NEURAL INFORMATION PROCESSING SYSTEMS

DGCBench: A Deep Graph Clustering Benchmark

Benyu Wu[1,2,*], *Yue Liu*[3,*], *Qiaoyu Tan*[4], *Xinwang Liu*[3]

Wei Du[1], *Jun Wang*[1], *Guoxian Yu*[1,+]

[1]*Shandong University*, [2]*China University of Mining and Technology*

[3]*National University of Defense Technology*, [4]*New York University Shanghai*

https://github.com/Marigoldwu/PyDGC

Motivation

Deep graph clustering (DGC) aims to partition graph nodes into distinct clusters through deep unsupervised learning.

Challenges

Fragmented datasets & Inconsistent metrics

Insufficient & Narrow evaluation

Lack of standardized protocols

Contributions

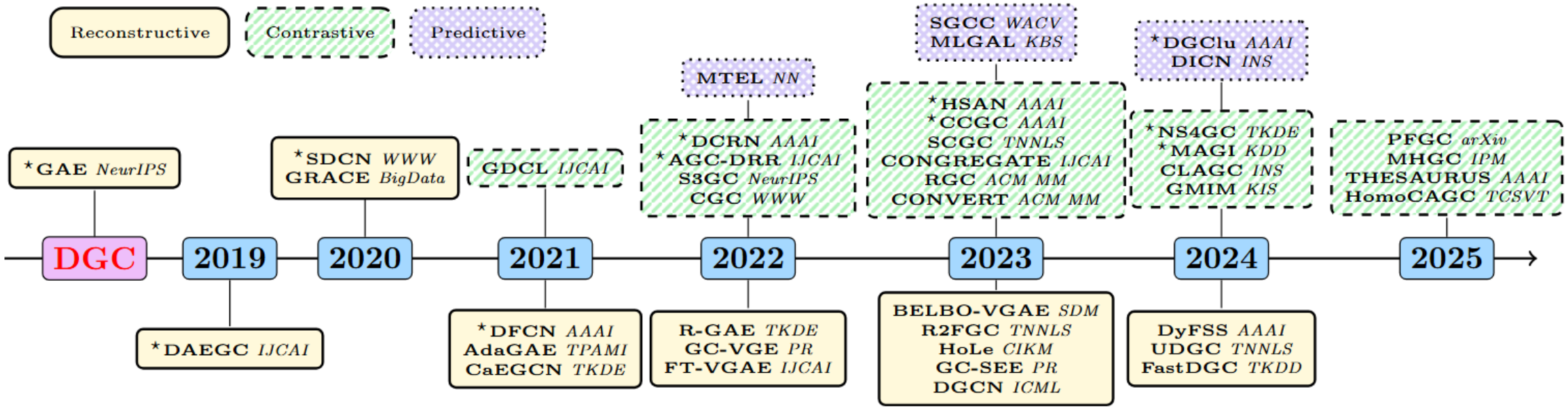
Unified Benchmark

Holistic & Multi-faceted Analysis

Open-source Toolkit

Benchmark Algorithms

Representative methods marked with "★" are benchmark, covering 3 mainstream paradigms: ***reconstructive***, ***contrastive***, and ***predictive***.

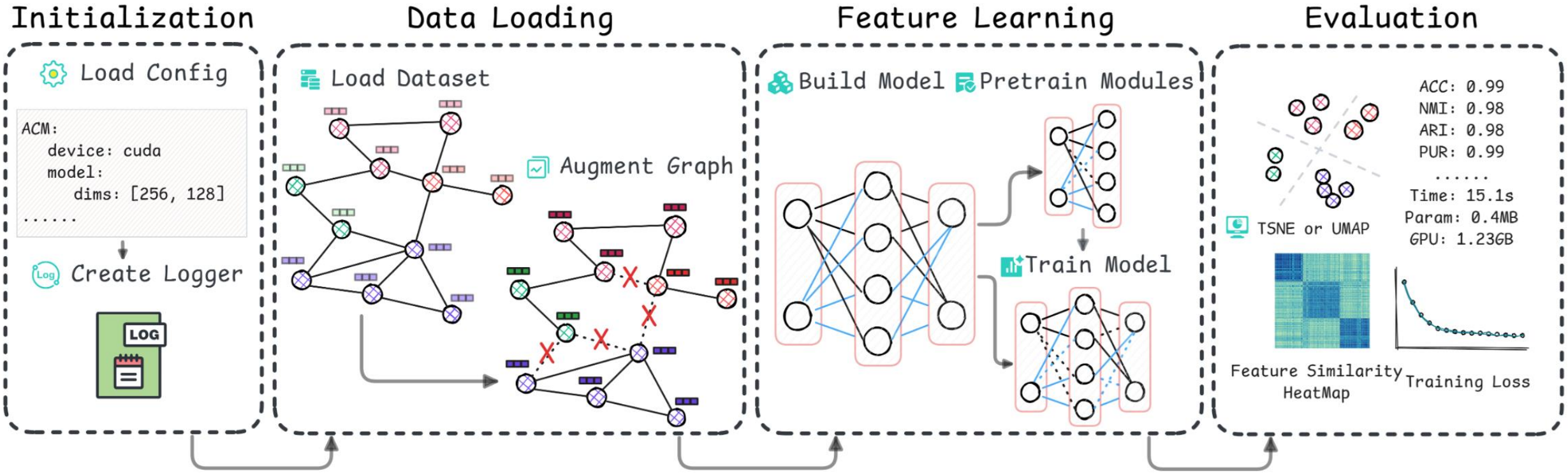


Benchmark Datasets

- Comprehensive **scale** coverage
- Diverse **feature** spaces
- Rich **data type** diversity

Type	Dataset	# Nodes	# Edges	# Features	# Classes	Homophily
Homogenous	Wiki	2,405	17,981	4,973	17	0.71
	Cora	2,708	5,429	1,433	7	0.81
	ACM	3,025	13,128	1,870	3	0.82
	Citeseer	3,327	9,104	3,703	6	0.74
	DBLP	4,057	3,528	334	4	0.80
	PubMed	19,717	88,648	500	3	0.80
	ARXIV	169,343	2,315,598	128	40	0.65
Heterogeneous	Blog	5,196	343,486	8,189	6	0.40
	Flickr	7,575	479,476	12,047	9	0.24
	Roman	22,662	65,854	300	18	0.05
*k*NN-Graph (*k*=3)	USPS	9,298	27,894	256	10	0.98
	HHAR	10,299	30,897	561	6	0.95

Pipeline



PyDGC

A Unified Library for Deep Graph Clustering

```python
from pydgc.utils import parse_arguments
from pydgc.pipelines import GAEPipeline

args = parse_arguments("CORA")
pipeline = GAEPipeline(args)
pipeline.run()
```

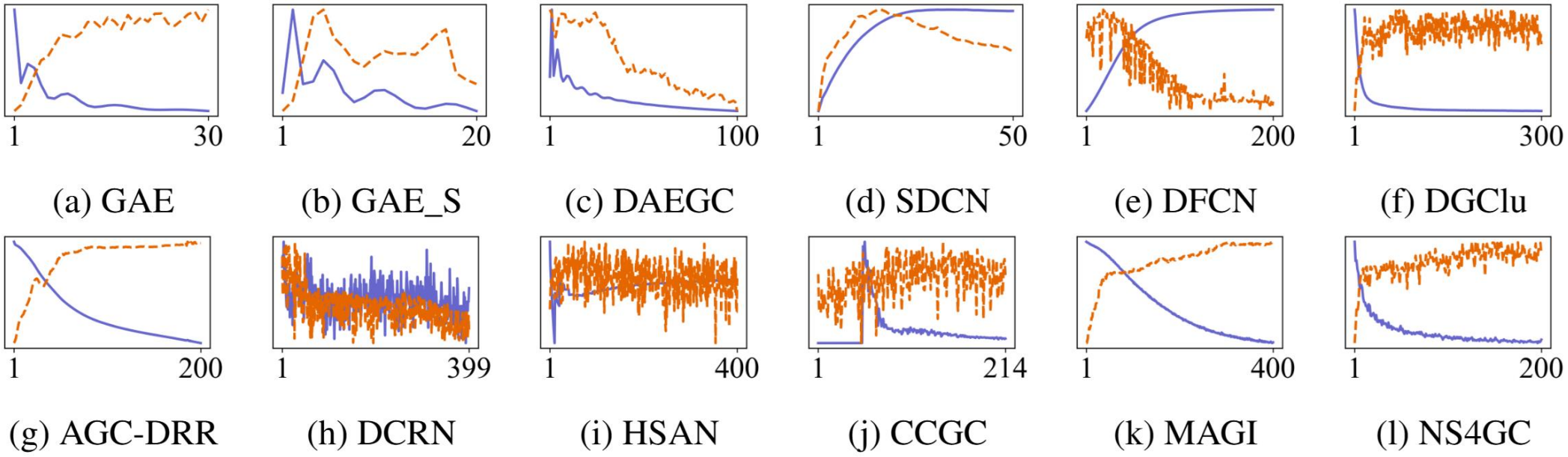
Benchmark Evaluations——Homophily Bottleneck

The benchmark algorithms generally suffer from a homophily bottleneck, and the training termination strategy based on fixed epochs tends to suboptimal clustering performance.

Method	Wiki		Cora		ACM		Citeseer		DBLP		PubMed	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
GAE	0.4444	0.4485	0.6556	0.5148	0.8796	0.6205	0.5821	0.3210	0.4700	0.1907	0.6446	0.2312
♯best	0.5301	0.4751	0.7166	0.5375	0.8803	0.6226	0.5891	0.3193	0.5101	0.1924	0.6480	0.2380
GAE_S	0.3913	0.3476	0.7014	0.5339	0.8507	0.6066	0.5644	0.3070	0.5165	0.2170	0.6636	0.2469
♯best	0.4321	0.4088	0.7078	0.5431	0.8574	0.6119	0.5717	0.3091	0.5276	0.2231	0.6685	0.2597
DAEGC	0.3925	0.3337	0.6681	0.4854	0.8436	0.5181	0.6019	0.3345	0.6140	0.2364	0.6533	0.2432
♯best	0.4069	0.3360	0.6860	0.5142	0.8968	0.6555	0.6143	0.3502	0.6553	0.2887	0.6533	0.2432
SDCN	0.1866	0.0303	0.5082	0.3175	0.8815	0.6373	0.5346	0.3019	0.6193	0.3176	0.5559	0.1960
♯best	0.1934	0.0338	0.5598	0.3689	0.9004	0.6743	0.5364	0.3010	0.6349	0.3368	0.6131	0.2311
DFCN	0.4372	0.4299	0.4885	0.4058	0.8996	0.6720	0.5550	0.3630	0.7148	0.3870	0.4765	0.0734
♯best	0.5249	0.4585	0.7020	0.5274	0.8996	0.6720	0.6836	0.4250	0.7524	0.4355	0.6742	0.3086
DGClu	0.5517	0.4574	0.5841	0.4827	0.7737	0.5217	0.4704	0.2418	0.3590	0.0586	0.5938	0.2079
♯best	0.5843	0.4839	0.7254	0.5593	0.9042	0.6872	0.5907	0.3027	0.5450	0.2117	0.7551	0.3430
AGC-DRR	0.4519	0.4040	0.6681	0.5211	0.9071	0.7086	0.6579	0.4152	0.7976	0.4887	0.6232	0.2523
♯best	0.4519	0.4040	0.6780	0.5153	0.9191	0.6835	0.6737	0.4048	0.8008	0.4942	0.6400	0.2372
DCRN	0.3472	0.2700	0.6471	0.5249	0.9183	0.7166	0.6831	0.4411	0.7053	0.4078	–	–
♯best	0.4072	0.3261	0.7193	0.5494	0.9185	0.7162	0.7042	0.4537	0.7372	0.4339	–	–
HSAN	0.4555	0.4534	0.6390	0.5243	0.5346	0.2338	0.5347	0.3483	0.6674	0.3885	–	–
♯best	0.5206	0.4782	0.7748	0.5912	0.8070	0.5014	0.6659	0.4150	0.7147	0.4230	–	–
CCGC	0.4682	0.4554	0.6170	0.4765	0.7062	0.4850	0.6011	0.3605	0.3382	0.0505	0.4592	0.0923
♯best	0.5354	0.4824	0.7396	0.5640	0.8910	0.6484	0.6918	0.4332	0.5495	0.2373	0.6545	0.3167
MAGI	0.2970	0.2726	0.7324	0.5575	0.8327	0.5345	0.6819	0.4334	0.6751	0.3770	0.6181	0.1912
♯best	0.5651	0.5046	0.7374	0.5619	0.9106	0.6972	0.6931	0.4437	0.7103	0.4075	0.6902	0.3024
NS4GC	0.4454	0.4263	0.7048	0.5683	0.7917	0.4871	0.6721	0.4332	0.7740	0.4544	0.6892	0.3128
♯best	0.5237	0.4980	0.7579	0.5977	0.8882	0.6367	0.6842	0.4351	0.7919	0.4823	0.7056	0.3284
SD		0.000	<0.004	<0.008	<0.012	<0.016	<0.020	<0.030	<0.050	<0.100	>=0.100	

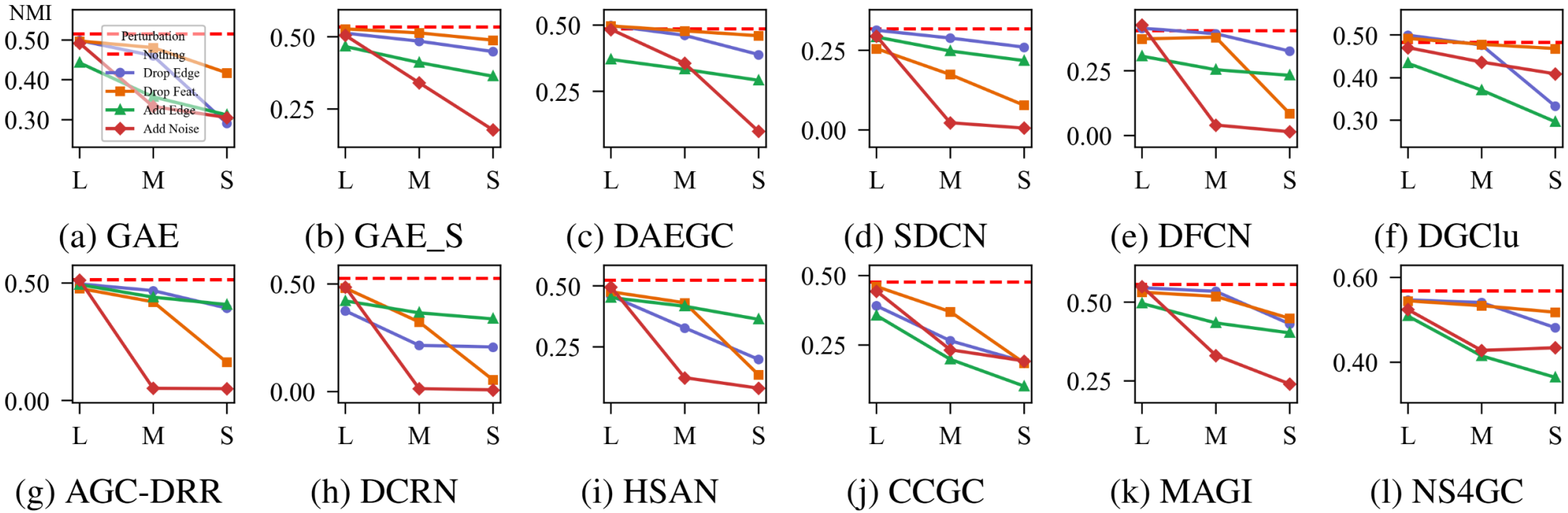
Benchmark Evaluations——Stability Deficiency

Existing algorithms generally **face stability challenges** characterized by "**attainable upper bounds but uncontrollable convergence paths**".



Benchmark Evaluations——Robustness Gap

While all DGC algorithms **degrade under moderate-to-severe noise**, some **unexpectedly demonstrate resilience or marginal performance improvements under light perturbations**.



Benchmark Evaluations——Efficiency Plateau

The comprehensive efficiency of most DGC algorithms is **even lower than that of the vanilla GAE**.

Metrics	Dataset	GAE	GAE_S	DAEGC	SDCN	DFCN	DGClu	AGC-DRR	DCRN	HSAN	CCGC	MAGI	NS4GC
NMI ↑	Cora	0.51	*0.53*	0.49	0.32	0.41	0.48	0.52	0.52	0.52	0.48	0.56	**0.57**
	Citeseer	0.32	0.31	0.33	0.30	0.36	0.24	0.40	**0.44**	0.35	0.36	0.43	*0.43*
	PubMed	0.23	0.25	*0.24*	0.20	-	0.21	-	-	-	0.09	0.19	**0.31**
Speed (it/s) ↑	Cora	56.44	36.52	43.87	32.60	19.73	**89.30**	2.60	19.73	12.29	33.55	80.12	*64.81*
	Citeseer	34.46	24.93	28.89	23.08	15.96	**77.47**	1.77	15.96	8.63	28.13	*44.31*	56.00
	PubMed	1.55	1.07	1.00	*4.88*	-	**17.98**	0.11	-	-	14.22	2.75	4.21
Time (s) ↓	Cora	**0.53**	0.55	2.28	*1.53*	20.27	3.36	76.91	20.27	32.55	11.92	4.99	3.09
	Citeseer	0.87	**0.80**	3.46	2.17	25.06	3.87	113.31	25.06	46.33	14.22	9.03	*0.89*
	PubMed	19.37	*18.76*	100.29	**10.26**	-	16.68	1802.37	-	-	62.10	145.47	47.46
GPU (MB) ↓	Cora	169.10	*172.39*	339.25	735.32	297.11	**157.62**	494.30	962.31	3526.05	2976.50	257.21	255.71
	Citeseer	248.16	*250.32*	529.43	817.16	422.78	**179.94**	669.59	1364.12	13226.41	2976.50	372.74	450.04
	PubMed	6023.24	*6027.89*	15702.99	6398.37	-	**3058.03**	17462.94	-	-	10140.18	8038.78	9591.20
Param (M) ↓	Cora	**0.37**	0.37	*0.37*	5.97	0.48	0.41	1.60	0.48	9.63	1.43	0.73	0.38
	Citeseer	0.95	0.95	0.95	9.38	**0.48**	0.99	1.60	**0.48**	11.49	3.70	2.16	*0.95*
	PubMed	**0.13**	0.13	*0.13*	4.57	0.48	0.17	1.60	0.48	20.22	0.50	0.19	0.19
Avg. ↓		**3.53**	*3.93*	5.80	7.40	7.73	4.13	9.20	7.73	10.33	7.87	5.27	3.67

Benchmark Evaluations——Scalability Challenges

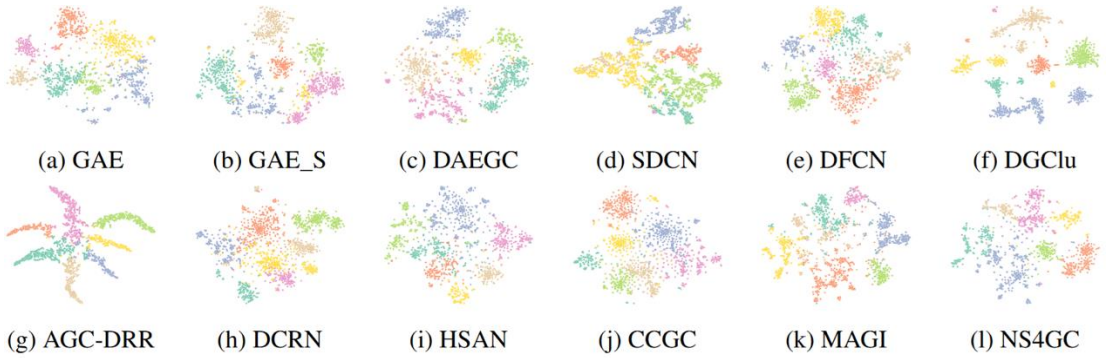
Current graph clustering methods **widely face scalability limitations**.

Method	$USPS_{k=3}$		$HHAR_{k=3}$		Blog		Flickr		Roman		ARXIV	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
GAE	0.5808	0.5251	0.4464	0.4396	0.4658	0.3016	0.2728	0.1596	0.1612	0.0966	–	–
♯best	0.6145	0.5246	0.5995	0.6010	0.5068	0.3308	0.3008	0.1753	0.1714	0.1009	–	–
GAE_S	0.6620	0.5908	0.4255	0.5017	0.4560	0.2825	0.2263	0.1256	0.1588	0.0867	–	–
♯best	0.6620	0.5907	0.4620	0.5185	0.4702	0.3039	0.2536	0.1309	0.1696	0.0901	–	–
DAEGC	0.6796	0.6618	0.5800	0.5748	0.4074	0.2363	0.2643	0.1243	–	–	–	–
♯best	0.6803	0.6597	0.5815	0.5790	0.4491	0.2737	0.2677	0.1262	–	–	–	–
SDCN	0.7320	0.7486	0.6921	0.7048	0.2404	0.0757	0.1730	0.0649	0.1461	0.0144	–	–
♯best	0.7366	0.7441	0.6998	0.7010	0.3260	0.1329	0.2281	0.1143	0.1507	0.0128	–	–
DFCN	0.7295	0.7506	0.6718	0.7301	0.3785	0.2679	0.3113	0.1858	0.1202	0.0624	–	–
♯best	0.7993	0.7662	0.7854	0.7731	0.6133	0.3921	0.3348	0.1904	0.2091	0.1324	–	–
DGClu	0.7128	0.6908	0.6915	0.7270	0.4506	0.2724	0.2317	0.1130	0.1154	0.0374	–	–
♯best	0.7313	0.6786	0.7044	0.7308	0.4706	0.2868	0.2642	0.1338	0.1638	0.0579	–	–
AGC-DRR	0.6733	0.6791	0.6520	0.6068	0.7568	0.6548	0.3139	0.2893	–	–	–	–
♯best	0.6900	0.6748	0.6600	0.6057	0.8369	0.5937	0.3649	0.2760	–	–	–	–
DCRN	0.2686	0.2229	0.3663	0.4696	0.5668	0.4248	0.1836	0.0894	–	–	–	–
♯best	0.3094	0.2748	0.3888	0.4689	0.8363	0.6683	0.2887	0.2055	–	–	–	–
HSAN	–	–	–	–	0.4540	0.3098	–	–	–	–	–	–
♯best	–	–	–	–	0.4900	0.3202	–	–	–	–	–	–
CCGC	0.3882	0.3620	0.3938	0.4193	0.2942	0.1095	0.1825	0.0753	0.1626	0.0932	–	–
♯best	0.5556	0.4852	0.5389	0.5470	0.3551	0.1516	0.1972	0.0051	0.1874	0.1135	–	–
MAGI	0.5703	0.5398	0.4504	0.4593	0.4172	0.2444	0.3121	0.1606	0.1790	0.0694	0.3903	0.4684
♯best	0.6650	0.5864	0.6927	0.6556	0.5129	0.3039	0.3446	0.1889	0.1966	0.1220	–	–
NS4GC	0.6957	0.6688	0.6056	0.5588	0.3636	0.2077	0.3151	0.0184	0.1442	0.0808	–	–
♯best	0.7346	0.6719	0.6367	0.5780	0.4398	0.2612	0.3729	0.2198	0.1724	0.1152	–	–
SD		0.000	<0.004	<0.008	<0.012	<0.016	<0.020	<0.030	<0.050	<0.100	>=0.100	

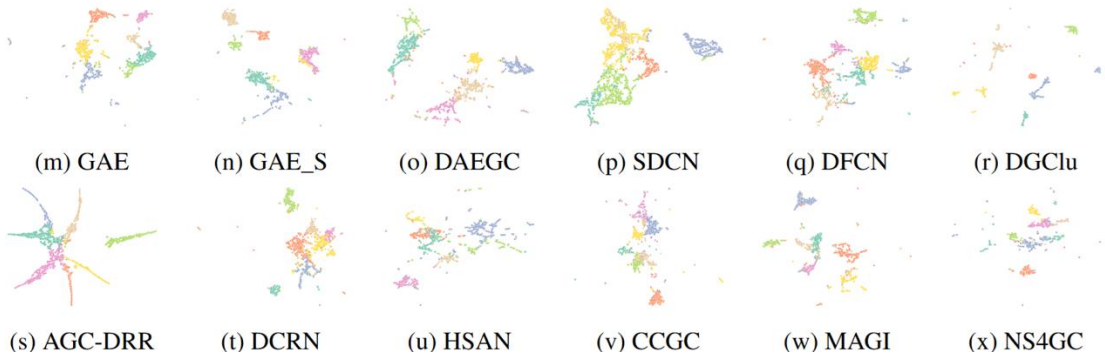
Benchmark Evaluations——Discriminability Limitations

DGC algorithms demonstrate **excellent cohesion within clusters** in embeddings but exhibit **notably insufficient inter-cluster separability**, particularly in distinguishing samples at the cluster boundaries.

	GAE	GAE_S	DAEGC	SDCN	DFCN	DGClu	AGC-DRR	DCRN	HSAN	CCGC	MAGI	NS4GC
HOM ↑	0.542	0.547	0.521	0.358	0.533	0.551	0.530	0.547	**0.593**	0.563	*0.567*	0.575
COM ↑	0.533	0.539	0.507	0.382	0.523	0.568	0.513	0.578	**0.590**	*0.565*	0.557	0.562
SC ↑	0.281	*0.347*	0.330	0.213	0.199	0.424	**0.526**	0.152	0.110	0.089	0.189	0.276
Avg. ↓	7.0	5.7	8.7	10.3	9.0	**3.0**	7.0	6.7	*4.3*	6.3	6.0	4.0



TSNE



UMAP

Conclusion & Future Directions

➢ Core Contribution:

Comprehensive Evaluation: Covers **12 benchmark methods** across **3 paradigms**, assessed via **6 dimensions** on **12 datasets** with diverse characteristics.

Tool Support: Provides a standardized experimental pipeline and an open-source tool, PyDGC.

➢ Key Research Directions for DGC

Homophily Bottleneck

Stability Deficiency

Robustness Gap

Efficiency Plateau

Scalability Challenges

Discriminability Limitations

Thank You!