Relation-Aware Diffusion for Heterogeneous Graphs with Partially Observed Features

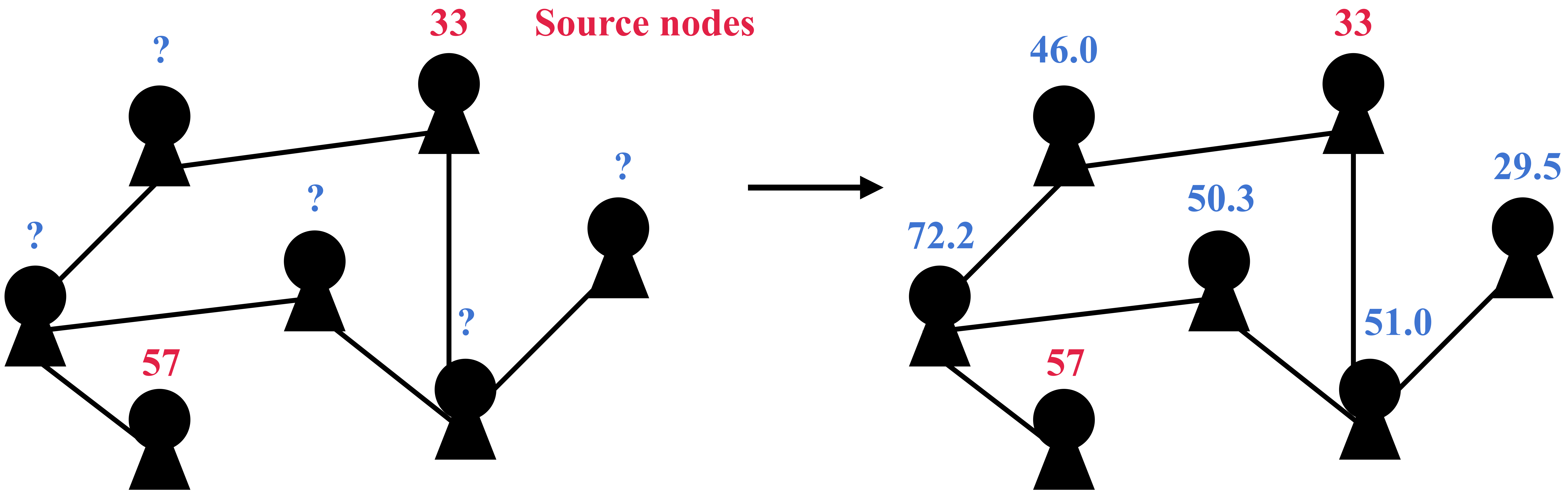
Daeho Um[1], Yoonji Lee[2], Jiwoong Park[3], Seuli Park[4], Yuneil Yeo[5], Seong Jin Ahn[6]

[1]AI Center, Samsung Electronics [2]Samsung Electronics, [3]Texas A&M University,
[4]University of Michigan, [5]UC Berkeley, [6]KAIST

Background

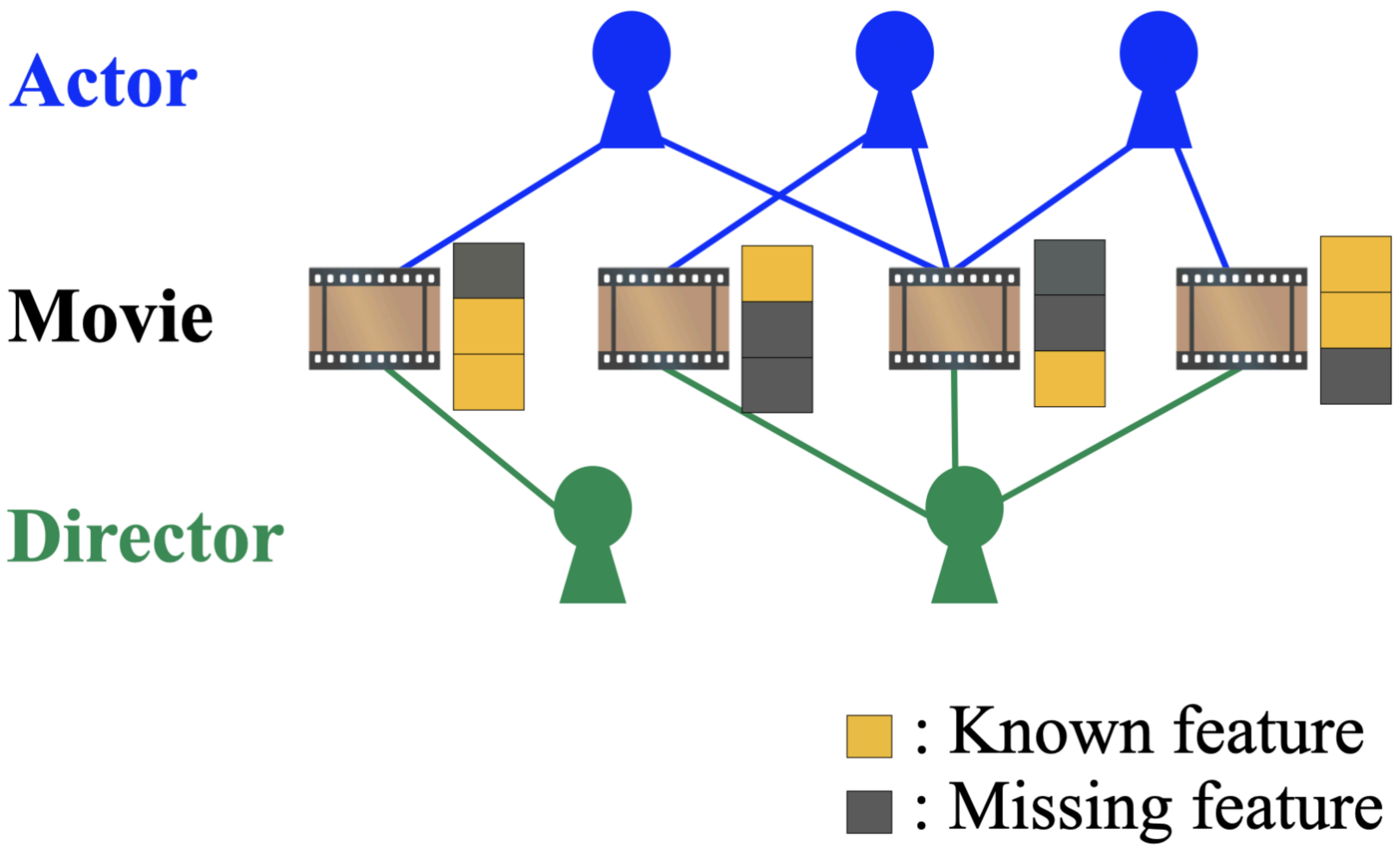
Diffusion-based feature imputation for graphs

- Diffusion-based feature imputation methods iteratively propagate known features along edges.
- At each step, each unknown feature updates its value by aggregating features from its neighboring nodes.

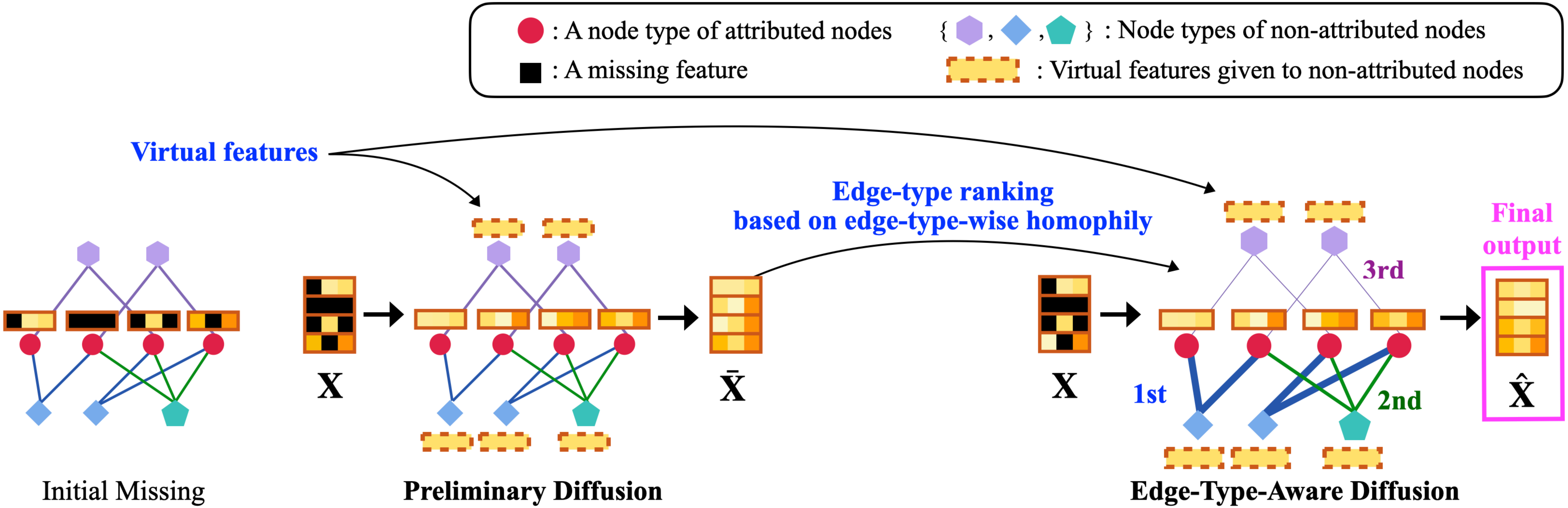


Motivation

- Heterogeneous graph
  - two challenges
  1) Node types without features block diffusion
  2) Multiple types of edges

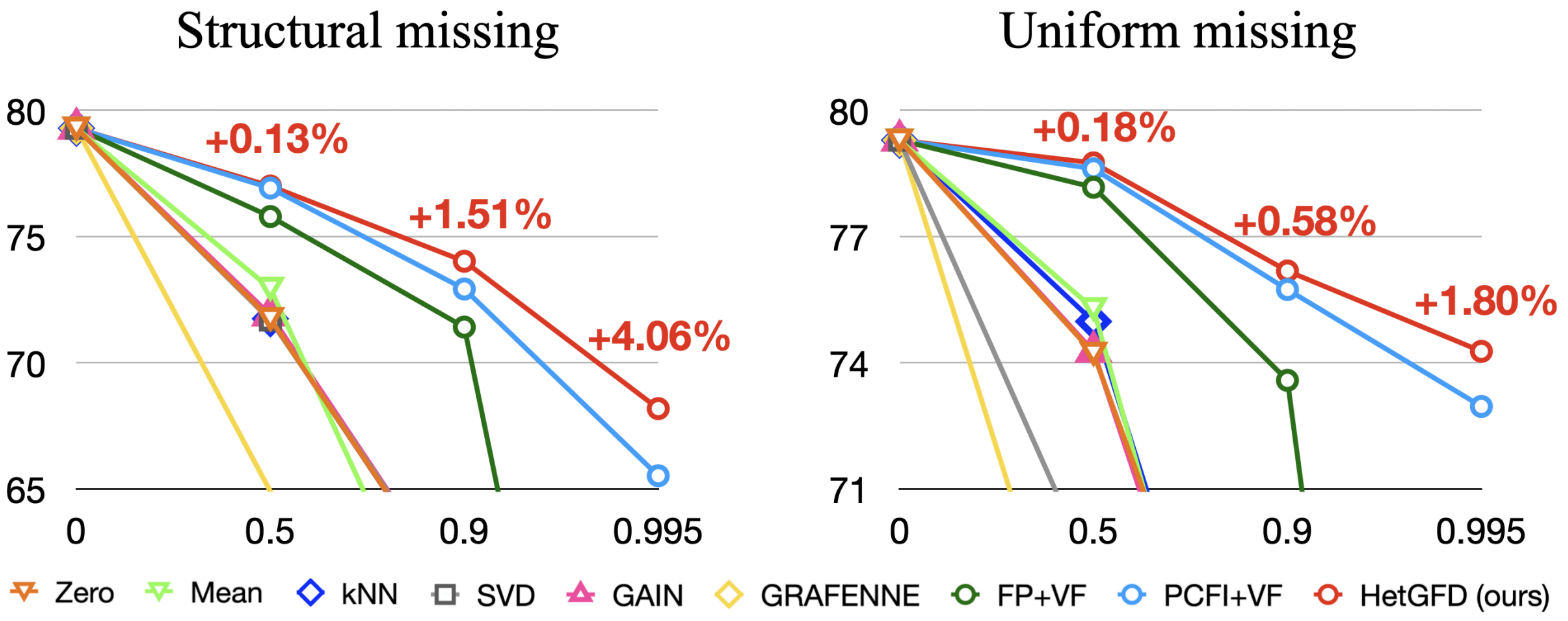


HetGFD
(Heterogeneous graph feature diffusion)



Experiments

Semi-supervised node classification



Experiments

Link prediction

Missing type	Method	ACM AUC	ACM AP	DBLP AUC	DBLP AP	IMDB AUC	IMDB AP
	Full features	76.25 ± 1.20	77.64 ± 1.07	71.52 ± 0.51	66.92 ± 0.66	92.47 ± 1.06	86.94 ± 1.61
Structural missing	Zero	71.65 ± 2.16	71.74 ± 3.55	72.49 ± 0.63	74.21 ± 0.60	92.48 ± 1.06	86.95 ± 1.60
	Mean	71.64 ± 1.30	71.66 ± 1.33	72.49 ± 0.63	74.20 ± 0.60	91.78 ± 1.13	85.80 ± 2.12
	kNN	72.04 ± 1.66	72.55 ± 2.11	71.96 ± 1.37	69.86 ± 1.89	91.10 ± 1.07	84.44 ± 1.97
	SVD	71.49 ± 1.77	72.29 ± 2.13	72.49 ± 0.63	74.21 ± 0.60	92.48 ± 1.06	86.95 ± 1.60
	GAIN	72.22 ± 1.19	73.21 ± 1.10	72.49 ± 0.63	74.20 ± 0.61	92.48 ± 1.06	86.95 ± 1.60
	GRAFENNE	74.87 ± 6.71	67.60 ± 5.87	90.14 ± 7.26	76.53 ± 7.12	82.38 ± 5.75	69.72 ± 4.60
	FP+VF	73.40 ± 0.75	74.03 ± 0.84	71.58 ± 0.85	70.01 ± 1.43	$\mathbf{92.50 \pm 1.04}$	$\mathbf{86.99 \pm 1.58}$
	PCFI+VF	73.41 ± 1.16	73.22 ± 1.18	71.37 ± 0.55	66.78 ± 0.74	91.71 ± 1.33	85.37 ± 2.08
	HetGFD (ours)	$\mathbf{78.25 \pm 1.34}$	$\mathbf{78.62 \pm 2.12}$	$\mathbf{91.94 \pm 0.67}$	$\mathbf{91.88 \pm 0.91}$	$\mathbf{92.50 \pm 1.04}$	$\mathbf{86.99 \pm 1.58}$
Uniform missing	Zero	70.69 ± 1.48	70.17 ± 3.07	72.48 ± 0.62	74.20 ± 0.60	$\mathbf{92.50 \pm 1.04}$	$\mathbf{86.99 \pm 1.58}$
	Mean	71.98 ± 1.02	72.02 ± 0.96	72.48 ± 0.62	74.20 ± 0.60	91.40 ± 1.14	85.33 ± 1.93
	kNN	71.02 ± 1.49	72.49 ± 2.46	72.72 ± 1.85	70.29 ± 3.73	91.15 ± 1.09	84.50 ± 2.04
	SVD	70.49 ± 2.11	70.70 ± 4.08	72.48 ± 0.62	74.20 ± 0.60	$\mathbf{92.50 \pm 1.04}$	$\mathbf{86.99 \pm 1.58}$
	GAIN	71.92 ± 0.92	73.17 ± 1.09	72.48 ± 0.62	74.20 ± 0.60	$\mathbf{92.50 \pm 1.04}$	$\mathbf{86.99 \pm 1.58}$
	GRAFENNE	74.76 ± 9.82	72.96 ± 9.71	63.78 ± 31.28	61.86 ± 28.14	80.69 ± 15.81	73.22 ± 14.33
	FP+VF	73.18 ± 0.96	73.77 ± 0.82	71.86 ± 1.66	70.03 ± 1.97	91.52 ± 1.15	85.67 ± 2.14
	PCFI+VF	74.94 ± 1.37	73.80 ± 1.63	70.76 ± 3.14	68.97 ± 3.85	91.54 ± 1.13	85.70 ± 2.08
	HetGFD (ours)	$\mathbf{76.96 \pm 1.74}$	$\mathbf{77.19 \pm 1.98}$	$\mathbf{92.17 \pm 0.56}$	$\mathbf{92.12 \pm 0.53}$	91.95 ± 1.72	86.72 ± 3.40

Experiments

Applicability to the biomedical domain

r_m	0	0.5	0.9	0.995
Zero	98.49 ± 0.13	78.74 ± 1.01	64.15 ± 1.18	62.20 ± 0.24
Mean	98.49 ± 0.13	64.40 ± 1.97	64.40 ± 1.97	62.14 ± 0.15
kNN	98.49 ± 0.13	78.74 ± 1.01	64.15 ± 1.18	62.20 ± 0.24
SVD	98.49 ± 0.13	79.30 ± 1.15	64.10 ± 1.23	62.23 ± 0.26
GAIN	98.49 ± 0.13	78.85 ± 1.09	64.13 ± 1.09	62.20 ± 0.24
GRAFENNE	83.76 ± 9.15	63.97 ± 1.87	63.15 ± 1.31	62.26 ± 0.00
FP+VF	98.49 ± 0.13	80.44 ± 2.34	64.78 ± 1.51	62.20 ± 0.24
PCFI+VF	98.49 ± 0.13	80.75 ± 1.68	65.22 ± 2.19	62.14 ± 0.32
HetGFD (ours)	98.49 ± 0.13	$\mathbf{81.57 \pm 1.04}$	$\mathbf{66.84 \pm 1.92}$	$\mathbf{63.20 \pm 0.37}$
Impr.	-	$\mathbf{+1.02\%}$	$\mathbf{+2.48\%}$	$\mathbf{+1.51\%}$

Conclusion

✓ To the best of our knowledge, this work is the first attempt to utilize diffusion-based feature imputation for heterogeneous graphs and to design relation-aware distance encoding.

✓ We further confirm that our virtual feature scheme effectively transfers the advantages of existing diffusion-based methods to the heterogeneous graph domain.

✓ We believe that our work will significantly contribute to solving missing data problems in various real-world scenarios that contain heterogeneity, due to the effectiveness and rapid imputation time of HetGFD.