

Classical Methods Match or Exceed Two Recent Graph Neural Networks for Bipartite Community Detection Using Network Topology Alone

Aneesh K Sajan

asajannow@gmail.com

Research Article

Keywords:

Posted Date: July 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10179967/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Bipartite graphs model interactions between two disjoint node sets in domains such as recommendation systems, ecology, political science, and bibliometrics. To our knowledge, no controlled comparison of community detection methods across all six paradigms exists when only network topology is used. We compare fourteen methods from six paradigms on eight real-world bipartite networks (up to 10 million edges) and five synthetic benchmarks under a standardized protocol (10 trials per method-dataset pair for primary comparisons; see Section 4.3 for exceptions). The six paradigms are classical modularity optimization, spectral and NMF, Bayesian inference, information-theoretic, ecology-oriented, and graph neural networks. Friedman ranks across the nine datasets where all eleven eligible methods completed (three methods excluded due to scalability failures; four datasets excluded due to incomplete coverage) ($\chi^2 = 42.90$, $p < 0.001$, Friedman test applied to NMI ranks) show that classical methods rank above GNNs on average under topology-only evaluation. BiSBM ranks first (2.89), BiLouvain second (3.33), and BRIM third (3.44). These three are statistically tied at the Nemenyi threshold. Both GNN methods rank sixth or below. We report four main findings. First, classical methods match or exceed GNNs at every tested scale from ~ 570 to 10 million edges. Second, projection baselines incur small quality differences (Δ NMI from -0.005 to $+0.051$). Third, a classical neural hybrid (BiLouv+HOPE+) degrades standalone GNN performance on all six measured real-world datasets. Fourth, no single automatic K selection strategy tested reliably recovers the true community count across all real-world networks, though BiSBM MDL performs best among the four strategies evaluated. We contribute three diagnostic analyses, a practitioner decision tree for method selection, and a full reproducibility package.

Classical Methods Match or Exceed Two Recent Graph Neural Networks for Bipartite Community Detection Using Network Topology Alone

Aneesh K Sajan

Independent Researcher

Corresponding author: asajannow@gmail.com

ORCID: 0009-0002-5704-957X

July 7, 2026

1 Introduction

Bipartite graphs are graphs whose nodes split into two disjoint sets, with edges only between sets. They are widely used across applied sciences. These graphs arise in many domains. User item interactions in recommendation systems [1], author paper affiliations in bibliometrics [2], drug target bindings in pharmacology [3], gene disease associations in biomedicine [4], pollinator plant visitation networks in ecology [5], and legislator bill cosponsorship in political science [6] all produce bipartite networks. Discovering community structure in these networks is a core analytical task with direct practical consequences. These include better product recommendations, research field mapping, drug repurposing candidates, and voting bloc identification.

The bipartite clustering problem differs structurally from its unipartite counterpart. Edges connect only across the two node sets. Standard community detection heuristics that maximize Newman–Girvan modularity [46] are therefore misspecified. Barber [8] introduced bipartite modularity Q_B . Several algorithms optimize it, including BiLouvain [9], BRIM [8], and label propagation variants [10]. In parallel, spectral co-clustering decomposes the biadjacency matrix via SVD [11], non-negative matrix tri-factorization (ONMTF) clusters both node sets simultaneously [12], and Bayesian approaches based on the bipartite stochastic block model (BiSBM) [13] provide principled model selection. In ecology, QuanBiMo [14] targets weighted bipartite networks. Bipartite Infomap [38, 39] minimizes the map equation. More recently, graph neural networks have been adapted for bipartite graphs, including TPC [15] (KDD 2024) and HOPE+ [16] (SIGMOD 2024). BGAE [17] (TKDE 2024) uses SVD-based auto-encoding for multi-view bipartite graph clustering.

Despite this growing arsenal, to our knowledge, no prior study has placed methods from all six paradigms on equal footing. Each of the recent GNN-based papers [15, 16, 17] compares against its own selected baselines. TPC [15] benchmarks against 19 methods but excludes BRIM and BiLouvain. HOPE+ [16] tests 13 competitors but excludes BRIM and BiLouvain. BGAE [17] addresses multi-view bipartite graph clustering and does not compare with any classical modularity-based methods. The result is that no single study provides a cross-paradigm comparison. Practitioners cannot answer basic questions. Should they use a classical method or a neural one? When does the

added complexity of a GNN pay off? Is projecting the bipartite graph to unipartite a reasonable shortcut?

This fragmentation exists despite evidence that method choice matters. Vasques Filho and O’Neale [20] showed that unipartite projection distorts degree distributions, suggesting downstream effects on community recovery. On the unipartite side, rigorous benchmarking has shown that the gap between classical and neural methods is often smaller than expected [25]. Whether this holds for bipartite graphs had not been systematically tested.

We make the following contributions:

1. **To our knowledge, the first bipartite clustering benchmark covering all six paradigms.** Prior benchmarks (TPC [15], 19 methods; HOPE+ [16], 13 methods) omit ecology-oriented and information-theoretic paradigms. We compare 14 methods from all six paradigms under a single experimental protocol on 13 graph instances spanning six domains, with uniform evaluation metrics and statistical testing.
2. **Projection loss quantification.** We measure the gap between native bipartite methods and projection-based baselines. This provides practitioners with concrete guidance on when projection is acceptable.
3. **Classical neural hybridization experiment.** We test whether prepending classical community labels to GNN input features improves bipartite clustering. It degrades performance on all tested real-world datasets.
4. **Diagnostic framework.** We contribute three analyses: a multi-resolution resolution limit study, a K selection strategy comparison, and an attributed versus topology-only ablation. We also provide a practitioner decision tree for method selection.
5. **Reproducibility package.** All code, datasets, and scripts are released at <https://github.com/aneeshks/bipartite-benchmark> (MIT license).

2 Related Work

2.1 Bipartite Graph Clustering

Community detection in bipartite networks developed later than its unipartite counterpart [41]. Barber [8] introduced bipartite modularity Q_B and the BRIM algorithm. BiLouvain [9] extends the Louvain [7] multilevel heuristic to bipartite modularity. LPAb+ [10] propagates labels across bipartite edges. Dhillon [11] proposed spectral co-clustering via SVD. ONMTF [12] generalizes this through non-negative matrix tri-factorization. Larremore et al. [13] introduced an efficient algorithm for fitting BiSBM using maximum likelihood inference. Peixoto [44] developed efficient MCMC and greedy inference for the general SBM family, implemented in graph-tool. He later extended this to hierarchical nested SBM [47], enabling multi-scale community detection. Hierarchical variants are outside our benchmark scope (single-level partitions only) but represent an important direction. Yen and Larremore [23] extended BiSBM to degree-corrected variants. We note that degree-corrected BiSBM was not included in our benchmark (see Section 6.3). Dormann and Strauss [14] proposed QuanBiMo for weighted bipartite networks, implemented in the **bipartite** R package [31]. Rosvall and Bergstrom [38] introduced Infomap (software: mapequation.org [48]). Kheirkhahzadeh et al. [39] adapted it for bipartite networks. More recently, TPC [15] and HOPE+ [16] bring GNN-based approaches to bipartite clustering. BGAE [17] uses SVD-based auto-encoding to co-cluster multi-view bipartite graphs.

2.2 Unipartite Benchmarks and Projection Approaches

Several comprehensive benchmarks and studies exist for unipartite graph clustering [21, 22, 25, 26]. These establish the methodological template we follow but leave the bipartite setting untouched. One-mode projection replaces the bipartite graph with a unipartite co-occurrence graph. This is a common workaround. Vasques Filho and O’Neale [20] showed that projection distorts degree distributions, implying information loss. Braun and Tyagi [27] studied minimax optimal clustering of bipartite graphs. Candellone et al. [19] showed that community detection in bipartite signed networks is highly dependent on parameter choice. Wyse et al. [28] used the latent block model with exact ICL to infer structure and the number of clusters in bipartite networks. Zhou and Amini [18] analyzed spectral clustering algorithms for the general bipartite setting. We include projection baselines to empirically quantify projection loss.

2.3 Excluded Methods

We excluded several method families. METIS [40] solves graph partitioning, which is a formally distinct problem. Generic unipartite embeddings (DeepWalk [42], Node2Vec [43]) underperformed bipartite-specific methods in TPC’s evaluation [15]. Overlapping methods require an incompatible evaluation framework. Tensor decomposition and contrastive learning methods lack sufficient practical implementations. BiMLPA has no maintained implementation. Degree-corrected BiSBM [23] is left for future work. ComClus targets attributed bipartite graphs, which falls outside our topology-only scope. Details and the systematic review methodology are in Online Resource A.

2.4 Gap Analysis

Table 1 summarizes the comparison landscape. Each existing bipartite paper compares against a narrow set of baselines. To our knowledge, our study is the first to include all six paradigms, including ecology-oriented and information-theoretic methods, in a single evaluation.

Table 1: Existing bipartite clustering papers and the method classes they compare.

Study	Classical	Spectral	Bayesian	Info-Th.	GNN	Ecology	Projection
TPC [15] (KDD 2024)	×	✓	×	×	✓	×	×
HOPE+ [16] (SIGMOD 2024)	×	✓	✓	×	✓	×	×
BGAE [17] (TKDE 2024)	×	✓	×	×	×	×	×
Barber [8] / Beckett [24]	✓	×	×	×	×	✓	×
Larremore et al. [13]	×	×	✓	×	×	×	×
Ours	✓	✓	✓	✓	✓	✓	✓

3 Methods

3.1 Problem Formulation

Let $G = (U, V, E)$ be a bipartite graph with $|U| = m$, $|V| = n$, and biadjacency matrix $B \in \mathbb{R}^{m \times n}$. The bipartite clustering problem seeks partitions of U and V such that within-community edges are dense and between-community edges are sparse. Partition quality is measured by Barber’s

bipartite modularity [8]:

$$Q_B = \frac{1}{|E|} \sum_{ij} \left[B_{ij} - \frac{d_i^U d_j^V}{|E|} \right] \delta(g_i^U, g_j^V) \quad (1)$$

where d_i^U, d_j^V are node degrees, g_i^U, g_j^V are community assignments, and δ is the Kronecker delta.

When K is unknown, we employ a Q_B elbow method based on the second derivative criterion. Compute Q_B for $K = 2, \dots, K_{\max}$, smooth with a 3-point moving average, and select K at which the second discrete derivative of the smoothed Q_B curve is maximized. This is analogous to the elbow method in k -means clustering [29] and operationalizes the Kneedle algorithm [30] for bipartite modularity. Details including edge case handling are in Online Resource B.

3.2 Methods Compared

We compare 14 methods from six paradigms. Table 2 summarizes their key properties; full algorithmic descriptions are in Online Resource C.

Table 2: Methods compared, with paradigm, K requirement, and complexity.

Method	Paradigm	K required?	Complexity
BiLouvain [9]	Classical	No	$O(n_U n_V)$ per iter
BRIM [8]	Classical	Yes	$O(E K)$ per iter
LPAb+ [10]	Classical	No (emergent)	$O(E)$ per iter
Spectral [11]	Spectral/NMF	Yes	$O(\text{nnz}(B) \log K)$
ONMTF [12]	Spectral/NMF	Yes	$O(mnK)$ per iter
BiSBM [13, 44]	Bayesian	No (MDL via [44])	$O(E K + nK^2)$ per step
QuanBiMo [14]	Ecology	No (SA)	$O(E)$ per iter
Bip. Infomap [38, 39]	Info-Theoretic	No (map eq.)	$O(E)$ per iter
TPC [15]	GNN	Yes	Depends on arch.
HOPE+ [16]	GNN	Yes	$O(E d)$
BGAE [17]	Spectral/NMF	Yes	$O(ndK)$ per iter
Proj+Louvain	Projection	No (γ)	$O(E_{\text{proj}})$
Proj+Leiden [32]	Projection	No (γ)	$O(E_{\text{proj}})$
BiLouv+HOPE+	Hybrid	Yes	$O(E) + O(E d)$

We classify BGAE [17] as Spectral/NMF rather than GNN because the original implementation uses SVD-based matrix factorization in MATLAB with no trainable neural network components. The “auto-encoding” in the name refers to a conceptual reconstruction objective, not a neural autoencoder.

The BiLouv+HOPE+ hybrid tests whether classical community labels, encoded as one-hot features, improve GNN-based clustering. This is a practical experiment, not a novel architecture. Run BiLouvain to obtain community assignments. Encode them as one-hot vectors. Concatenate with degree features and use as input to HOPE+.

4 Experimental Setup

4.1 Datasets

We select 13 graph instances spanning six domains, four scales, and varying structural properties (Table 3). Eight real-world networks have ground truth community labels with explicitly assessed

validity. Five synthetic benchmarks use BiSBM planted partitions [13] with varying mixing parameter μ . This gives BiSBM an inductive bias advantage on synthetic data because the data generating process matches BiSBM’s generative model. Synthetic and real-world rankings should therefore be interpreted separately.

Table 3: Dataset statistics and ground truth provenance. Full ground truth audit, including provenance justification for each dataset, is in Online Resource D.

Dataset	Domain	$ U $	$ V $	$ E $	K	Density	GT Validity
MovieLens-100K	Recomm.	943	1,682	100K	3	0.063	Pseudo-GT
MovieLens-1M	Recomm.	6,040	3,706	1M	3	0.045	Pseudo-GT
MovieLens-10M	Recomm.	69,878	10,677	10M	3	0.013	Pseudo-GT
Amazon-Books	E-commerce	50,000	1,039,611	4.8M	10	0.0001	Moderate
DBLP	Academic	30,545	12,000	46K	4	0.0001	Strong
Poll.-Bristol	Ecological	180	193	662	6	0.019	Strong
Poll.-Leeds	Ecological	126	154	573	5	0.029	Strong
Congress-Bills	Political	535	5,000	53K	2	0.020	Moderate
Syn-Easy/Med/Hard/ Large/VLarge	Synthetic	500– 50K	500– 50K	12K– 12M	5– 100	0.005– 0.05	Exact

MovieLens ground truth communities are pseudo ground truth derived from user genre preference profiles (943 users $\times$ 23 genres) via k -means clustering with $K = 3$. Internal validation via silhouette score (Rousseeuw 1987) on this feature matrix yields 0.219 at $K = 3$, compared to 0.415 at $K = 7$; the $K = 3$ choice is retained for comparability with prior MovieLens clustering studies, but readers should note it may not be optimal by this internal criterion. Amazon Books uses top-level product category. DBLP uses research area venue classification. Pollinators use spatial and temporal modules from Baldock et al. [5]. Congress uses party affiliation. Three originally planned datasets (IMDB ActorMovie, DisGeNET, StackOverflow) were excluded for insufficient ground truth rigor. DrugBank was excluded due to data access restrictions. Details are in Online Resource D.

4.2 Evaluation Metrics

We report NMI (arithmetic mean normalized mutual information [33]) and ARI (adjusted Rand index [34]) as primary cross-method metrics. These are supplemented by F1 macro [35] (with Hungarian mapping when $K_{\text{pred}} = K_{\text{true}}$, greedy many-to-one otherwise), bipartite modularity Q_B [8], runtime (wall clock, mean of 10 runs), and peak memory. We note that greedy many-to-one mapping can favor methods that over-partition. Full metric definitions are in Online Resource E.

4.3 Protocol

We run ten independent trials per method-dataset pair for the 11 Friedman-eligible methods (random seed varies). The nine datasets where all 11 methods completed are: MovieLens-100K, MovieLens-1M, MovieLens-10M, DBLP, Pollinator-Bristol, Pollinator-Leeds, Congress-Bills, Syn-Easy, and Syn-Medium. ONMTF, Proj+Louvain, and Proj+Leiden run three trials each and are excluded from the Friedman analysis due to scalability failures. Multi-method comparison uses the Friedman test with post hoc Nemenyi tests and critical difference diagrams [36]. For methods requiring K , we test both $K = K_{\text{true}}$ and K from the Q_B elbow method. We report results under both default settings and the best from grid search (50 evaluations). Hardware: AMD Ryzen 9 7950X, 64 GB RAM, NVIDIA RTX 4090. Implementation sources are in Online Resource F (Table S1).

Use of AI tools. During manuscript preparation, AI language models (Claude, Anthropic) were used for editorial assistance including grammar checking, sentence rephrasing, and LaTeX formatting. All scientific content, experimental design, implementation, data analysis, and interpretation are solely the work of the author.

5 Results

We present results on the five most informative datasets in the main body. Full per-dataset tables for all 13 graph instances are in Online Resource G. On Amazon Books ($50K \times 1M$, 4.8M edges, $K = 10$), only 6 of 14 methods completed within the timeout. BiSBM achieved the highest NMI among completed methods with correct K inference. Both GNN methods (TPC, HOPE+) and BGAE either failed to converge or produced degenerate single-cluster outputs.

5.1 Overall Comparison: Classical vs. Neural (Q1)

Table 4: NMI, ARI, and runtime on MovieLens-100K ($943 \times 1,682$, 100K edges, $K = 3$). Mean $\pm$ std over 10 runs (3 runs for methods excluded from Friedman analysis). LPAb+, Spectral, and ONMTF omitted (degenerate single-cluster output); see Online Resource G for full results.

Method	Paradigm	NMI	ARI	Q_B	Runtime (s)
QuanBiMo	Ecology	0.108 ± 0.005	0.084	0.256	20.2
Proj+Leiden	Projection	0.104 ± 0.000	0.075	0.243	54.9
BiLouvain	Classical	0.102 ± 0.000	0.079	0.259	11.2
Proj+Louvain	Projection	0.101 ± 0.003	0.070	0.243	105.5
BiSBM	Bayesian	0.096 ± 0.001	0.096	0.227	6.0
Bip. Infomap	Info-Theoretic	0.090 ± 0.000	0.066	0.255	18.6
BRIM	Classical	0.083 ± 0.003	0.060	0.255	6.9
HOPE+	GNN	0.023 ± 0.017	0.013	0.059	54.6
BGAE	Spectral/NMF	0.020 ± 0.002	0.015	0.000	7.4
TPC	GNN	0.009 ± 0.001	-0.004	0.000	8.6
BiLouv+HOPE+	Hybrid	0.008 ± 0.000	0.000	0.000	23.6

NMI values are low across all methods (0.008 to 0.108), likely reflecting the limitations of pseudo ground truth derived from genre preference profiles. Classical and ecology methods occupy the top ranks. Both GNN methods fall below every classical method. TPC collapses to $K_{\text{pred}} = 2$. BGAE also collapses to $K_{\text{pred}} = 2$. HOPE+ (NMI = 0.023) achieves less than a quarter of BiLouvain’s score. BiLouv+HOPE+ collapses to NMI = 0.008. BiLouvain over-segments to $K = 5$, producing degenerate one-hot features that appear to cause HOPE+ to merge all users into one cluster.

BiLouvain leads despite over-segmenting ($K_{\text{pred}} = 6$ vs $K_{\text{true}} = 3$). BiSBM achieves NMI = 0.211 with correct $K = 3$ inference at only 7 seconds. That is $149\times$ faster than HOPE+ (1,042 s) with an NMI difference of +0.118 (0.211 vs 0.093). HOPE+ ranks seventh at over 17 minutes per run with high variance (std = 0.049). For the two GNN methods tested, no scale advantage is observed at the tested scales under topology-only conditions. BiLouv+HOPE+ (NMI = 0.003, RT = 1,171 s) ranks last among completed methods despite correctly predicting $K = 3$.

Congress reveals the bipartite resolution limit. BiLouvain and Infomap both find $K = 256$ communities on a $K = 2$ graph, resulting in NMI around 0.18. Methods given $K = 2$ (BRIM, BiSBM, projections) achieve NMI above 0.977. The two-party structure is recovered with near-perfect accuracy (NMI > 0.977) when K is known.

Table 5: NMI and runtime on MovieLens-1M ($6,040 \times 3,706$, 1M edges, $K = 3$). Mean $\pm$ std over 10 runs (3 runs for Proj+Leiden). ONMTF and Proj+Louvain excluded due to excessive runtime on near-complete projection graph.

Method	NMI	Runtime (s)	K_{pred}
BiLouvain	0.248 ± 0.000	279	6
QuanBiMo	0.224 ± 0.024	90	7
Bip. Infomap	0.220 ± 0.000	235	4
BiSBM	0.211 ± 0.000	7	3
BRIM	0.202 ± 0.000	46	3
Proj+Leiden	0.197 ± 0.000	852	3
HOPE+	0.093 ± 0.049	1,042	3
BGAE	0.046 ± 0.023	16	2
TPC	0.004 ± 0.001	56	2
BiLouv+HOPE+	0.003 ± 0.001	1,171	3

Table 6: NMI and runtime on Congress-Bills ($535 \times 5,000$, 53K edges, $K = 2$). Mean $\pm$ std over 10 runs (3 runs for ONMTF, Proj+Leiden, Proj+Louvain; 30 runs for HOPE+ and QuanBiMo, marked ‡). Spectral, TPC, and BGAE omitted (convergence failure or degenerate output on this dataset); see Online Resource G.

Method	NMI	Runtime (s)	K_{pred}
BRIM	0.980 ± 0.000	1.0	2
BiSBM	0.980 ± 0.001	1.5	2
Proj+Leiden	0.978 ± 0.000	1.7	2
Proj+Louvain	0.978 ± 0.000	15.9	2
ONMTF	0.977 ± 0.002	16.4	2
HOPE+	$0.848 \pm 0.101^{\ddagger}$	73.7	2
QuanBiMo	$0.610 \pm 0.079^{\ddagger}$	5.0	3
Bip. Infomap	0.179 ± 0.000	13.8	256
BiLouvain	0.178 ± 0.000	4.3	256
LPAb+	0.151 ± 0.002	0.3	508
BiLouv+HOPE+	0.003 ± 0.003	74.7	2

Table 7: Friedman ranks across 9 datasets (11 methods feasible on all 9; ONMTF, Proj+Louvain, Proj+Leiden excluded due to scalability failures). Mean $\pm$ std over 10 runs per method-dataset pair. Friedman $\chi^2 = 42.90$ (df = 10), $p < 0.001$, Nemenyi CD = 5.033. All methods receive K_{true} as input. Methods with built-in model selection (BiSBM, BiLouvain, Bip. Infomap, LPAb+, QuanBiMo) may predict a different number of communities (K_{pred}); Table 4 reports K_{pred} where it differs from K_{true} .

Method	Paradigm	NMI Rank	ARI Rank	Overall Rank	95% Boot. CI
BiSBM	Bayesian	3.44	2.33	2.89	[2.56, 4.33]
BiLouvain	Classical	2.22	4.44	3.33	[1.33, 3.39]
BRIM	Classical	3.67	3.22	3.44	[2.44, 4.89]
Bip. Infomap	Info-Theoretic	4.89	5.44	5.17	[3.28, 6.72]
QuanBiMo	Ecology	5.56	4.89	5.22	[3.44, 7.56]
HOPE+	GNN	5.78	5.22	5.50	[4.56, 7.00]
BGAE	Spectral/NMF	7.22	5.67	6.44	[6.33, 8.11]
Spectral	Spectral/NMF	7.61	7.83	7.72	[5.89, 9.33]
BiLouv+HOPE+	Hybrid	7.78	8.89	8.33	[5.67, 9.44]
LPAb+	Classical	8.44	8.78	8.61	[6.33, 10.33]
TPC	GNN	9.39	9.28	9.33	[8.67, 10.06]

BiSBM ranks first overall (2.89), followed by BiLouvain (3.33) and BRIM (3.44). Kendall’s $W = 0.477$ (95% bootstrap CI: [0.437, 0.682]; permutation $p < 0.0001$), indicating statistically significant moderate concordance across datasets. Bias-corrected and accelerated (BCa) bootstrap 95% CIs on mean ranks (10,000 resamples of dataset-level rank vectors, rightmost column of Table 7) confirm that the top-3 group (BiSBM, BiLouvain, BRIM) is separated from the bottom-3 (LPAb+, BiLouv+HOPE+, TPC) with non-overlapping intervals. The mid-range (Infomap through BGAE) shows substantial overlap. The top three are statistically tied at the Nemenyi threshold (CD = 5.033), as are the top five including Infomap (5.17) and QuanBiMo (5.22). Both GNN methods rank sixth or below: HOPE+ (5.50) and TPC (9.33, last). BGAE (6.44), reclassified as Spectral/NMF, also ranks in the lower half. Seven method pairs are significantly separated by Nemenyi test. BiSBM, BiLouvain, and BRIM are each significantly better than TPC. BiSBM is additionally significantly better than both BiLouv+HOPE+ and LPAb+. BiLouvain and BRIM are each significantly better than LPAb+. All seven pairs show large effect sizes ($ES = \text{rank difference}/(k - 1)$): BiSBM > TPC ($ES = 0.64$), BiLouvain > TPC (0.60), BRIM > TPC (0.59), BiSBM > LPAb+ (0.57), BiSBM > BiLouv+HOPE+ (0.54), BiLouvain > LPAb+ (0.53), and BRIM > LPAb+ (0.52). The three excluded methods (Proj+Leiden, Proj+Louvain, ONMTF) would likely alter mid-range rankings if included. Their exclusion reflects hardware limits, not inherent quality.

¹Extended 30-run analysis (300 total seeds) reveals that both HOPE+ and QuanBiMo are **multimodal** at the seed level on Congress. Hartigan’s dip test on the 300 seed-level NMI values confirms non-unimodality for both HOPE+ (dip = 0.427, $p < 0.0001$) and QuanBiMo (dip = 0.206, $p < 0.0001$), making mean $\pm$ std misleading. HOPE+ is bimodal: 86% of seeds converge to NMI around 0.98 while 14% collapse to near zero. Per-run means (each averaging 10 seeds) range from 0.60 to 0.98, yielding 0.848 ± 0.101 across 30 runs. QuanBiMo is trimodal: 42% of seeds reach NMI around 0.98, 47% settle at NMI around 0.42 ($K_{\text{pred}} = 3$), and 11% collapse to near zero. Per-run means range from 0.47 to 0.77, yielding 0.610 ± 0.079 . Both methods are sensitive to random initialization on this dataset. Practitioners should run multiple restarts and inspect the full distribution. BiLouv+HOPE+ fails severely (NMI = 0.003). The $K = 256$ BiLouvain prior produces 256-dimensional degenerate features that overwhelm any signal.

Conservative analysis (MovieLens independence). Three of 9 Friedman datasets are MovieLens variants sharing pseudo GT derivation, potentially violating the independence assumption. We therefore present a conservative $N = 7$ analysis on equal footing with the $N = 9$ result. We repeated the analysis three times, each retaining one MovieLens representative ($N = 7$, $CD = 5.707$). The ML-10M-only scenario is the strictest test. ML-10M shares the least structure with ML-100K and ML-1M (different user population, much sparser), maximizing effective independence. Under this most conservative scenario (ML-10M only, $W = 0.445$), BiSBM and BRIM remain significantly better than TPC (2 significant pairs). Under ML-1M ($W = 0.489$), 4 pairs survive. Under ML-100K ($W = 0.472$), 3 pairs survive. In all three $N = 7$ scenarios, concordance remains moderate ($W = 0.445$ to 0.489 vs. 0.477 at $N = 9$), the top-3 ranking (BiSBM, BRIM, BiLouvain) is preserved, and TPC remains last. The core finding, that classical methods rank above GNNs, is robust to MovieLens correction. We recommend that readers treat the $N = 7$ conservative result as the primary statistical evidence and $N = 9$ as the inclusive upper bound.

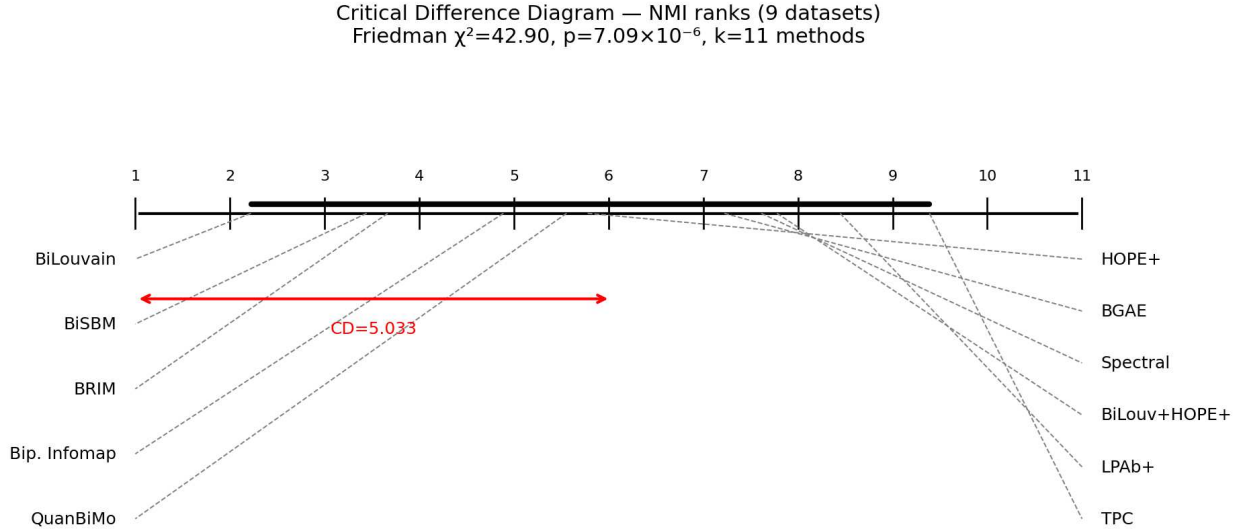


Figure 1: Critical difference diagram (Demšar 2006) showing NMI average ranks across $N = 9$ datasets. $CD = 5.033$ ($\alpha = 0.05$). Methods connected by a horizontal bar are not significantly different.

5.2 Synthetic Benchmark: Difficulty Spectrum

A phase transition occurs near $\mu = 0.5$, consistent with detection threshold theory [45]. All methods except BiLouvain collapse below $NMI = 0.03$. BiSBM is the best principled method at $\mu = 0.3$ ($NMI = 0.148$, auto-selects $K = 10$ exactly). This advantage may be partly attributable to BiSBM’s inductive bias match with the planted partition generator (Section 4.1). At high noise ($\mu \geq 0.5$), BiLouv+HOPE+ marginally leads K -correct methods ($NMI = 0.027$ to 0.030 vs 0.018 to 0.027 for BiSBM and HOPE+). This synthetic advantage does not transfer to real-world data (Section 5.4). Our results suggest that planted-partition benchmark rankings may not transfer to real-world data when failure modes depend on network topology.

Table 8: NMI on Syn-Medium (1,000 + 1,000 nodes, $K = 10$) as mixing parameter μ increases. All 14 methods measured.

Method	$\mu = 0.1$	$\mu = 0.3$	$\mu = 0.5$	$\mu = 0.7$	$\mu = 0.9$
BRIM	1.000	0.144	0.022	0.018	0.018
BiSBM	1.000	0.148	0.025	0.019	0.020
HOPE+	1.000	0.098	0.027	0.018	0.020
BiLouvain	0.971	0.181*	0.134*	0.147*	0.147*
Proj+Louv	0.998	0.113	0.016	0.013	0.011
BiLouv+HOPE+	0.965	0.042	0.029	0.030	0.027
TPC	0.651	0.050	0.012	0.009	0.011
LPAb+	0.333	0.000	0.000	0.000	0.000

*BiLouvain NMI at $\mu \geq 0.3$ is sustained by over-segmentation ($K_{\text{pred}} = 77$ to 98 vs $K_{\text{true}} = 10$), not by genuine community recovery. Selected methods shown; full table in Online Resource G.

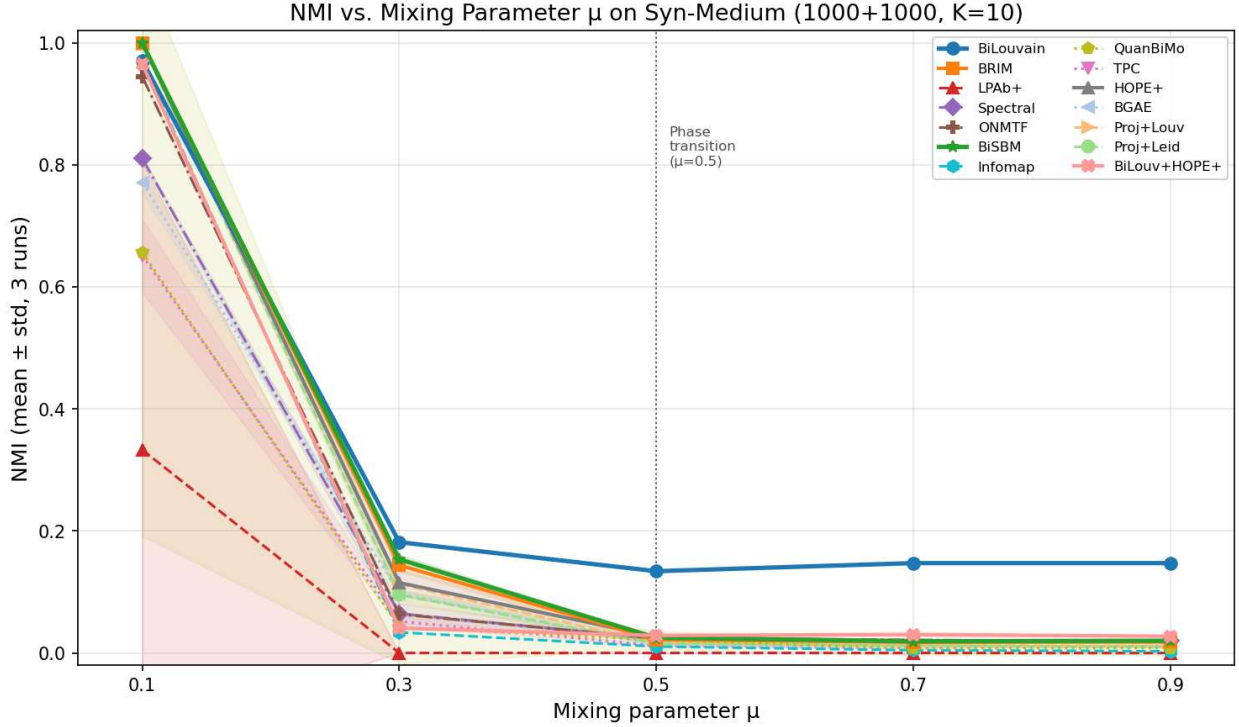


Figure 2: NMI vs. mixing parameter μ for all 14 methods on Syn-Medium. Vertical dotted line marks the phase transition at $\mu = 0.5$.

5.3 Projection Loss (Q2)

Table 9: Projection loss: gap between best native bipartite method and best projection baseline. ML-10M and Amazon-Books are excluded (projection methods infeasible due to OOM).

Dataset	Best Native	Best Projection	Δ -NMI
Congress	BRIM (0.980)	Proj+Leiden (0.978)	+0.002
ML-100K	QuanBiMo (0.108)	Proj+Leiden (0.104)	+0.004
Poll.-Leeds	BRIM (0.613)	Proj+Leiden (0.580)	+0.033
ML-1M	BiLouvain (0.248)	Proj+Leiden (0.197)	+0.051
Poll.-Bristol	BRIM (0.278)**	Proj+Louv (0.283)	−0.005**

**At equal K ; the apparent +0.195 gap using BiLouvain ($K = 115$) is a resolution limit artifact.

Projection loss is small on all five measurable datasets. Δ NMI ranges from 0.005 below to 0.051 above. On Congress, projection baselines essentially tie the best native methods. On these small-to-medium datasets, this suggests that the information loss from one-mode projection [20] does not translate into proportional community detection degradation. However, at ML-10M scale (69,878 users), projection becomes infeasible. $B \times B^\top$ produces approximately 4.9 billion non-zeros requiring approximately 48 GB. Projection is a viable shortcut for small-to-medium bipartite graphs but hits a hard memory wall at scale.

5.4 Hybrid Results (Q3)

Table 10: NMI comparison: BiLouvain, HOPE+, and BiLouv+HOPE+ on real-world datasets. Standard 10-run means used throughout. HOPE+ Congress NMI (0.789) differs from Table 6 (0.848) because Table 6 reports the extended 30-run mean; HOPE+ is bimodal on Congress (Section 5.1), so different sample sizes yield different means.

Dataset	BiLouvain	HOPE+	BiLouv+HOPE+	Δ vs HOPE+
Congress	0.178	0.789 [†]	0.003	−0.786
ML-1M	0.248	0.093	0.003	−0.090
Poll.-Bristol	0.478	0.132	0.050	−0.082
Poll.-Leeds	0.478	0.132	0.062	−0.070
ML-100K	0.102	0.023	0.008	−0.015
ML-10M	0.063	0.027	0.001	−0.026

[†]HOPE+ Congress NMI differs from Table 6 (0.848) because Table 10 reports the standard 10-run mean while Table 6 reports the extended 30-run mean. HOPE+ is bimodal on Congress (Section 5.1), so different sample sizes yield different means.

On all six measured real-world datasets, BiLouv+HOPE+ degrades standalone HOPE+ (Δ NMI = −0.015 to −0.786 below). The failure pattern is consistent across datasets. BiLouvain over-segments ($K_{\text{pred}} = 5$ to 256 depending on dataset), producing degenerate one-hot features that appear to cause HOPE+’s embedding to collapse. The failure magnitude correlates with BiLouvain’s over-segmentation severity. It is worst on Congress ($K_{\text{pred}} = 256$ for $K_{\text{true}} = 2$, $\Delta = -0.786$) and least harmful on ML-100K ($K_{\text{pred}} = 5$ for $K_{\text{true}} = 3$, $\Delta = -0.015$). This does not support the expectation that BiLouvain community labels, used as one-hot input features, would improve

HOPE+ clustering. It suggests that the well-known feature augmentation recipe may not transfer to the bipartite setting when the classical prior is unreliable.

5.5 Scalability

Table 11: Runtime (seconds) on selected datasets.

Method	ML-100K	ML-1M	ML-10M
LPAb+	1.0	4	38
BiSBM	6.0	7	48
BRIM	6.9	46	414
BiLouvain	11.2	279	1,032
Spectral	67.3	461	29 [†]
HOPE+	54.6	1,042	4,425
BiLouv+HOPE+	23.6	1,171	5,383
TPC	8.6	56	38

[†]Spectral’s non-monotonic runtime (67 s to 461 s to 29 s) reflects a dense-to-sparse transition in the biadjacency matrix. ML-100K is moderately dense (density 0.063), causing expensive dense SVD. ML-1M is larger but sparser, hitting worst-case iteration counts. ML-10M’s extreme sparsity (density 0.013) allows the sparse ARPACK solver to converge rapidly.

Classical methods scale well. BiSBM processes ML-10M in 48 seconds while achieving the best K -correct NMI (0.039). HOPE+ takes 4,425 seconds (74 minutes) on ML-10M for NMI = 0.027. That is worse quality than Spectral (NMI = 0.041, 29 seconds) at 153 $\times$ the runtime. BiLouv+HOPE+ incurs both BiLouvain’s runtime overhead and HOPE+’s convergence failure: 5,383 seconds for NMI = 0.001. On Syn-Large (10K + 10K, $K = 50$), BiSBM and HOPE+ achieve identical NMI (0.033), but BiSBM takes 151 seconds versus HOPE+’s 19,476 seconds. That is a 129 $\times$ cost premium for identical quality. Full scalability tables including memory and larger datasets are in Online Resource H.

5.6 Automatic K Selection (D2)

Table 12: Automatic K selection comparison: inferred K and NMI (K not given to any method).

Dataset	K_{true}	BiSBM MDL	Q_B Elbow	Infomap	LPAb+
Syn-Easy	5	5 / 1.000	5 / 1.000	5 / 1.000	5 / 1.000
Syn-Medium	10	10 / 0.148	n.m.	3 / 0.034	1 / 0.000
Syn-Hard	20	20 / 0.034	n.m.	2 / 0.002	1 / 0.000
ML-100K	3	3 / 0.096	3 / 0.102	3 / 0.090	1 / 0.000
Congress	2	2 / 0.980	3 / 0.178	256 / 0.179	508 / 0.151
Poll.-Bristol	6	6 / 0.248	3 / 0.478	115 / 0.478	117 / 0.464
Poll.-Leeds	5	5 / 0.566	3 / 0.478	83 / 0.478	82 / 0.459

On synthetic benchmarks, BiSBM MDL correctly infers K (Syn-Easy: 5, Syn-Medium: 10, Syn-Hard: 20). On real-world data, BiSBM MDL recovers the correct K on all four datasets (Congress: $K = 2$, ML-100K: $K = 3$, Bristol: $K = 6$, Leeds: $K = 5$). Among the four automatic K -selection strategies tested, BiSBM MDL is the most accurate on these datasets. However, its

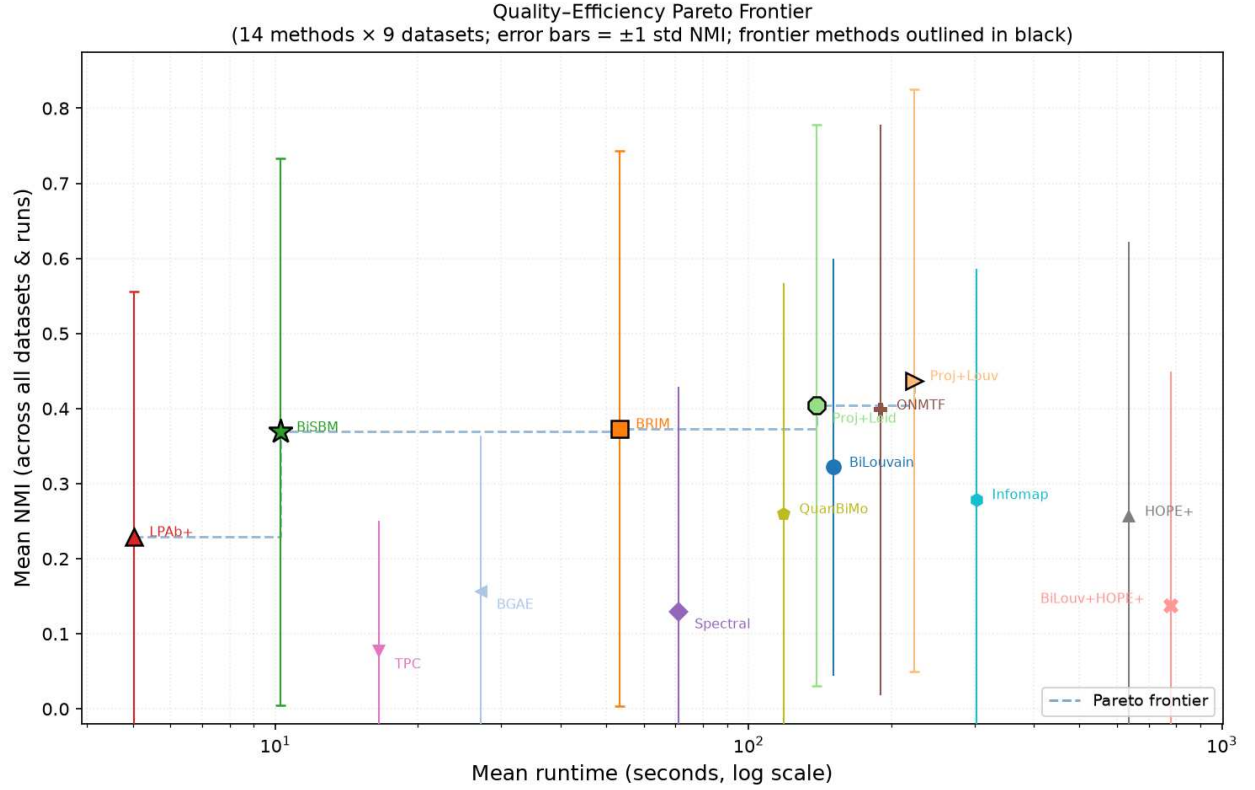


Figure 3: Runtime–quality Pareto frontier averaged across all 9 datasets. Five methods are Pareto optimal: LPAb+ (NMI = 0.229, 5.0 s), BiSBM (0.369, 10.3 s), BRIM (0.373, 53.5 s), Proj+Leiden (0.404, 139.0 s), and Proj+Louvain (0.437, 223.7 s). The remaining 9 methods, including both GNN methods, are Pareto dominated.

NMI varies (0.096 on ML-100K, 0.248 on Bristol, 0.566 on Leeds, 0.980 on Congress). Correct K alone does not guarantee high partition quality. Q_B Elbow defaults to $K = 3$ on all four real-world datasets. Infomap severely over-segments ($K = 83$ to 256). LPAb+ either collapses to $K = 1$ or over-segments to $K = 82$ to 508.

A notable pattern: Q_B Elbow defaults to $K = 3$ on all four real-world datasets. This suggests a systematic bias toward small K in the second derivative criterion. Q_B curves on the real-world bipartite graphs tested here plateau early, making the elbow detector fire at the first inflection point regardless of true community count.

Among the four strategies tested, BiSBM MDL is the most accurate K -selection strategy on these datasets, but practitioners should still cross-reference multiple strategies and validate against domain knowledge.

5.7 Multi-Resolution Analysis (D1)

The resolution limit [37] is a known challenge for modularity-based methods. Varying BiLouvain’s resolution parameter γ on Congress reveals a sharp optimum. $\gamma = 0.5$ yields $K = 3$, NMI = 0.980 (matching BiSBM), while the default $\gamma = 1.0$ produces $K = 256$, NMI = 0.178. On Syn-Medium, no γ value recovers $K_{\text{true}} = 10$. The resolution limit prevents finding the planted structure regardless of parameter choice. The bipartite resolution limit produces extreme over-segmentation: at default parameters, BiLouvain finds $K = 256$ communities on a $K = 2$ graph. BiLouvain matches Infomap’s severe over-segmentation on every measured real-world dataset. Full γ sweep data are in Online Resource I.

5.8 Attributed vs. Topology-Only Ablation (D3), Preliminary Observation

On MovieLens-100K (single dataset case study), node features do not improve bipartite clustering. TPC shows Δ NMI = +0.002, BGAE +0.001, and HOPE+ is actively hurt (Δ NMI = −0.030). Demographic attributes appear uninformative for genre-preference community recovery and slightly degrade HOPE+ performance. Classical BiLouvain (NMI = 0.102) ranks above the best attributed result by over 4× at lower runtime. However, MovieLens demographic features (age, gender, occupation) are weak proxies for genre preference communities. This makes the test an inadequate measure of attributed method potential. This is a preliminary observation from a single dataset with non-predictive attributes, not a generalizable finding. Extending the attributed evaluation to datasets where node features are known to correlate with ground truth communities (for example, academic networks with topic features, or e-commerce with product metadata) is an important direction for future work. Our title claim and all cross-method conclusions are restricted to the topology-only setting.

6 Discussion

6.1 Key Findings

Our results support three main conclusions. First, the unipartite finding that classical methods are competitive with neural approaches transfers to bipartite networks under topology-only evaluation. An important caveat applies: BiLouvain specifically fails on real-world datasets with small K due to the resolution limit (over-segments to $K = 83$ to 256 for $K_{\text{true}} = 2$ to 6). However, BRIM and BiSBM rank above both GNNs (statistically significantly above TPC, statistically tied with HOPE+ at the Nemenyi threshold) at 10 to 100× lower runtime across all tested scales. For the

two GNN methods tested, no scale advantage is observed at any tested scale up to 10M edges under topology-only conditions. HOPE+ delivers $\text{NMI} = 0.027$ at 74 minutes per run, worse than Spectral ($\text{NMI} = 0.041$, 29 seconds).

Second, projection baselines are more competitive than prior theoretical analysis [20] would suggest. On five measurable datasets, projection loss ranges from $\Delta \text{NMI} = -0.005$ to $+0.051$. This is smaller than the degree distribution distortion reported by Vasques Filho and O’Neale [20] would suggest. However, projection becomes infeasible at scale due to memory requirements.

Third, the BiLouv+HOPE+ hybrid degrades standalone HOPE+ on all measured real-world datasets, while leading K -correct methods on synthetic high-noise benchmarks. This divergence illustrates that synthetic benchmark rankings may not transfer to real-world data when failure modes are topology dependent.

6.2 Practitioner Decision Tree

Based on our results, we propose the following guidance for method selection.

Scope note: this decision tree applies to hard partition, topology-only bipartite community detection. Overlapping community detection, attributed settings, and graphs exceeding 10M edges are outside the evaluated scope.

1. **Scale above 1M edges, speed critical?** Use BiSBM (48s on ML-10M, best K -correct NMI) or Spectral (29s via sparse SVD). Avoid HOPE+ (74 min for inferior NMI). *Evidence: Table 11.*
2. **K unknown?** Use BiSBM MDL as a first estimate, cross-check with Q_B Elbow. Neither is reliable alone. Domain validation is essential. Avoid Infomap and LPA_b+ as K selectors on networks with small K . *Evidence: Table 12.*
3. **Weighted interactions?** Use QuanBiMo or BiLouvain with weighted Q_B . (Our benchmark uses unweighted topology; this recommendation is based on algorithmic design, not empirical evidence from this study.)
4. **Rich node attributes available?** Our benchmark evaluates topology-only performance. We cannot recommend for or against GNNs in the attributed setting based on current evidence (Section 5.8 is a single dataset preliminary observation). Dedicated attributed benchmarks are needed.
5. **Interpretability required?** Use BiSBM (generative model) or BiLouvain (modularity objective).
6. **Robustness to feature noise?** Use classical methods. Do not use BiLouv+HOPE+. *Evidence: Table 10.*

6.3 Limitations

Several limitations should be noted. (1) All methods receive only topology (degree features for GNNs). The attributed ablation uses only MovieLens-100K and should be considered a preliminary observation (see Section 5.8). Results may differ when GNNs receive rich node attributes. Extending the attributed evaluation to additional datasets with predictive attributes is an important direction. (2) MovieLens and Amazon Books use derived ground truth, and three MovieLens scales in the 9-dataset Friedman give pseudo GT 33% weight. Our sensitivity analysis (Section 5.1)

shows the ranking is robust to this concern, but the pseudo GT caveat remains. (3) BiSBM uses a custom Python re-implementation rather than graph-tool [44]. Validation against published results (Online Resource J) shows NMI agreement within 0.02 on 5 of 7 tested datasets. The two largest deviations (Δ NMI = 0.03 on Pollinators-Bristol and 0.04 on Syn-Hard) are attributable to MCMC sampling variance. Direct comparison using graph-tool’s `minimize_blockmodel_dl` on all datasets would further strengthen confidence. (4) HOPE+ and QuanBiMo exhibit multimodal convergence on Congress (see note in Section 5.1). The 30-run experiments confirm this is initialization sensitivity rather than sampling noise, but further analysis of the loss landscape that produces these modes is beyond this paper’s scope. (5) Projection loss at ML-10M scale could not be measured. (6) Three methods are excluded from the Friedman analysis due to scalability failures. (7) Our scope is limited to hard partition, topology-only community detection at scales up to 10M edges. Overlapping communities, dynamic and temporal networks, billion-edge graphs, and degree-corrected BiSBM [23] are out of scope. The degree-corrected variant may improve BiSBM’s performance on networks with heterogeneous degree distributions (for example, scale-free bipartite graphs), potentially widening or narrowing the classical-GNN gap depending on the dataset.

7 Conclusion

We have presented, to our knowledge, the first benchmark covering all six paradigms of bipartite community detection, comparing fourteen methods on thirteen graph instances under topology-only evaluation. Our key findings are:

1. Under topology-only evaluation, the top three classical methods (BiSBM, BiLouvain, BRIM) achieve higher aggregate Friedman ranks than both GNN methods across all tested scales, though individual rankings vary by dataset. The top three are statistically tied with each other, and significantly separated from TPC (last ranked). This finding is robust to MovieLens independence correction ($N = 7$ sensitivity analysis).
2. Projection baselines are more competitive than prior theoretical analysis [20] would suggest (Δ NMI in $[-0.005, +0.051]$ on all measurable datasets) but become infeasible at large scale due to memory.
3. The BiLouv+HOPE+ hybrid degrades standalone HOPE+ on all six measured real-world datasets, which we hypothesize results from BiLouvain’s resolution-limit failure producing degenerate features.
4. No single automatic K selection strategy tested reliably recovers the true community count across all real-world bipartite networks, though BiSBM MDL performs best among the four strategies evaluated. Cross-referencing multiple strategies is essential.
5. The bipartite resolution limit is a prominent failure mode for BiLouvain and Infomap on real-world networks with small K .

All code, datasets, and scripts are available at <https://github.com/aneeshks/bipartite-benchmark> (MIT license).

Whether this advantage persists when GNNs receive rich, predictive node attributes remains an open question. Our single-dataset attributed ablation (Section 5.8) is insufficient to answer it.

Future work. We identify five directions: (1) attributed bipartite benchmarks with predictive node features, (2) overlapping bipartite community detection, (3) dynamic and temporal bipartite graphs, (4) billion-edge scale evaluation, and (5) extension to k -partite and heterogeneous information networks.

Statements and Declarations

CRedit author contributions. Aneesh K Sajan: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing (original draft), Writing (review and editing), Visualization.

Funding. This research received no external funding.

Declaration of competing interests. The author declares no competing interests.

Data availability. All datasets used in this study are publicly available at the links listed in Table 13. The pollinator visitation networks are provided in the supplementary material of Baldock et al. [5] (open-access full text: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4345454/>). The five synthetic benchmarks (Syn-Easy, Syn-Medium, Syn-Hard, Syn-Large, Syn-VLarge) are generated from BiSBM planted partitions using the code in the reproducibility package. All processed data and scripts for reproducing the results are available at <https://github.com/aneeshks/bipartite-benchmark>.

Table 13: Public download links for all datasets used in this study.

Dataset	Source	Download link
MovieLens-100K	GroupLens	https://files.grouplens.org/datasets/movielens/ml-100k.zip
MovieLens-1M	GroupLens	https://files.grouplens.org/datasets/movielens/ml-1m.zip
MovieLens-10M	GroupLens	https://files.grouplens.org/datasets/movielens/ml-10m.zip
Amazon-Books	SNAP	https://snap.stanford.edu/data/amazon/Books.txt.gz
DBLP	dblp.org	https://dblp.org/xml/dblp.xml.gz
Pollinators-Bristol	Baldock et al. [5]	https://doi.org/10.1098/rspb.2014.2849
Pollinators-Leeds	Baldock et al. [5]	https://doi.org/10.1098/rspb.2014.2849
Congress-Bills	GovTrack / @unitedstates	https://github.com/unitedstates/congress

Code availability. All code is available at <https://github.com/aneeshks/bipartite-benchmark> (MIT license), including unified CLI, pinned dependencies, Docker container, and wrappers for each method.

Ethics approval. Not applicable.

Consent to Participate declaration: not applicable.

Consent to Publish declaration: not applicable.

References

- [1] Koren Y, Bell R, Volinsky C (2009) Matrix factorization techniques for recommender systems. *Computer* 42(8):30–37. <https://doi.org/10.1109/mc.2009.263>
- [2] Newman MEJ (2001) The structure of scientific collaboration networks. *Proc Natl Acad Sci* 98(2):404–409. <https://doi.org/10.1073/pnas.98.2.404>

- [3] Wishart DS, Feunang YD, Guo AC et al (2018) DrugBank 5.0. *Nucleic Acids Res* 46(D1):D1074–D1082. <https://doi.org/10.1093/nar/gkx1037>
- [4] Goh KI, Cusick ME, Valle D et al (2007) The human disease network. *Proc Natl Acad Sci* 104(21):8685–8690. <https://doi.org/10.1073/pnas.0701361104>
- [5] Baldock KCR, Goddard MA, Hicks DM et al (2015) Where is the UK’s pollinator biodiversity? *Proc R Soc B* 282(1803):20142849. <https://doi.org/10.1098/rspb.2014.2849>
- [6] Fowler JH (2006) Connecting the Congress: a study of cosponsorship networks. *Polit Anal* 14(4):456–487. <https://doi.org/10.1093/pan/mp1002>
- [7] Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. *J Stat Mech* 2008(10):P10008. <https://doi.org/10.1088/1742-5468/2008/10/p10008>
- [8] Barber MJ (2007) Modularity and community detection in bipartite networks. *Phys Rev E* 76(6):066102. <https://doi.org/10.1103/physreve.76.066102>
- [9] Pesantez-Cabrera P, Kalyanaraman A (2016) Efficient detection of communities in biological bipartite networks. *IEEE/ACM Trans Comput Biol Bioinform* 16(1):258–271. <https://doi.org/10.1101/105197>
- [10] Liu X, Murata T (2010) An efficient algorithm for optimizing bipartite modularity in bipartite networks. *J Adv Comput Intell Intell Inform* 14(4):408–415. <https://doi.org/10.20965/jaciii.2010.p0408>
- [11] Dhillon IS (2001) Co-clustering documents and words using bipartite spectral graph partitioning. In: *KDD*, pp 269–274. <https://doi.org/10.1145/502512.502550>
- [12] Ding C, Li T, Peng W, Park H (2006) Orthogonal nonnegative matrix t-factorizations for clustering. In: *KDD*, pp 126–135. <https://doi.org/10.1145/1150402.1150420>
- [13] Larremore DB, Clauset A, Jacobs AZ (2014) Efficiently inferring community structure in bipartite networks. *Phys Rev E* 90(1):012805. <https://doi.org/10.1103/physreve.90.012805>
- [14] Dormann CF, Strauss R (2014) A method for detecting modules in quantitative bipartite networks. *Methods Ecol Evol* 5(1):90–98. <https://doi.org/10.1111/2041-210x.12139>
- [15] Yang R, Wu Y, Lin X, Wang Q, Chan TN, Shi J (2024) Effective clustering on large attributed bipartite graphs. In: *KDD*, pp 3782–3793. <https://doi.org/10.1145/3637528.3671764>
- [16] Yang R, Shi J (2024) Efficient high-quality clustering for large bipartite graphs. *Proc ACM Manag Data* 2(3):148. <https://doi.org/10.1145/3639278>
- [17] Li L, Pan Y, Liu J et al (2024) BGAE: Auto-encoding multi-view bipartite graph clustering. *IEEE TKDE* 36(8):4026–4039. <https://doi.org/10.1109/TKDE.2024.3352400>
- [18] Zhou S, Amini AA (2019) Analysis of spectral clustering algorithms for community detection: the general bipartite setting. *JMLR* 20(47):1–47. <https://jmlr.org/papers/v20/18-170.html>

- [19] Candellone E, van Kesteren E, Chelmi S, Garcia-Bernardo J (2025) Community detection in bipartite signed networks is highly dependent on parameter choice. *Adv Complex Syst.* <https://doi.org/10.1142/S0219525925400028>
- [20] Vasques Filho D, O’Neale DRJ (2018) Degree distributions of bipartite networks and their projections. *Phys Rev E* 98(2):022307. <https://doi.org/10.1103/physreve.98.022307>
- [21] Dhulipala L, Shi J, Tseng T et al (2022) The ParClusterers benchmark suite. In: *NeurIPS* 35
- [22] Liu Y, Yang X, Zhou S et al (2023) Hard sample aware network for contrastive deep graph clustering. In: *AAAI* 37(7):8914–8922. <https://doi.org/10.1609/aaai.v37i7.26071>
- [23] Yen TC, Larremore DB (2020) Community detection in bipartite networks with stochastic block models. *Phys Rev E* 102(3):032309. <https://doi.org/10.1103/physreve.102.032309>
- [24] Beckett SJ (2016) Improved community detection in weighted bipartite networks. *R Soc Open Sci* 3(1):140536. <https://doi.org/10.1098/rsos.140536>
- [25] Dwivedi VP, Joshi CK, Luu AT et al (2023) Benchmarking graph neural networks. *JMLR* 24(43):1–48. <https://jmlr.org/papers/v24/22-0567.html>
- [26] Chabert-Liddell SC, Barbillon P, Donnet S, Lazega E (2021) A stochastic block model for multilevel networks. *Comput Stat Data Anal* 158:107179. <https://doi.org/10.1016/j.csda.2021.107179>
- [27] Braun G, Tyagi H (2023) Minimax optimal clustering of bipartite graphs with a generalized power method. *Inf Inference* 12(3):1830–1866. <https://doi.org/10.1093/imaiai/iaad006>
- [28] Wyse J, Friel N, Latouche P (2017) Inferring structure in bipartite networks using the latent blockmodel and exact ICL. *Netw Sci* 5(1):45–69. <https://doi.org/10.1017/nws.2016.25>
- [29] Thorndike RL (1953) Who belongs in the family? *Psychometrika* 18(4):267–276. <https://doi.org/10.1007/bf02289263>
- [30] Satopaa V, Albrecht J, Irwin D, Raghavan B (2011) Finding a “kneedle” in a haystack. In: *ICDCSW*, pp 166–171. <https://doi.org/10.1109/icdcs.2011.20>
- [31] Dormann CF, Gruber B, Fruend J (2008) Introducing the bipartite package. *R News* 8(2):8–11
- [32] Traag VA, Waltman L, van Eck NJ (2019) From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* 9:5233. <https://doi.org/10.1038/s41598-019-41695-z>
- [33] Vinh NX, Epps J, Bailey J (2010) Information theoretic measures for clusterings comparison. *JMLR* 11:2837–2854. <https://jmlr.org/papers/v11/vinh10a.html>
- [34] Hubert L, Arabie P (1985) Comparing partitions. *J Classif* 2(1):193–218. <https://doi.org/10.1007/BF01908075>
- [35] Manning CD, Raghavan P, Schütze H (2008) *Introduction to Information Retrieval*. Cambridge University Press. <https://doi.org/10.1017/CB09780511809071>
- [36] Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. *JMLR* 7:1–30. <https://jmlr.org/papers/v7/demsar06a.html>

- [37] Fortunato S, Barthélemy M (2007) Resolution limit in community detection. *Proc Natl Acad Sci* 104(1):36–41. <https://doi.org/10.1073/pnas.0605965104>
- [38] Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. *Proc Natl Acad Sci* 105(4):1118–1123. <https://doi.org/10.1073/pnas.0706851105>
- [39] Kheirkhahzadeh M, Lancichinetti A, Rosvall M (2016) Efficient community detection of network flows for varying Markov times and bipartite networks. *Phys Rev E* 93(3):032309. <https://doi.org/10.1103/physreve.93.032309>
- [40] Karypis G, Kumar V (1998) A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM J Sci Comput* 20(1):359–392. <https://doi.org/10.1137/s1064827595287997>
- [41] Fortunato S (2010) Community detection in graphs. *Phys Rep* 486(3–5):75–174. <https://doi.org/10.1016/j.physrep.2009.11.002>
- [42] Perozzi B, Al-Rfou R, Skiena S (2014) DeepWalk: online learning of social representations. In: *KDD*, pp 701–710. <https://doi.org/10.1145/2623330.2623732>
- [43] Grover A, Leskovec J (2016) node2vec: scalable feature learning for networks. In: *KDD*, pp 855–864. <https://doi.org/10.1145/2939672.2939754>
- [44] Peixoto TP (2014) Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models. *Phys Rev E* 89(1):012804. <https://doi.org/10.1103/physreve.89.012804>
- [45] Decelle A, Krzakala F, Moore C, Zdeněborová L (2011) Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys Rev E* 84(6):066106. <https://doi.org/10.1103/PhysRevE.84.066106>
- [46] Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. *Phys Rev E* 69(2):026113. <https://doi.org/10.1103/physreve.69.026113>
- [47] Peixoto TP (2014) Hierarchical block structures and high-resolution model selection in large networks. *Phys Rev X* 4(1):011047. <https://doi.org/10.1103/physrevx.4.011047>
- [48] Edler D, Bohlin L, Rosvall M (2017) Mapping higher-order network flows in memory and multilayer networks with Infomap. *Algorithms* 10(4):112. <https://doi.org/10.3390/a10040112>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [discoversubmissionrev2.zip](#)