HCD Simulations Write Up

Audrey Fu Lab

2024-03-07

Data Simulation

Simulating the network

We adopt a top-down approach to simulate hierarchical networks, considering various simulation parameters such as graph sparsity, noise, and the architecture of the super-level graph(s), including small-world, scale-free, and random graph networks (Watts and Strogatz 1998; Barabási and Bonabeau 2003).

Our simulations focus on basic hierarchies comprising one or two hierarchical layers. Two-layer networks mirror classical community detection on graphs, where our aim is to recover the true community labels from a given graph. Meanwhile, three-layer networks present a more intricate scenario, where the bottom layer of the hierarchy contains two levels of community structure. Here, the top level corresponds to the nodes at the uppermost layer of the hierarchy, and the middle level consists of communities nested within the top-level communities. The objective with these networks is to identify both sets of community partitions.

In each hierarchy, for fully connected networks, we initiate by simulating n_{top} top-level nodes, adhering to a directed small-world, random graph, or scale-free network architecture (Watts and Strogatz 1998; Barabási and Bonabeau 2003). In cases where the network is disconnected, we simply simulate n_{top} disconnected nodes. For networks with three hierarchical layers, we then generate a subnetwork of n_{middle} nodes from each top-layer node, adhering to the network structure utilized at the top level. If the network is fully connected, we apply a probability p_{between} to the nodes from different top-level communities being connected.

The final step in all hierarchies is to generate the nodes in the observed (bottom) layer of the hierarchy. For each top-layer or middle-layer node, we generate a subnetwork of n_{bottom} nodes under the same subnetwork structure as the previous layers, and we apply a probability p_{between} for nodes from different communities to share an edge.

Simulating gene expression

Once we simulate a hierarchical graph, we utilize this hierarchy to generate the node-feature matrix, which depicts the expression of N genes across p samples. Here, N denotes the number of nodes in the observed (bottom) layer of the hierarchy, and its range is governed by $a^{\ell+1} < N < a \times b^{\ell}$, where ℓ signifies the number of hierarchical layers.

We simulate the node-feature matrix using the topological order the observed level graph. We start by generating the features of nodes that have no parental input. We refer to these nodes as origin nodes. All origin nodes are simulated from a normal distribtion with mean 0 and standard deviation σ. All other nodes are simulated from a normal distribution centered at the mean of their parent nodes and with standard deviation σ.

Datasets

We consider three sets of hierarchical networks which represent varying difficulty levels for inference:

1. **Complex networks** - - used for final simulation assessment - **Table 1-3**
2. **Intermediate networks** - used for investigative model tuning and performance assessment - **Table 4**
3. **Simple networks** - used for code implementation and debugging - **Table 5**

Application to Intermediate Networks

A summary of the intermediate networks can be found in **Table 4**. The intermediate networks dataset consists of three layer networks of small world, scale free, and random graph architectures that are less complex then the three layer networks in the **Complex networks** dataset. Each of these networks has 5 super layer nodes, 15 middle layer nodes and approximately 300 bottom layer nodes. We primarily use this dataset to investigate the behavior of the HCD method when applied to 3-layer network.

Preliminary Findings

Tables

Table 1: Summary statistics for all small world networks in the complex networks datset

Value	Network1	Network2	Network3	Network4	Network5	Network6	Network7	Network8
Subgraph type	small world	small world	small world	small world	small world	small world	small world	small world
Connection type	disc	disc	disc	disc	full	full	full	full
Layers	2	2	3	3	2	2	3	3
Standard deviation	0.1	0.5	0.1	0.5	0.1	0.5	0.1	0.5
Nodes per layer	(10, 63)	(10, 63)	(10, 63, 1604)	(10, 63, 1604)	(10, 63)	(10, 63)	(10, 63, 1604)	(10, 63, 1604)
Edges per layer	(0, 63)	(0, 63)	(0, 63, 2011)	(0, 63, 2031)	(45, 115)	(45, 109)	(45, 114, 1604)	(45, 111, 1604)
Subgraph probability	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Sample size	500	500	500	500	500	500	500	500
Modularity (top)	0.898	0.898	0.898	0.898	0.447	0.477	0.766	0.771
Average node degree top	1	1	1.254	1.266	1.825	1.73	1.433	1.439
Avg connections within top communities	6.3	6.3	201.1	203.1	6.3	6.3	199.3	201.3
Avg. connections between top communities	0	0	0	0	0.578	0.511	3.389	3.278
Modularity (middle)	NA	NA	0.762	0.758	NA	NA	0.667	0.663
Average node degree middle	NA	NA	1.254	1.266	NA	NA	1.433	1.439
Avg connections within middle communities	NA	NA	24.825	24.968	NA	NA	24.937	24.873
Avg connections between middle communities	NA	NA	0.114	0.117	NA	NA	0.186	0.19

Table 2: Summary statistics for all scale free networks in the complex networks datset

Value	Network1	Network2	Network3	Network4	Network5	Network6	Network7	Network8
Subgraph type	scale free	scale free	scale free	scale free	scale free	scale free	scale free	scale free
Connection type	disc	disc	disc	disc	full	full	full	full
Layers	2	2	3	3	2	2	3	3
Standard deviation	0.1	0.5	0.1	0.5	0.1	0.5	0.1	0.5
Nodes per layer	(10, 58)	(10, 58)	(10, 58, 1450)	(10, 58, 1450)	(10, 58)	(10, 58)	(10, 58, 1450)	(10, 58, 1450)
Edges per layer	(0, 74)	(0, 74)	(0, 74, 6700)	(0, 74, 6670)	(45, 120)	(45, 120)	(45, 123, 1450)	(45, 122, 1450)
Subgraph probability	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Sample size	500	500	500	500	500	500	500	500
Modularity (top)	0.89	0.89	0.892	0.893	0.513	0.513	0.854	0.849
Average node degree top	1.276	1.276	4.621	4.6	2.069	2.069	4.781	4.843
Avg connections within top communities	7.4	7.4	670	667	7.4	7.4	665.9	671.4
Avg. connections between top communities	0	0	0	0	0.511	0.511	3.033	3.422
Modularity (middle)	NA	NA	0.906	0.91	NA	NA	0.875	0.864
Average node degree middle	NA	NA	4.621	4.6	NA	NA	4.781	4.843
Avg connections within middle communities	NA	NA	107.069	107.069	NA	NA	107.069	107.069
Avg connections between middle communities	NA	NA	0.148	0.139	NA	NA	0.218	0.246

Table 3: Summary statistics for all random graph networks in the complex networks datset

Value	Network1	Network2	Network3	Network4	Network5	Network6	Network7	Network8
Subgraph type	random graph	random graph	random graph	random graph	random graph	random graph	random graph	random graph
Connection type	disc	disc	disc	disc	full	full	full	full
Layers	2	2	3	3	2	2	3	3
Standard deviation	0.1	0.5	0.1	0.5	0.1	0.5	0.1	0.5
Nodes per layer	(10, 45)	(10, 45)	(10, 45, 725)	(10, 45, 725)	(10, 45)	(10, 45)	(10, 45, 725)	(10, 45, 725)
Edges per layer	(0, 32)	(0, 32)	(0, 32, 678)	(0, 32, 665)	(45, 77)	(45, 77)	(45, 78, 725)	(45, 78, 725)
Subgraph probability	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Sample size	500	500	500	500	500	500	500	500
Modularity (top)	0.883	0.883	0.886	0.885	0.313	0.313	0.758	0.721
Average node degree top	0.711	0.711	0.935	0.917	1.711	1.711	1.04	1.09
Avg connections within top communities	3.2	3.2	67.8	66.5	3.2	3.2	65.5	65.7
Avg. connections between top communities	0	0	0	0	0.5	0.5	1.1	1.478
Modularity (middle)	NA	NA	0.783	0.803	NA	NA	0.703	0.669
Average node degree middle	NA	NA	0.935	0.917	NA	NA	1.04	1.09
Avg connections within middle communities	NA	NA	12.156	12.222	NA	NA	12.178	12.156
Avg connections between middle communities	NA	NA	0.066	0.058	NA	NA	0.104	0.123

Table 4: Summary statistics for intermediate difficulty simulated networks.

Value	Network1	Network2	Network3	Network4	Network5	Network6
Subgraph type	small world	small world	scale free	scale free	random graph	random graph
Connection type	disc	full	disc	full	disc	full
Layers	3	3	3	3	3	3
Standard deviation	0.1	0.1	0.1	0.1	0.1	0.1
Nodes per layer	(5, 15, 300)	(5, 15, 300)	(5, 15, 300)	(5, 15, 300)	(5, 12, 167)	(5, 12, 167)
Edges per layer	(0, 15, 354)	(10, 25, 300)	(0, 10, 966)	(10, 20, 300)	(0, 7, 133)	(10, 17, 167)
Subgraph probability	0.05	0.05	0.05	0.05	0.05	0.05
Sample size	500	500	500	500	500	500
Modularity (top)	0.799	0.715	0.78	0.751	0.791	0.665
Average node degree top	1.18	1.34	3.22	3.32	0.796	0.886
Avg connections within top communities	70.8	73.6	193.2	193.2	26.6	26
Avg. connections between top communities	0	1.7	0	1.5	0	0.9
Modularity (middle)	0.781	0.679	0.873	0.845	0.787	0.696
Average node degree middle	1.18	1.34	3.22	3.32	0.796	0.886
Avg connections within middle communities	20	20	61.333	61.333	9.667	9.667
Avg connections between middle communities	0.257	0.486	0.219	0.362	0.129	0.242

Table 5: Summary statistics for simple simulated networks. These networks contain fewer than 100 nodes at the observed level and only cover small world subgraph architecture

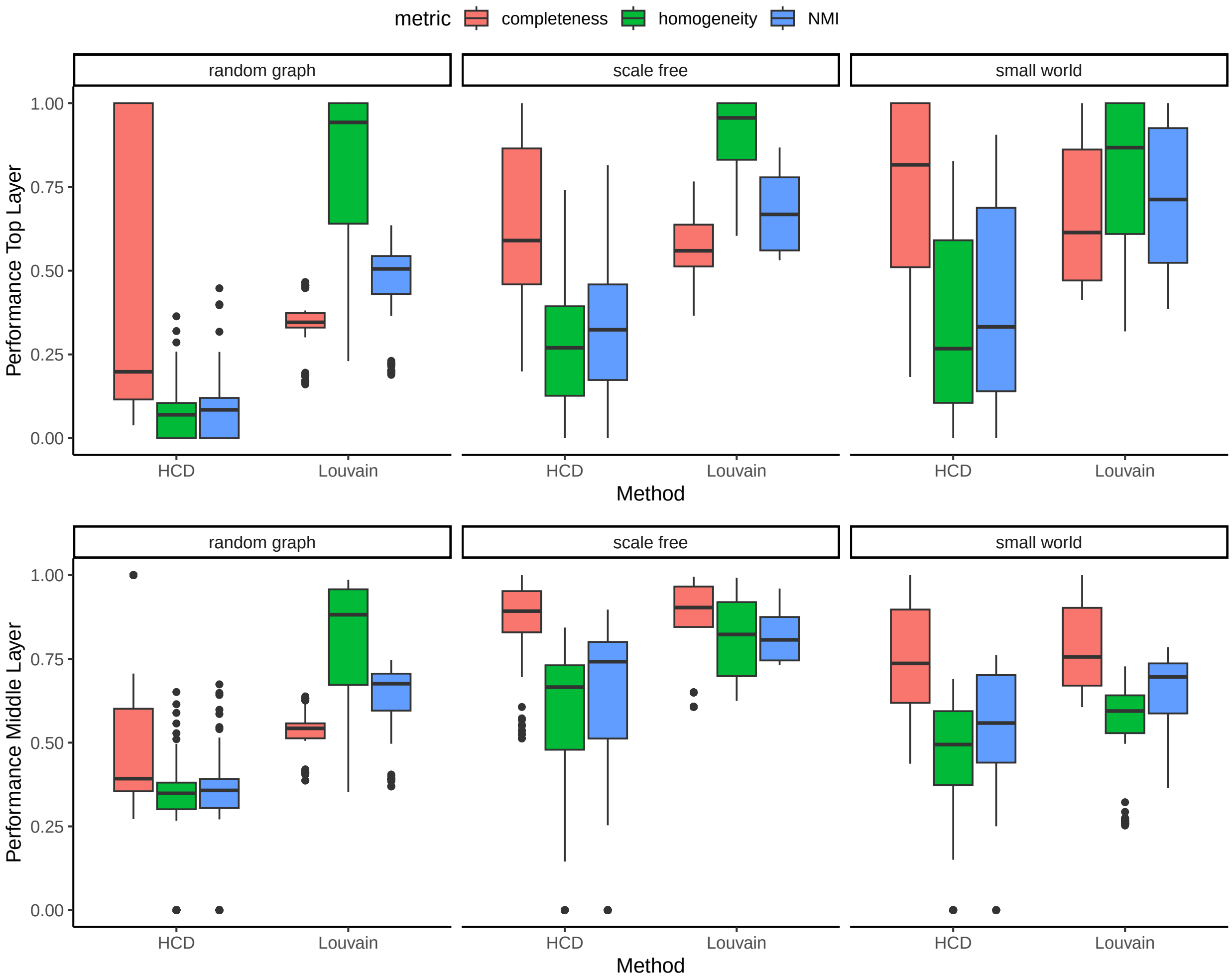
Value	Network1	Network2	Network3	Network4
Subgraph type	small world	small world	small world	small world
Connection type	disc	disc	full	full
Layers	2	3	2	3
Standard deviation	0.1	0.1	0.1	0.1
Nodes per layer	(2, 6)	(2, 6, 18)	(2, 6)	(2, 6, 18)
Edges per layer	(0, 6)	(0, 6, 24)	(1, 7)	(1, 7, 18)
Subgraph probability	0.05	0.05	0.05	0.05
Sample size	500	500	500	500
Modularity (top)	0.5	0.5	0.357	0.46
Average node degree top	1	1.333	1.167	1.389
Avg connections within top communities	3	12	3	12
Avg. connections between top communities	0	0	0.5	0.5
Modularity (middle)	NA	0.583	NA	0.553
Average node degree middle	NA	1.333	NA	1.389
Avg connections within middle communities	NA	3	NA	3
Avg connections between middle communities	NA	0.2	NA	0.233

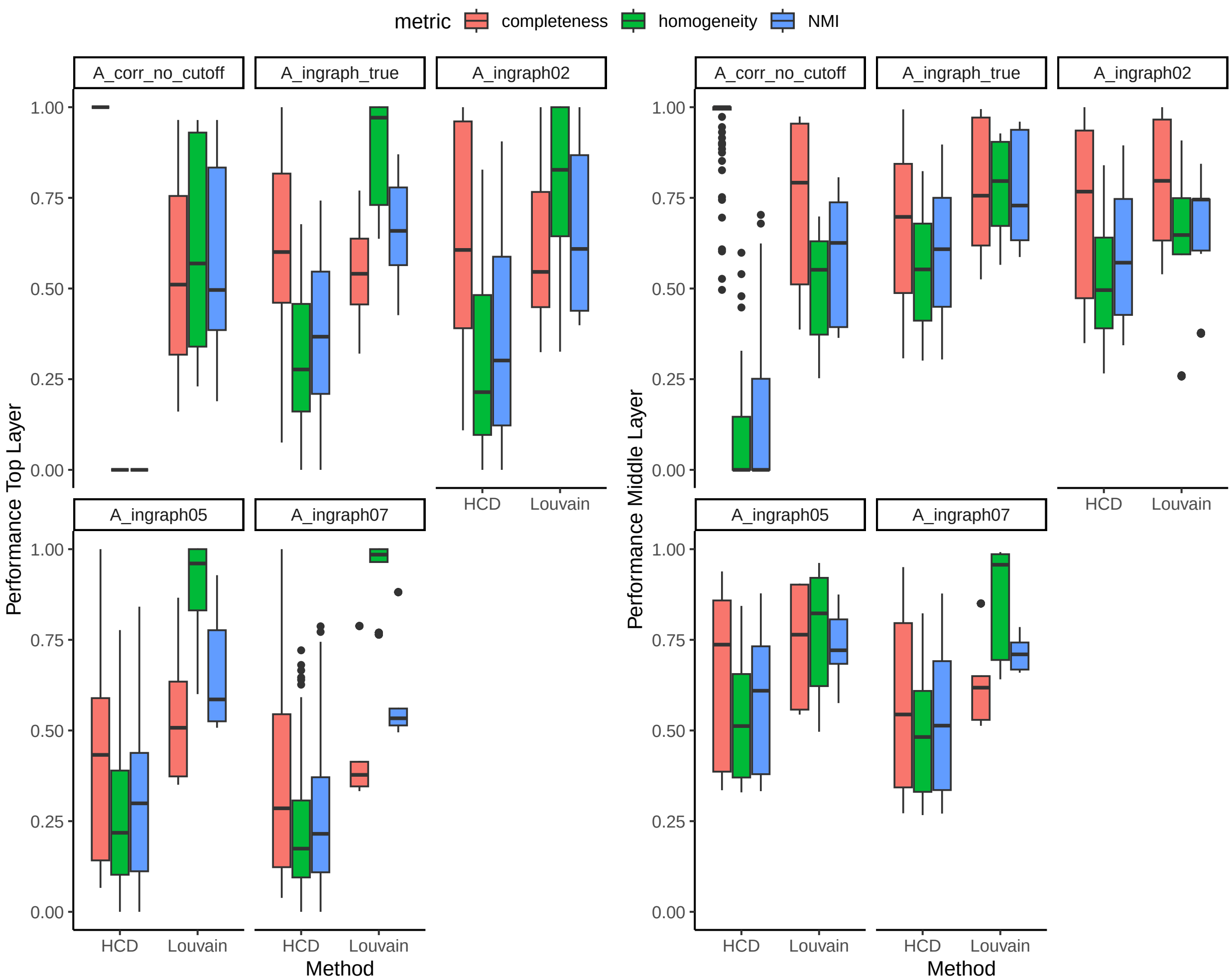
Table 6: Simulation settings for intermediate difficulty networks. Each row represents a single simulation scenario applied to all 6 simulated networks given in Table 1

Input Graph	Graph Recon. Loss	Attr. Recon. Loss	Modularity Weight	Clust. Weight
A_ingraph_true	1 = on	False (on)	1 = on	1 (middle), 1 (top)
A_corr_no_cutoff	1 = on	False (on)	1 = on	1 (middle), 1 (top)
A_ingraph02	1 = on	False (on)	1 = on	1 (middle), 1 (top)
A_ingraph05	1 = on	False (on)	1 = on	1 (middle), 1 (top)
A_ingraph07	1 = on	False (on)	1 = on	1 (middle), 1 (top)
A_ingraph_true	0 = off	False (on)	1 = on	1 (middle), 1 (top)
A_corr_no_cutoff	0 = off	False (on)	1 = on	1 (middle), 1 (top)
A_ingraph02	0 = off	False (on)	1 = on	1 (middle), 1 (top)
A_ingraph05	0 = off	False (on)	1 = on	1 (middle), 1 (top)
A_ingraph07	0 = off	False (on)	1 = on	1 (middle), 1 (top)
A_ingraph_true	1 = on	True (off)	1 = on	1 (middle), 1 (top)
A_corr_no_cutoff	1 = on	True (off)	1 = on	1 (middle), 1 (top)
A_ingraph02	1 = on	True (off)	1 = on	1 (middle), 1 (top)
A_ingraph05	1 = on	True (off)	1 = on	1 (middle), 1 (top)
A_ingraph07	1 = on	True (off)	1 = on	1 (middle), 1 (top)
A_ingraph_true	0 = off	True (off)	1 = on	1 (middle), 1 (top)
A_corr_no_cutoff	0 = off	True (off)	1 = on	1 (middle), 1 (top)
A_ingraph02	0 = off	True (off)	1 = on	1 (middle), 1 (top)
A_ingraph05	0 = off	True (off)	1 = on	1 (middle), 1 (top)
A_ingraph07	0 = off	True (off)	1 = on	1 (middle), 1 (top)
A_ingraph_true	1 = on	False (on)	0 = off	1 (middle), 1 (top)
A_corr_no_cutoff	1 = on	False (on)	0 = off	1 (middle), 1 (top)
A_ingraph02	1 = on	False (on)	0 = off	1 (middle), 1 (top)
A_ingraph05	1 = on	False (on)	0 = off	1 (middle), 1 (top)
A_ingraph07	1 = on	False (on)	0 = off	1 (middle), 1 (top)
A_ingraph_true	0 = off	False (on)	0 = off	1 (middle), 1 (top)
A_corr_no_cutoff	0 = off	False (on)	0 = off	1 (middle), 1 (top)
A_ingraph02	0 = off	False (on)	0 = off	1 (middle), 1 (top)
A_ingraph05	0 = off	False (on)	0 = off	1 (middle), 1 (top)
A_ingraph07	0 = off	False (on)	0 = off	1 (middle), 1 (top)
A_ingraph_true	1 = on	True (off)	0 = off	1 (middle), 1 (top)
A_corr_no_cutoff	1 = on	True (off)	0 = off	1 (middle), 1 (top)
A_ingraph02	1 = on	True (off)	0 = off	1 (middle), 1 (top)
A_ingraph05	1 = on	True (off)	0 = off	1 (middle), 1 (top)
A_ingraph07	1 = on	True (off)	0 = off	1 (middle), 1 (top)
A_ingraph_true	0 = off	True (off)	0 = off	1 (middle), 1 (top)
A_corr_no_cutoff	0 = off	True (off)	0 = off	1 (middle), 1 (top)
A_ingraph02	0 = off	True (off)	0 = off	1 (middle), 1 (top)
A_ingraph05	0 = off	True (off)	0 = off	1 (middle), 1 (top)
A_ingraph07	0 = off	True (off)	0 = off	1 (middle), 1 (top)
A_ingraph_true	1 = on	False (on)	1 = on	0.1 (middle), 1e-4 (top)
A_corr_no_cutoff	1 = on	False (on)	1 = on	0.1 (middle), 1e-4 (top)
A_ingraph02	1 = on	False (on)	1 = on	0.1 (middle), 1e-4 (top)
A_ingraph05	1 = on	False (on)	1 = on	0.1 (middle), 1e-4 (top)
A_ingraph07	1 = on	False (on)	1 = on	0.1 (middle), 1e-4 (top)

A_ingraph_true	0 = off	False (on)	1 = on	0.1 (middle), 1e-4 (top)
A_corr_no_cutoff	0 = off	False (on)	1 = on	0.1 (middle), 1e-4 (top)
A_ingraph02	0 = off	False (on)	1 = on	0.1 (middle), 1e-4 (top)
A_ingraph05	0 = off	False (on)	1 = on	0.1 (middle), 1e-4 (top)
A_ingraph07	0 = off	False (on)	1 = on	0.1 (middle), 1e-4 (top)
A_ingraph_true	1 = on	True (off)	1 = on	0.1 (middle), 1e-4 (top)
A_corr_no_cutoff	1 = on	True (off)	1 = on	0.1 (middle), 1e-4 (top)
A_ingraph02	1 = on	True (off)	1 = on	0.1 (middle), 1e-4 (top)
A_ingraph05	1 = on	True (off)	1 = on	0.1 (middle), 1e-4 (top)
A_ingraph07	1 = on	True (off)	1 = on	0.1 (middle), 1e-4 (top)
A_ingraph_true	0 = off	True (off)	1 = on	0.1 (middle), 1e-4 (top)
A_corr_no_cutoff	0 = off	True (off)	1 = on	0.1 (middle), 1e-4 (top)
A_ingraph02	0 = off	True (off)	1 = on	0.1 (middle), 1e-4 (top)
A_ingraph05	0 = off	True (off)	1 = on	0.1 (middle), 1e-4 (top)
A_ingraph07	0 = off	True (off)	1 = on	0.1 (middle), 1e-4 (top)
A_ingraph_true	1 = on	False (on)	0 = off	0.1 (middle), 1e-4 (top)
A_corr_no_cutoff	1 = on	False (on)	0 = off	0.1 (middle), 1e-4 (top)
A_ingraph02	1 = on	False (on)	0 = off	0.1 (middle), 1e-4 (top)
A_ingraph05	1 = on	False (on)	0 = off	0.1 (middle), 1e-4 (top)
A_ingraph07	1 = on	False (on)	0 = off	0.1 (middle), 1e-4 (top)
A_ingraph_true	0 = off	False (on)	0 = off	0.1 (middle), 1e-4 (top)
A_corr_no_cutoff	0 = off	False (on)	0 = off	0.1 (middle), 1e-4 (top)
A_ingraph02	0 = off	False (on)	0 = off	0.1 (middle), 1e-4 (top)
A_ingraph05	0 = off	False (on)	0 = off	0.1 (middle), 1e-4 (top)
A_ingraph07	0 = off	False (on)	0 = off	0.1 (middle), 1e-4 (top)
A_ingraph_true	1 = on	True (off)	0 = off	0.1 (middle), 1e-4 (top)
A_corr_no_cutoff	1 = on	True (off)	0 = off	0.1 (middle), 1e-4 (top)
A_ingraph02	1 = on	True (off)	0 = off	0.1 (middle), 1e-4 (top)
A_ingraph05	1 = on	True (off)	0 = off	0.1 (middle), 1e-4 (top)
A_ingraph07	1 = on	True (off)	0 = off	0.1 (middle), 1e-4 (top)
A_ingraph_true	0 = off	True (off)	0 = off	0.1 (middle), 1e-4 (top)
A_corr_no_cutoff	0 = off	True (off)	0 = off	0.1 (middle), 1e-4 (top)
A_ingraph02	0 = off	True (off)	0 = off	0.1 (middle), 1e-4 (top)
A_ingraph05	0 = off	True (off)	0 = off	0.1 (middle), 1e-4 (top)
A_ingraph07	0 = off	True (off)	0 = off	0.1 (middle), 1e-4 (top)

Figures


metric
completeness
homogeneity
NMI
random graph
scale free
small world
Performance Top Layer
1.00
0.75
0.50
0.25
0.00
HCD
Louvain
Method
Performance Middle Layer
Method


metric
completeness
homogeneity
NMI
A_corr_no_cutoff
A_ingraph_true
A_ingraph02
A_ingraph05
A_ingraph07
Performance Top Layer
Performance Middle Layer
1.00
0.75
0.50
0.25
0.00
HCD
Louvain
Method

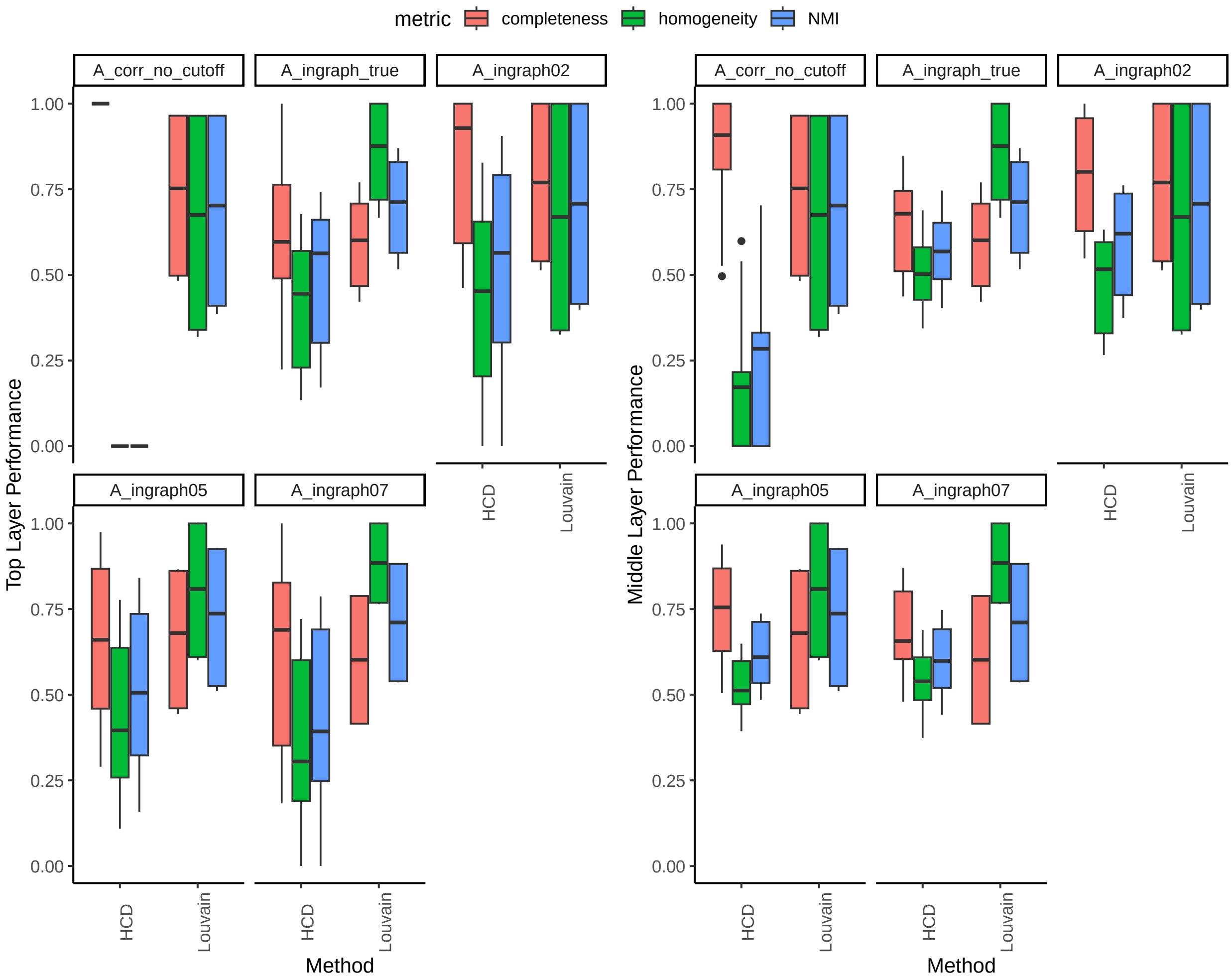


Figure 1: Small world graphs

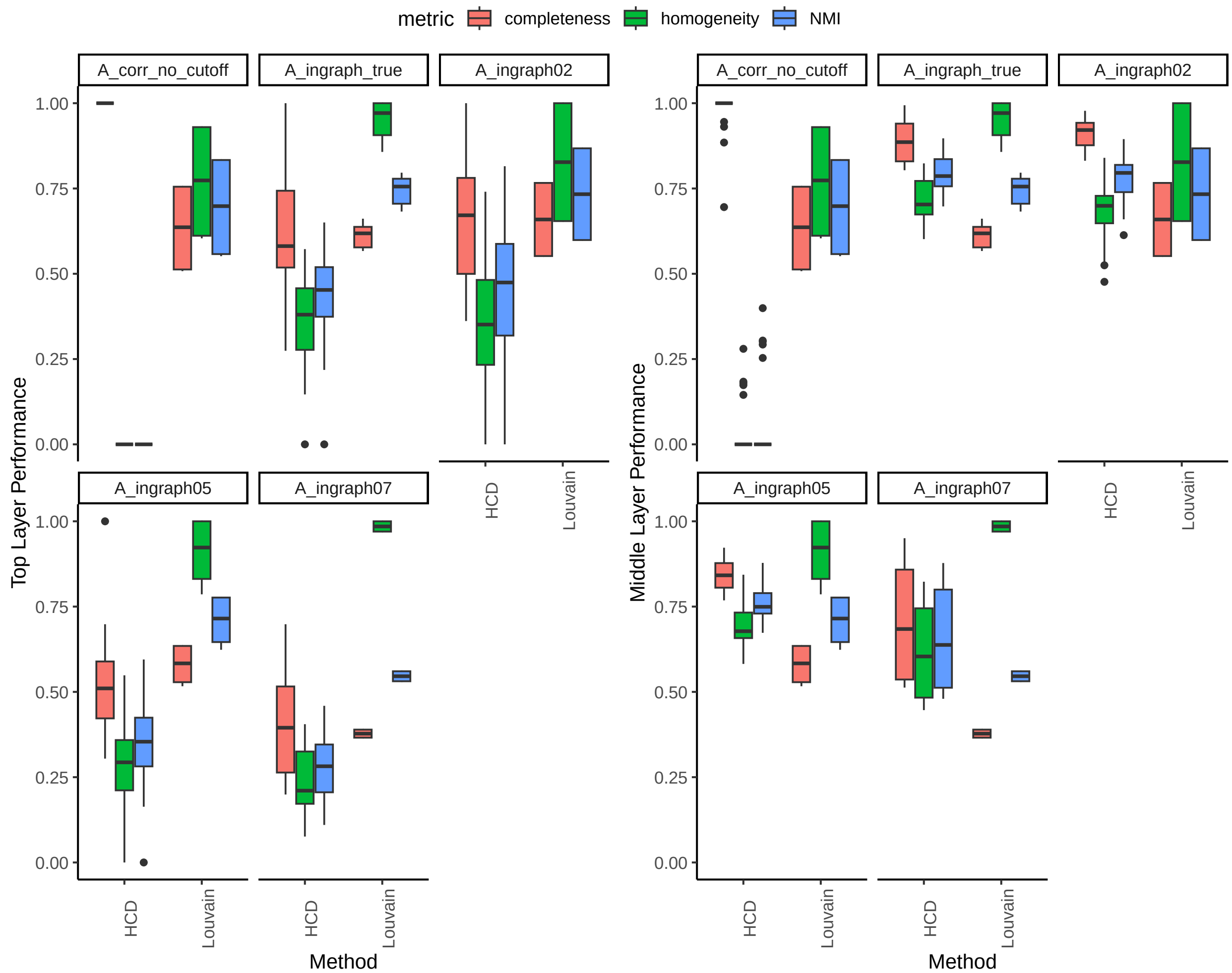


Figure 2: Scale free graphs

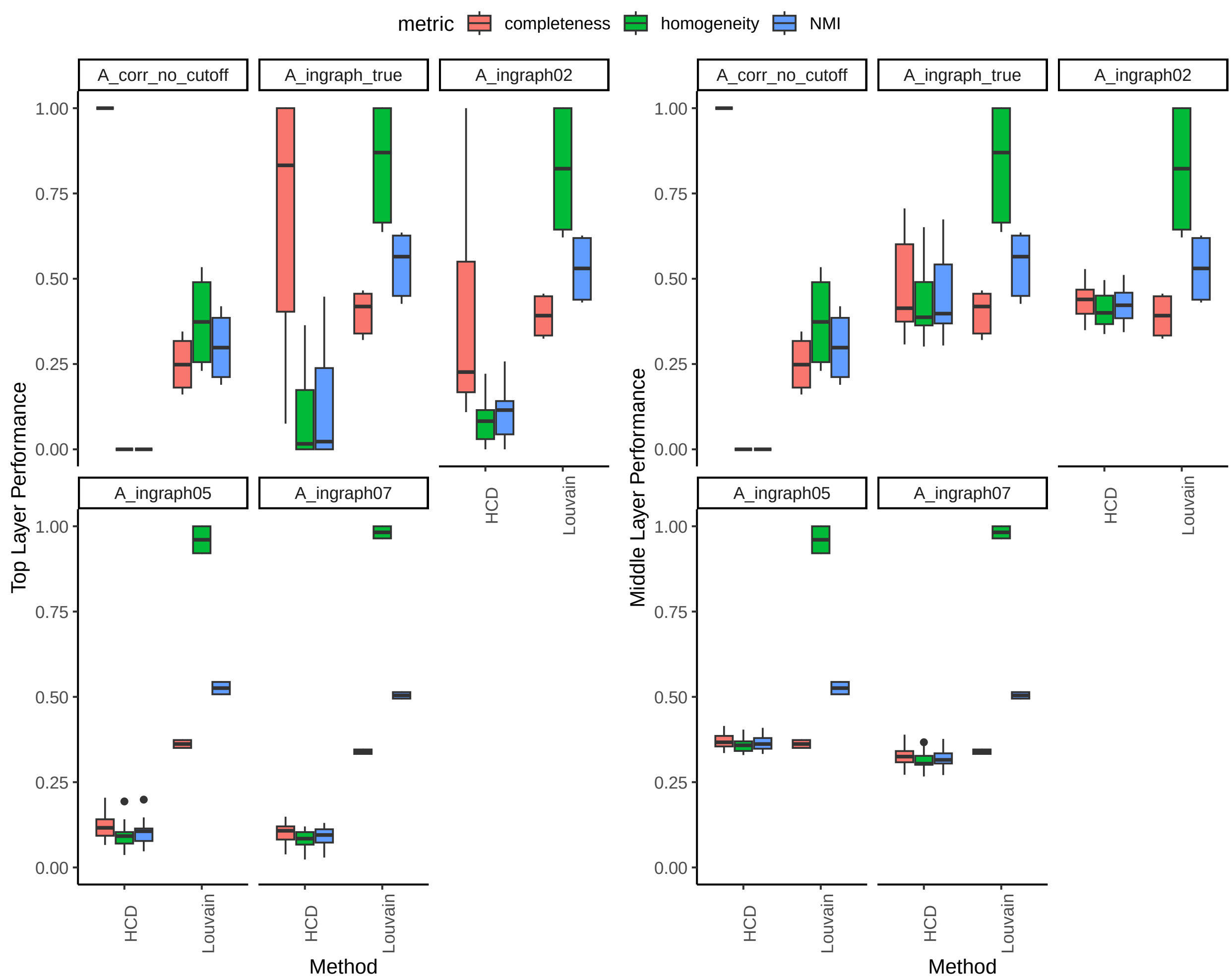


Figure 3: random graphs

References

Barabási, Albert-László, and Eric Bonabeau. 2003. "Scale-Free Networks." *Scientific American* 288 (5): 60–69.

Watts, Duncan J, and Steven H Strogatz. 1998. "Collective Dynamics of 'Small-World'networks." *Nature* 393 (6684): 440–42.