CMStatistics 2020 – Virtual conference

Session EG012: Contributions in computational and methodological statistics

Spectral clustering on spherical coordinates under the degree-corrected stochastic blockmodel

Imperial College London

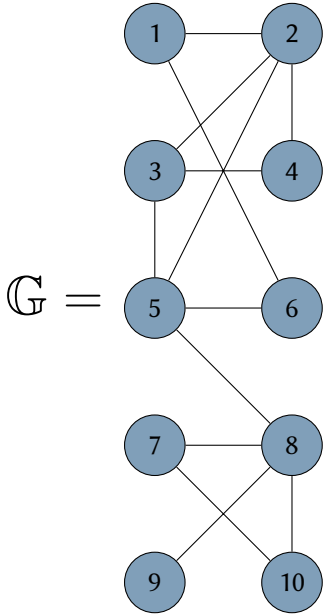
Francesco Sanna Passino[†], Nick Heard[†], Patrick Rubin-Delanchy[‡]

[†]Department of Mathematics, Imperial College London

[‡]School of Mathematics, University of Bristol

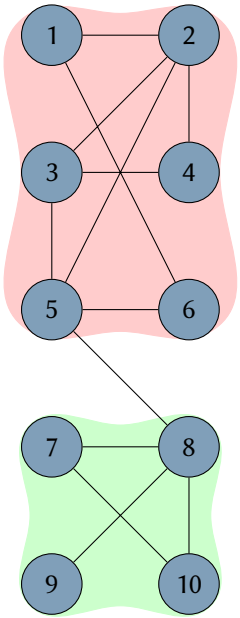
francesco.sanna-passino16@imperial.ac.uk

20th December, 2020



$$\mathbb{G} = \Rightarrow \mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \end{bmatrix}$$

$\mathbb{G} =$



$\Rightarrow \mathbf{A} =$

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \end{bmatrix}$$

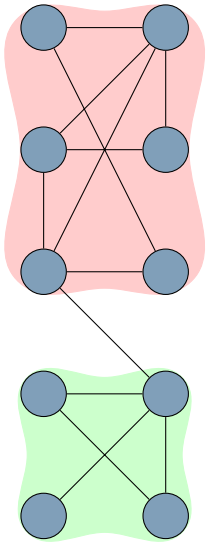
Stochastic blockmodels and degree corrections

- Consider an **undirected graph** with **symmetric adjacency matrix** $\mathbf{A} \in \{0,1\}^{n \times n}$.
- The **stochastic blockmodel (SBM)** is the classical model for community detection in graphs. Given a matrix $\mathbf{B} \in [0,1]^{K \times K}$ of within-community probabilities, the probability of a link depends on the **community allocations** $z_i, z_j \in \{1, \dots, K\}$ of the two nodes:

$$\mathbb{P}(A_{ij} = 1) = B_{z_i z_j}.$$

- The **degree-corrected stochastic blockmodel (DCSBM)** extends the SBM, allowing for **heterogeneous within-community degree distributions**. The probability of a link is corrected using parameters $\rho_i, \rho_j \in [0,1]$:

$$\mathbb{P}(A_{ij} = 1) = \rho_i \rho_j B_{z_i z_j}.$$



Generalised random dot product graphs

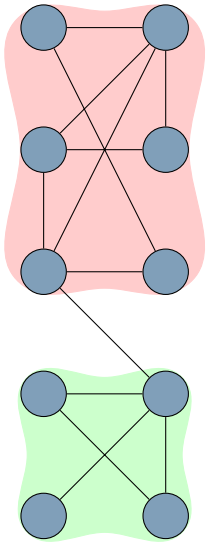
- In **generalised random dot product graphs** (GRDPG, Rubin-Delanchy et al., 2017), the probability of a link between two nodes is expressed as the inner product between two d-dimensional latent positions $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathbb{X}$:

$$\mathbb{P}(A_{ij} = 1) = \boldsymbol{x}_i^\top \mathbf{I}(d_+, d_-)\boldsymbol{x}_j,$$

  where $\mathbf{I}(d_+, d_-) = \mathrm{diag}(1, \ldots, 1, -1, \ldots, -1)$ with d_+ ones and d_- minus ones, such that $d_+ + d_- = d$.
- Requirement: $0 \leq \boldsymbol{x}^\top \mathbf{I}(d_+, d_-)\boldsymbol{y} \leq 1 \; \forall \, \boldsymbol{x}, \boldsymbol{y} \in \mathbb{X}$.
- SBMs and DCSBMs can be interpreted as **special cases** of random dot product graphs. For the SBM, if $\boldsymbol{x}_i = \boldsymbol{\mu}_{z_i}$, then:

$$\mathbb{P}(A_{ij} = 1) = B_{z_i z_j} = \boldsymbol{\mu}_{z_i}^\top \mathbf{I}(d_+, d_-)\boldsymbol{\mu}_{z_j}.$$

  In this framework: $d = \mathrm{rank}(\mathbf{B}) \leq K$.
- A similar result holds for the DCSBM, setting $\boldsymbol{x}_i = \rho_i \boldsymbol{\mu}_{z_i}$.



Spectral embedding

- The **latent positions** can be **consistently estimated** via spectral embedding.
- The **adjacency spectral embedding (ASE)** of $\mathbf{A}$ in $\mathbb{R}^d$ is:

$$\hat{\mathbf{X}} = [\hat{\boldsymbol{x}}_1, \dots, \hat{\boldsymbol{x}}_n]^\top = \hat{\mathbf{\Gamma}} |\hat{\mathbf{\Lambda}}|^{1/2} \in \mathbb{R}^{n \times d},$$

  where $\hat{\mathbf{\Lambda}}$ is a $d \times d$ diagonal matrix containing the d largest eigenvalues in magnitude of $\mathbf{A}$, and $\hat{\mathbf{\Gamma}}$ is a $n \times d$ matrix containing the corresponding eigenvectors.
- **ASE-CLT** – Taking a graph with N nodes with d known, and restricting the attention to the first n nodes, with $n < N$, then:

$$N^{1/2}(\mathbf{Q}_N \hat{\boldsymbol{x}}_i - \boldsymbol{x}_i) \longrightarrow \mathbb{N}_d\{\mathbf{0}, \mathcal{S}(\boldsymbol{x}_i)\}$$

  in distribution as $N \to \infty$, independently for $i = 1, \dots, n$, where $\mathbf{Q}_N$ is a matrix from the indefinite orthogonal group $\mathbb{O}(d_+, d_-)$, $\mathbb{N}_d(\cdot)$ is the d-dimensional multivariate normal distribution, and $\mathcal{S}(\boldsymbol{x}_i)$ can be analytically computed (Rubin-Delanchy et al., 2017).
- For SBMs: $\mathbf{Q}_N \hat{\boldsymbol{x}}_i \approx \mathbb{N}_d\{\boldsymbol{\mu}_{z_i}, \mathcal{S}(\boldsymbol{\mu}_{z_i})\}$ (*Gaussian point clouds*).
- For DCSBMs: $\mathbf{Q}_N \hat{\boldsymbol{x}}_i \approx \mathbb{N}_d\{\rho_i \boldsymbol{\mu}_{z_i}, \mathcal{S}(\rho_i \boldsymbol{\mu}_{z_i})\}$ (*rays through the origin*).

A synthetic example

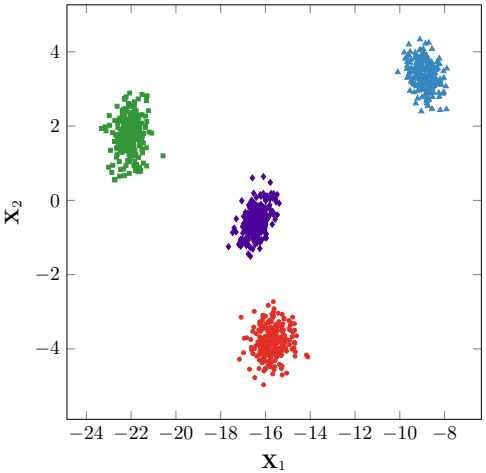


Figure 1. Scatterplot of the 2-dimensional ASE for a simulated SBM with $d = K = 4$, $\mathbf{B} \sim \text{Uniform}(0, 1)^{K \times K}$, and 100 nodes per community.

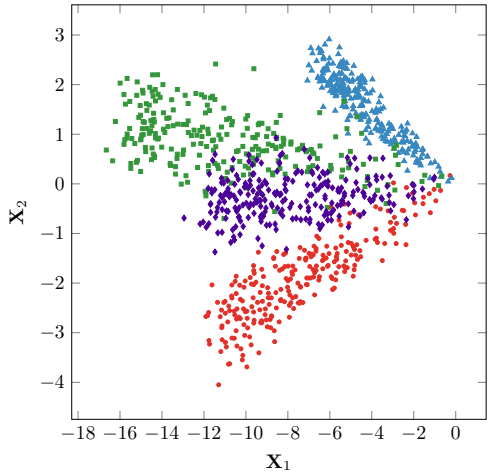


Figure 2. Scatterplot of the 2-dimensional ASE for a simulated DCSBM, identical to the SBM in Figure 1, corrected with $\rho_i \sim \text{Beta}(2, 1)$.

Spectral estimation of the stochastic blockmodel

- Based on the ASE-CLT, the following algorithm is appropriate for estimating SBMs:

Algorithm: Spectral estimation of the SBM (Rubin-Delanchy et al., 2017)

Input: adjacency matrix $\mathbf{A}$, dimension d, and number of communities $K \geq d$.

1 compute spectral embedding $\hat{\mathbf{X}} = [\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_n]^\top$ into $\mathbb{R}^d$,

2 fit a Gaussian mixture model with K components.

Result: return cluster centres $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K \in \mathbb{R}^d$ and node memberships $z_1, \ldots, z_n$.

- In practice: d and K are estimated **sequentially**. Issues:
  - Sequential approach is **sub-optimal**: the estimate of K depends on choice of d.
  - Theoretical results only hold for d **fixed and known**.
  - Distributional assumptions when d is misspecified are **not available**.
- Furthermore, Gaussian mixture modelling on the DCSBM embedding is **not** appropriate.
  - Possible solution (Ng, Jordan, and Weiss, 2001): k-means on the **row-normalised embedding** $\tilde{\boldsymbol{x}}_i = \hat{\boldsymbol{x}}_i / \|\hat{\boldsymbol{x}}_i\|$. This is problematic, because $\tilde{\boldsymbol{x}}_i$ is constrained to have unit norm.
- **This talk discusses a novel framework for spectral clustering under the DCSBM, and for simultaneously jointly estimate d and K.**

A synthetic example

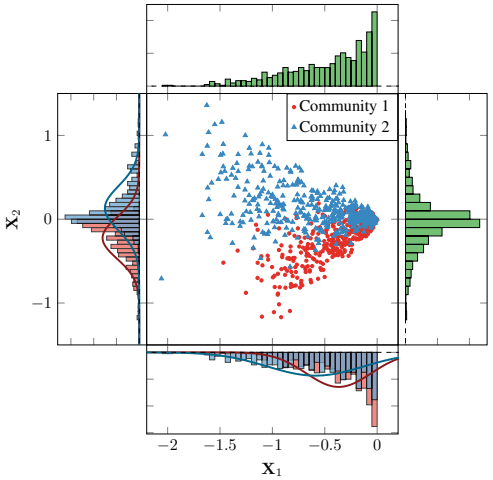


Figure 3. Scatterplot of the 2-dimensional **ASE** for a simulated DCSBM with $d = K = 2$, $B_{11} = 0.1, B_{12} = B_{21} = 0.05$ and $B_{22} = 0.15$, and 500 nodes per community, corrected with $\rho_i \sim \text{Beta}(2, 1)$.

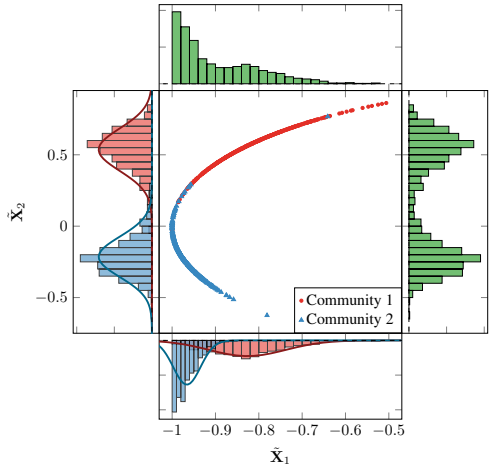


Figure 4. Scatterplot of the 2-dimensional **row-normalised ASE** for the simulated DCSBM in Figure 3.

A model for DCSBM embeddings

- Proposed solution: parametric model on the **spherical coordinates** of the embedding.
- For the m-dimensional ASE $\mathbf{X} \in \mathbb{R}^{n \times m}$, let $\mathbf{X}_{:d}$ and $\boldsymbol{x}_{i,:d}$ denote respectively the first d columns of the matrix and d elements of the vector, and $\mathbf{X}_{d:}$ and $\boldsymbol{x}_{i,d:}$ represent the remaining $m - d$ components.
- Consider a m-dimensional vector $\boldsymbol{x} \in \mathbb{R}^m$. The m Cartesian coordinates $\boldsymbol{x} = (x_1, \dots, x_m)$ can be converted in $m-1$ spherical coordinates $\boldsymbol{\theta} = (\theta_1, \dots, \theta_{m-1})$ on the unit m-sphere using a mapping $f_m : \mathbb{R}^m \to [0, 2\pi)^{m-1}$ such that $f_m : \boldsymbol{x} \mapsto \boldsymbol{\theta}$, where:

$$\theta_1 = \begin{cases} \arccos(x_2 / \|\boldsymbol{x}_{:2}\|) & x_1 \geq 0, \\ 2\pi - \arccos(x_2 / \|\boldsymbol{x}_{:2}\|) & x_1 < 0, \end{cases}$$

$$\theta_j = 2 \arccos(x_{j+1} / \|\boldsymbol{x}_{:j+1}\|), \ j = 2, \dots, m-1.$$

- From the $(m+1)$-dimensional adjacency embedding $\mathbf{X} \in \mathbb{R}^{n \times (m+1)}$, define its transformation $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n]^\top \in [0, 2\pi)^{n \times m}$, such that $\boldsymbol{\theta}_i = f_{m+1}(\boldsymbol{x}_i), \ i = 1, \dots, n$.

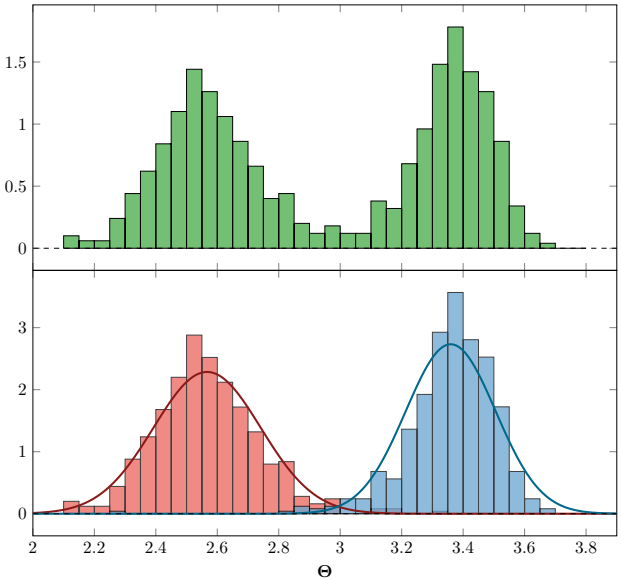


Figure 5. Scatterplot of the **transformed ASE** $\boldsymbol{\Theta}$ for the simulated DCSBM in Figure 3.

"Gaussianisation" of the ASE

A model on spherical coordinates for DCSBM spectral embeddings

- Let $\boldsymbol{\Theta}_{:d}$ and $\boldsymbol{\theta}_{i,:d}$ denote respectively the first d columns of the matrix and d elements of the vector, and $\boldsymbol{\Theta}_{d:}$ and $\boldsymbol{\theta}_{i,d:}$ the remaining $m-d$ components.
- For selection of d and K, the ***"overshooting"*** approach of Sanna Passino and Heard, 2020, and Yang et al., 2020, is followed: choose an arbitrarily large integer $m < n$ and obtain an **extended** transformed embedding $\boldsymbol{\Theta} \in \mathbb{R}^{n \times m}$ using ASE.
- For a given pair (d, K), the transformed ASE $\boldsymbol{\Theta}$ is assumed to have the distribution:

$$\boldsymbol{\theta}_i | d, z_i, \boldsymbol{\vartheta}_{z_i}, \boldsymbol{\Sigma}_{z_i}, \boldsymbol{\sigma}^2_{z_i} \sim \mathbb{N}_m \left(\begin{bmatrix} \boldsymbol{\vartheta}_{z_i} \\ \pi \mathbf{1}_{m-d} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{z_i} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\sigma}^2_{z_i} \mathbf{I}_{m-d} \end{bmatrix} \right), \tag{1}$$

  where $\boldsymbol{\vartheta}_{z_i} \in [0, 2\pi)^d$ represents a community-specific mean angle, $\mathbf{1}_m$ is a m-dimensional vector of ones, $\boldsymbol{\Sigma}_{z_i}$ is a $d \times d$ full covariance matrix, and $\boldsymbol{\sigma}^2_k = (\sigma^2_{k,d+1}, \ldots, \sigma^2_{k,m})$ is a vector of positive variances.
- The pair (d, K) is chosen using BIC, for m **fixed** (Yang et al., 2020).
- The conjecture for the likelihood in (1) mirrors the model for Cartesian coordinates of Sanna Passino and Heard, 2020.

Empirical model validation

- $N = 1000$ simulations of a GRDPG-DCSBM with $n = 1500$, $d = K = 3$;
- $\mathbf{B} \sim \text{Uniform}(0,1)^{K \times K}$ fixed across all N simulations, communities of equal size;
- $\rho_i \sim \text{Beta}(2,1)$.

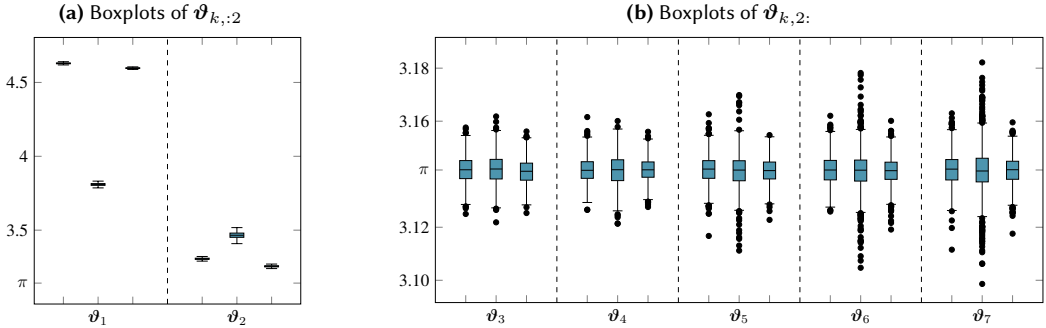


Figure 6. Boxplots for $N = 1{,}000$ simulations of a DCSBM with $n = 1{,}500$ nodes, $K = 3$, equal number of nodes allocated to each group, and $\mathbf{B} \sim \text{Uniform}(0,1)^{K \times K}$, corrected by $\rho_i \sim \text{Beta}(2,1)$.

Empirical model validation

- $N = 1000$ simulations of a GRDPG-DCSBM with $n = 1500$, $d = K = 3$;
- $\mathbf{B} \sim \text{Uniform}(0,1)^{K \times K}$ fixed across all N simulations, communities of equal size;
- $\rho_i \sim \text{Beta}(2,1)$.

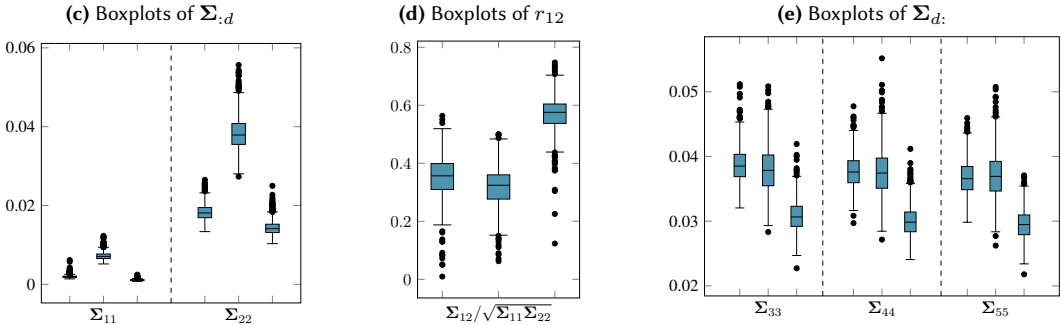
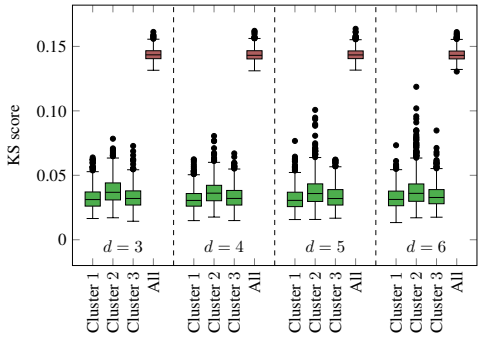


Figure 6. Boxplots for $N = 1{,}000$ simulations of a DCSBM with $n = 1{,}500$ nodes, $K = 3$, equal number of nodes allocated to each group, and $\mathbf{B} \sim \text{Uniform}(0,1)^{K \times K}$, corrected by $\rho_i \sim \text{Beta}(2,1)$.

Empirical model validation

- $N = 1000$ simulations of a GRDPG-DCSBM with $n = 1500$, $d = K = 3$;
- $\mathbf{B} \sim \text{Uniform}(0,1)^{K \times K}$ fixed across all N simulations, communities of equal size;
- $\rho_i \sim \text{Beta}(2,1)$.

(f) Kolmogorov-Smirnov scores for Gaussian fit in $\mathbf{\Theta}_d$:



(g) Boxplots of $r_{k\ell}$ for the redundant components

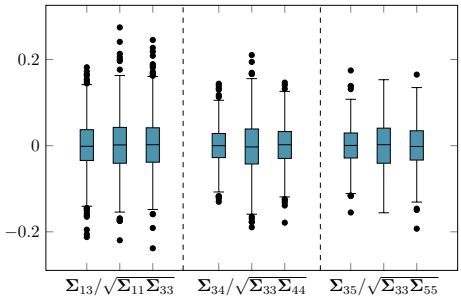


Figure 6. Boxplots for $N = 1{,}000$ simulations of a DCSBM with $n = 1{,}500$ nodes, $K = 3$, equal number of nodes allocated to each group, and $\mathbf{B} \sim \text{Uniform}(0,1)^{K \times K}$, corrected by $\rho_i \sim \text{Beta}(2,1)$.

Imperial College NetFlow data

- The model is extended to **directed** and **bipartite** graphs, via SVD embedding.
- For clustering the source nodes in a bipartite graph, construct the embedding $\mathbf{X} = \mathbf{U}\mathbf{D}^{1/2}$ using singular values and the corresponding left singular vectors.
- Bipartite graph of HTTP (port 80) and HTTPS (port 443) connections from machines hosted in computer labs at ICL.
- $439 \times 60{,}635$ nodes, 717,912 links.
- Observation period: 1–31 January 2020.
- Departments can be used as labels.
  - Chemistry,
  - Civil & Environmental Engineering,
  - Mathematics,
  - School of Medicine.

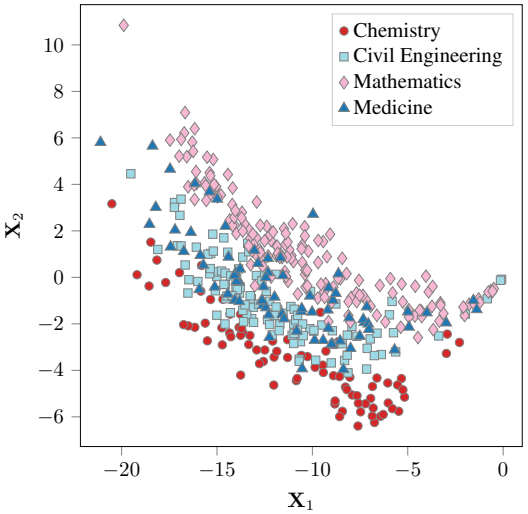


Figure 7. Scatterplot of $\mathbf{X}_{:2}$, coloured by department.

Imperial College NetFlow data

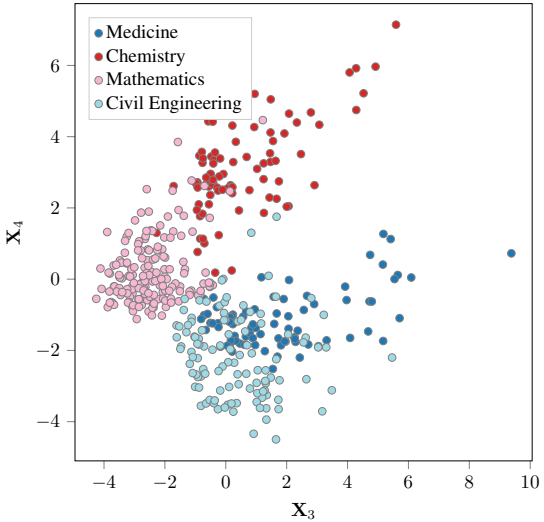


Figure 8. Scatterplot of $\mathbf{X}_3$ and $\mathbf{X}_4$, coloured by department.

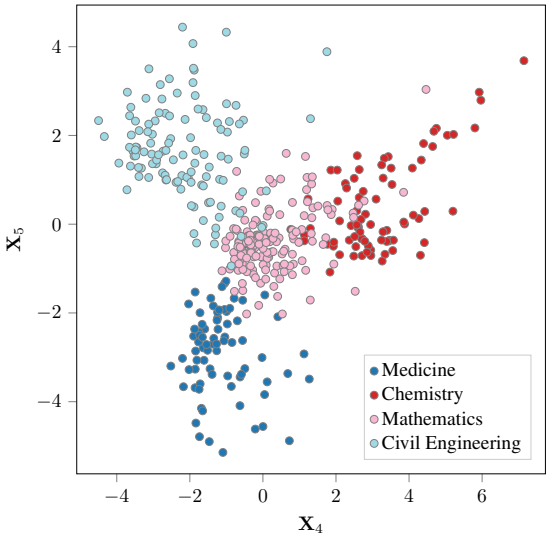


Figure 9. Scatterplot of $\mathbf{X}_4$ and $\mathbf{X}_5$, coloured by department.

Imperial College NetFlow data

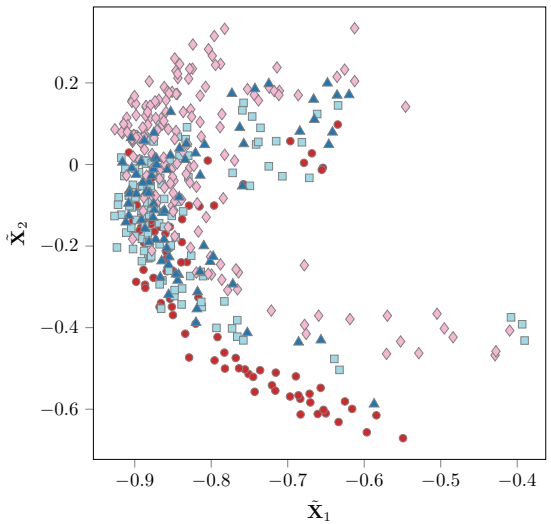


Figure 10. Scatterplot of $\tilde{\mathbf{X}}_{:2}$ for $m = 30$.

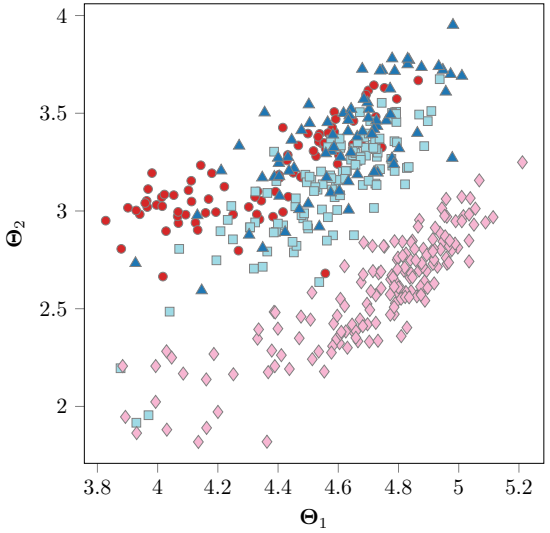


Figure 11. Scatterplot of $\mathbf{\Theta}_{:2}$ for $m = 30$.

Imperial College NetFlow data

	$m = 30$			$m = 50$		
	$\mathbf{X}$	$\tilde{\mathbf{X}}$	$\mathbf{\Theta}$	$\mathbf{X}$	$\tilde{\mathbf{X}}$	$\mathbf{\Theta}$
Estimated (d, K)	$(28, 5)$	$(8, 7)$	$(15, 4)$	$(29, 4)$	$(8, 7)$	$(15, 4)$
Adjusted Rand Index (ARI)	0.441	0.736	0.938	0.359	0.743	0.938

Table 1. Estimates of (d, K) and ARIs for the embeddings $\mathbf{X}$, $\tilde{\mathbf{X}}$ and $\mathbf{\Theta}$ for $m \in \{30, 50\}$.

- Estimates for $\mathbf{X}$ and $\tilde{\mathbf{X}}$ are obtained using the model of Sanna Passino and Heard, 2020,
- Using $\mathbf{\Theta}$, the correct value of K is estimated (corresponding to the number of departments),
- Using $\mathbf{\Theta}$, only 9 **nodes** are misclassified,
- The constraint of unit row-norm on $\tilde{\mathbf{X}}$ causes issues in the estimation of K,
- Estimates appear to be stable for different values of m.

Conclusion

- Spectral clustering in DCSBMs:
  - Model on spherical coordinates for simultaneous selection of K and d under a GRDPG interpretation,
  - Gaussian mixture model on spherical coordinates (with constraints) based on adjacency spectral embedding,
  - Allow for initial misspecification of the arbitrarily large parameter m, then refine estimate d,
  - The transformation to spherical coordinates appears to "Gaussianise" the ASE,
  - Easy to extend to directed and bipartite graphs.
- For more details, see Sanna Passino, Heard, and Rubin-Delanchy, 2020, arXiv: 2011.04558.

References

- Ng, A. Y., M. I. Jordan, and Y. Weiss (2001). "On Spectral Clustering: Analysis and an Algorithm". In: *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*. MIT Press, pp. 849–856.
- Rubin-Delanchy, P. et al. (2017). "A statistical interpretation of spectral embedding: the generalised random dot product graph". In: *arXiv e-prints*. arXiv: 1709.05506.
- Sanna Passino, F. and N. A. Heard (2020). "Bayesian estimation of the latent dimension and communities in stochastic blockmodels". In: *Statistics and Computing* 30.5, pp. 1291–1307.
- Sanna Passino, F., N. A. Heard, and P. Rubin-Delanchy (2020). "Spectral clustering on spherical coordinates under the degree-corrected stochastic blockmodel". In: *arXiv e-prints*. arXiv: 2011.04558.
- Yang, C. et al. (2020). "Simultaneous dimensionality and complexity model selection for spectral graph clustering". In: *Journal of Computational and Graphical Statistics* (to appear).