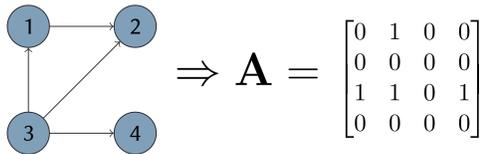


1. Latent position models and RDPGs

- Consider the **adjacency matrix** $\mathbf{A} = \{A_{ij}\} \in \{0, 1\}^{n \times n}$ of a **graph**, where $A_{ij} = 1$ if node i connects to node j , and $A_{ij} = 0$ otherwise.



- Graph adjacency matrices can be modelled via **latent position models** (LPMs; Hoff et al., 2002):

$$\mathbf{x}_i \stackrel{iid}{\sim} F \rightarrow \mathbb{P}(A_{ij} = 1 | \mathbf{x}_i, \mathbf{x}_j) = \kappa(\mathbf{x}_i, \mathbf{x}_j) \rightarrow$$



- LPMs are built on a powerful idea: expressing **edge-specific probabilities** through **latent node features** $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d$, using a **kernel function** $\kappa: \mathcal{X} \times \mathcal{X} \rightarrow [0, 1]$.
- If the **inner product** is used, the model is known as **random dot product graph** (RDPG; Athreya et al., 2018):

$$A_{ij} | \mathbf{x}_1, \dots, \mathbf{x}_n \sim \text{Bernoulli}(\mathbf{x}_i^\top \mathbf{x}_j).$$

- RDPGs include many popular network models:
 - Stochastic blockmodels** (SBMs): $\mathbf{x}_i = \boldsymbol{\mu}_{z_i}$ for a community $z_i \in \{1, \dots, K\}$, giving a between-community connection probability $B_{kl} = \boldsymbol{\mu}_k^\top \boldsymbol{\mu}_l$;
 - Degree-corrected SBMs**: $\mathbf{x}_i = \rho_i \boldsymbol{\mu}_{z_i}$ for $z_i \in \{1, \dots, K\}$ and degree-correction $\rho_i \in (0, 1)$.
- In RDPGs, the latent positions are **estimated via spectral decomposition** of the adjacency matrix.

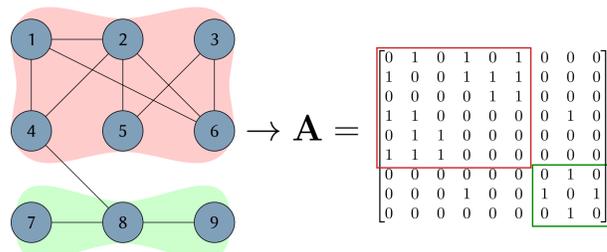
Definition 1: Adjacency spectral embedding (ASE)

For an integer $d \in \{1, \dots, n\}$ and a binary *symmetric* adjacency matrix $\mathbf{A} \in \{0, 1\}^{n \times n}$, the d -dimensional adjacency spectral embedding $\hat{\mathbf{X}} = [\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_n]^\top$ of \mathbf{A} is

$$\hat{\mathbf{X}} = \Gamma \Lambda^{1/2} \in \mathbb{R}^{n \times d},$$

where Λ is a $d \times d$ diagonal matrix containing the absolute values of the d largest eigenvalues in magnitude, and Γ is a $n \times d$ matrix containing corresponding eigenvectors.

- For **directed** graphs, the **singular value decomposition** (SVD) is used.
- In practice, spectral embeddings often exhibit **manifold structure** (Rubin-Delanchy, 2020).
- Spectral graph clustering** consists in unsupervised detection of groups of nodes from spectral embeddings.



- There are two **simultaneous challenges** in graph clustering via spectral embedding under the RDPG:

Manifold structure } Group-specific manifolds.
Group structure }

- Manifold structure** is accounted for by **latent structure models** (LSM, Athreya et al., 2021):

- The latent positions $\mathbf{x}_i \stackrel{iid}{\sim} F$ are determined by draws from an underlying univariate distribution G on $[0, 1]$, inducing F on a **univariate submanifold** $\mathcal{S} \subset \mathbb{R}^d$.
- The distribution F on \mathcal{S} is the distribution of the transformation $\mathbf{f}(\theta)$ of a univariate random variable $\theta \sim G$, where $\mathbf{f}: [0, 1] \rightarrow \mathcal{S}$ is a function mapping θ to \mathcal{S} .
- In simple terms, each node is assigned a draw θ_i from the underlying distribution G , representing *how far* along \mathcal{S} the corresponding latent position lies:

$$\mathbf{x}_i = \mathbf{f}(\theta_i).$$

Acknowledgements

This work is funded by the **Microsoft Security AI** research grant "Understanding the enterprise: Host-based event prediction for automatic defence in cyber-security".

2. Proposed methodology: latent structure blockmodels (LSBMs)

- Add a **group structure** to LSMs \rightarrow **latent structure blockmodels** (LSBMs).
- Each node is assigned a **latent membership** $z_i \in \{1, \dots, K\}$, with probabilities η_1, \dots, η_K , with $\eta_k \geq 0$ and $\sum_{k=1}^K \eta_k = 1$.
- Each community is associated with a **different one-dimensional structural support submanifold** $\mathcal{S}_k \subset \mathbb{R}^d$, $k = 1, \dots, K$. Implicitly, $F = \sum_{k=1}^K \eta_k F_k$ is a mixture distribution with components F_1, \dots, F_K supported on $\mathcal{S}_1, \dots, \mathcal{S}_K$.
- Assuming community allocations $\mathbf{z} = (z_1, \dots, z_n)$, the latent positions are obtained as

$$\mathbf{x}_i | z_i \sim F_{z_i}, \quad i = 1, \dots, n,$$

where F_{z_i} is the distribution of the community-specific transformation $\mathbf{f}_{z_i}(\theta)$ of a shared univariate random variable $\theta \sim G$.

- G is common to *all* the nodes, and the pair (θ_i, z_i) , where $\theta_i \sim G$, determines the latent position \mathbf{x}_i through \mathbf{f}_{z_i} , such that:

$$\mathbf{x}_i = \mathbf{f}_{z_i}(\theta_i).$$

3. A Bayesian model for LSBMs

$$\begin{aligned} \hat{\mathbf{x}}_i | \theta_i, \mathbf{f}_{z_i}, \sigma_{z_i}^2 &\sim \mathcal{N}(\mathbf{f}_{z_i}(\theta_i), \sigma_{z_i}^2 \mathbf{I}_d), \quad i = 1, \dots, n, \\ \theta_i &\sim \mathcal{N}(\mu_\theta, \sigma_\theta^2), \quad i = 1, \dots, n, \\ f_{k,j} | \sigma_{k,j}^2 &\sim \text{GP}(0, \sigma_{k,j}^2 \xi_{k,j}), \quad k = 1, \dots, K, \quad j = 1, \dots, d, \\ \sigma_{k,j}^2 &\sim \text{IG}(a_0, b_0), \quad k = 1, \dots, K, \quad j = 1, \dots, d, \\ z_i | \boldsymbol{\eta}, K &\sim \text{Categorical}(\boldsymbol{\eta}), \quad i = 1, \dots, n, \\ \boldsymbol{\eta} | K &\sim \text{Dirichlet}(\nu/K, \dots, \nu/K), \\ K &\sim \text{Geometric}(\omega), \end{aligned}$$

where $a_0, b_0, \nu, \omega, \sigma_\theta^2 \in \mathbb{R}_+$, $\mu_\theta \in \mathbb{R}$, and $\xi_{k,j}$ is a positive semi-definite kernel function.

4. Posterior inference

- After marginalisation of $(f_{k,j}, \sigma_{k,j}^2)$ and $\boldsymbol{\eta}$, inference is limited to the community allocations \mathbf{z} and parameters θ .
- The posterior distribution $p(\mathbf{z}, \theta, K | \hat{\mathbf{X}})$ is analytically intractable; inference is performed using **MCMC methods**.
 - Resample the community allocations \mathbf{z} ;
 - Resample the latent parameters θ ;
 - Split-merge communities;
 - Add or remove an empty community;
 - If prior on kernels is used: resample kernels.
- The kernel function is usually assumed to be in inner product form, with Zellner's g -prior on the scaling matrix.

5. Examples of LSBMs: SBMs, DCSBMs and quadratic LSBMs

- Kernels – SBM: $\xi_{k,j}(\theta, \theta') = \Delta_{k,j}$, $\Delta_{k,j} \in \mathbb{R}_+$; DCSBM: $\xi_{k,j}(\theta, \theta') = \theta \theta' \Delta_{k,j}$; Quadratic: $\xi_{k,j}(\theta, \theta') = (\theta, \theta') \Delta_{k,j} (\theta', \theta')^\top$.

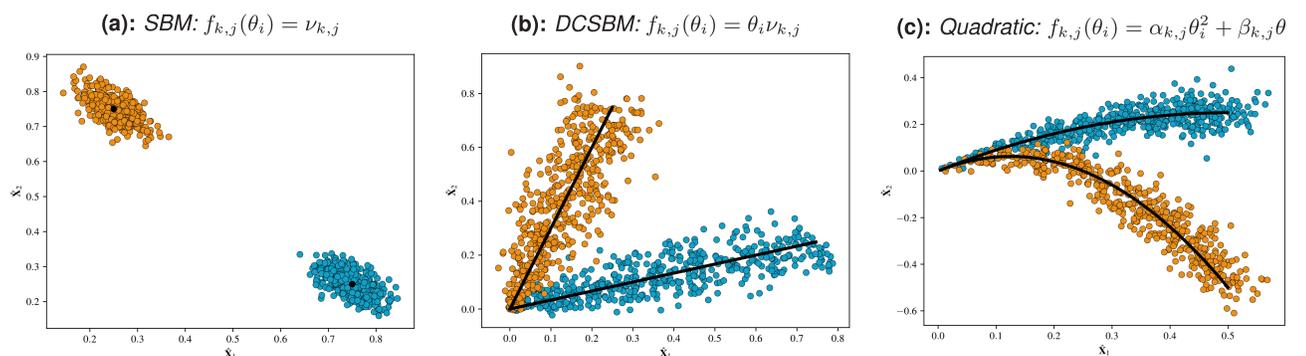


Figure 1: Scatterplots of the two-dimensional ASE of simulated graphs arising from simple models, and true underlying latent curves (in black). For each graph, $n = 1000$ with $K = 2$ communities of equal size. For (a) and (b), $\nu_1 = [3/4, 1/4]$, $\nu_2 = [1/4, 3/4]$ and $\theta_i \sim \text{Beta}(1, 1)$. For (c), $\alpha_k = [-1, -4]$, $\beta_k = [1, 1]$, $\gamma_k = [0, 0]$ and $\theta_i \sim \text{Beta}(2, 1)$.

Method	LSBM($\hat{\mathbf{X}}$)	GMM($\hat{\mathbf{X}}$)	GMM($\tilde{\mathbf{X}}$)	SCSC($\hat{\mathbf{X}}$)	PGP($\hat{\mathbf{X}}$)	HLouvain	HClust($\hat{\mathbf{X}}$)
SBM	1.0	1.0	1.0	1.0	1.0	1.0	1.0
DCSBM	0.853	0.802	0.838	0.887	0.842	0.827	0.411
Quadratic LSBM	0.838	0.620	0.712	0.636	0.691	0.582	0.101

Table 1: ARI for communities estimated using LSBM and alternative methodologies on the embeddings in Figure 1.

6. Results on three networks

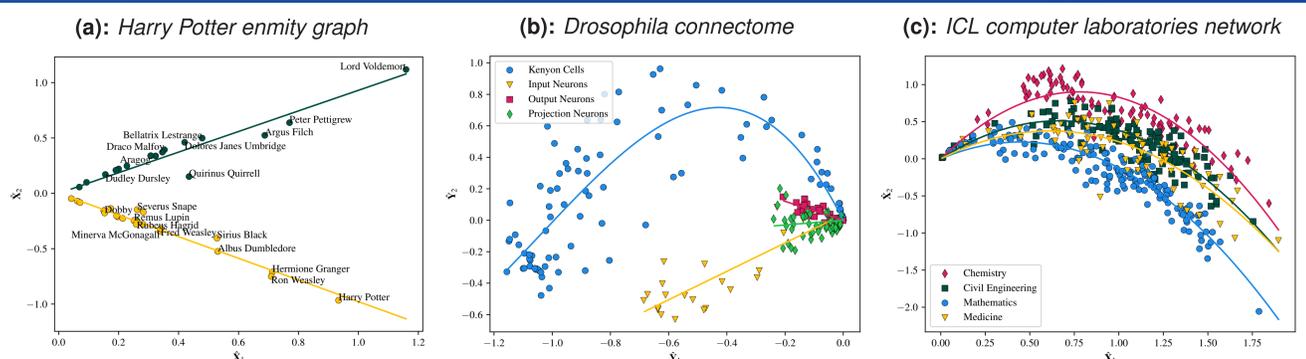


Figure 2: Two-dimensional embeddings, estimated communities and true labels for three real-world dataset.

Method	LSBM($\hat{\mathbf{X}}$)	GMM($\hat{\mathbf{X}}$)	GMM($\tilde{\mathbf{X}}$)	SCSC($\hat{\mathbf{X}}$)	PGP($\hat{\mathbf{X}}$)	HLouvain	HClust($\hat{\mathbf{X}}$)
Drosophila connectome	0.875	0.599	0.585	0.667	0.555	0.087	0.321
ICL computer laboratories	0.940	0.659	0.766	0.921	0.895	0.602	0.139

Table 2: ARI for communities estimated using LSBM and alternative methodologies on the Drosophila and ICL laboratories networks.

References

- Athreya, A. et al. (2018). "Statistical Inference on Random Dot Product Graphs: a Survey". Journal of Machine Learning Research 18, 1–92.
- Athreya, A. et al. (2021). "On Estimation and Inference in Latent Structure Random Graphs". Statistical Science 36.1, 68–88.
- Hoff, P. D. et al. (2002). "Latent space approaches to social network analysis". Journal of the American Statistical Association 97, 1090–1098.
- Rubin-Delanchy, P. (2020). "Manifold structure in graph embeddings". Advances in Neural Information Processing Systems 33, 11687–11699.

Paper and python library 1sbm

Paper



python library 1sbm

