Graph embedding workshop

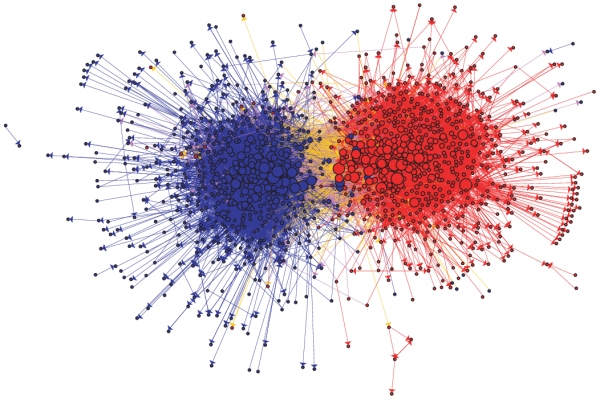
Alexander Modell, Nick Heard, Annie Gray, Karl Rohe

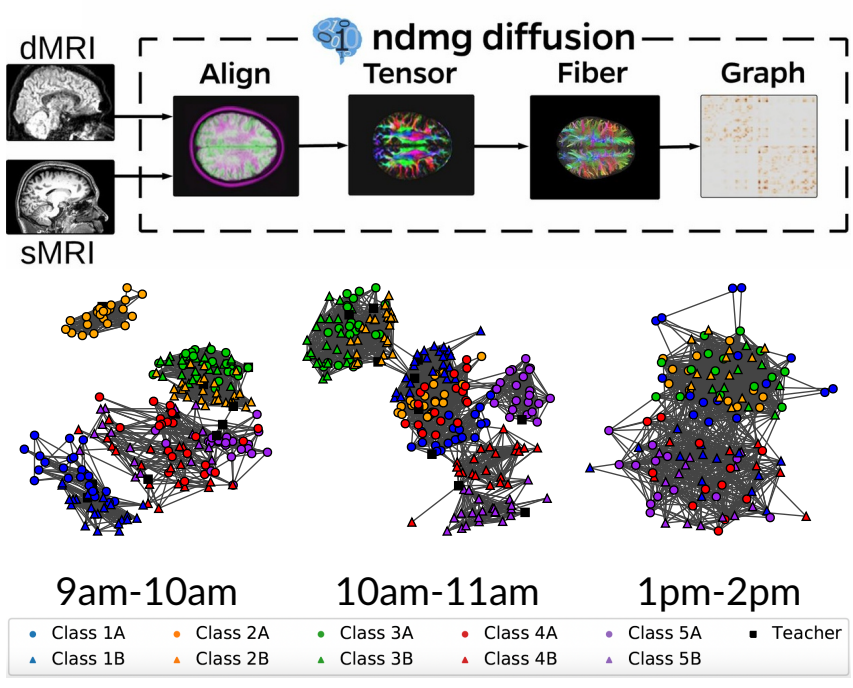
Plan for today

Time	Session
10am	Introduction to spectral embedding *Alexander Modell, Imperial*
11am	Manifold structure in graph embeddings *Annie Gray, Alan Turing Institute*
11:30am	Variational inference for Graph Embeddings *Nick Heard, Imperial*
12:00pm	Lunch break
1:00pm	Vintage Factor Analysis with Varimax Performs Statistical Inference *Karl Rohe, University of Wisconsin-Madison*
1:30pm	Workshop intro: organise my a wedding *Alexander Modell, Imperial*
2:00pm	Workshop

Network data describes relationships and/or interactions between entities

- Friendships / enmities
- Websites connected by hyperlinks
- Biological interactions (between proteins, genes, haplotypes, drugs, diseases, side-effects)
- Linguistic relationships (between words, documents, authors etc.)
- fMRI scans
- Physical interactions
- Trade data
- Online social network data (retweets, comments, likes)





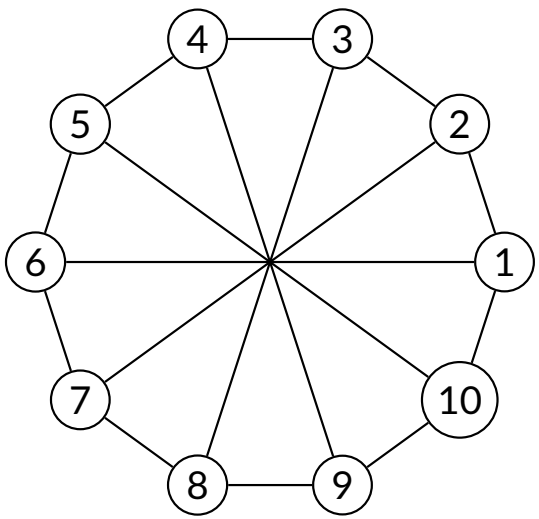
Graphs provide a mathematical representation of network data

A *graph* is a set of nodes labelled $1, \ldots, n$ and a set of numerical relationships $\{a_{ij} \in \mathbb{R}\}_{i<j}$.

Graphs provide a mathematical representation of network data

A *graph* is a set of nodes labelled $1, \ldots, n$ and a set of numerical relationships $\{a_{ij} \in \mathbb{R}\}_{i<j}$.

The matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{n \times n}$ is called the *adjacency matrix* of the graph.

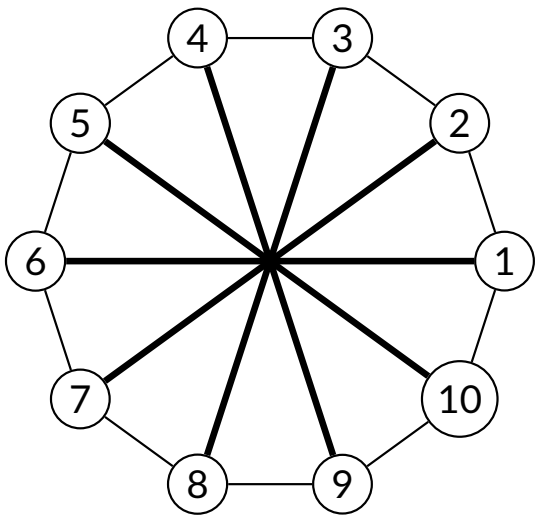


0	1	0	0	0	1	0	0	0	1
1	0	1	0	0	0	1	0	0	0
0	1	0	1	0	0	0	1	0	0
0	0	1	0	1	0	0	0	1	0
0	0	0	1	0	1	0	0	0	1
1	0	0	0	1	0	1	0	0	0
0	1	0	0	0	1	0	1	0	0
0	0	1	0	0	0	1	0	1	0
0	0	0	1	0	0	0	1	0	1
1	0	0	0	1	0	0	0	1	0

Graphs provide a mathematical representation of network data

A *graph* is a set of nodes labelled $1, \ldots, n$ and a set of numerical relationships $\{a_{ij} \in \mathbb{R}\}_{i<j}$.

The matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{n \times n}$ is called the *adjacency matrix* of the graph.



0	1	0	0	0	**5**	0	0	0	1
1	0	1	0	0	0	**5**	0	0	0
0	1	0	1	0	0	0	**5**	0	0
0	0	1	0	1	0	0	0	**5**	0
0	0	0	1	0	1	0	0	0	**5**
5	0	0	0	1	0	1	0	0	0
0	**5**	0	0	0	1	0	1	0	0
0	0	**5**	0	0	0	1	0	1	0
0	0	0	**5**	0	0	0	1	0	1
1	0	0	0	**5**	0	0	0	1	0

Singular values and vectors

Singular values and vectors

- Let $\mathbf{A} \in \mathbb{R}^{n \times m}$ be a real-valued matrix.

Singular values and vectors

- Let $\mathbf{A} \in \mathbb{R}^{n \times m}$ be a real-valued matrix.
- Its singular values and vectors satisfy the equation

$$\mathbf{A} v_i = \sigma_i u_i, \qquad i = 1, \ldots, r \leq \min\{n, m\}$$

with $\sigma_i > 0$.

Singular values and vectors

- Let $\mathbf{A} \in \mathbb{R}^{n \times m}$ be a real-valued matrix.
- Its singular values and vectors satisfy the equation

$$\mathbf{A} v_i = \sigma_i u_i, \qquad i = 1, \ldots, r \leq \min\{n, m\}$$

  with $\sigma_i > 0$.
- We can choose a system of left and right singular vectors such that

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0, \qquad u_i^\top u_j = v_i^\top v_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases}$$

Singular values and vectors

- Let $\mathbf{A} \in \mathbb{R}^{n \times m}$ be a real-valued matrix.
- Its singular values and vectors satisfy the equation

$$\mathbf{A} v_i = \sigma_i u_i, \qquad i = 1, \ldots, r \leq \min\{n, m\}$$

  with $\sigma_i > 0$.
- We can choose a system of left and right singular vectors such that

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0, \qquad u_i^\top u_j = v_i^\top v_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases}.$$

- Then $\mathbf{A}$ can be written as the *singular value decomposition* (SVD)

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$$

Singular values and vectors

- Let $\mathbf{A} \in \mathbb{R}^{n \times m}$ be a real-valued matrix.
- Its singular values and vectors satisfy the equation

$$\mathbf{A} v_i = \sigma_i u_i, \qquad i = 1, \ldots, r \leq \min\{n, m\}$$

  with $\sigma_i > 0$.
- We can choose a system of left and right singular vectors such that

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0, \qquad u_i^\top u_j = v_i^\top v_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases}.$$

- Then $\mathbf{A}$ can be written as the *singular value decomposition* (SVD)

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$$

When $\mathbf{A}$ is symmetric, either $u_i = v_i$ or $u_i = -v_i$.

Singular vectors as graph embeddings

Singular vectors as graph embeddings

Suppose $\mathbf{A}$ has the singular value decomposition $\mathbf{A} = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$ with $\sigma_1 \geq \cdots \geq \sigma_n$.

Singular vectors as graph embeddings

Suppose $\mathbf{A}$ has the singular value decomposition $\mathbf{A} = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$ with $\sigma_1 \geq \cdots \geq \sigma_n$.

The adjacency spectral embedding of the graph into $\mathbb{R}^d$, denoted $X_1, \ldots, X_n \in \mathbb{R}^d$, is given by the rows of the matrix

$$\mathbf{X} = \begin{pmatrix} X_1^\top \\ \vdots \\ X_n^\top \end{pmatrix} := \begin{pmatrix} \sigma_1^{1/2} u_1 & \cdots & \sigma_n^{1/2} u_r \end{pmatrix}$$

obtained by stacking the scaled left singular vectors $\sigma_1^{1/2} u_1, \ldots, \sigma_n^{1/2} u_r$ in columns.

Singular vectors as information extractors

Singular vectors as information extractors

- The *rank* of a matrix $\mathbf{M}$ is the smallest integer d such that there exist matrices $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{Y} \in \mathbb{R}^{m \times d}$ such that

$$\mathbf{M} = \mathbf{X}\mathbf{Y}^{\top}.$$

Singular vectors as information extractors

- The *rank* of a matrix $\mathbf{M}$ is the smallest integer d such that there exist matrices $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{Y} \in \mathbb{R}^{m \times d}$ such that

$$\mathbf{M} = \mathbf{X}\mathbf{Y}^\top.$$

- Consider the problem of finding the best rank-d approximation of $\mathbf{A}$, i.e. such a matrix $\mathbf{M} = \mathbf{X}\mathbf{Y}^\top$ which minimises

$$\|\mathbf{M} - \mathbf{A}\|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} (m_{ij} - a_{ij})^2.$$

Singular vectors as information extractors

- The *rank* of a matrix $\mathbf{M}$ is the smallest integer d such that there exist matrices $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{Y} \in \mathbb{R}^{m \times d}$ such that

$$\mathbf{M} = \mathbf{X}\mathbf{Y}^\top.$$

- Consider the problem of finding the best rank-d approximation of $\mathbf{A}$, i.e. such a matrix $\mathbf{M} = \mathbf{X}\mathbf{Y}^\top$ which minimises

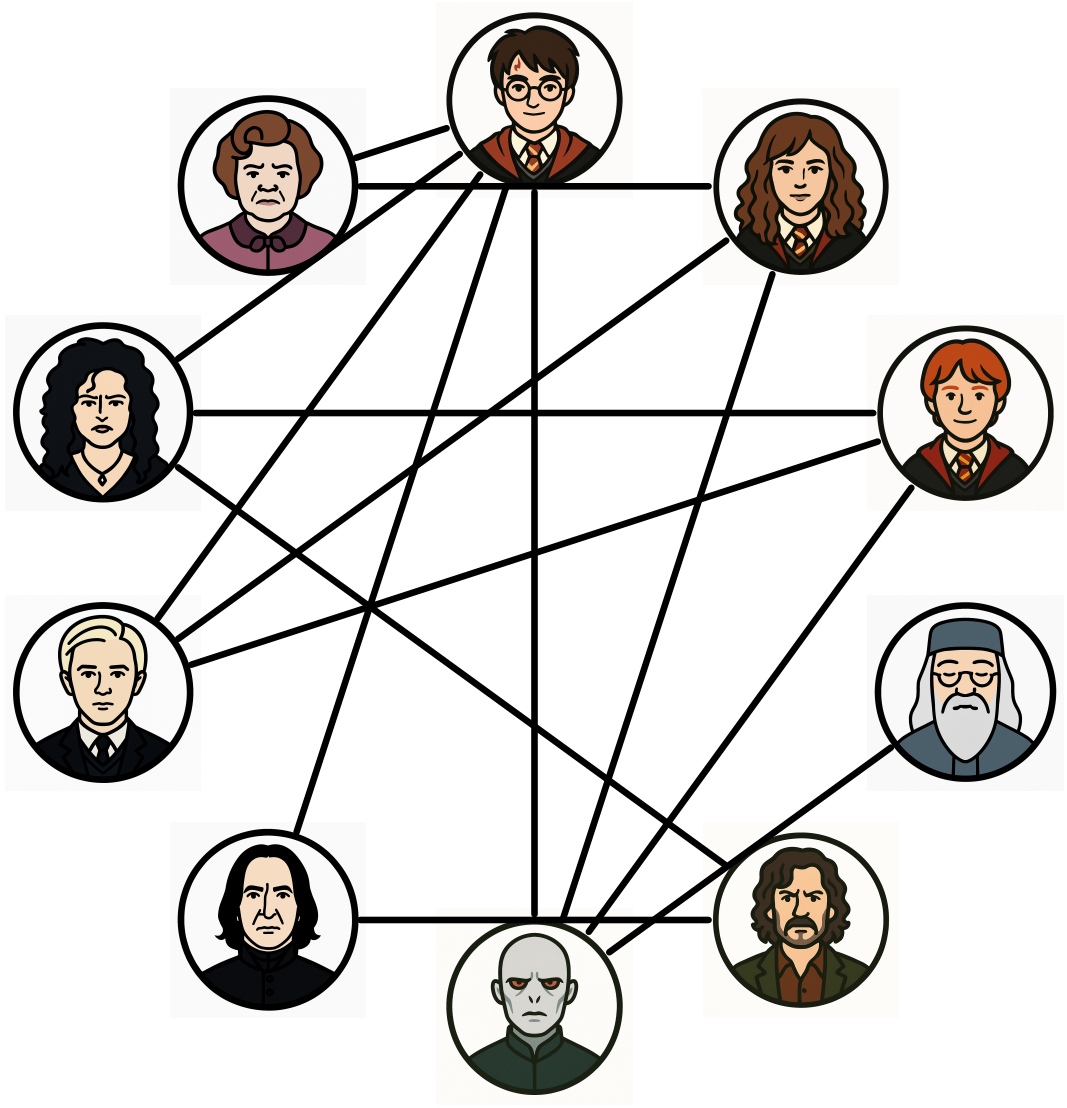
$$\|\mathbf{M} - \mathbf{A}\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}(m_{ij} - a_{ij})^2.$$

The best rank-d approximation of $\mathbf{A}$ is given by the first d singular values and vectors of $\mathbf{A}$:

$$\mathbf{M} = \sum_{i=1}^{d} \sigma_i u_i v_i^\top =: \mathbf{X}\mathbf{Y}^\top$$

where $\mathbf{X} = [\sigma_1^{1/2}u_1, \ldots, \sigma^{1/2}u_d] \in \mathbb{R}^{n \times d}$ and $\mathbf{Y} = [\sigma_1^{1/2}v_1, \ldots, \sigma_d^{1/2}v_d] \in \mathbb{R}^{m \times d}$.

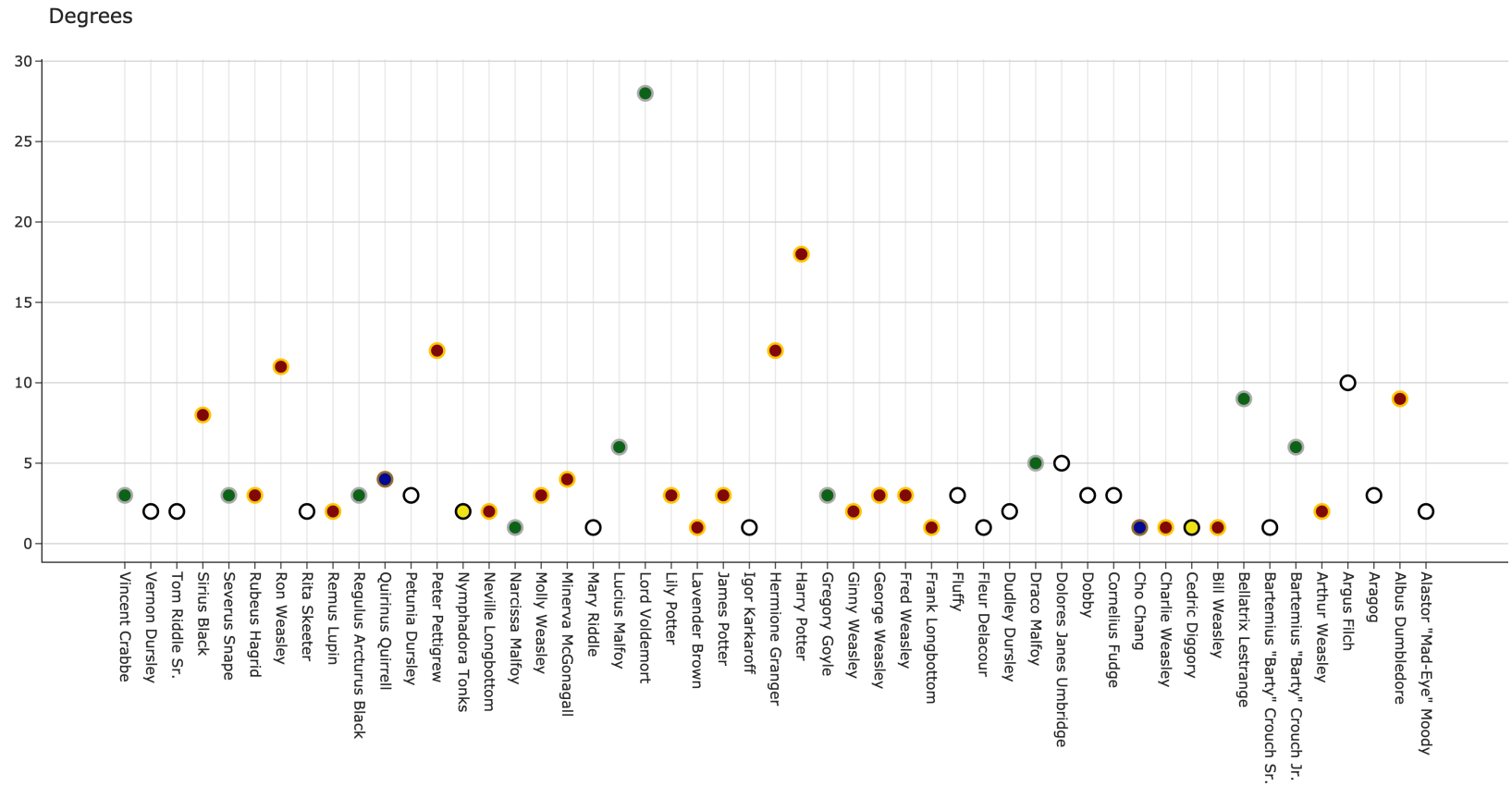
Example: enmities in *Harry Potter*



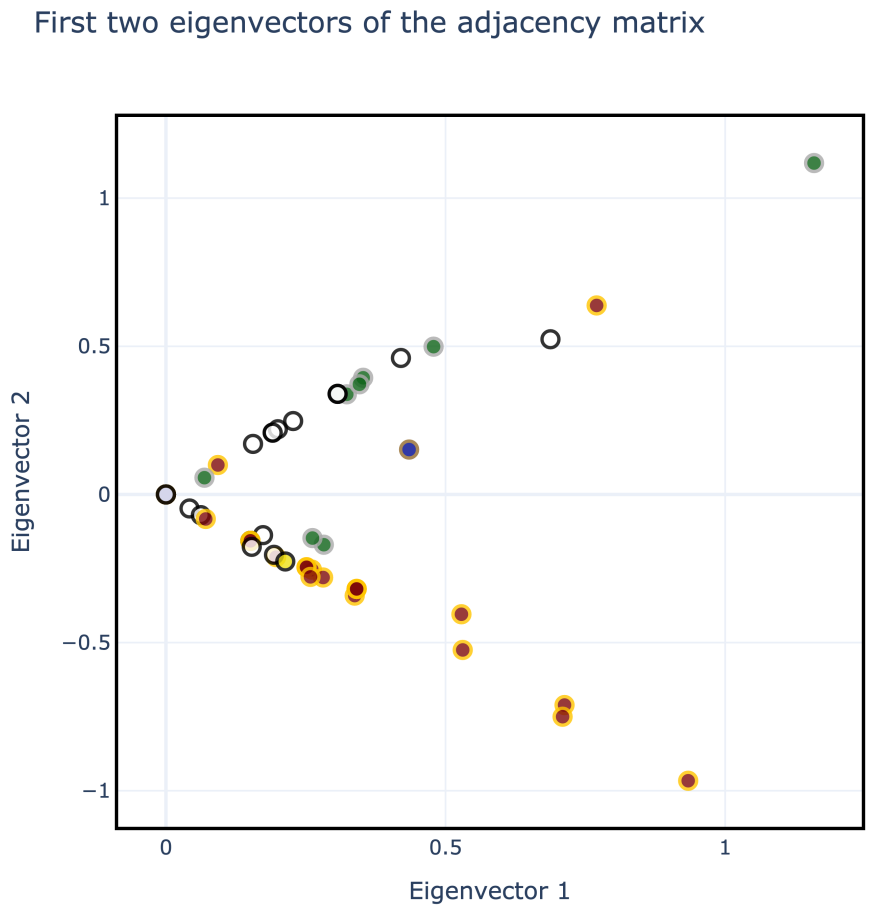
65 characters, 111 enmities.

`github.com/efekarakus/potter-network`

Number of enmities between characters in *Harry Potter*



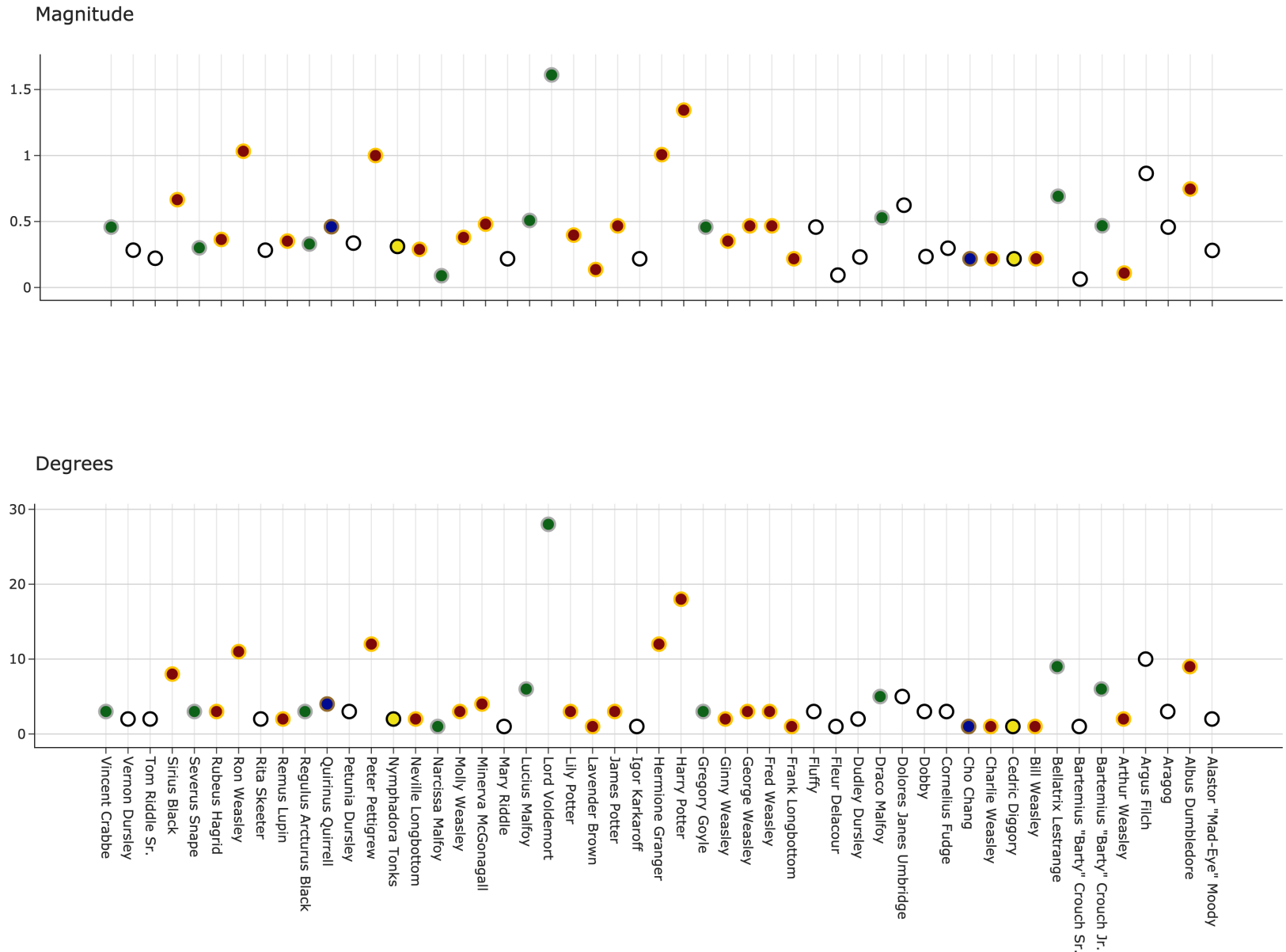
Singular vectors of the adjacency matrix



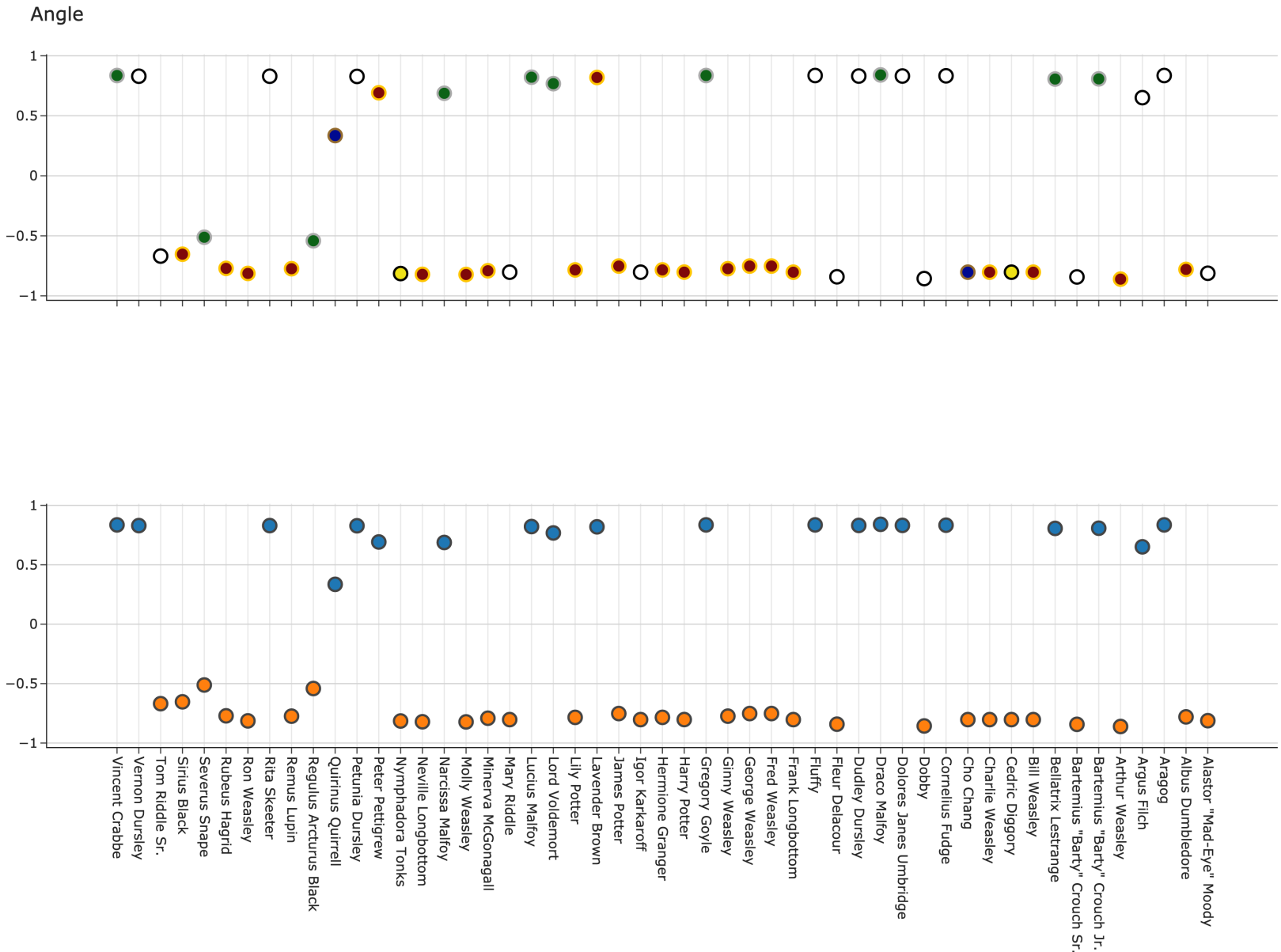
Intuition:

- magnitude captures number of connections;
- angle captures types of connections?

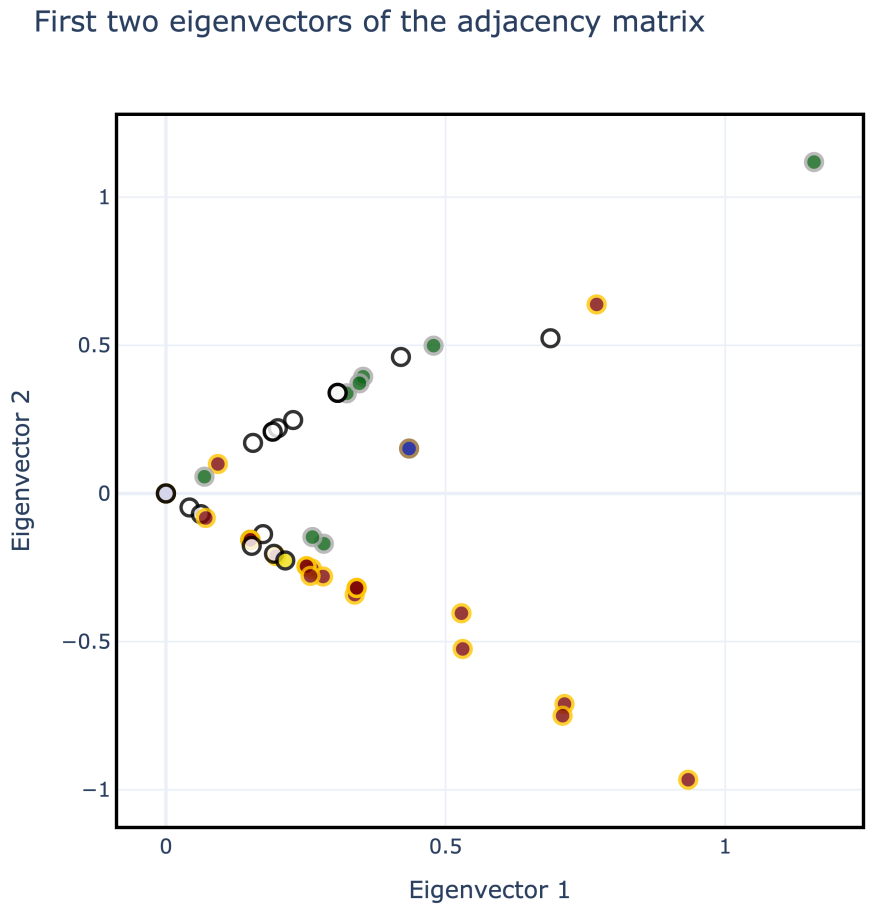
Singular vectors of the adjacency matrix



Eigenvectors of the adjacency matrix



How show we interpret the geometry in an adjacency spectral embedding?



Intuition:

- magnitude captures number of connections;
- angle captures types of connections?

Random dot product graph model

A graph $\mathbf{A}$ is said to follow a *random dot product graph model* if there exists a matrix of latent positions

$$\mathbf{X}_\star = \begin{pmatrix} X_{\star,1}^\top \\ \vdots \\ X_{\star,n}^\top \end{pmatrix} \in \mathbb{R}^{n \times d}$$

such that $\mathbf{A}$ has independent Bernoulli entries with probabilities

$$\mathbf{P} := \mathbb{E}(\mathbf{A}) = \mathbf{X}_\star \mathbf{X}_\star^\top.$$

Identifiability: The latent positions $\mathbf{X}_\star$ identifiable up to a orthognonal rotation, i.e. $\mathbf{P} = \mathbf{X}_\star \mathbf{X}_\star^\top = \mathbf{X}_\star(\mathbf{O}^\top \mathbf{O})\mathbf{X}_\star^\top = (\mathbf{X}_\star \mathbf{O}^\top)(\mathbf{X}_\star \mathbf{O}^\top)^\top$ for any rotation matrix $\mathbf{O}$ such that $\mathbf{O}^\top \mathbf{O} = \mathbf{I}$.

Spectral embedding estimates the latent positions of an RDPG

Spectral embedding estimates the latent positions of an RDPG

Let $X_{\star,1}, \ldots, X_{\star,n} \overset{\text{iid}}{\sim} F$ where $\mathcal{X} := \text{supp}(F) \subset \mathbb{R}^d$ is bounded and $\Delta := \mathbb{E}_{x \sim F}(xx^\top)$ has full rank.

Spectral embedding estimates the latent positions of an RDPG

Let $X_{\star,1}, \ldots, X_{\star,n} \overset{\text{iid}}{\sim} F$ where $\mathcal{X} := \text{supp}(F) \subset \mathbb{R}^d$ is bounded and $\Delta := \mathbb{E}_{x \sim F}(xx^\top)$ has full rank.

Suppose $\mathbf{A}$ follows an RDPG model with latent positions $\mathbf{X}_\star$. Then,

Spectral embedding estimates the latent positions of an RDPG

Let $X_{\star,1}, \ldots, X_{\star,n} \overset{\text{iid}}{\sim} F$ where $\mathcal{X} := \text{supp}(F) \subset \mathbb{R}^d$ is bounded and $\Delta := \mathbb{E}_{x \sim F}(xx^\top)$ has full rank.

Suppose $\mathbf{A}$ follows an RDPG model with latent positions $\mathbf{X}_\star$. Then,

Consistency: There exists a rotation matrix $\mathbf{O}$ such that

$$\max_{i \in \{1,\ldots,n\}} \|\mathbf{O}X_i - X_{\star,i}\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{\log n}{n}}\right)$$

Spectral embedding estimates the latent positions of an RDPG

Let $X_{\star,1}, \ldots, X_{\star,n} \overset{\text{iid}}{\sim} F$ where $\mathcal{X} := \text{supp}(F) \subset \mathbb{R}^d$ is bounded and $\Delta := \mathbb{E}_{x \sim F}(xx^\top)$ has full rank.

Suppose $\mathbf{A}$ follows an RDPG model with latent positions $\mathbf{X}_\star$. Then,

Consistency: There exists a rotation matrix $\mathbf{O}$ such that

$$\max_{i \in \{1,\ldots,n\}} \|\mathbf{O}X_i - X_{\star,i}\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{\log n}{n}}\right)$$

Asymptotic Gaussianity: Conditional on $X_{\star,i} = x$, there exists a rotation matrix $\mathbf{O}$ such that

$$n^{1/2}(\mathbf{O}X_i - X_{\star,i}) \to \mathcal{N}(0, \mathbf{\Sigma}(x))$$

where $\mathbf{\Sigma}(x) = \mathbb{E}_{\xi \sim F}\left\{x^\top \xi (1 - x^\top \xi)\xi\xi^\top\right\}$.

Stochastic block model

A graph $\mathbf{A}$ is said to follow a *stochastic block model* (SBM) if there exists

- community labels $z_1, \ldots, z_n \in \{1, \ldots, K\}$,
- a matrix of probabilities $\mathbf{B} \in [0,1]^{K \times K}$

such that

$$\mathbf{A}_{ij} \overset{\text{iid}}{\sim} \text{Bernoulli}(\mathbf{B}_{z_i, z_j}), \qquad i < j.$$

Stochastic block model

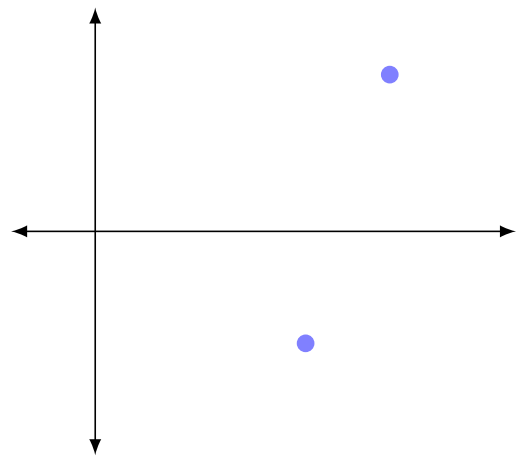
A graph $\mathbf{A}$ is said to follow a *stochastic block model* (SBM) if there exists

- community labels $z_1, \ldots, z_n \in \{1, \ldots, K\}$,
- a matrix of probabilities $\mathbf{B} \in [0,1]^{K \times K}$

such that

$$\mathbf{A}_{ij} \overset{\text{iid}}{\sim} \text{Bernoulli}(\mathbf{B}_{z_i, z_j}), \qquad i < j.$$

An SBM is a special case of an RDPG model with latent positions $X_{\star,i} = v_{z_i}$ where $v_1, \ldots, v_K \in \mathbb{R}^d$ satisfy $v_k^\top v_l = \mathbf{B}_{kl}$.



Stochastic block model

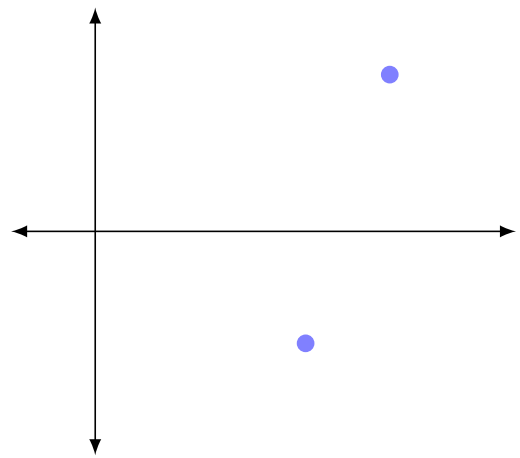
A graph $\mathbf{A}$ is said to follow a *stochastic block model* (SBM) if there exists

- community labels $z_1, \ldots, z_n \in \{1, \ldots, K\}$,
- a matrix of probabilities $\mathbf{B} \in [0,1]^{K \times K}$

such that

$$\mathbf{A}_{ij} \overset{\text{iid}}{\sim} \text{Bernoulli}(\mathbf{B}_{z_i, z_j}), \qquad i < j.$$

An SBM is a special case of an RDPG model with latent positions $X_{\star,i} = v_{z_i}$ where $v_1, \ldots, v_K \in \mathbb{R}^d$ satisfy $v_k^\top v_l = \mathbf{B}_{kl}$.



Each node in the same community has the same expected degree.

Degree-corrected stochastic block model

A graph $\mathbf{A}$ is said to follow a *degree-corrected stochastic block model* (DCSBM) if there exists

- node-specific weights $w_1, \ldots, w_n \in [0, \infty)$,
- community labels $z_1, \ldots, z_n \in \{1, \ldots, K\}$,
- a matrix of probabilities $\mathbf{B} \in [0,1]^{K \times K}$

such that

$$\mathbf{A}_{ij} \overset{\text{iid}}{\sim} \text{Bernoulli}(w_i w_j \mathbf{B}_{z_i, z_j}), \qquad i < j.$$

Degree-corrected stochastic block model

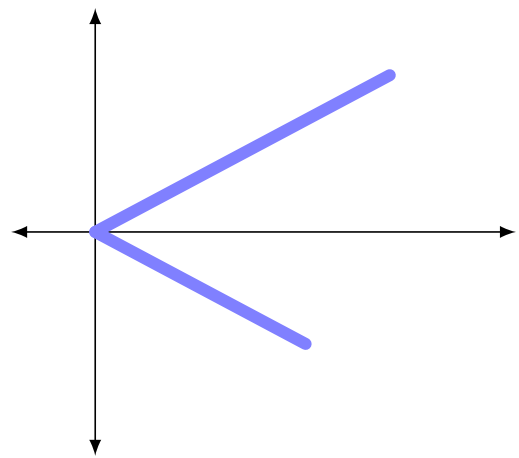
A graph $\mathbf{A}$ is said to follow a *degree-corrected stochastic block model* (DCSBM) if there exists

- node-specific weights $w_1, \ldots, w_n \in [0, \infty)$,
- community labels $z_1, \ldots, z_n \in \{1, \ldots, K\}$,
- a matrix of probabilities $\mathbf{B} \in [0,1]^{K \times K}$

such that

$$\mathbf{A}_{ij} \overset{\text{iid}}{\sim} \text{Bernoulli}(w_i w_j \mathbf{B}_{z_i, z_j}), \qquad i < j.$$

An DCSBM is a special case of an RDPG model with latent positions $X_{\star,i} = w_i v_{z_i}$ where $v_1, \ldots, v_K \in \mathbb{R}^d$ satisfy $v_k^\top v_l = \mathbf{B}_{kl}$.



Degree-corrected stochastic block model

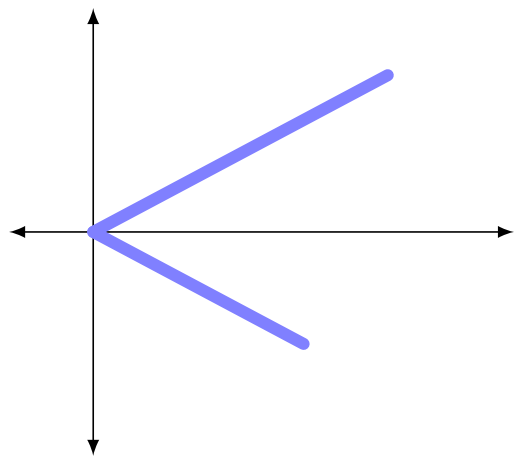
A graph $\mathbf{A}$ is said to follow a *degree-corrected stochastic block model* (DCSBM) if there exists

- node-specific weights $w_1, \ldots, w_n \in [0, \infty)$,
- community labels $z_1, \ldots, z_n \in \{1, \ldots, K\}$,
- a matrix of probabilities $\mathbf{B} \in [0,1]^{K \times K}$

such that

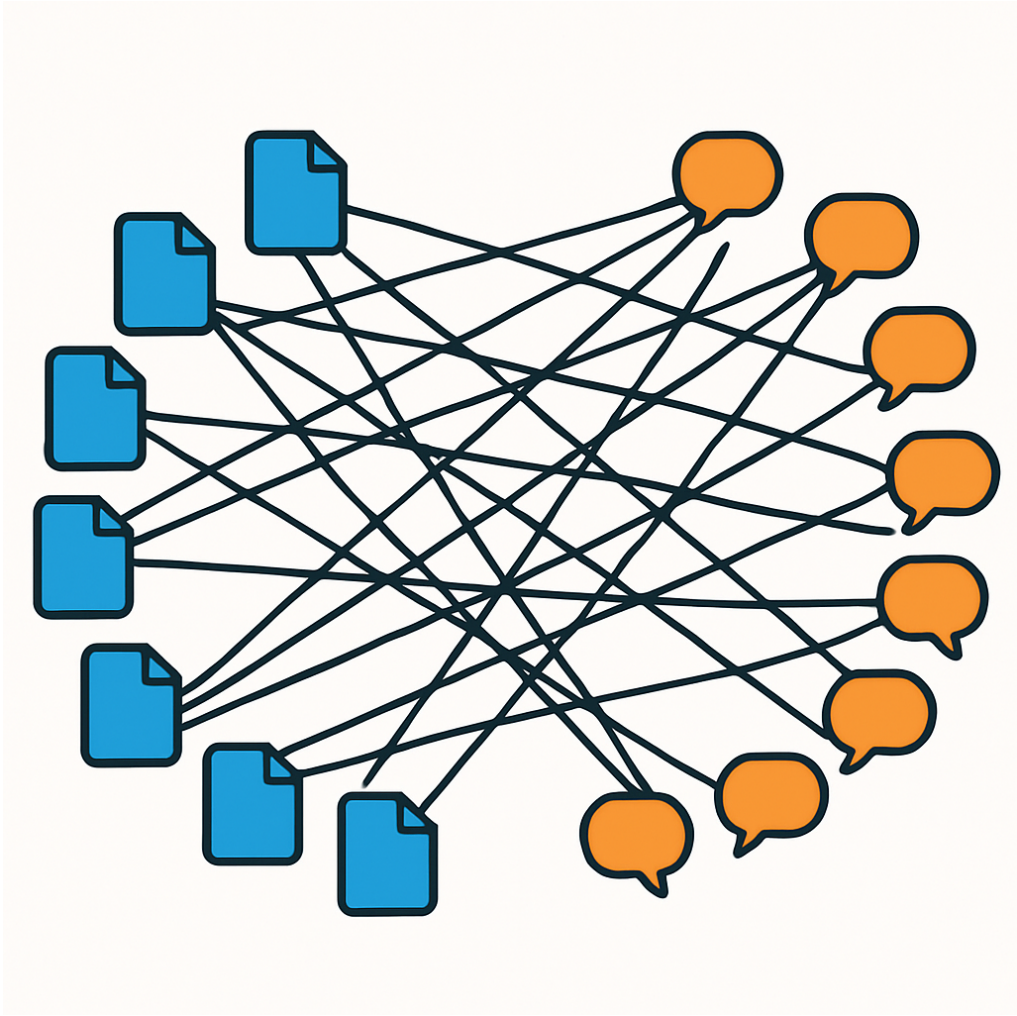
$$\mathbf{A}_{ij} \overset{\text{iid}}{\sim} \text{Bernoulli}(w_i w_j \mathbf{B}_{z_i, z_j}), \qquad i < j.$$

An DCSBM is a special case of an RDPG model with latent positions $X_{\star,i} = w_i v_{z_i}$ where $v_1, \ldots, v_K \in \mathbb{R}^d$ satisfy $v_k^\top v_l = \mathbf{B}_{kl}$.



What about when there's no clear community structure?.

Example: document-term bipartite network



1,800 new articles from the Associated Press, 7,338 terms.

$\mathbf{A}_{ij} = 1$ if article i contains term j, and 0 otherwise.

Degree-corrected mixed-membership stochastic block model

A graph $\mathbf{A}$ is said to follow a *degree-corrected mixed-membership stochastic block model* (DCMM-SBM) if there exists

- node-specific weights $w_1, \ldots, w_n \in [0, \infty)$,
- community allocation vectors $\pi_1, \ldots, \pi_n \in \mathbb{S}^{K-1}$,
- a matrix of probabilities $\mathbf{B} \in [0,1]^{K \times K}$

such that

$$\mathbf{A}_{ij} \overset{\text{iid}}{\sim} \text{Bernoulli}(w_i w_j \, \pi_i^\top \mathbf{B} \pi_j), \qquad i < j.$$

Degree-corrected mixed-membership stochastic block model

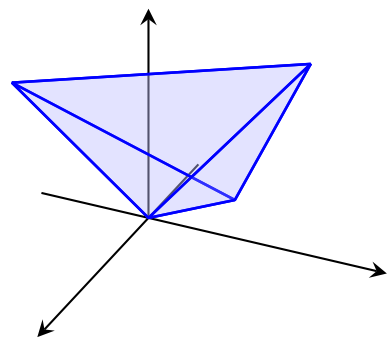
A graph $\mathbf{A}$ is said to follow a *degree-corrected mixed-membership stochastic block model* (DCMM-SBM) if there exists

- node-specific weights $w_1, \ldots, w_n \in [0, \infty)$,
- community allocation vectors $\pi_1, \ldots, \pi_n \in \mathbb{S}^{K-1}$,
- a matrix of probabilities $\mathbf{B} \in [0,1]^{K \times K}$

such that

$$\mathbf{A}_{ij} \overset{\text{iid}}{\sim} \text{Bernoulli}(w_i w_j \, \pi_i^\top \mathbf{B} \pi_j), \qquad i < j.$$

An DCMM-SBM is a special case of an RDPG model with latent positions $X_{\star,i} = w_i \sum_{k=1}^{K} \pi_{ik} v_k$ where $v_1, \ldots, v_K \in \mathbb{R}^d$ satisfy $v_k^\top v_l = \mathbf{B}_{kl}$.



The normalized Laplacian matrix

Recall that $\mathbf{M} = \mathbf{X}\mathbf{Y}^\top$ minimises the criterion

$$\|\mathbf{M} - \mathbf{A}\|_F^2 = \sum_{i,j=1}^{n} (\mathbf{M}_{ij} - \mathbf{A}_{ij})^2 \tag{1}$$

The normalized Laplacian matrix

Recall that $\mathbf{M} = \mathbf{X}\mathbf{Y}^\top$ minimises the criterion

$$\|\mathbf{M} - \mathbf{A}\|_F^2 = \sum_{i,j=1}^{n} (\mathbf{M}_{ij} - \mathbf{A}_{ij})^2 \tag{1}$$

This optimisation problem is most heavily influenced by *high-degree* nodes.

The normalized Laplacian matrix

Recall that $\mathbf{M} = \mathbf{X}\mathbf{Y}^\top$ minimises the criterion

$$\|\mathbf{M} - \mathbf{A}\|_F^2 = \sum_{i,j=1}^{n} (\mathbf{M}_{ij} - \mathbf{A}_{ij})^2 \tag{1}$$

This optimisation problem is most heavily influenced by *high-degree* nodes.

The normalised Laplacian $\mathbf{L}$ is the matrix with entries

$$(\mathbf{L})_{ij} = \frac{\mathbf{A}_{ij}}{\sqrt{d_i d_j}}$$

where $d_i = \sum_{j=1}^{n} \mathbf{A}_{ij}$ is the degree of node i and $\tau > 0$ is a regularisation parameter.

The normalized Laplacian matrix

Recall that $\mathbf{M} = \mathbf{X}\mathbf{Y}^\top$ minimises the criterion

$$\|\mathbf{M} - \mathbf{A}\|_F^2 = \sum_{i,j=1}^{n} (\mathbf{M}_{ij} - \mathbf{A}_{ij})^2 \tag{1}$$

This optimisation problem is most heavily influenced by *high-degree* nodes.

The normalised Laplacian $\mathbf{L}$ is the matrix with entries

$$(\mathbf{L})_{ij} = \frac{\mathbf{A}_{ij}}{\sqrt{d_i d_j}}$$

where $d_i = \sum_{j=1}^{n} \mathbf{A}_{ij}$ is the degree of node i and $\tau > 0$ is a regularisation parameter.

Replacing $\mathbf{A}$ with $\mathbf{L}$ in (1) increases the influence of *low-degree* nodes in the optimsation.

The regularized Laplacian

The regularized Laplacian $\mathbf{L}_\tau$ is a matrix which interpolates between $\mathbf{L}$ and $\mathbf{A}$.

The regularized Laplacian

The regularized Laplacian $\mathbf{L}_\tau$ is a matrix which interpolates between $\mathbf{L}$ and $\mathbf{A}$.

The normalised Laplacian $\mathbf{L}$ is the matrix with entries

$$(\mathbf{L}_\tau)_{ij} = \frac{(1+\tau)\mathbf{A}_{ij}}{\sqrt{(d_i+\tau)(d_j+\tau)}}$$

where $d_i = \sum_{j=1}^{n} \mathbf{A}_{ij}$ is the degree of node i and $\tau > 0$ is a regularisation parameter.

The regularized Laplacian

The regularized Laplacian $\mathbf{L}_\tau$ is a matrix which interpolates between $\mathbf{L}$ and $\mathbf{A}$.

The normalised Laplacian $\mathbf{L}$ is the matrix with entries

$$(\mathbf{L}_\tau)_{ij} = \frac{(1+\tau)\mathbf{A}_{ij}}{\sqrt{(d_i+\tau)(d_j+\tau)}}$$

where $d_i = \sum_{j=1}^{n} \mathbf{A}_{ij}$ is the degree of node i and $\tau > 0$ is a regularisation parameter.

Note that $\mathbf{L}_0 = \mathbf{L}$ and $\mathbf{L}_\infty := \lim_{\tau\to\infty} \mathbf{L}_\tau = \mathbf{A}$.

Manifold structure in graph embeddings

Manifold structure in graph embeddings

Patrick Rubin-Delanchy
University of Bristol
patrick.rubin-delanchy@bristol.ac.uk

Matrix factorisation and the interpretation of geodesic distance

Nick Whiteley
University of Bristol
nick.whiteley@bristol.ac.uk

Annie Gray
University of Bristol
annie.gray@bristol.ac.uk

Patrick Rubin-Delanchy
University of Bristol
patrick.rubin-delanchy@bristol.ac.uk

The Origins of Representation Manifolds in Large Language Models

Alexander Modell
Department of Mathematics
Imperial College London
a.modell@imperial.ac.uk

Patrick Rubin-Delanchy
School of Mathematics
University of Edinburgh
prd@ed.ac.uk

Nick Whiteley
School of Mathematics
University of Bristol
nick.whiteley@bristol.ac.uk

Latent position models

Latent position models

Under the Random Dot Product Graph, we assume that there exists latent positions $X_1, \ldots, X_n \in \mathbb{R}^d$ and that

$$\mathbf{A}_{ij} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\mathbf{P}_{ij}), \qquad \mathbf{P}_{ij} = X_i^\top X_j.$$

Latent position models

Under the Random Dot Product Graph, we assume that there exists latent positions $X_1, \ldots, X_n \in \mathbb{R}^d$ and that

$$\mathbf{A}_{ij} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\mathbf{P}_{ij}), \qquad \mathbf{P}_{ij} = X_i^\top X_j.$$

Under this assumption, theory (informally) tells us that there exists an orthogonal matrix $\mathbf{O}$ such that $\mathbf{O}\widehat{X}_i \to X_i$ with high-probability, and in an asymptotically Gaussian manner.

Latent position models

Under the Random Dot Product Graph, we assume that there exists latent positions $X_1, \ldots, X_n \in \mathbb{R}^d$ and that

$$\mathbf{A}_{ij} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\mathbf{P}_{ij}), \qquad \mathbf{P}_{ij} = X_i^\top X_j.$$

Under this assumption, theory (informally) tells us that there exists an orthogonal matrix $\mathbf{O}$ such that $\mathbf{O}\widehat{X}_i \to X_i$ with high-probability, and in an asymptotically Gaussian manner.

What about if we replace the dot product with a different link function?

Latent position models

Under the Random Dot Product Graph, we assume that there exists latent positions $X_1, \ldots, X_n \in \mathbb{R}^d$ and that

$$\mathbf{A}_{ij} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\mathbf{P}_{ij}), \qquad \mathbf{P}_{ij} = X_i^\top X_j.$$

Under this assumption, theory (informally) tells us that there exists an orthogonal matrix $\mathbf{O}$ such that $\mathbf{O}\widehat{X}_i \to X_i$ with high-probability, and in an asymptotically Gaussian manner.

What about if we replace the dot product with a different link function?

Definition (Latent position model)

Let $f : \mathcal{Z} \times \mathcal{Z} \to [0, 1]$ be a symmetric function. A graph $\mathbf{A}$ is said to follow a *latent position model* if

$$\mathbf{A}_{ij} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\mathbf{P}_{ij}), \qquad \mathbf{P}_{ij} = f(Z_i, Z_j).$$

Some examples

Let's simulate some example graphs from this model and take a look at their spectral embeddings.

- Sociability kernel: $f(x, y) = 1 - \exp(-2xy)$
- Gaussian radial basis kernel: $f(x, y) = \exp(-\|x - y\|^2/(2\sigma^2))$

Why does manifold structure emerge in spectral embeddings of latent position model graphs?

Let's consider if there exists some $X_1, \ldots, X_n \in \mathbb{R}^r$ such that

$$\mathbf{P}_{ij} = f(Z_i, Z_j) \approx X_i^\top X_j.$$

Then, then the theory from random dot product graphs might suggest that (up to a rotation), the spectral embeddings estimate $X_1, \ldots, X_n$.

Why does manifold structure emerge in spectral embeddings of latent position model graphs?

Let's consider if there exists some $X_1, \ldots, X_n \in \mathbb{R}^r$ such that

$$\mathbf{P}_{ij} = f(Z_i, Z_j) \approx X_i^\top X_j.$$

Then, then the theory from random dot product graphs might suggest that (up to a rotation), the spectral embeddings estimate $X_1, \ldots, X_n$.

Theorem (Mercer's theorem)

Suppose that $(\mathcal{Z}, d, \mu)$ is a metric measure space and $f : \mathcal{Z} \times \mathcal{Z} \to [0, \kappa)$ is a continuous, bounded, positive semi-definite kernel. Then there exists a system of eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots > 0$ and corresponding orthonormal eigenfunctions $u_1, u_2, \ldots \in L^2(\mu)$ such that

$$f(x, y) = \sum_{i=1}^{\infty} \lambda_i u_i(x) u_i(y), \qquad x, y \in \mathcal{Z}$$

where the infinite sum converges uniformly and absolutely.

Feature maps

Feature maps

We define the following infinite-dimensional feature maps $\phi : \mathcal{Z} \to \ell_2$ from the eigendecomposition of f:

$$\phi(z) := \begin{pmatrix} \lambda_1^{1/2} u_1(z) & \lambda_2^{1/2} u_2(z) & \lambda_3^{1/2} u_3(z) & \cdots \end{pmatrix}$$

Feature maps

We define the following infinite-dimensional feature maps $\phi : \mathcal{Z} \to \ell_2$ from the eigendecomposition of f:

$$\phi(z) := \begin{pmatrix} \lambda_1^{1/2} u_1(z) & \lambda_2^{1/2} u_2(z) & \lambda_3^{1/2} u_3(z) & \cdots \end{pmatrix}$$

The eigendecomposition of f can be rewritten using the feature maps:

$$f(x, y) = \sum_{i=1}^{\infty} \lambda_i u_i(x) u_i(y) = \sum_{i=1}^{\infty} \phi_i(x) \phi_i(y) = \langle \phi(x), \phi(y) \rangle_2$$

Feature maps

We define the following infinite-dimensional feature maps $\phi : \mathcal{Z} \to \ell_2$ from the eigendecomposition of f:

$$\phi(z) := \begin{pmatrix} \lambda_1^{1/2} u_1(z) & \lambda_2^{1/2} u_2(z) & \lambda_3^{1/2} u_3(z) & \cdots \end{pmatrix}$$

The eigendecomposition of f can be rewritten using the feature maps:

$$f(x, y) = \sum_{i=1}^{\infty} \lambda_i u_i(x) u_i(y) = \sum_{i=1}^{\infty} \phi_i(x) \phi_i(y) = \langle \phi(x), \phi(y) \rangle_2$$

Letting $X_i = \phi(Z_i)$, we have that $\mathbf{P}_{ij} = \langle X_i, X_j \rangle_2$.

Truncated feature maps

Truncated feature maps

We typically perform spectral embedding into some small dimension, whereas these feature maps we have defined are infinite dimensional. How do we make sense of this?

Truncated feature maps

We typically perform spectral embedding into some small dimension, whereas these feature maps we have defined are infinite dimensional. How do we make sense of this?

Let's consider the truncated feature map $\widetilde{\phi} : \mathcal{Z} \to \mathbb{R}^d$:

$$\widetilde{\phi}(z) = \begin{pmatrix} \lambda_1^{1/2} u_1(z) & \lambda_2^{1/2} u_2(z) & \cdots & \lambda_d^{1/2} u_d(z) \end{pmatrix}$$

and set $\widetilde{X}_i = \widetilde{\phi}(Z_i)$.

Truncated feature maps

We typically perform spectral embedding into some small dimension, whereas these feature maps we have defined are infinite dimensional. How do we make sense of this?

Let's consider the truncated feature map $\widetilde{\phi} : \mathcal{Z} \to \mathbb{R}^d$:

$$\widetilde{\phi}(z) = \begin{pmatrix} \lambda_1^{1/2} u_1(z) & \lambda_2^{1/2} u_2(z) & \cdots & \lambda_d^{1/2} u_d(z) \end{pmatrix}$$

and set $\widetilde{X}_i = \widetilde{\phi}(Z_i)$.

Then we can show that

$$\left\| \widetilde{\phi} - \phi \right\|_{L^2(\mu)} = \sum_{i=d+1}^{\infty} \lambda_i$$

where I've assumed that $\widetilde{\phi}(z)$ is padded with infinitely many zeroes so it lives in ℓ_2.

Truncated feature maps

We typically perform spectral embedding into some small dimension, whereas these feature maps we have defined are infinite dimensional. How do we make sense of this?

Let's consider the truncated feature map $\widetilde{\phi} : \mathcal{Z} \to \mathbb{R}^d$:

$$\widetilde{\phi}(z) = \begin{pmatrix} \lambda_1^{1/2} u_1(z) & \lambda_2^{1/2} u_2(z) & \cdots & \lambda_d^{1/2} u_d(z) \end{pmatrix}$$

and set $\widetilde{X}_i = \widetilde{\phi}(Z_i)$.

Then we can show that

$$\left\| \widetilde{\phi} - \phi \right\|_{L^2(\mu)} = \sum_{i=d+1}^{\infty} \lambda_i$$

where I've assumed that $\widetilde{\phi}(z)$ is padded with infinitely many zeroes so it lives in ℓ_2.

This is small whenever d is large enough relative to the eigenvalue decay of f.

Truncated feature maps

If this is the case, then we have that

$$\mathbf{P}_{ij} = \langle X_i, X_j \rangle_2 \approx \widetilde{X}_i^\top \widetilde{X}_j.$$

So, we might expect that our random dot product graph theory applies, at least approximately, to these latent positions under a latent position model.

Geometry of feature maps

Suppose that

A1. For all $x, y \in \mathcal{Z}$, there exists $z \in \mathcal{Z}$ such that $f(x, z) \neq f(y, z)$.

A2. f is twice continuously differentiable and the matrix $\mathbf{H}_z$ with entries

$$(\mathbf{H}_z)_{ij} := \frac{\partial f^2}{\partial x_i \partial y_j}|_{(z,z)}, \qquad i, j = 1, \ldots, d.$$

is positive definite for all $z \in \mathcal{Z}$.

Geometry of feature maps

Suppose that

A1. For all $x, y \in \mathcal{Z}$, there exists $z \in \mathcal{Z}$ such that $f(x, z) \neq f(y, z)$.

A2. f is twice continuously differentiable and the matrix $\mathbf{H}_z$ with entries

$$(\mathbf{H}_z)_{ij} := \frac{\partial f^2}{\partial x_i \partial y_j}|_{(z,z)}, \qquad i, j = 1, \ldots, d.$$

is positive definite for all $z \in \mathcal{Z}$.

Then, by *Whiteley et al. (2021)* we have the following result.

Theorem (homeomorphism)

Suppose **A1** *and* **A2** *hold. Then ϕ is a bi-Lipschitz homeomorphism between $\mathcal{Z}$ and $\mathcal{M} := \phi(\mathcal{Z})$.*

Interpreting geodesic distances

Suppose that f is a function of the distance in the metric space, i.e.

A3. There exists a twice continuously diffentiable function $g : [0, \infty) \to [0, \kappa)$ with $g'(0) < 0$ such that

$$f(x, y) = g(d(x, y)^2), \qquad x, y \in \mathcal{Z}.$$

Interpreting geodesic distances

Suppose that f is a function of the distance in the metric space, i.e.

A3. There exists a twice continuously diffentiable function $g : [0, \infty) \to [0, \kappa)$ with $g'(0) < 0$ such that

$$f(x, y) = g(d(x, y)^2), \qquad x, y \in \mathcal{Z}.$$

Then, by *Modell et al. (2025)* (see also *Whiteley et al. (2021)*),

Theorem (isometry)

Let η be a path on $\mathcal{Z}$ of finite length and let γ be corresponding path on $\mathcal{M} := \phi(\mathcal{Z})$, then assuming **A1** *and* **A3**,

$$L(\gamma) = \sqrt{-2g'(0)}L(\eta)$$

where $L(\cdot)$ denote the length of a path.

Estimating geodesic distances in practice

Algorithm 1 Isomap procedure

input p-dimensional points $\hat{X}_1, \ldots, \hat{X}_n$

1: Compute the neighbourhood graph of radius ϵ: a weighted graph connecting i and j, with weight $\|\hat{X}_i - \hat{X}_j\|$, if $\|\hat{X}_i - \hat{X}_j\| \leq \epsilon$

2: Compute the matrix of shortest paths on the neighbourhood graph, $\hat{\mathbf{D}}_{\mathcal{M}} \in \mathbb{R}^{n \times n}$

3: Apply classical multidimensional scaling (CMDS) to $\hat{\mathbf{D}}_{\mathcal{M}}$ into $\mathbb{R}^d$

return d-dimensional points $\hat{Z}_1, \ldots, \hat{Z}_n$

Simulated example from *Whiteley et al. (2021)*

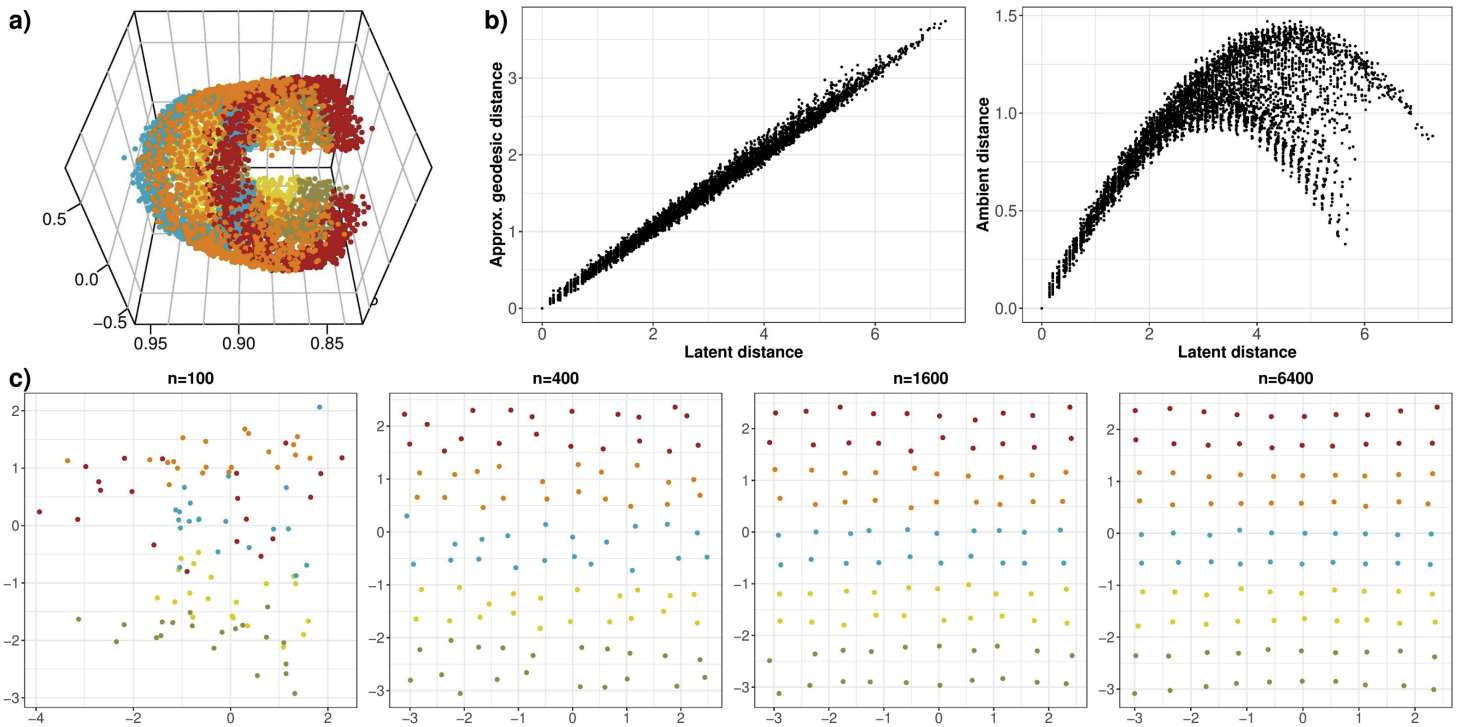


Figure 2: Simulated data example. a) Spectral embedding (first 3 dimensions), b) comparison of latent, approximate geodesic, and ambient distance and c) latent position recovery in the dense regime by spectral embedding followed by Isomap for increasing n. To aid visualisation, all plots in c) display a subset of 100 estimated positions corresponding to true positions on a sub-grid which is common across n. Estimated positions are coloured according to their true y-coordinate.

Manifold structure in large language model embeddings?

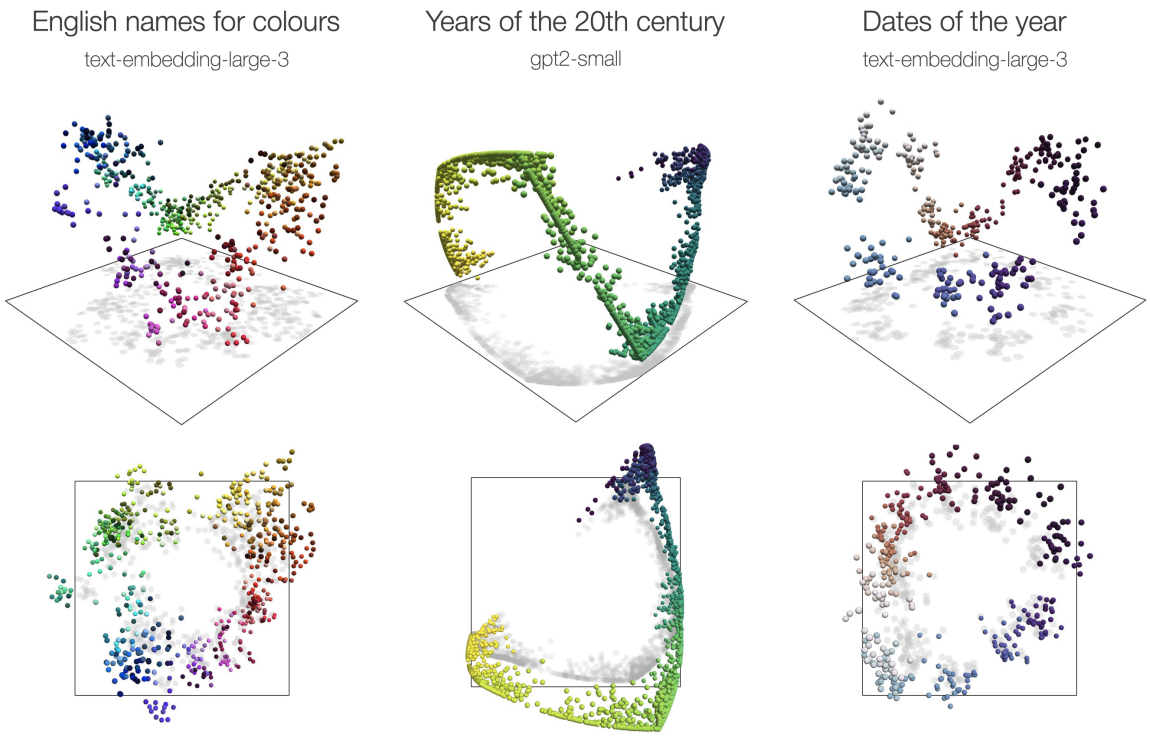


Figure 1: Representation manifolds in large language models: colours, years and dates. The first and third example show text embeddings obtained from OpenAI's `text-embedding-large-3` model from prompts relating to English names for colours and dates of the year, respectivly. The second example shows token activations from layer 7 of `GPT2-small`, which were studied in Engels et al. (2025). The token activations were processed via an SAE to extract a feature corresponding to years of the twentieth century as in Engels et al. (2025), and normalized to have norm one. For each example, we perform principal component analysis (PCA) to reduce the dimension to three and display the resulting point clouds from two perspectives. The embeddings of English names for colours are displayed in their respective colour value. Years are coloured from blue (1900) through green to yellow (1999), and dates are coloured from white (1st Janurary) through blue to black (1st July) through red and back to white.

Workshop: organize James and Emmy's wedding!

`amodell.me/nest-workshop.html`