STAT293 Project on "**SIMPLE**: Statistical inference on membership profiles in large networks"

Funmi Olawole & Yuxin Liu

Network data:

- Social networks
- Biological networks
- Financial transaction networks
- Citation networks

Traditionally, research in network analysis has focused heavily on community detection — identifying groups of nodes that are more densely connected. However, a key question has been largely ignored:

"How confident are we that two nodes truly belong to the same community?"

Introduction

Motivation

"How confident are we that two nodes truly belong to the same community?"

Why important? For example:

- If you estimate that two stocks belong to the same market community, how statistically significant is this conclusion?
- If two DNA samples from tree logs are grouped closely in a similarity network, can this be used as statistically valid evidence in legal proceedings?
- If two users in a social network appear to share the same interest group, is that similarity real or just noise?

Introduction

What SIMPLE aims to solve

The goal of **SIMPLE** is to provide a valid statistical hypothesis test for any pair of nodes:

$$H_0: \ \boldsymbol{\pi}_i = \boldsymbol{\pi}_j \text{ v.s. } H_a: \ \boldsymbol{\pi}_i \neq \boldsymbol{\pi}_j,$$

where $\boldsymbol{\pi}_i \in \mathbb{R}^K$ denotes the community membership probability vector of node i.

SIMPLE provides:

- A pairwise test statistic for membership similarity.
- A way to compute p-values for how likely two nodes share the same membership.

Methodology

We observe an **undirected network** with adjacency matrix

$$X = (x_{ij})_{1 \leq i,\ j \leq n} \in \mathbb{R}^{n \times n},$$

where $x_{ij} = 1$ if there is an edge between nodes i and j, and 0 otherwise.

Each node i has a **membership probability vector** over K latent communities:

$$\boldsymbol{\pi}_i = \big(\pi_i(1), \ldots, \pi_i(K)\big), \qquad \sum_{k=1}^{K} \pi_i(k) = 1,$$

where

$$P(\text{node } i \text{ belongs to community } \mathcal{C}_k) = \pi_i(k), \qquad k = 1, \ldots, K.$$

Methodology

The observed adjacency matrix X is modeled as:

$$X = H + W,$$

where:

- H is the mean (probability) matrix, low-rank with rank K,
- W is a noise matrix with mean zero and independent entries on and above the diagonal.

The mean matrix H admits an eigen-decomposition:

$$H = VDV^\top,$$

where:

- $D = \mathrm{diag}(d_1, \ldots, d_K)$ contains the nonzero eigenvalues,
- $V = (v_1, \ldots, v_K) \in \mathbb{R}^{n \times K}$ contains the corresponding eigenvectors.

Methodology

Degree-Corrected Mixed Membership Model (DCMM)

Connection Probability

For two distinct nodes $i \neq j$, the probability of observing an edge under the DCMM model is:

$$H_{ij} = P(x_{ij} = 1) = \theta_i \theta_j \sum_{k=1}^{K} \sum_{\ell=1}^{K} \pi_i(k)\, \pi_j(\ell)\, p_{k\ell}.$$

- $\theta_i > 0$: degree heterogeneity parameter of node i.
- $\pi_i(k)$: membership proportion of node i in community k.
- $p_{k\ell}$: probability that a *typical* member of community k connects to a typical member of community ℓ.

Matrix Form

Writing the model compactly,

$$H = \Theta \, \Pi \, P \, \Pi^\top \Theta,$$

where:

- $\Theta = \mathrm{diag}(\theta_1, \ldots, \theta_n)$: degree heterogeneity matrix
- $\Pi = (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_n)^\top \in \mathbb{R}^{n \times K}$: membership matrix
- $P = (p_{k\ell}) \in \mathbb{R}^{K \times K}$: community connectivity matrix

The MMSB model arises naturally as a special case of the DCMM model. To obtain MMSB, simply remove degree heterogeneity by setting:

$$\Theta = \sqrt{\theta}\, I_n,$$

for some constant $\theta > 0$.

Then each node has the same degree parameter, and the connection probability reduces to:

$$H_{ij} = P(x_{ij} = 1) = \theta \sum_{k=1}^{K} \sum_{\ell=1}^{K} \pi_i(k)\, \pi_j(\ell)\, p_{k\ell}.$$

Matrix form:

$$H = \theta \Pi P \Pi^\top.$$

SIMPLE Test 1 (without degree heterogeneity)

- Designed for the Mixed Membership Model (MMSB).

$$H = \theta \Pi P \Pi^{\top}.$$

SIMPLE Test 2 (with degree heterogeneity)

- Designed for the general DCMM model.

$$H = \Theta \, \Pi \, P \, \Pi^{\top} \Theta,$$

Methodology

SIMPLE Test 1 (without degree heterogeneity)

Key Spectral Insight

Let

$$H = VDV^\top$$

be the eigendecomposition of the mean matrix.

- If two nodes have identical memberships:

$$\pi_i = \pi_j \quad \Rightarrow \quad V(i) = V(j).$$

- In practice, we observe only X, so we use its empirical top-K eigenvectors:

$$\hat{V} = (\hat{v}_1, \ldots, \hat{v}_K).$$

Thus, comparing rows of $\hat{V}$ directly tests whether two nodes share the same membership profile.

Methodology

SIMPLE Test 1 (without degree heterogeneity)

Test Statistic

Form the row difference:

$$d_{ij} = \hat{V}(i) - \hat{V}(j).$$

Construct statistics:

$$T_{ij} = d_{ij}^\top \Sigma_1^{-1} d_{ij},$$

where Σ_1 is the asymptotic covariance of d_{ij}.

Under the null hypothesis:

$$T_{ij} \xrightarrow{d} \chi^2_K.$$

Key Idea: Use Ratios to Remove Degree Effects

Let $\hat{v}_1$ be the leading eigenvector of X, and $\hat{v}_k$ be the others.

Define the ratio:

$$Y(i,k) = \frac{\hat{v}_k(i)}{\hat{v}_1(i)}, \qquad k = 2, \ldots, K.$$

These ratios cancel out the node-specific degree factor θ_i, because:

$$\frac{v_k(i)}{v_1(i)} \quad \text{depends only on } \pi_i.$$

Thus under the null,

$$\pi_i = \pi_j \quad \Rightarrow \quad Y_i = Y_j,$$

where

$$Y_i = (Y(i,2), \ldots, Y(i,K))^\top \in \mathbb{R}^{K-1}.$$

Methodology

SIMPLE Test 2 (with degree heterogeneity)

Test Statistic

Form the difference:

$$r_{ij} = Y_i - Y_j.$$

Construct statistics:

$$G_{ij} = r_{ij}^\top \Sigma_2^{-1} r_{ij},$$

where Σ_2 is the asymptotic covariance of r_{ij}.

Under the null hypothesis:

$$G_{ij} \xrightarrow{d} \chi^2_{K-1}.$$

Methodology

Estimation of unknown parameters

- $\hat{K}$

$$\hat{K} = \left| \left\{ \hat{d}_i : \hat{d}_i^2 > 2.01(\log n)\, \check{d}_n,\ i \in [n] \right\} \right|$$

- Σ_1

$$\Sigma_1 = \frac{1}{d_a d_b} \left\{ \sum_{t \in \{i,j\}} \sum_{l=1}^{n} \sigma_{tl}^2\, v_a(l) v_b(l) \ - \ \sigma_{ij}^2 \left[\, v_a(j) v_b(i) + v_a(i) v_b(j) \,\right] \right\}$$

- Σ_2

$$\cdots$$

Simulation - Mixed Membership Stochastic Blockmodel

Network size

- $n = 1500$ *or* 3000
- K = 3 Communities

Membership Structure

- Null case: $\pi_i = \pi_j = (1, 0, 0)$
- Alternative case $\pi_j = (1, -\theta, \theta, 0)$
- Signal strength parameter: $\theta \in 0.1, ..., 0.9$

Connectivity Matrix

- $X_{ij} \sim Bernoulli(\pi_i^T P \pi_j)$
- $P = \rho B$, with $\rho = 0.2$ SIMPLE Test Statistic

$$T_{ij} = (V_i - V_j)^T \hat{\Sigma_1^{-1}} (V_i - V_j)$$

- Threshold: $\chi^2_{3,0.95}$

Simulation - Mixed Membership Stochastic Blockmodel

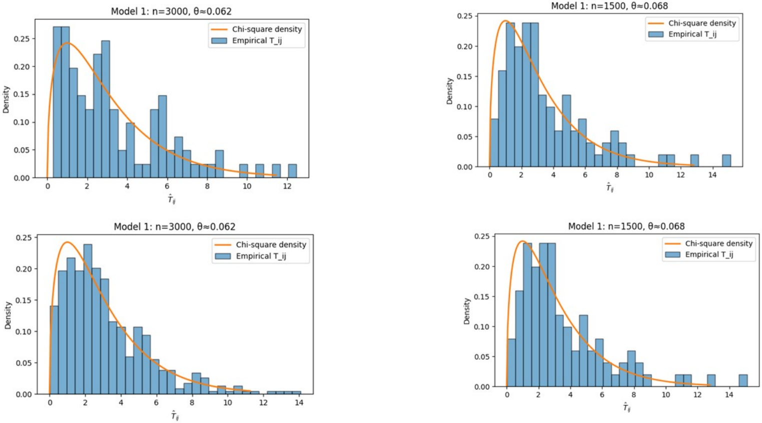


Figure: Density Plot of Model 1 SIMPLE Test Statistics

Simulation - Mixed Membership Stochastic Blockmodel

Model 1	θ	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Rep = 100	Size	0.070	0.030	0.030	0.010	0.010	0.000	0.000	0.010
	Power	0.850	0.900	1	1	1	1	1	1
Rep = 500	Size	0.046	0.022	0.022	0.016	0.004	0.010	0.008	0.004
	Power	0.896	0.982	1	1	1	1	1	1

Table: Power and Size of SIMPLE Statistics for Simulated network when n = 3000

Model 1	θ	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Rep = 100	Size	0.030	0.090	0.050	0.060	0.040	0.030	0.020	0.000
	Power	0.760	0.900	0.950	0.980	1	1	1	1
Rep = 500	Size	0.052	0.046	0.038	0.024	0.044	0.018	0.022	0.018
	Power	0.708	0.884	0.972	0.998	1	1	1	1

Table: Power and Size of SIMPLE Statistics for Simulated network when n = 1500

Simulation - Mixed Membership Stochastic Blockmodel

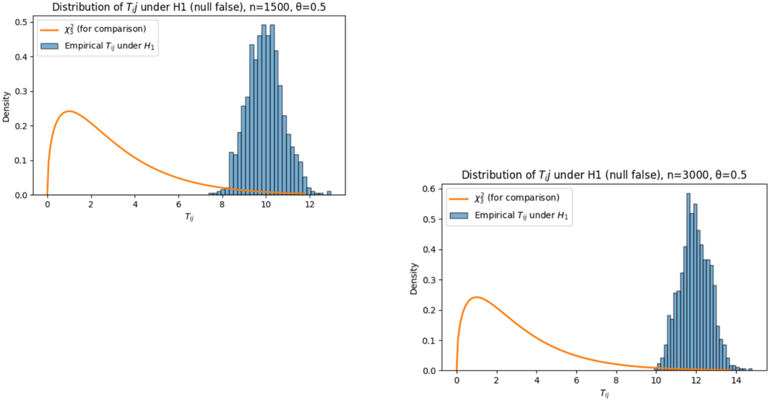


Figure: SIMPLE Test Statistics under Alternative hypothesis

Simulation - Degree-Corrected Mixed Membership Model

Network Structure

- $n = 1500$ *or* 3000
- $K = 3$ communities
- Degree parameters: $\theta_i \sim Uniform(0.5, 1.5)$

Membership Configuration

- Null pair: both nodes in pure class $\pi_i = \pi_j = (1, 0, 0)$
- Alternative pair: pure vs mixed $\pi_j = (0, 1, 0)$
  Signal strength: $r^2 = 0.1 - 0.9$
- Edge generation

$$\theta_i \theta_j \pi_i^T P \pi_j$$

- Connectivity scaled by sparsity $\rho = 0.2$

Simulation - Degree-Corrected Mixed Membership Model

Embedding and Test Statistics

- Ratio Embedding:

$$Y_i(K) = \frac{v_{k+1}(i)}{v_i(i)}$$

- SIMPLE Statistic:

$$G_{ij} = (\sqrt{n}(Y_i - Y_j))^T \hat{\Sigma_2^{-1}} (\sqrt{n}(Y_i - Y_j))$$

- Threshold: $\chi^2_{2,0.95}$
- $\hat{\Sigma_2}$ estimated from pure null group nodes

Simulation - Degree-Corrected Mixed Membership Model

Model 2	r^2	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Rep = 100	Size	0.190	0.200	0.240	0.190	0.260	0.190	0.190	0.250
	Power	0.630	0.880	0.930	1	1	1	1	1
Rep = 500	Size	0.216	0.274	0.196	0.220	0.214	0.198	0.210	0.234
	Power	0.874	0.956	0.992	0.996	1	1	1	1

Table: Power and Size of SIMPLE Statistics for Simulated network when n = 3000

Model 2	r^2	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Rep = 100	Size	0.170	0.180	0.210	0.220	0.220	0.250	0.210	0.230
	Power	0.720	0.860	0.880	0.960	1	0.990	1	1
Rep = 500	Size	0.206	0.214	0.222	0.214	0.264	0.190	0.222	0.212
	Power	0.692	0.836	0.938	0.962	0.982	0.996	1	1

Table: Power and Size of SIMPLE Statistics for Simulated network when n = 1500

Simulation - Degree-Corrected Mixed Membership Model

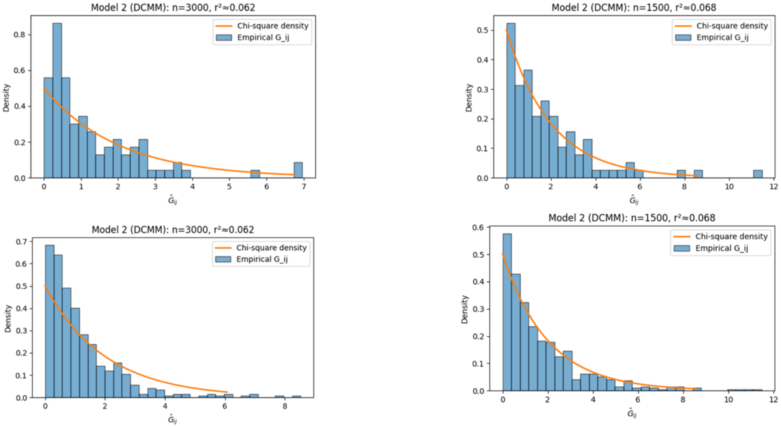
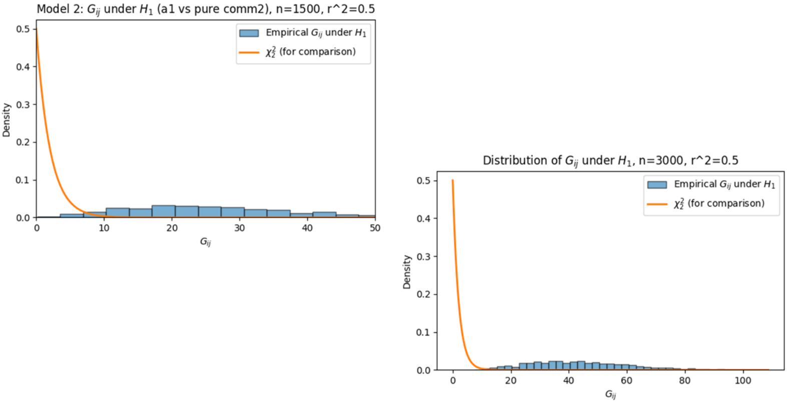


Figure: Density plot for Model 2 SIMPLE test Statistics

Simulation - Degree-Corrected Mixed Membership Model



Model 2: G_{ij} under H_1 (a1 vs pure comm2), n=1500, r^2=0.5

Distribution of G_{ij} under H_1, n=3000, r^2=0.5

Figure: SIMPLE Test Statistics for Model 2 under the Alternative Hypothesis

Sensitivity Analysis

- Assess robustness of the SIMPLE test beyond the parameter settings used in the article.
- Examine how size and power change when key parameters are varied:
- Signal strength (r^2)
- Network sparsity (ρ)
- Degree heterogeneity (Model 2)
- Validate theoretical claims by checking whether empirical behavior matches expected limiting results.
- Identify limitations or instabilities, especially under degree-corrected models.
- Ensure reproducibility and confirm that results are not artifacts of a specific parameter choice.

Sensitivity Analysis - Parameter Extension

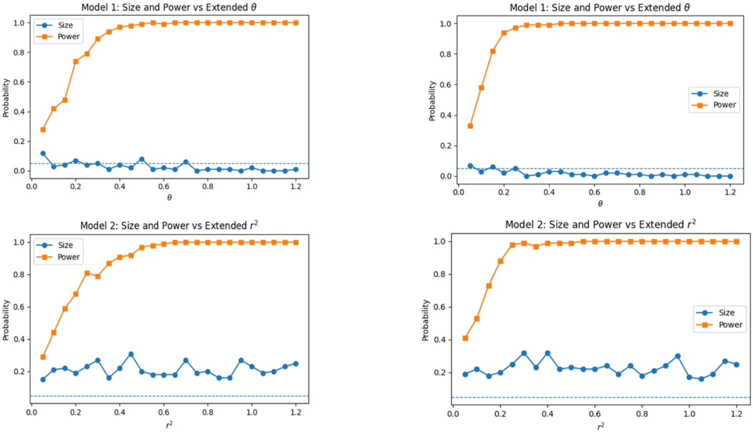


Figure: Plot of size and Power under Parameter Extension

Sensitivity Analysis - Varying Sparsity Or Density

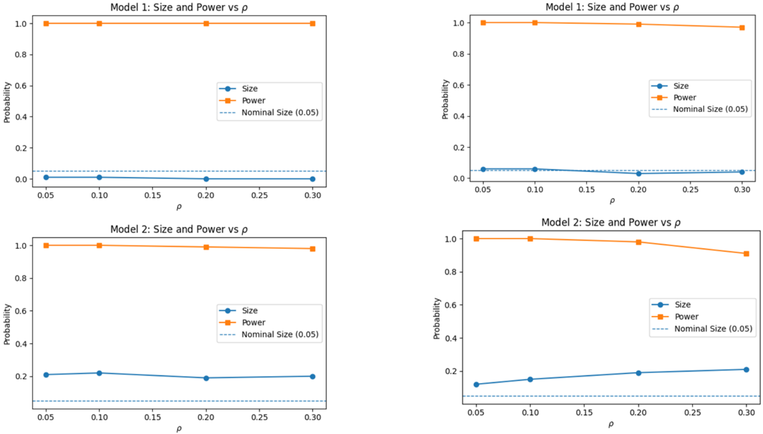
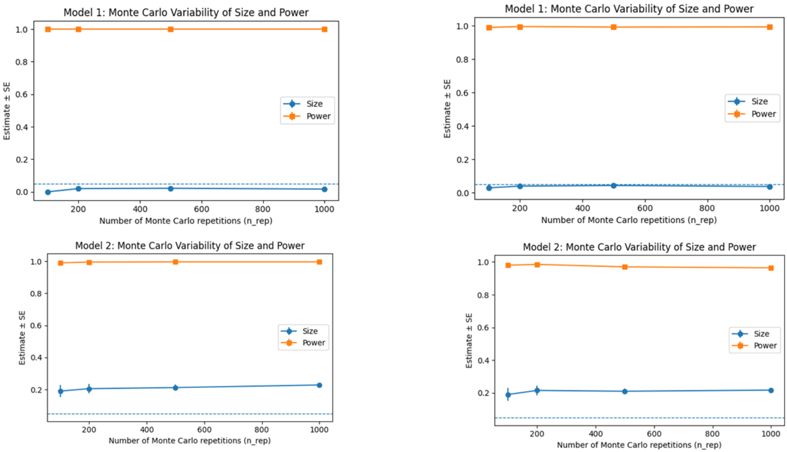


Figure: Size and Power under Varying Sparsity

Sensitivity Analysis - Monte-Carlo Variability



Model 1: Monte Carlo Variability of Size and Power

Model 1: Monte Carlo Variability of Size and Power

Model 2: Monte Carlo Variability of Size and Power

Model 2: Monte Carlo Variability of Size and Power

Figure: Size and Power with Monte-Carlo Variability

Real Data Application

Airport Network

Source: OpenFlights global airport and route database

- 14,110 airports worldwide
- 67,663 recorded flight routes
- Each airport includes name, city, country, IATA/ICAO codes, latitude and longitude

Network construction:

- Filter airports that appear in at least one route
- Build an undirected adjacency matrix based on route connections
- Final network: **3,218 airports (nodes)** and **18,858 edges**

Data Source: https://github.com/jpatokal/openflights/tree/master

Real Data Application

Airport Network

Pre-classifying Airports via Spectral Clustering

We apply spectral clustering to the airport network to create an intuitive grouping structure for the real-data illustration.

- Helps visualize the global network's structural patterns.
- Provides a convenient way to check SIMPLE p-values within and across data-driven communities.
- Serves only as a supportive tool for interpretation, not as a required step in the SIMPLE methodology.

Real Data Application

Airport Network

Spectral Clustering Procedure

- Compute the top-K eigenvectors of the adjacency matrix A
- Row-normalize the $n \times K$ eigenvector matrix
- Run K-means on the spectral embedding
- Assign each airport to one of K preliminary clusters

Real Data Application

Airport Network

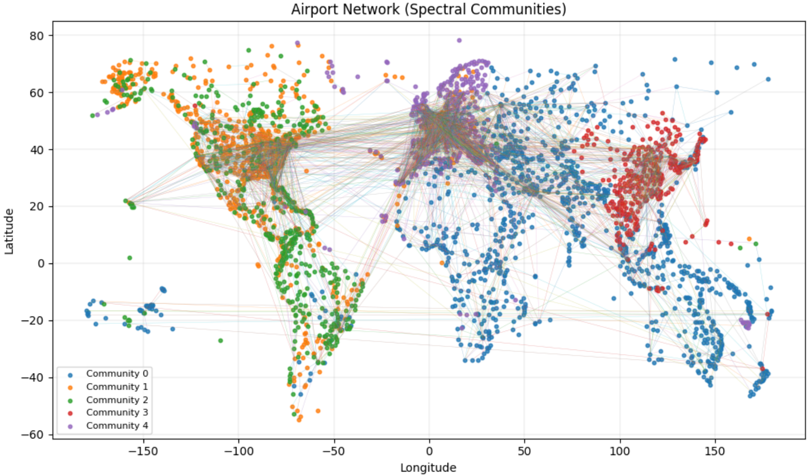


Figure: Cluster assignments of the airport network obtained by spectral clustering with $K = 5$. Each dot represents an airport, colored according to its spectral community.

Real Data Application

Airport Network

Across-cluster airports:

	TA (C0)	CVRA (C1)	WA (C2)	NSTA (C3)	BA (C4)
Toliara Airport (C0)	1.0000	0.7914	0.0000	0.9949	0.7210
Chippewa Valley Regional Airport (C1)	0.7915	1.0000	0.0000	0.7309	0.5860
Wold-Chamberlain Airport (C2)	0.0000	0.0000	1.0000	0.0000	0.0000
Nakhon Si Thammarat Airport (C3)	0.9949	0.7309	0.0000	1.0000	0.6188
Badajoz Airport (C4)	0.7210	0.5860	0.0000	0.6188	1.0000

Table: P-values for 1 randomly selected airport per cluster on Model 1 with $K = 5$.

	TA (C0)	CVRA (C1)	WA (C2)	NSTA (C3)	BA (C4)
Toliara Airport (C0)	1.0000	0.0001	0.0000	0.0000	0.7587
Chippewa Valley Regional Airport (C1)	0.0001	1.0000	0.1474	0.0000	0.0000
Wold-Chamberlain Airport (C2)	0.0000	0.1474	1.0000	0.0000	0.0000
Nakhon Si Thammarat Airport (C3)	0.0000	0.0000	0.0000	1.0000	0.0000
Badajoz Airport (C4)	0.7587	0.0000	0.0000	0.0000	1.0000

Table: P-values for 1 randomly selected airport per cluster on Model 2 with $K = 5$.

Real Data Application

Airport Network

Within-cluster airports:

	SEMA	SFXA	PA	LAA	JDRA
Senadora Eunice Micheles Airport	1.0000	1.0000	0.9949	0.9794	0.4409
São Félix do Xingu Airport	1.0000	1.0000	0.9948	0.9793	0.4404
Perales Airport	0.9949	0.9948	1.0000	0.9993	0.6250
Los Alamos Airport	0.9794	0.9793	0.9993	1.0000	0.5793
Jardines Del Rey Airport	0.4409	0.4404	0.6250	0.5793	1.0000

Table: P-values for 5 randomly selected airports in Cluster 2 on Model 1 with $K = 5$.

	SEMA	SFXA	PA	LAA	JDRA
Senadora Eunice Micheles Airport (C2)	1.0000	0.0092	0.0933	0.1269	0.0001
São Félix do Xingu Airport (C2)	0.0092	1.0000	0.0001	0.0559	0.0000
Perales Airport (C2)	0.0933	0.0001	1.0000	0.0003	0.2725
Los Alamos Airport (C2)	0.1269	0.0559	0.0003	1.0000	0.0000
Jardines Del Rey Airport (C2)	0.0001	0.0000	0.2725	0.0000	1.0000

Table: P-values for 5 randomly selected airports in Cluster 2 on Model 2 with $K = 5$.

Real Data Application

Airport Network

Conclusions for case $K = 5$

- **Across-cluster airports:** P-values using SIMPLE (both models) are mostly close to 0, meaning airports in different regions have very different connectivity/membership patterns.
- **Within-cluster airports:** The 5-airport tests inside cluster 2 show mixed p-values, ranging from near 0 to moderate. Some airports share similar connectivity, others are structurally different even within the same broad region.

Real Data Application

Airport Network

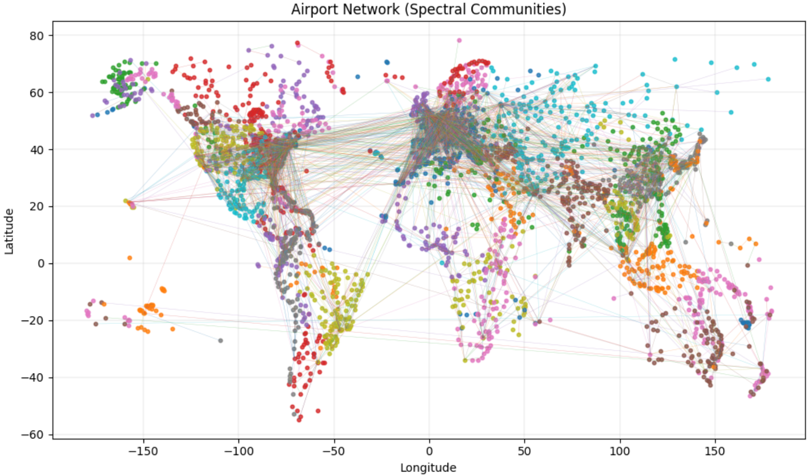


Figure: Cluster assignments of the airport network obtained by spectral clustering with $K = 50$. Each dot represents an airport, colored according to its spectral community.

Real Data Application

Airport Network

Across-cluster airports:

	MA (C0)	MU (C1)	BA (C2)	TCA (C3)	LBPIA (C4)
Manchester Airport (C0)	1.0000	0.0000	0.0000	0.0000	0.0000
Mulu Airport (C1)	0.0000	1.0000	1.0000	1.0000	0.0000
Buckland Airport (C2)	0.0000	1.0000	1.0000	1.0000	0.0000
Treasure Cay Airport (C3)	0.0000	1.0000	1.0000	1.0000	0.0000
Lester B. Pearson International Airport (C4)	0.0000	0.0000	0.0000	0.0000	1.0000

Table: P-values for 1 randomly selected airport per cluster on Model 1 with $K = 50$.

	MA (C0)	MU (C1)	BA (C2)	TCA (C3)	LBPIA (C4)
Manchester Airport (C0)	1.0000	0.0000	0.0000	0.0000	0.0000
Mulu Airport (C1)	0.0000	1.0000	0.0000	0.0000	0.0000
Buckland Airport (C2)	0.0000	0.0000	1.0000	0.0000	0.0000
Treasure Cay Airport (C3)	0.0000	0.0000	0.0000	1.0000	0.0000
Lester B. Pearson International Airport (C4)	0.0000	0.0000	0.0000	0.0000	1.0000

Table: P-values for 1 randomly selected airport per cluster on Model 2 with $K = 50$.

Real Data Application

Airport Network

Within-cluster airports:

	TA	EA	AA	BATGF	NA
Teller Airport	1.0000	1.0000	1.0000	0.0000	1.0000
Emmonak Airport	1.0000	1.0000	1.0000	0.0000	1.0000
Ambler Airport	1.0000	1.0000	1.0000	0.0000	1.0000
Boise Air Terminal/Gowen Field	0.0000	0.0000	0.0000	1.0000	0.0000
Napaskiak Airport	1.0000	1.0000	1.0000	0.0000	1.0000

Table: P-values for 5 randomly selected airports in Cluster 2 on Model 1 with $K = 50$.

	TA	EA	AA	BATGF	NA
Teller Airport (C2)	1.0000	0.0000	0.0000	0.0000	0.0000
Emmonak Airport (C2)	0.0000	1.0000	0.0000	0.0000	0.0000
Ambler Airport (C2)	0.0000	0.0000	1.0000	0.0000	0.0000
Boise Air Terminal/Gowen Field (C2)	0.0000	0.0000	0.0000	1.0000	0.0000
Napaskiak Airport (C2)	0.0000	0.0000	0.0000	0.0000	1.0000

Table: P-values for 5 randomly selected airports in Cluster 2 on Model 2 with $K = 50$.

Real Data Application

Airport Network

Conclusions for case $K = 50$

- **Across-cluster airports:** Shows even stronger contrast. Many cross-cluster p-values are exactly 0.0000. Model 2 gives 0.0000 for all cross-cluster.
- **Within-cluster airports:** Even airports within the same cluster often have p-values of 0 (Model 2) or sometimes 1 (Model 1 in some cases).

Real Data Application

Airport Network

- Spectral clustering gives a useful initial partition of the network. However, SIMPLE reveals that even within spectral clusters, many airports differ significantly.
- This behavior may also reflect the fact that the pre-classification obtained from spectral clustering encodes a different structural notion than the membership similarity assessed by SIMPLE.

Conclusions

- SIMPLE provides a reliable framework for testing whether two nodes share the same latent membership profile in large networks.
- Simulation studies show high power, controlled size, and stable performance under varying signal strengths, sparsity levels, and degree heterogeneity.
- In the real-world airport network, SIMPLE indicates structural differences across and within clusters, providing an inference to clustering results.
- Overall, SIMPLE is a robust and flexible tool for statistical inference on membership profiles in complex network data.

Reference

Jianqing Fan, Yingying Fan, Xiao Han, and Jinchi Lv. Simple: Statistical inference on membership profiles in large networks, 2021.

Thank you