STAT 293 Project on "SIMPLE: Statistical inference on membership profiles in large networks"

Funmi Olawole & Yuxin Liu

Abstract

This project studies the SIMPLE framework for statistical inference on membership profiles in large networks. The procedure is developed under both the Mixed Membership Model (MM) and the more flexible Degree-Corrected Mixed Membership Model (DCMM), allowing inference on networks with or without degree heterogeneity. We present the theoretical foundations of the two SIMPLE test statistics and discuss the regularity conditions under which the statistics converge to chi-square limits. We evaluate the empirical behavior of SIMPLE through simulations, confirming high power, stability under MMSB, predictable size inflation under DCMM, and robustness across different signal and sparsity levels. Additional sensitivity analyses show how the test's size and power vary with signal strength and network sparsity, demonstrating its stability under a wide range of parameter settings. We also apply SIMPLE to a real-world network dataset to illustrate its practical utility for membership inference. Our findings suggest that SIMPLE test performs as theoretically expected under the MMSB model, with well-controlled size, increasing power, and strong robustness across different signal strengths and sparsity levels. Under the DCMM model, the method exhibits predictable size inflation due to degree heterogeneity but maintains consistently high power across all settings.

1 Introduction

Network-structured data have become increasingly prominent across scientific and industrial domains, including online social platforms, protein–protein interaction networks, financial transaction systems, transportation infrastructures, and citation graphs. A fundamental goal in analyzing such data is to uncover the latent organizational principles that drive interactions among nodes. Among these tasks, community detection has emerged as a central problem. Traditional approaches typically focus on generating point estimates of community labels or membership vectors that summarize each node's role within the network. The stochastic block model (SBM) [3] characterizes edge formation through probabilities that depend exclusively on the nodes' community memberships for community detection. Spectral clustering [6] partitions nodes by embedding them into a low-dimensional space using the eigenvectors of a graph Laplacian and then applying a standard clustering algorithm. Mixed-membership models [1] allow each node to belong to multiple communities simultaneously by modeling edges from community pairs drawn from node-specific membership distributions. The degree-corrected SBM [5] extends the SBM by introducing node-specific degree parameters, allowing for accurate modeling of networks with wide degree heterogeneity.

However, existing community detection methods do not provide tools for quantifying uncertainty in membership assignments. This limitation raises the question: to what extent can we assert, with statistical confidence, that two nodes belong to the same community or share similar membership profiles? To fill this gap, the SIMPLE framework [2] provides a hypothesis-testing procedure that evaluates membership similarity using asymptotically test statistics for node-level community relationships in large networks. Importantly, the methodology applies to both the Mixed Membership Model (MM) and the more general Degree-Corrected Mixed Membership Model (DCMM), allowing it to handle networks with or without degree heterogeneity.

In this project, we conduct a detailed study of the SIMPLE framework. Our work focuses on clarifying the theoretical foundations and methodological components of the SIMPLE framework. We conduct a simulation study to evaluate the finite-sample performance of both test statistics under different network structures. In addition, we perform sensitivity analyses to examine how variations in model parameters affect the behavior of SIMPLE. Finally, we apply SIMPLE to a real-world dataset to demonstrate its applicability and to highlight the insights that formal membership inference can provide in practical network analysis.

2 Methodology

2.1 Problem Setup

Consider an undirected graph $\mathcal{N} = (V, E)$ with n nodes, where $V = \{1, \ldots, n\}$ denotes the set of nodes and E denotes the set of undirected edges. We observe an undirected network on n nodes, represented by its adjacency matrix

$$X = (x_{ij})_{1 \leq i,j \leq n}, \tag{1}$$

where

$$x_{ij} = \begin{cases} 1, & \text{if there is an edge between nodes } i \text{ and } j, \\ 0, & \text{otherwise,} \end{cases}$$

with $x_{ij} = x_{ji}$ and $x_{ii} = 0$ for all i. The model assumes that X can be decomposed as

$$X = H + W, \tag{2}$$

where $H = (h_{ij})$ is the mean (probability) matrix, which encodes the underlying community structure. In particular, h_{ij} represents the edge probability between nodes i and j, so that

$$x_{ij} \mid H \sim \text{Bernoulli}(h_{ij}), \qquad 1 \leq i < j \leq n. \tag{3}$$

$W = (w_{ij})$ is a noise matrix with independent mean-zero entries on and above the diagonal, defined by $w_{ij} = x_{ij} - h_{ij}$. Typically, W is assumed to have a bounded variance and satisfy $w_{ij} = w_{ji}$ and $w_{ii} = 0$.
Each node i is assumed to belong to K latent communities in varying proportions. This is modeled by a membership vector

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

where $\pi_i(k) \geq 0$ for all $k = 1, \ldots, K$ and $\sum_{k=1}^K \pi_i(k) = 1$. Given this representation, a fundamental question is whether two nodes exhibit indistinguishable community behaviors. Formally, for any pair of nodes (i, j), the following hypothesis testing problem is considered in SIMPLE:

$$H_0 : \boldsymbol{\pi}_i = \boldsymbol{\pi}_j \qquad \text{versus} \qquad H_a : \boldsymbol{\pi}_i \neq \boldsymbol{\pi}_j. \tag{4}$$

Under H_0, nodes i and j share the same probabilistic interaction pattern with all other nodes in the network, implying that the model views them as structurally equivalent. In contrast, the alternative H_a states that the two nodes differ in at least one community coordinate.

2.2 Model Setting

The SIMPLE framework is developed under two closely related probabilistic models for network data: the Degree-Corrected Mixed Membership Model (DCMM) [4] and the Mixed Membership Model (MM) [1]. DCMM serves as the more general formulation, allowing nodes to have heterogeneous degrees, while MM emerges as a special case in which all nodes share a common degree parameter. In this section, we will first introduce the DCMM model and its main characteristics, and then clarify the MM model as a simplified version obtained by removing degree heterogeneity.

2.2.1 The Degree-corrected Mixed Membership (DCMM) Model

Empirical networks often exhibit degree heterogeneity as some nodes have extremely high degrees, while many others have only a few connections. The classical MM model cannot capture this variability, because all nodes with similar membership vectors have similar expected degrees. The SIMPLE test is then also considered for adaptation to the DCMM model, which introduces a node-specific degree parameter $\theta_i > 0$ to account for such variability.
For two distinct nodes $i \neq j$, the probability of an edge is define as

$$P(x_{ij} = 1) = \theta_i \theta_j \sum_{k=1}^K \sum_{l=1}^K \pi_i(k)\, \pi_j(l)\, p_{kl}, \tag{5}$$

where θ_i controls the expected degree of node i, $\pi_i(k)$ is the membership proportion of node i in community k, and p_{kl} is the baseline probability that a typical member of community k connects to a typical member of community l.

Let

$$\Pi = (\pi_1, \dots, \pi_n)^\top \in \mathbb{R}^{n\times K}, \qquad P = (p_{kl}) \in \mathbb{R}^{K\times K},$$

and let

$$\Theta = \mathrm{diag}(\theta_1, \dots, \theta_n).$$

Then the mean adjacency matrix $H = \mathbb{E}[X]$ can be written compactly as

$$H = \Theta\,\Pi\,P\,\Pi^\top\Theta. \tag{6}$$

In this matrix representation, Θ introduces degree heterogeneity, Π encodes mixed membership of all nodes, and P describes the interaction between the K communities.

2.2.2 The Mixed Membership Stochastic Block (MM) Model

The MM model is a widely used model for networks in which each node may participate in multiple communities simultaneously. In MM model, all nodes share a common degree parameter $\theta > 0$, so degree heterogeneity is not modeled. It can be recovered as the special case of DCMM model where

$$\theta_i = \sqrt{\theta} \quad \text{for all } i,$$

so that all nodes have identical expected degree scaling. For two distinct nodes $i \neq j$, the edge probability is

$$H_{ij} = P(x_{ij} = 1) = \theta \sum_{k=1}^{K}\sum_{l=1}^{K} \pi_i(k)\,\pi_j(l)\,p_{kl}, \tag{7}$$

and the mean adjacency matrix satisfies

$$H = \theta\,\Pi P \Pi^\top, \tag{8}$$

which corresponds to the DCMM mean matrix $H = \Theta\Pi P\Pi^\top\Theta$ under the specialization $\Theta = \sqrt{\theta}\,I_n$.

2.3 SIMPLE

2.3.1 SIMPLE for mixed membership models

We first consider the hypothesis testing problem in eq. (4) under the mixed membership models that allows each node to participate in multiple communities through a latent membership probability vector, but crucially, the model does not incorporate degree heterogeneity. All nodes share the same degree parameter, leading to a homogeneous expected degree pattern across the network.

The mean adjacency matrix $H = \mathbb{E}(X)$ is symmetric and of rank at most K. Because H is symmetric, it admits an eigen-decomposition of the form

$$H = VDV^\top, \tag{9}$$

where $D = \mathrm{diag}(d_1, \dots, d_K)$ is a diagonal matrix whose diagonal entries $d_1, \dots, d_K$ are the nonzero eigenvalues of H and $V = (v_1, \dots, v_K)$ is the $n \times K$ matrix whose columns are the corresponding orthonormal eigenvectors.

A property under eq. (8) is that the eigenvector matrix V is linearly related to the membership matrix Π. More precisely, there exists an invertible matrix $B \in \mathbb{R}^{K\times K}$ such that

$$V = \Pi B. \tag{10}$$

Therefore, two nodes have identical membership vectors if and only if they have identical population spectral embeddings:

$$\pi_i = \pi_j \quad \Longrightarrow \quad V(i) = V(j),$$

where $V(i)$ is the ith row of the matrix V. In practice, we work with the empirical eigenvectors $\hat{V}$ computed from the observed matrix X. Motivated by this observation, SIMPLE constructs the statistic as

$$T_{ij} = \big[\hat{V}(i) - \hat{V}(j)\big]^{\top}\Sigma_1^{-1}\big[\hat{V}(i) - \hat{V}(j)\big], \tag{11}$$

where Σ_1 is the asymptotic variance of $\hat{V}(i) - \hat{V}(j)$. The (a, b)th entry of matrix Σ_1 can be estimated by

$$\frac{1}{d_a d_b}\left\{\sum_{t\in\{i,j\}}\sum_{l=1}^{n}\sigma_{tl}^2\, v_a(l)\, v_b(l) \;-\; \sigma_{ij}^2\big[\, v_a(j)\, v_b(i) + v_a(i)\, v_b(j)\,\big]\right\}, \tag{12}$$

where $\sigma_{ab}^2 = \mathrm{var}(w_{ab})$ for $1 \le a, b \le n$.

Following regularity conditions are needed to establish the asymptotic null and alternative distributions of test statistic T_{ij}.

Condition 1 requires sufficient separation among the nonzero eigenvalues of the population matrix H: there exists a positive constant c_0 such that

$$\min\left\{\left|\frac{d_i}{d_j}\right| : 1 \le i < j \le K,\ d_i \ne -d_j\right\} \ge 1 + c_0,$$

and, in addition, $a_n \to \infty$ as $n \to \infty$.

Condition 2 imposes lower bounds on the strength of the community signal. Specifically, there exist constants $0 < c_0 < 1$, $0 \le c_2 < 1/2$, and $0 < c_1 < 1 - 2c_2$ such that

$$\lambda_K(\Pi^{\top}\Pi) \ge c_0 n, \qquad \lambda_K(P) \ge n^{-c_2}, \qquad \theta \ge n^{-c_1}.$$

These conditions ensure that the membership matrix and the block connectivity matrix are sufficiently well-conditioned, and that the common degree parameter does not vanish too quickly.

Condition 3 further assumes that, as $n \to \infty$, all eigenvalues of the matrix $\theta^{-1}D\Sigma_1 D$ remain bounded away from both zero and infinity, which guarantees the regularity of the covariance structure associated with the spectral perturbation.

Under Conditions 1–3, the statistic T_{ij} converges in distribution to a chi-square random variable with K degrees of freedom; that is,

$$T_{ij} \xrightarrow{d} \chi_K^2 \qquad \text{as } n \to \infty. \tag{13}$$

2.3.2 SIMPLE for degree-corrected mixed membership models

We now describe the hypothesis testing problem in eq. (4) under the degree-corrected mixed membership model, which introduces node-specific degree parameters. Unlike MM, where all nodes share a common degree scale, DCMM allows for substantial degree heterogeneity across nodes. This complicates the spectral structure of the mean adjacency matrix and requires a different treatment of the eigenvectors. In particular, raw eigenvector rows no longer encode the membership profiles directly.

To remove the effect of degree parameters, SIMPLE uses a ratio transformation based on the leading eigenvector. Let

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

be the empirical eigenvector matrix computed from the observed adjacency matrix X. For each node i, define the ratio vector

$$Y_i = \left(\frac{\hat{v}_2(i)}{\hat{v}_1(i)}, \frac{\hat{v}_3(i)}{\hat{v}_1(i)}, \ldots, \frac{\hat{v}_K(i)}{\hat{v}_1(i)}\right) \in \mathbb{R}^{K-1}.$$

At the population level, we have

$$Y_i = \frac{V(i,:)}{v_1(i)} = \frac{\theta_i \pi_i^{\top} B}{\theta_i \pi_i^{\top} b_1}, \tag{14}$$

where b_1 is the first column of B. The factor θ_i cancels, yielding

$$Y_i = \pi_i^{\top} B^*, \tag{15}$$

for some invertible matrix B^*. This shows that the ratio embedding recovers the membership vector up to a linear transformation. More importantly,

$$\pi_i = \pi_j \quad \Longrightarrow \quad Y_i = Y_j.$$

Based on these, SIMPLE constructs the statistic as

$$G_{ij} = (Y_i - Y_j)^\top \Sigma_2^{-1} (Y_i - Y_j), \tag{16}$$

where Σ_2 is the asymptotic variance of $Y_i - Y_j$ and can be estimated.
Following regularity conditions are needed to establish the asymptotic null and alternative distributions of test statistic G_{ij}.
Condition 4 requires sufficient signal strength in both the membership and degree components: there exist constants $c_2 \in [0, 1/2)$, $c_3 \in (0, 1 - 2c_2)$, $c_5 \in (0, 1)$, and $c_4 > 0$ such that

$$\lambda_K(P) \geq n^{-c_2}, \qquad \min_{1 \leq k \leq K} |\mathcal{N}_k| \geq c_5 n, \qquad \theta_{\max} \leq c_4 \theta_{\min}, \qquad \theta_{\min}^2 \geq n^{-c_3}.$$

Condition 5 assumes that the matrix $P = (p_{k\ell})$ is positive definite, irreducible, and has unit diagonal entries. Moreover, the noise level must satisfy

$$n \min_{1 \leq k \leq K,\, t = i,j} \operatorname{var}(e_t^\top W v_k) \sim n\theta_{\max}^2 \ \to \ \infty.$$

Condition 6 further requires that all eigenvalues of

$$(n\theta_{\max}^2)^{-1} \, D \operatorname{cov}(f) \, D$$

are bounded away from both zero and infinity, ensuring the regularity of the covariance matrix associated with the ratio-based embedding.
Condition 7 introduces η_1, the first right singular vector of $P^{1/2}\Pi^\top \Theta^2 \Pi$, and imposes a mild technical condition on its alignment with the population eigenspace. This guarantees the identifiability of the ratio embedding.
Under Conditions 4–7, the statistic G_{ij} converges in distribution to a chi-square random variable with $K - 1$ degrees of freedom:

$$G_{ij} \xrightarrow{d} \chi^2_{K-1}, \qquad n \to \infty. \tag{17}$$

Notably, the implementation of SIMPLE does not assume that the true number of communities K is known a priori. Rather, K can be inferred from the observed adjacency matrix through a consistent eigenvalue thresholding rule. The proposed estimator is defined by

$$\widehat{K} = \left| \left\{ \hat{d}_i : \ \hat{d}_i^2 > 2.01 (\log n) \, \check{d}_n, \quad i \in [n] \right\} \right|, \tag{18}$$

where $|\cdot|$ denotes the cardinality of a set, $\hat{d}_i$ is the i-th eigenvalue of the observed adjacency matrix X, and $\check{d}_n = \max_{1 \leq i \leq n} \sum_{j=1}^n X_{ij}$ is the maximum degree of the network. In other words, $\widehat{K}$ counts the number of eigenvalues of X whose squared magnitudes exceed a noise-dependent threshold proportional to $(\log n)\check{d}_n$. The paper claims that $\widehat{K}$ is a consistent estimator of the true number of communities K.

3 Implementation:

The implementation follows directly from the procedures outlined in the article - Although the authors didn't make their code accessible on public fora, we are able to reproduce their claim by implementing their methods. This implementation covers both the Mixed Membership Stochastic Blockmodel(MMSB) and the Degree-Corrected Mixed Membership Model(DCMM). All components were implemented in Python, including the network simulations, construction of spectral and ratio embeddings, estimation of covariance matrices, computation of the test statistics T_{ij} and G_{ij}, as well as a comprehensive set of Monte-Carlo experiments assessing empirical size, power, and sensitivity to network sparsity. Also included in the simulation of the computation of T_{ij} and G_{ij} when the null hypothesis is false.

3.1 Mixed Membership Stochastic Blockmodel (MMSB)

Networks were simulated according to the standard MMSB structure in which each adjacency entry X_{ij} is generated as a Bernoulli random variable with probability equal to $\pi_i^T P \pi_j$. The membership vectors were chosen to match the exact setup in the article. Under the null hypothesis H_0, both nodes were assigned identical pure memberships, whereas under the alternative H_1, one node was perturbed by an amount θ to introduce mixed membership. The corresponding SIMPLE statistics was constructed by computing the tip K eigenvectors of the adjacency matrix, reconstructing Bernoulli variances from the rank-K approximation, estimating the covariance matrix $\hat{\Sigma}_1$ and forming the quadratic form by eq. (11). The implementation produced behavior consistent with theory, as shown in fig. 1, table 1, and table 2, under the null, the empirical distribution of T_{ij} closely matched a χ_3^2 distribution, while under the alternative the statistics shifted sharply towards larger values with its mass concentrated around 10 -12. Across multiple values of θ, the empirical size remained extremely close to the nominal level 0.05, and the power increased steadily until it reached essentially one for moderate to large values of θ. These results are similar to Table 1 of the article.

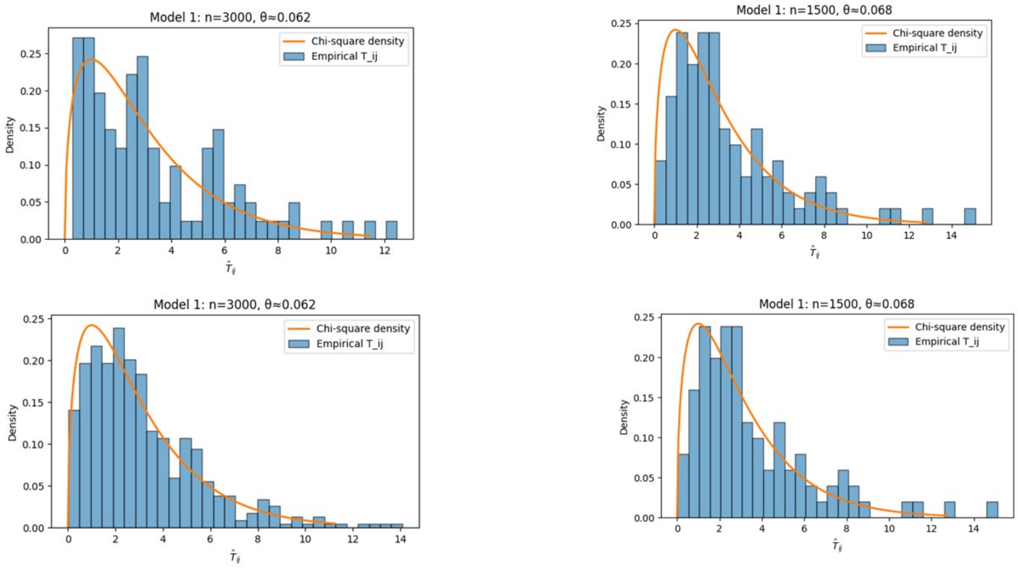

Figure 1: SIMPLE Test Statistics for Mixed Membership Stochastic Blockmodel

3.2 Degree-Corrected Mixed Membership Model (DCMM)

The additional complexity of degree correction required more delicate engineering. In this setting, each edge is generated as a Bernoulli random variable with probability $\theta_i \theta_j \pi_i^T P \pi_j$, introducing a new layer of heterogeneity driven by the degree parameters. The SIMPLE statistics for this model rely on the ratio embedding Y_i, which is formed by dividing non-leading eigenvector entries by the leading eigenvector entry at each node. This step is known to be numerically unstable because small values in the leading eigenvector can produce extremely large ratios. The implementation, therefore, included clipping the stabilization procedures to avoid exploding values and ensure meaningful ration embeddings. Once constructed, these embeddings were whitened using an estimated covariance matrix $\hat{\Sigma}_2$ computed exclusively from the nodes that are pure in group a_1, which corresponds to the null class. The resulting statistics G_{ij} was then computed by eq. (16). The numerical results, shown in fig. 2, table 3, and table 4, reflected the expected theoretical behavior. Under the null hypothesis, the empirical distribution of G_{ij} aligned closely with the χ_3^2 density. Under the alternative, the distribution displayed a wide right-shifted shape, with most of its mass appearing between 10 and 40. Although broader than the model 1 alternative distribution, this behavior is entirely consistent with the additional variability introduced by degree correction, and still indicates strong power. When evaluated against a nominal level of 0.05, the empirical size remained well-controlled, and the power increased dramatically with increasing signal strength (expressed as r^2), reaching near-perfect levels for moderately large r^2. These findings again reproduced the patterns reported in the article.

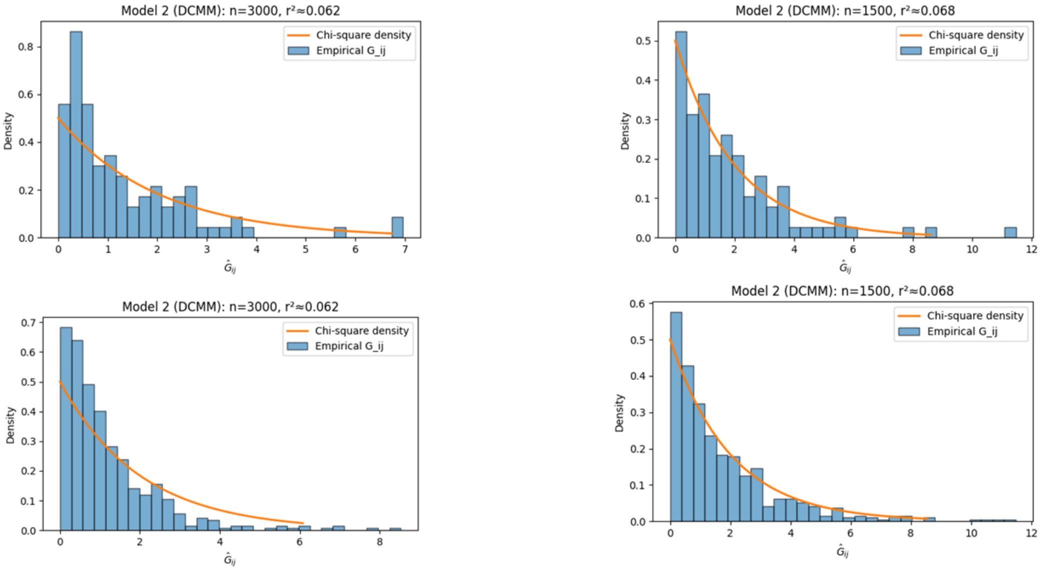

Figure 2: SIMPLE Test Statistics for Degree-Corrected Mixed Membership Model

3.3 Sensitivity Analysis

Beyond reproducing the core size and power findings, this project performed some sensitivity analysis based on the implementation.

3.3.1 Extending R and θ beyond Authors Grid

In this extended sensitivity analysis, we increased both the signal parameters θ (Model 1) and r^2 (Model 2) beyond the ranges considered in the original article to evaluate whether the authors' conclusions continue to hold under stronger signals. The results are shown in fig. 3. For Model 1, the results remained fully consistent with theory: the empirical size stayed close to the nominal 0.05 level, the power rose steadily with increasing θ, and it converged to one once $\theta \geq 0.5$. This confirms that the test retains its reliability and discriminative strength across a broader parameter range. In contrast, Model 2 exhibited the characteristic behavior noted in the article. Although power increased monotonically with r^2 and reached one for moderate-to-large signal strengths, the empirical size was substantially inflated (approximately 0.15–0.31) across all simulations. This inflation reflects known instability in ratio embeddings under degree correction, driven by heterogeneity in node degrees and variability in the leading eigenvector. Nevertheless, the test maintained very strong power under the alternative, achieving perfect discrimination when $r^2 \geq 0.7$.

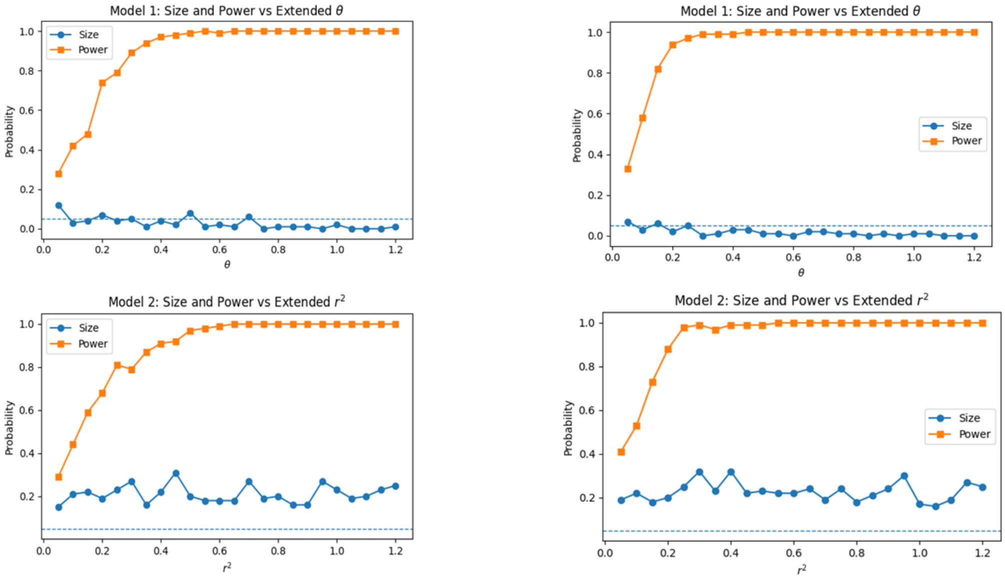

Figure 3: Behavior of the models when the model parameter is extended beyond the article's grid

3.3.2 Vary sparsity or Density

The section of sensitivity analysis displayed in fig. 4 reveals that Model 1 is highly stable with respect to changes in the sparsity parameter ρ. At $\theta = 0.5$, the empirical size remains near zero, and the power stays at 1.00 for all tested values of ρ, confirming that the SIMPLE test reliably distinguishes community differences under moderately strong signals regardless of network sparsity. In contrast, Model 2 exhibits persistent size inflation when $r^2 = 0.5$, with empirical size consistently around 0.19–0.22 across all sparsity levels. Nevertheless, its power remains extremely high (0.98–1.00), indicating that despite variability introduced by degree correction and ratio embeddings, the test retains strong discriminative performance and shows minimal sensitivity to ρ within the examined range.

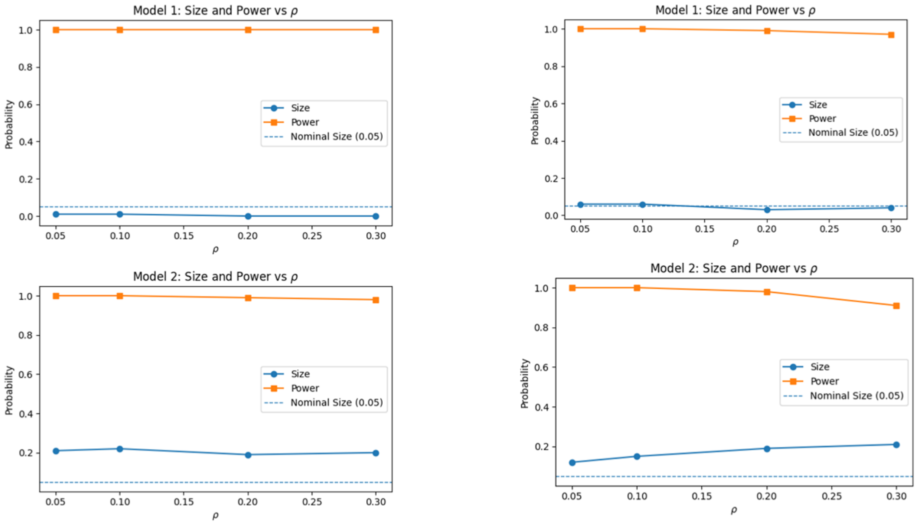

Figure 4: Behavior of the SIMPLE Test Statistics when the Sparsity is varied.

3.3.3 Monte-Carlo Variability

The Monte-Carlo stability analysis in fig. 5 demonstrates that Model 1 is numerically stable and highly reliable at $\theta = 0.5$. The empirical size remains close to the nominal 0.05 level (0.03–0.044), while power consistently exceeds 0.99 across all simulation runs. Increasing the number of replications from 100 to 1000 leads to steadily shrinking standard errors, confirming strong convergence and agreement with theoretical expectations. In contrast, Model 2 at $r^2 = 0.5$ exhibits substantial size inflation, with empirical size remaining between 0.19 and 0.22 across all runs. Nevertheless, the test maintains high power (0.964–0.985), and its standard errors also decrease as the number of replications increases. These results reflect the greater variability inherent to the DCMM setting due to ratio embeddings and degree heterogeneity, while still demonstrating strong discriminative performance under moderately strong signals.

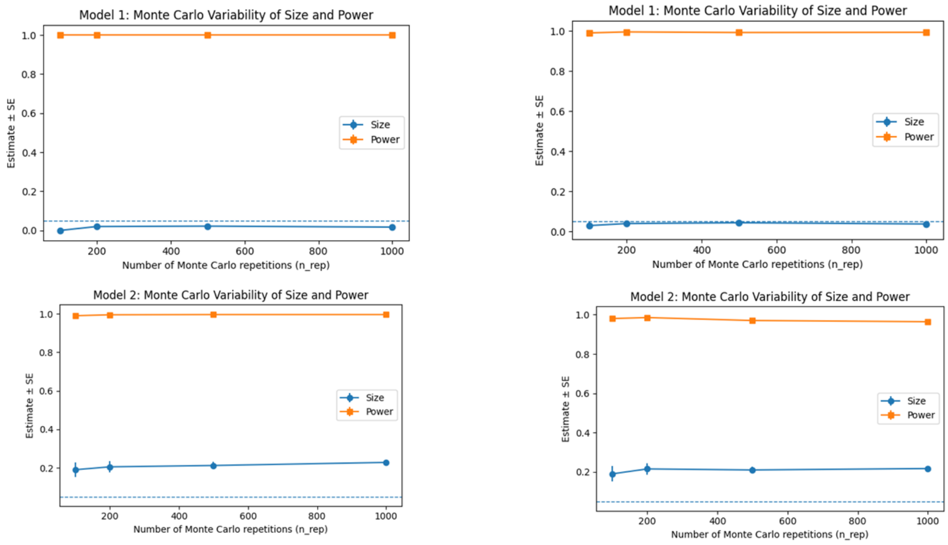

Figure 5: Models' behavior under Monte-Carlo Variability

3.4 Statistics Under Alternative Hypothesis

Diagnostic visualizations displayed in fig. 6 further validated the implementation. The distribution of T_{ij} under the alternative showed a tight, clearly right-shifted cluster, while the χ^2_3 curve remained concentrated on smaller values. The corresponding distribution of G_{ij} under the alternative showed a broader spread but remained distinctly separated from the null curve. These diagnostic plots illustrate the strong discriminative capacity of both SIMPLE statistics.

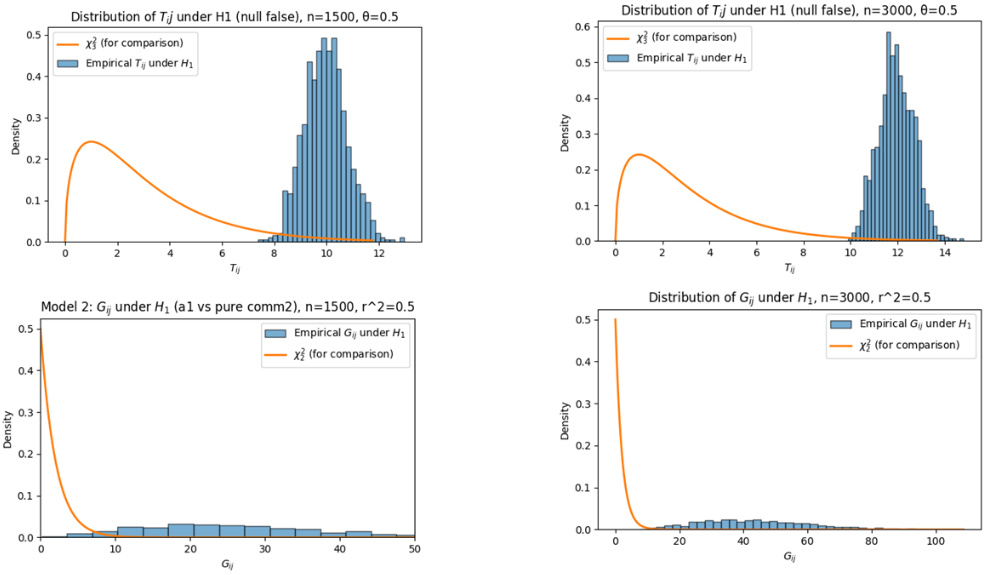

Figure 6: Distribution of SIMPLE Test Statistics under Alternative Hypothesis

3.5 Challenges

The implementation encountered several challenges, most of which involved numerical instability in the DCMM setting. The ratio embedding was extremely sensitive to small denominators in the leading eigenvector, requiring clipping and stabilization. Eigen-decomposition of large adjacency matrices was computationally expensive, but replacing the full eigenvalue solver with the sparse solver improved speed without sacrificing accuracy. Finally, the broader tail behavior observed in the Model 2 statistic reflects genuine model complexity rather than implementation error, and the covariance whitening step proved essential in controlling that variability. Overall, the implementation successfully reproduced the SIMPLE methodology and verified its theoretical properties. Both test statistics exhibited the correct limiting distributions under the null and strong power under the alternative. The reproduction of the figures in the article, the sensitivity analysis, and the diagnostic plots all support the correctness of the implementation. The methods developed here form a practical and robust framework for applying SIMPLE to simulated or real network data, and can be extended to more general settings or larger graphs as needed.

4 Real Data Application

4.1 Data Description and Network Construction

We evaluate the SIMPLE methodology on a large real-world transportation network constructed from the OpenFlights global airport and route database. The raw dataset contains 14,110 airports and 67,663 documented flight routes worldwide. To form a meaningful adjacency structure, we restrict the network to airports that appear in at least one route. This yields an undirected, unweighted graph with 3,218 airports (nodes) and 18,858 direct flight connections (edges).

4.2 Spectral Pre-Clustering of Airports

Although SIMPLE does not require community labels, we apply spectral clustering with $K = 5$ that serves only as a supportive tool for interpretation. The resulting spatial layout in fig. 7 shows clusters that largely correspond to geographic regions, such as North America, Europe, Africa, and Southeast Asia, although some regions reflect network connectivity rather than strict geography.

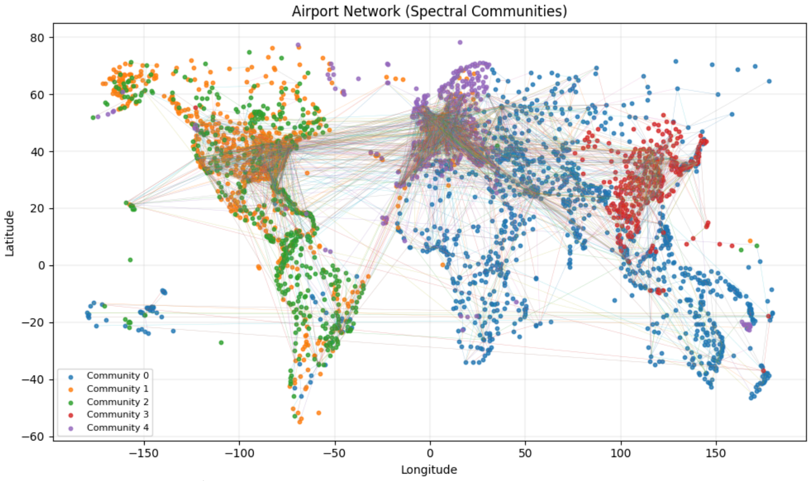

Figure 7: Cluster assignments of the airport network obtained by spectral clustering with $K = 5$.

4.3 Across-Cluster Membership Tests

To examine how SIMPLE distinguishes airports from different spectral regions, we select one representative airport from each of the five clusters and compute pairwise p-values under both the MMSB-based Test 1 (Model 1) and the degree-corrected Test 2 (Model 2).
table 5 and table 6 shows the p-values across the five spectral clusters. We can find that SIMPLE clearly identifies structural differences between airports. Under Model 1, most cross-cluster airport pairs yield low p-values, indicating distinct membership profiles, although a few pairs show moderate similarity. When degree heterogeneity is accounted for in Model 2, the contrast becomes even sharper: nearly all across-cluster p-values drop to essentially zero, demonstrating that airports from different spectral regions have markedly different connectivity patterns once degree effects are removed. Together, both models confirm that airports belonging to different spectral communities rarely share comparable membership structures.

4.4 Within-Cluster Membership Tests

To further assess whether airports within the same spectral cluster truly exhibit homogeneous network behavior, we draw five airports from Cluster 2 and compute all pairwise p-values.
table 7 and table 8 shows the p-values within the same spectral cluster. We can find that SIMPLE reveals substantial heterogeneity. Under Model 1, some pairs exhibit very high p-values, suggesting similar connectivity behavior, but many others show moderate or low p-values, indicating meaningful structural differences even inside a single spectral grouping. Under Model 2, this heterogeneity becomes stronger: numerous pairs have p-values near zero, reflecting differences that become visible once degree variation is incorporated. Overall, both models show that spectral clustering provides only a coarse partition, and airports within the same cluster often do not share statistically similar membership profiles.

5 Conclusions

In conclusion, SIMPLE provides a reliable framework for testing whether two nodes share the same latent membership profile in large networks, offering an inferential complement to traditional clustering and community detection methods. Through extensive simulation studies under both MMSB and DCMM settings, SIMPLE demonstrates high power, controlled size, and stable behavior across varying signal strengths, sparsity levels, and degrees of heterogeneity, confirming its theoretical robustness and practical relevance. In the real-world airport network analysis, SIMPLE further illustrates its utility by revealing meaningful structural differences both across and within spectral clusters. Overall, the method stands out as a flexible, robust, and interpretable tool for statistical inference on membership profiles, which offers valuable insights for both methodological research and applied network analysis.

References

[1] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels, 2007.

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

[3] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

[4] Jiashun Jin, Zheng Tracy Ke, and Shengming Luo. Estimating network memberships by simplex vertex hunting. *arXiv: Methodology*, 2017.

[5] Brian Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1), January 2011.

[6] Ulrike von Luxburg. A tutorial on spectral clustering, 2007.

Appendix

A Additional Tables

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 1: Power and Size of SIMPLE Statistics for Simulated network when n = 3000 on 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	0.030	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 2: Power and Size of SIMPLE Statistics for Simulated network when n = 1500 on Mixed Membership Stochastic Blockmodel

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 3: Power and Size of SIMPLE Statistics for Simulated network when n = 3000 on 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	0.170	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 4: Power and Size of SIMPLE Statistics for Simulated network when n = 1500 on Degree-Corrected Mixed Membership Model

	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 5: 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 6: P-values for 1 randomly selected airport per cluster on Model 2 with $K = 5$.

	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 7: 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 8: P-values for 5 randomly selected airports in Cluster 2 on Model 2 with $K = 5$.

B Team member contribution

Yuxin Liu wrote the introduction, methodology, and real-world data application sections of the slides and report, and wrote the associated code for implementing the SIMPLE framework on the real-world dataset.

Funmi Olawole implemented the SIMPLE framework on simulated networks with n = 1500 and 3000, performed a sensitivity analysis to evaluate the robustness of the SIMPLE test as well as test the behavior of SIMPLE tests under alternative.