-degree nodes and super high-degree nodes perform worse than those normal nodes
- DegUIL outperforms PALE across all degree groups, validating its effectiveness in handling long-tail issues.

Conclusions

- Problem
  - performance bottlenecks: tail nodes, super head nodes
  - Long-tailed UIL with GNNs
- Algorithm: DegUIL
  - Degree-aware model
  - Training two modules to correct the neighborhood bias without additional attributes
  - learning high-quality node embeddings for tail nodes' alignment
- Evaluations
  - significant advantages in dealing with complex networks

Thanks!