IMPERIAL

MSci Research Project

Bayesian Inference for Latent Structures in Text-Attributed Networks

Author:
Benjamin Pellow

Supervisor:
Francesco Sanna Passino

Submitted in partial fulfillment of the requirements for the MSci in Mathematics at Imperial College London

June 9, 2025

Abstract

We propose a Bayesian model for clustering nodes in text-attributed networks by jointly modelling graph-based community structure and node-level textual attributes. Each node is represented by a low-dimensional spectral embedding and a sparse bag-of-words vector, with clustering performed via a collapsed Gibbs sampler that integrates multivariate Gaussian likelihoods with Dirichlet-Multinomial models. To balance the relative contribution of graph structure and text, the sampler includes tunable modality reweighting parameters. The model is evaluated on synthetic data to test robustness under varying levels of embedding separation, topic distinctiveness, and class imbalance, and is further applied to the Cora citation network, where it demonstrates improved recovery of latent structure over graph-only or text-only approaches. Initialising cluster assignments and prior covariances using Gaussian mixture fits enhances convergence and stabilises inference. The results highlight the value of hybrid probabilistic models in multimodal network data, especially when one modality, such as text in the Cora dataset, is weak, noisy, or semantically overlapping.

Acknowledgments

I would like to thank my supervisor, Dr Francesco Sanna Passino, for his invaluable support and guidance throughout this project. Our weekly meetings were always insightful, and I greatly appreciated his thoughtful feedback and patience in helping me navigate both conceptual and technical challenges.

Plagiarism statement

The work contained in this thesis is my own work unless otherwise stated.

Signature: Benjamin Pellow
Date: June 9, 2025

Contents

Chapter 1

Introduction

Understanding the structure underlying network data is a central goal in modern data science. Many real-world networks-such as social networks, citation networks, and communication patterns are believed to be governed by lower-dimensional latent geometries. This intuition is formalised in the manifold hypothesis (Whiteley et al., 2025), which asserts that data which appear high-dimensional often lie close to a lower-dimensional manifold within that space.

However, many methods rely exclusively on the graph structure, neglecting auxiliary information often available at the node level, such as text, tags, or metadata. In this project, we address this limitation by proposing a novel Bayesian model that performs joint clustering in the spectral embedding space while incorporating node-level, textual attributes via topic modelling. This allows us to infer latent structures that are coherent both structurally and semantically.

To assess the model's effectiveness, we conduct experiments on both simulated networks and the Cora citation dataset, a well-known benchmark in network science where nodes represent scientific publications and edges denote citation relationships. Each node is further associated with a sparse bag-of-words feature vector capturing the content of the corresponding paper. This setting provides an ideal test case for evaluating our method's ability to recover clusters that are meaningful in terms of both graph structure and semantics. While the textual signals in Cora are relatively weak and noisy, our results suggest that integrating graph topology with node-level textual covariates can offer meaningful improvements over using graph structure alone, particularly in datasets with clearer thematic separation or richer node annotations.

This project makes two main methodological contributions. First, we develop a collapsed Gibbs sampler for clustering in text-attributed networks, combining Gaussian likelihoods over node embeddings with Dirichlet-Multinomial models for textual attributes. Second, we introduce tunable reweighting parameters that balance the influence of structure and content, allowing the model to adapt to modality quality. These innovations are evaluated in a series of simulations and applied to the Cora citation network dataset, where we demonstrate the importance of modality weighting and prior choice in recovering ground truth structure. In our implementation, we initialise prior covariances using the within-cluster variances from a Gaussian Mixture Model (GMM) fitted to the embeddings, ensuring that prior scales reflect the observed data variability.

Chapter 2

Background

2.1 Network representation

We can mathematically specify a network $\mathcal{G} = (V, E)$ as a set of nodes $V = \{1, 2, \ldots, n\}$ and a set of edges $E \subseteq V \times V$, such that $(i, j) \in E$ if i is connected to j. Throughout this paper we will focus on *undirected* networks, where $(i, j) \in E \implies (j, i) \in E$. A network can be fully characterised by its adjacency matrix $\mathbf{A} \in \{0, 1\}^{n \times n}$, where $A_{ij} = \mathbb{1}_E\{(i, j)\}$, where $\mathbb{1}_E\{\cdot\}$ denotes the indicator function.

2.1.1 Text attributed networks

In many applications, nodes in a network are associated not only with links to other nodes but also with descriptive textual features. We extend the definition of a network to a *text-attributed network* by including a matrix of node-level textual attributes. Formally, we define

$$\mathcal{G} = (V, E, \mathbf{W}),$$

where $\mathbf{W} \in \mathbb{N}_0^{n \times |\mathcal{V}|}$ is a matrix of bag-of-words counts, with each row $\mathbf{w}_i$ representing the textual data associated with node i, over a vocabulary $\mathcal{V}$ of size $|\mathcal{V}|$.

This representation allows us to jointly analyse both the structure of the graph and the content associated with each node, forming the basis for multimodal clustering approaches explored in this report.

2.2 Random dot product graphs

A natural and tractable model for purely structural network data in which edges are assumed to arise from latent geometric structure, consistent with the manifold hypothesis (Whiteley et al., 2025), is the Random Dot Product Graph (RDPG; Young and Scheinerman (2007)), a special case of the more general Latent Position Model (LPM; Hoff et al. (2002)). In the LPM framework, each node is associated with a latent position in a metric space $\mathcal{X}$ (the latent space). Edges are generated independently given these positions, with the edge probability p_{ij} between nodes i and j determined by a symmetric kernel function $\kappa(\mathbf{x}_i, \mathbf{x}_j)$, that typically reflects similarity or proximity.

The RDPG arises when this kernel takes the form of an inner product, $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top \mathbf{x}_j$, thus modelling edge probabilities as a function of geometric similarity in Euclidean space. Each

node i is associated with a latent position $\mathbf{x}_i \in \mathbb{R}^d$ drawn from a latent space $\mathcal{X} \subseteq \mathbb{R}^d$, where $\mathcal{X}$ is defined such that $\mathbf{x}^\top \mathbf{y} \in [0, 1]$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ i.e. $\mathcal{X}$ is defined as the unit d-ball intersected with the positive quadrant:

$$\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d_{\geq 0} : ||\mathbf{x}||_2 \leq 1\}$$

which ensures valid Bernoulli probabilities. This construction and its statistical properties are discussed in detail in Athreya et al. (2018). Under this model edges form independently with

$$A_{ij} \sim \text{Bern}(\mathbf{X}_i^\top \mathbf{X}_j)$$

The RDPG thus provides a flexible and interpretable framework for modelling unattributed networks where edges are observed but no additional features (such as node-level text) are available. To model networks with node-level textual attributes, as defined in Section 2.1.1, it is necessary to extend the RDPG framework by incorporating additional techniques, such as those introduced in Section 2.5.

When we make the simplifying assumption that each node belongs to one of K discrete clusters, and that all nodes in the same cluster share the same latent position denoted $\mu_k \in \mathbb{R}^d$ for cluster $k \in \mathbb{N}$, the random dot product graph reduces to a model known as the *Stochastic Block Model* (SBM) (Holland et al., 1983).

Formally, if node i is assigned to cluster $z_i \in \{1, \ldots, K\}$, we model its latent position as

$$X_i = \mu_{z_i}.$$

Then, as before, edges form independently according to

$$A_{ij} \sim \text{Bernoulli}(X_i^\top X_j) = \text{Bernoulli}(\mu_{z_i}^\top \mu_{z_j}).$$

This leads to a block-structured adjacency matrix, where the probability of an edge depends only on the cluster memberships of the nodes. The SBM is therefore a special case of the RDPG with latent positions constrained to lie on a finite set of fixed vectors. SBMs can be parametrised by a matrix $\mathbf{B} \in [0, 1]^{K \times K}$ where each entry B_{kl} is the probability of connection between nodes of cluster $k \in \mathbb{N}$ and nodes of cluster $l \in \mathbb{N}$.

2.3 Estimating latent positions via spectral methods

Given an observed adjacency matrix, a central challenge is to recover estimates of the latent positions $x_1, \ldots, x_n$. For RDPGs, these latent positions can only be identified up to an orthogonal transformation $\mathbf{Q} \in \mathbb{O}(d)$ the orthogonal group with dimension d defined as

$$\mathbb{O}(d) = \left\{\mathbf{Q} \in \mathbb{R}^{d \times d} \ : \ \mathbf{Q}^\top \mathbf{Q} = \mathbf{I}_d\right\}$$

This non-identifiability arises because the the edge probabilities only depend on inner products between latent positions, which are preserved under orthogonal transformations:

$$x^\top y = x^\top \mathbf{Q}^\top \mathbf{Q} y = (\mathbf{Q}x)^\top (\mathbf{Q}y)$$

A common approach to estimating the latent positions is through spectral embedding methods, which project the nodes into a low-dimensional Euclidean space based on the spectral properties of the graph. While these embeddings are identifiable only up to an unknown orthogonal transformation, they often preserve meaningful geometric structure and can reveal underlying patterns such as clustering or community organisation.

2.3.1 Adjacency spectral embedding

Definition

An effective choice, particularly for RDPGs due to theoretical guarantees is the Adjacency Spectral Embedding (ASE), defined below

Adjacency Spectral Embedding (ASE). Let $\mathbf{A} \in \{0,1\}^{n\times n}$ be a symmetric adjacency matrix, and let $d \in \{1,\dots,n\}$. The adjacency spectral embedding of dimension d is defined using the eigendecomposition of $\mathbf{A}$.

Since $\mathbf{A}$ is real and symmetric, it admits a spectral decomposition of the form:

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top,$$

where $\mathbf{U} \in \mathbb{R}^{n\times n}$ is an orthogonal matrix of eigenvectors and $\mathbf{\Lambda} \in \mathbb{R}^{n\times n}$ is a diagonal matrix of corresponding eigenvalues.

Let $\mathbf{\Lambda}_d$ be the diagonal matrix containing the d largest (in absolute value) eigenvalues of $\mathbf{A}$, and let $\mathbf{U}_d$ be the corresponding matrix of eigenvectors. The d-dimensional adjacency spectral embedding (ASE) of $\mathbf{A}$ is then given by:

$$\hat{\mathbf{X}} = \mathbf{U}_d\mathbf{\Lambda}_d^{1/2} \in \mathbb{R}^{n\times d}.$$

Each row $\hat{\mathbf{x}}_i \in \mathbb{R}^d$ of $\hat{\mathbf{X}}$ provides a d-dimensional embedding of node i. This embedding captures geometric structure in the graph and, under the RDPG model, provides a consistent estimate (up to orthogonal transformation) of the true latent position of node i.

Asymptotic Normality and Central Limit Theorem

The asymptotic distribution of the estimated positions can also be characterised (Athreya et al., 2018). In particular, the ASE satisfies a central limit theorem, which describes how the fluctuations of the embedded points around the true latent positions behave in large graphs.

ASE Central Limit Theorem. Let $\mathbf{A}^{(n)} \in \{0,1\}^{n\times n}$ be a sequence of symmetric adjacency matrices generated from a random dot product graph (RDPG) with latent positions $\mathbf{X}^{(n)} = (x_1^{(n)},\dots,x_n^{(n)})^\top$, and let $\hat{\mathbf{X}}^{(n)}$ denote the d-dimensional ASE of $\mathbf{A}^{(n)}$.

For any fixed integer $m > 0$, any fixed latent positions $x_1,\dots,x_m \in \mathbb{R}^d$, and for any $u_1,\dots,u_m \in \mathbb{R}^d$, there exists a sequence of orthogonal matrices $\mathbf{Q}_n \in \mathbb{O}(d)$ such that

$$\lim_{n\to\infty} \mathbb{P}\left\{ \bigcap_{i=1}^{m} \sqrt{n}\left(\mathbf{Q}_n\hat{x}_i^{(n)} - x_i^{(n)}\right) \le u_i \,\middle|\, x_i^{(n)} = x_i \right\} = \prod_{i=1}^{m} \Phi\left(u_i, \Sigma(x_i)\right), \tag{2.1}$$

where $\Phi(u,\Sigma(x))$ denotes the cumulative distribution function of a d-dimensional multivariate normal distribution with mean zero and covariance matrix $\Sigma(x)$

Informally, for large n, the embedded points converge in distribution:

$$\mathbf{Q}_n\hat{x}_i^{(n)} \overset{d}{\approx} \mathcal{N}(x_i, \tfrac{1}{n}\Sigma(x_i)),$$

i.e., the estimated positions follow an approximately normal distribution centred at the true latent positions, with standard deviation proportional to $1/\sqrt{n}$. This result relies on the fact that the multivariate normal distribution is preserved under orthogonal transformations: if

$\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$, then $\mathbf{QX} \sim \mathcal{N}(\mathbf{Q}\mu, \mathbf{Q}\Sigma\mathbf{Q}^\top)$. This is a special property of the Gaussian distribution. For other distributions, rotating the random variable can change its shape or introduce dependencies. The CLT result holds in part because the Gaussian is stable under such rotations, making it a natural limiting distribution in this setting.

Empirical illustration

To illustrate the central limit behaviour of the ASE, we simulate a sequence of SBMs with two distinct latent positions $\mu_1 = (0.25, 0.5)$ and $\mu_2 = (0.5, 0.25)$ and increase the number of nodes n. Figure 2.1 shows the raw embeddings produced by the ASE for each graph, while Figure 2.2 displays the same embeddings after alignment via a Procrustes transformation (Schönemann, 1966). This transformation finds the optimal orthogonal matrix that best aligns the estimated embeddings with the ground truth by minimising the Frobenius norm of their difference. In our case, it removes the rotational non-identifiability inherent in spectral embeddings by applying the closest orthogonal transformation.

As predicted by the central limit theorem, we observe that the estimated positions become increasingly concentrated around the true latent positions as n grows. After alignment, the clusters exhibit approximately Gaussian structure, consistent with the asymptotic normality described in the ASE CLT.

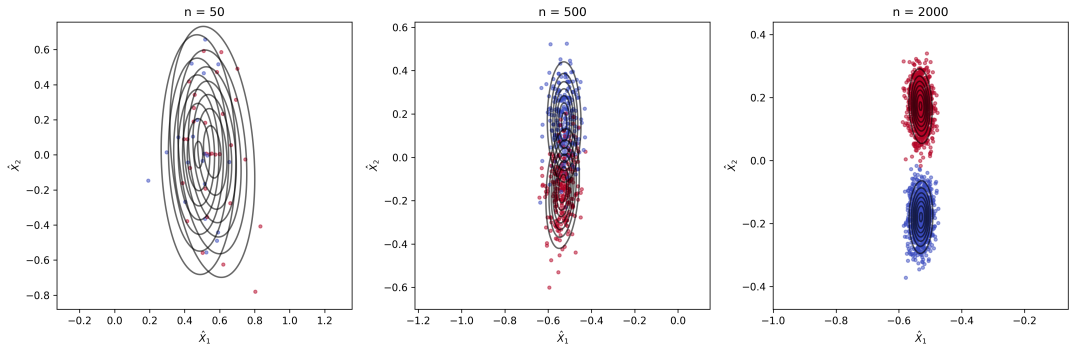


Figure 2.1 ASE of a RDPG generated with latent positions $\mu_1 = (0.25, 0.5), \mu_2 = (0.5, 0.25)$, increasing the number of nodes n

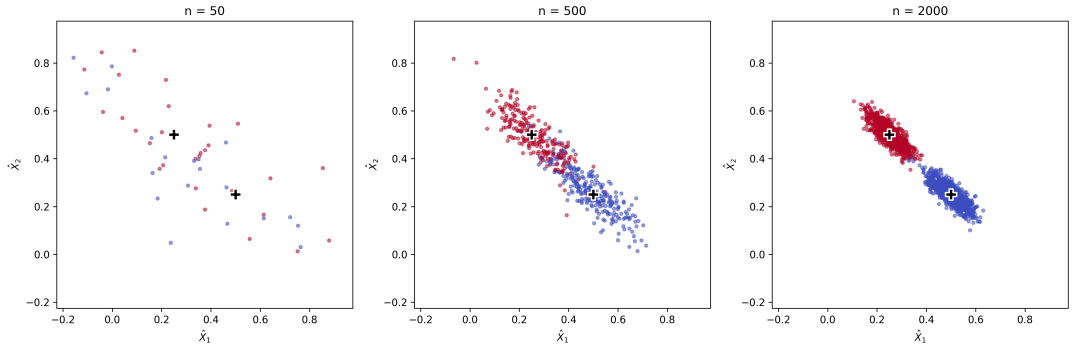


Figure 2.2 After applying Procrustes alignment to the estimated latent positions, the clusters are clearly seen to follow Gaussian distributions centred at μ_1 and μ_2 (indicated by black plus signs). As the sample size increases, the estimated positions more closely align with the true cluster centres, as expected.

2.3.2 Laplacian embedding

An alternative to the ASE is to use the eigenvectors of the graph Laplacian to form the Laplacian spectral embedding (LSE). This approach transforms the adjacency matrix to a form that accounts for node degrees, making it particularly well-suited to sparse networks or graphs with significant degree heterogeneity.

Laplacian Spectral Embedding (LSE). Let $\mathbf{A} \in \{0,1\}^{n\times n}$ be a symmetric adjacency matrix, and let $\mathbf{D}$ be the diagonal degree matrix with entries $D_{ii} = \sum_{j\neq i} A_{ij}$. The *symmetric normalised Laplacian* is defined as

$$\mathcal{L}(\mathbf{A}) = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}.$$

Given a positive integer $d \in \{1,\dots,n\}$, the LSE of $\mathbf{A}$ into $\mathbb{R}^d$ is defined by

$$\check{\mathbf{X}} = \mathbf{U}_{\mathcal{L}(\mathbf{A})}\tilde{\mathbf{S}}^{1/2}_{\mathcal{L}(\mathbf{A})},$$

where $\mathbf{U}_{\mathcal{L}(\mathbf{A})} \in \mathbb{R}^{n\times d}$ contains the eigenvectors corresponding to the d largest eigenvalues in magnitude of $\mathcal{L}(\mathbf{A})$, and $\tilde{\mathbf{S}}_{\mathcal{L}(\mathbf{A})} \in \mathbb{R}^{d\times d}$ is the diagonal matrix of these eigenvalues.

Note on Terminology. In spectral clustering and graph signal processing, the *symmetric normalised Laplacian* is typically defined as

$$\mathcal{L}_{\text{std}} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2},$$

and embeddings are formed from the eigenvectors corresponding to the *smallest* eigenvalues of $\mathcal{L}_{\text{std}}$.

However, in the context of random dot product graphs (RDPGs), particularly in Athreya et al. (2018), the term "Laplacian spectral embedding" refers to the eigendecomposition of the *normalised adjacency matrix*:

$$\mathcal{L}(\mathbf{A}) = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2},$$

as we use here. This matrix is not a Laplacian in the classical sense, but it preserves a similar spectral structure and admits a central limit theorem in the RDPG setting. Throughout this report, we follow the convention in Athreya et al. (2018) and refer to $\mathcal{L}(\mathbf{A})$ as the basis for the LSE.

Asymptotic Normality and Central Limit Theorem

The LSE also satisfies a central limit theorem under the random dot product graph model. Unlike ASE, the LSE estimates do not converge directly to the latent positions, but rather to a scaled version that reflects the influence of degree-normalisation.

LSE Central Limit Theorem.

Let $\mathbf{A}^{(n)} \in \{0,1\}^{n\times n}$ be a sequence of adjacency matrices generated from a random dot product graph (RDPG), where the latent positions $\mathbf{X}^{(n)} = (x_1^{(n)},\dots,x_n^{(n)})^\top$ are drawn independently from a distribution F over $\mathbb{R}^d$. Let $\check{\mathbf{X}}^{(n)}$ denote the d-dimensional LSE of $\mathbf{A}^{(n)}$.

Then there exists a sequence of orthogonal matrices $\mathbf{Q}_n \in \mathbb{O}(d)$ such that, for any fixed integer $m > 0$, fixed latent positions $x_1,\dots,x_m \in \mathbb{R}^d$, and vectors $u_1,\dots,u_m \in \mathbb{R}^d$,

$$\lim_{n\to\infty} \mathbb{P}\left\{ \bigcap_{i=1}^{m} n\left(\mathbf{Q}_n\check{x}_i^{(n)} - \frac{x_i}{\sqrt{\sum_{j=1}^n x_i^\top x_j}} \right) \leq u_i \,\middle|\, x_i^{(n)} = x_i \right\} = \prod_{i=1}^{m} \Phi\left(u_i, \tilde{\Sigma}(x_i)\right), \quad (2.2)$$

where $\Phi(u, \tilde{\Sigma}(x))$ denotes the CDF of a d-dimensional multivariate normal distribution with mean zero and covariance matrix $\tilde{\Sigma}(x)$. The matrix $\tilde{\Sigma}(x)$ depends not only on the local latent position x but also on global properties of the distribution F, such as the expected degree structure and normalised second moments of the latent space.

The full derivation and form of $\tilde{\Sigma}(x)$ are provided in Athreya et al. (2018), where it is shown to involve additional terms that account for the transformation induced by the normalised Laplacian operator.

2.4 Non-spectral alternatives

In addition to spectral embeddings, we also consider non-spectral alternatives such as *Graph Convolutional Network* (GCN) embeddings and *node2vec*. While not central to our model, these embeddings serve as useful baselines or enhancements, particularly in settings where spectral methods underperform due to data sparsity or high heterogeneity.

Graph Convolutional Networks (GCNs) (Kipf and Welling, 2017) learn node embeddings by iteratively aggregating information from a node's local neighbourhood. At each layer, a node updates its representation by combining its own features with those of its neighbours, enabling the model to incorporate both attribute and structural information. Conceptually, GCNs can be viewed as a natural generalisation of convolutional neural networks (CNNs) to non-Euclidean data - where the regular grid of an image is replaced by the irregular structure of a graph. Mathematically, each GCN layer implements a first-order approximation of a spectral graph convolution, acting as a localised smoothing filter over the graph topology. Stacking multiple layers allows information to propagate across multi-hop neighbourhoods, enabling the model to capture increasingly global structure.

Node2vec (Grover and Leskovec, 2016) provides another non-spectral alternative, learning node embeddings by simulating biased random walks and applying the skip-gram model - a shallow neural network trained to predict nearby nodes in the walk sequence. This approach captures both structural equivalence and local neighbourhood similarity, offering a flexible means of encoding graph information without requiring eigendecomposition.

While our primary focus remains on spectral embeddings, incorporating alternatives such as *GCNs* and *node2vec* allows us to evaluate how different embedding strategies affect downstream inference in the hybrid network-text setting. We prioritise spectral methods like ASE and LSE because they come with strong theoretical guarantees under the random dot product graph (RDPG) model, most notably central limit theorems (2.1), (2.2), which describe the asymptotic distribution of the estimated latent positions. These results provide a principled foundation for uncertainty quantification, likelihood-based modelling, and rigorous statistical inference (Athreya et al., 2018) in the embedding space. In contrast, while GCNs and node2vec often perform well in practice, especially when node features are informative, they lack comparable distributional guarantees, making them less suitable for probabilistic modelling or theoretical analysis. Spectral embeddings therefore offer a more interpretable and statistically grounded framework for this report.

2.5 Text Modelling and Topic Distributions

Textual features associated with nodes can be used independently to infer latent group structure. A standard approach is to model the word count vector $\mathbf{w}_i$ of each node using probabilistic topic models. For instance, Latent Dirichlet Allocation (LDA) (Blei et al., 2003) represents each

document as a mixture over latent topics, with each topic corresponding to a distribution over words drawn from a Dirichlet prior.

In settings where each node is assumed to belong to a single cluster, a simpler alternative is to assign a cluster-specific word distribution $\phi_k \sim \text{Dirichlet}(\eta/|\mathcal{V}|, \ldots, \eta/|\mathcal{V}|)$ and model word counts via:

$$\mathbf{w}_i \mid z_i = k \sim \text{Multinomial}(M_i, \phi_k),$$

where $M_i = \sum_v^{|\mathcal{V}|} w_{iv}$ is the total word count for node i, (see, for example Sanna Passino et al., 2025).

This Dirichlet-Multinomial framework enables clustering based purely on textual similarity, capturing thematic structure in the vocabulary distribution across nodes. It has seen wide use in applications such as document classification and content-based recommendation, independent of any underlying network.

In this report, we treat such text-based inference as one component of a broader framework, to be later combined with graph-based latent position models for joint clustering.

Chapter 3

Methods

This chapter introduces and motivates the Bayesian model used for the joint clustering of node embeddings and associated textual features. Building on the background provided in Chapter 2, where we discussed spectral embeddings such as ASE and LSE, we now formalise our approach and describe the inference techniques employed.

Our model assumes a latent community structure that governs both the distribution of embeddings and the semantic content of associated documents. It builds upon the Bayesian framework introduced by Sanna Passino and Heard (2020), which performs inference over latent clusters using adjacency spectral embeddings, Normal-Inverse-Wishart priors, and collapsed Gibbs sampling.

While their focus is primarily on inferring the latent embedding dimension alongside clustering, we treat the embedding dimension as fixed and instead extend the framework to incorporate node-level text data. Specifically, we model documents using cluster-specific multinomial distributions over a shared vocabulary. The embedding component of our model is designed to accommodate different spectral embeddings discussed earlier, allowing us to compare how ASE and LSE interact with the textual signal during clustering.

3.1 Model Overview

3.1.1 Likelihoods

Embedding likelihood

For simplicity, let $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n) \in \mathbb{R}^{n \times d}$, denote the d-dimensional embeddings obtained from a pre-processing step (such as adjacency spectral embedding or normalised Laplacian embedding). The dimension d is fixed in advance and will be dicussed in Chapter 4, where we explore scree plots and empirical criteria for selection.

Conditional on the latent cluster assignment $z_i \in \{1, \dots, K\}$, we model $\mathbf{x}_i$ as drawn from a cluster specific Gaussian:

$$(\mathbf{x}_i \mid z_i = k) \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

where $\boldsymbol{\mu}_k \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{d \times d}$.

Text likelihood

For the textual information, we first enumerate the full vocabulary $\mathcal{V}$, allowing each unique word type to be labelled with an index $j = 1, \ldots, |\mathcal{V}|$, where $|\mathcal{V}|$ is the total size of the vocabulary. Each node i is then associated with a bag-of-words count vector, denoted $\mathbf{w}_i \in \mathbb{N}_0^{|\mathcal{V}|}$. The j-th component of this vector, w_{ij}, represents the count of the j-th word type from the vocabulary in the abstract associated with node i.

Let $M_i = \sum_{j=1}^{|\mathcal{V}|} w_{ij}$ be the total number of words (tokens) in the abstract for node i. Conditional on the latent cluster assignment $z_i = k$ and this total word count M_i, the vector of word counts $\mathbf{w}_i$ is modelled as drawn from a cluster-specific Multinomial distribution:

$$(\mathbf{w}_i \mid z_i = k, M_i) \sim \text{Multinomial}(M_i, \boldsymbol{\phi}_k)$$

where $\boldsymbol{\phi}_k = (\phi_{k,1}, \ldots, \phi_{k,|\mathcal{V}|})$ is the vector of word probabilities for cluster k. Each $\phi_{k,j}$ is the probability of observing word type j in an abstract belonging to cluster k, and this vector lies on the probability simplex, i.e., $\boldsymbol{\phi}_k \in \Delta^{|\mathcal{V}|}$ (meaning $\phi_{k,j} \geq 0$ for all j and $\sum_{j=1}^{|\mathcal{V}|} \phi_{k,j} = 1$).

3.1.2 Prior distributions

In order to complete the Bayesian model specification, we place conjugate priors on the latent variables and component parameters. This facilitates tractable inference via collapsed Gibbs sampling, as the posterior conditionals are analytically computable.

Cluster proportions

We place a symmetric Dirichlet prior with hyper-parameter $\boldsymbol{\gamma} = (\gamma/K, \ldots, \gamma/K)$ on the cluster assignment probabilities $\boldsymbol{\psi} \in \Delta^K$:

$$p(\boldsymbol{\psi} \mid \boldsymbol{\gamma}) = \text{Dirichlet}_K(\boldsymbol{\psi} \mid \gamma/K, \ldots, \gamma/K)$$

. This prior encourages a balanced (but flexible) allocation of nodes across clusters. The hyper-parameter γ influences the expected distribution of cluster sizes as follows:

- For $\gamma_k < 1$, the prior favours sparse allocations where one or a few clusters dominate,
- For $\gamma_k = 1$, the prior is uniform over the simplex Δ^K, giving equal weight to all possible proportions of cluster sizes,
- For $\gamma_k > 1$, the prior favours dense allocations where cluster assignments are evenly distributed.

Cluster embedding parameters

For the mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$ of each of the cluster-specific distributions of the spectral embeddings, we use a Normal-Inverse-Wishart (NIW) prior with hyper-parameters $\boldsymbol{\beta} = (\mathbf{m}_0, \kappa_0, \nu_0 + d - 1, \mathbf{S}_0)$:

$$p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \mid \boldsymbol{\beta}) = \text{NIW}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \mid \mathbf{m}_0, \kappa_0, \nu_0 + d - 1, \mathbf{S}_0),$$

This is the conjugate to the multivariate normal distribution, which allows the mean and covariance $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ to be updated jointly in closed form.

Cluster word distributions

For the multinomial likelihood over bag-of-words vectors, we place a symmetric Dirichlet prior with hyper-parameter $\boldsymbol{\eta} = (\eta/|\mathcal{V}|, \ldots, \eta/|\mathcal{V}|)$ on each cluster's word distribution:

$$p(\boldsymbol{\phi}_k|\boldsymbol{\eta}) \sim \text{Dirichlet}_{|\mathcal{V}|}\left(\boldsymbol{\phi}_k \mid \frac{\eta}{|\mathcal{V}|}, \ \ldots, \ \frac{\eta}{|\mathcal{V}|}\right),$$

where $\eta > 0$ governs the prior concentration. This choice is conjugate to the multinomial likelihood, enabling closed-form posterior updates.

3.1.3 Model Summary

Model Component	**Distribution / Definition**
d-dimensional node embeddings	$\mathbf{x}_i, \quad i = 1, \ldots, n$
Cluster assignment	$z_i \mid \boldsymbol{\psi} \sim \text{Categorical}(\boldsymbol{\psi}), \quad \boldsymbol{\psi} = (\psi_1, \ldots, \psi_K)$
Node embedding given cluster	$(\mathbf{x}_i \mid z_i = k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad k = 1, \ldots, K$
Word counts in abstract i	$\mathbf{w}_i = (w_{i1}, \ldots, w_{i\|\mathcal{V}\|})$, where w_{ij} is count of word type j.
	$M_i = \sum_{j=1}^{\|\mathcal{V}\|} w_{ij}$ (total words in abstract i)
Word count distribution for abstract i given cluster	$(\mathbf{w}_i \mid z_i = k, M_i, \boldsymbol{\phi}_k) \sim \text{Multinomial}(M_i, \boldsymbol{\phi}_k)$, where $\boldsymbol{\phi}_k = (\phi_{k1}, \ldots, \phi_{k\|\mathcal{V}\|})$
Cluster mean and covariance	$(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \sim \text{NIW}(\mathbf{m}_0, \kappa_0, \nu_0 + d - 1, \mathbf{S}_0)$
Word probability vector for cluster k	$\boldsymbol{\phi}_k \sim \text{Dirichlet}_{\|\mathcal{V}\|}(\eta/\|\mathcal{V}\|, \ldots, \eta/\|\mathcal{V}\|)$
Cluster assignment probabilities	$\boldsymbol{\psi} \sim \text{Dirichlet}_K(\gamma/K, \ldots, \gamma/K)$
Observations	$\mathcal{D} = \{\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n, \mathbf{W} = \{\mathbf{w}_i\}_{i=1}^n\}$
Parameters	$\boldsymbol{\Theta} = \{\mathbf{z}, \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K, \{\boldsymbol{\phi}_k\}_{k=1}^K, \boldsymbol{\psi}\}$

Table 3.1 Summary of the Bayesian model defined above.

3.2 Inference Techniques

The goal of inference in this model is to recover the posterior distribution over latent cluster assignments $\mathbf{z}$, conditioned on observed node embeddings $\mathbf{X}$ and text attributes $\mathbf{W}$. Due to the joint discrete-continuous structure of the model comprising discrete cluster labels and continuous embedding vectors, standard gradient-based inference methods are not applicable. Instead, we employ collapsed Gibbs sampling, which integrates out nuisance parameters to reduce variance and improve mixing (Liu, 1994).

In our model setup, all latent parameters aside from the cluster assignments $\mathbf{z}$ are treated as nuisance parameters. These include the cluster proportions $\boldsymbol{\psi}$, the text distributions $\{\boldsymbol{\phi}_k\}_{k=1}^K$, and the embedding parameters $\{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^K$. By choosing conjugate priors, we are able to integrate them out analytically, resulting in a more efficient collapsed sampler that targets the posterior over $\mathbf{z}$ directly and reduces the dimensionality of the sampling space.

3.2.1 Collapsed Posterior Updates

The key equation for updating z_i in the collapsed Gibbs sampler is:

$$
\begin{aligned}
p(z_i = k \mid \mathbf{z}_{-i}, \mathbf{X}, \mathbf{W}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma}) \propto\ & p(\mathbf{x}_i \mid \mathbf{x}_{-i}, z_i = k, \mathbf{z}_{-i}, \boldsymbol{\beta}) \\
& \times\ p(\mathbf{w}_i \mid \mathbf{w}_{-i}, z_i = k, \mathbf{z}_{-i}, \boldsymbol{\eta}) \\
& \times\ p(z_i = k \mid \mathbf{z}_{-i}, \boldsymbol{\gamma})
\end{aligned} \tag{3.1}
$$

Each component in (3.1) is derived analytically by marginalising over the associated parameters. Full derivations are provided in Appendix A. For notational simplicity, we omit explicit conditioning on hyper-parameters in the expressions below, since these remain fixed throughout inference. The resulting expressions used during sampling are as follows:

$$
p(\mathbf{x}_i \mid \mathbf{x}_{-i}, z_i = k, \mathbf{z}_{-i}) = t_{\nu_k}\left(\mathbf{x}_i \mid \mathbf{m}_k, \frac{\kappa_k + 1}{\kappa_k \nu_k}\mathbf{S}_k\right) \tag{3.2}
$$

$$
p(\mathbf{w}_i \mid \mathbf{w}_{-i}, z_i = k, \mathbf{z}_{-i}) = \frac{1}{\prod_{r=0}^{M_i - 1}\left(\eta + C_{-i}^{(k)} + r\right)} \prod_{v:\, w_{iv} > 0} \frac{\Gamma\left(c_{v,-i}^{(k)} + w_{iv} + \eta/|\mathcal{V}|\right)}{\Gamma\left(c_{v,-i}^{(k)} + \eta/|\mathcal{V}|\right)} \tag{3.3}
$$

$$
p(z_i = k \mid \mathbf{z}_{-i}) = \frac{N_{k,-i} + \gamma/K}{n - 1 + \gamma} \tag{3.4}
$$

where $\mathbf{c}_{-i}^{(k)}$ denotes the vector of word counts for cluster k, excluding node i, and $C_{-i}^{(k)} = \sum_j c_{j,-i}^{(k)}$ is the corresponding total word count.

3.2.2 Collapsed Gibbs Sampler Algorithm

Algorithm 1: Random permutation collapsed Gibbs sampler for joint clustering model.

Input: Embeddings $\mathbf{X}$, textual covariates $\mathbf{W}$, number of clusters K, initial cluster assignments $\mathbf{z}^{(0)}$
Output: Cluster assignments $\mathbf{z}$
for M *iterations* **do**
  Generate a random permutation π of $\{1, \ldots, n\}$
  for $i = 1$ *in* π **do**
    Remove node i's contribution from the sufficient statistics of cluster z_i.
    for $k = 1$ *to* K **do**
      Calculate $p(\mathbf{x}_i \mid \mathbf{x}_{-i}, z_i = k, \mathbf{z}_{-i}, \boldsymbol{\beta})$ using (3.2).
      Calculate $p(\mathbf{w}_i \mid \mathbf{w}_{-i}, z_i = k, \mathbf{z}_{-i}, \boldsymbol{\eta})$ using (3.3).
      Calculate $p(z_i = k \mid \mathbf{z}_{-i}, \boldsymbol{\gamma})$ using (3.4).
      Combine:

$$
\begin{aligned}
p(z_i = k \mid \mathbf{z}_{-i}, \mathbf{X}, \mathbf{W}) \propto\ & p(\mathbf{x}_i \mid \mathbf{x}_{-i}, z_i = k, \mathbf{z}_{-i}) \\
& \times\ p(\mathbf{w}_i \mid \mathbf{w}_{-i}, z_i = k, \mathbf{z}_{-i}) \\
& \times\ p(z_i = k \mid \mathbf{z}_{-i})
\end{aligned}
$$

    Sample k' from $p(z_i \mid \mathbf{z}_{-i}, \mathbf{X}, \mathbf{W})$ after normalising.
    Add node i's contribution to the sufficient statistics of the newly assigned cluster $z_i = k'$.

Variants of Gibbs sampling algorithm

Several scan strategies are available for Gibbs sampling, including *sequential scan*, *random scan*, and *random permutation* updates. In this work, we adopt **random permutation Gibbs sampling**, where a new random ordering of the variables is used at each iteration. This avoids potential artefacts associated with fixed update orders, such as poor mixing or order-induced bias, while maintaining desirable properties like ergodicity and aperiodicity (Roberts and Sahu, 1997).

Numerical Stability

All calculations are performed in the log domain to prevent numerical underflow, particularly when evaluating small probabilities. In the Gibbs sampler, the unnormalised log-probabilities for each cluster k are first computed:

$$\log p_k = \log p(\mathbf{x}_i \mid \dots) + \log p(\mathbf{w}_i \mid \dots) + \log p(z_i = k \mid \dots)$$

To convert these back into probabilities, we use the *log-sum-exp* trick when normalising:

$$p_k = \frac{\exp(\log p_k)}{\sum_j \exp(\log p_j)} = \frac{\exp(\log p_k - m)}{\sum_j \exp(\log p_j - m)}$$

where $m = \max_j \log p_j$. Subtracting m improves numerical stability by ensuring that all exponentials are evaluated at non-positive values, avoiding overflow.

3.2.3 Monitoring Convergence

In order to track the progress of the Gibbs sampler and ensure adequate mixing, we monitor the joint marginal likelihood of the observed data $\mathcal{D}$. This includes the marginal likelihoods of the embeddings $\mathbf{X}$, textual covariates $\mathbf{W}$, and cluster assignments $\mathbf{z}$. These quantities jointly capture how well the current configuration explains the data, and the plausibility of the cluster assignments under the prior. Tracking the joint marginal likelihood over iterations allows us to evaluate convergence and identify issues such as poor mixing or mode switching.

$$p(\mathbf{X}, \mathbf{W}, \mathbf{z} \mid \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\beta}) = p(\mathbf{X} \mid \mathbf{z}, \boldsymbol{\beta}) p(\mathbf{W} \mid \mathbf{z}, \boldsymbol{\eta}) p(\mathbf{z} \mid \boldsymbol{\gamma}) \tag{3.5}$$

Each term in (3.5) can be computed in closed form by marginalising over the conjugate priors, as described in A.4.

3.2.4 Inferring communities

To recover a consensus clustering from the posterior samples of cluster assignments $\{\mathbf{z}^{(s)}\}_{s=1}^{M}$, we estimate the posterior similarity matrix $\hat{\pi}$, where each entry $\hat{\pi}_{ij}$ represents the empirical posterior probability that nodes i and j belong to the same cluster. Let M denote the number of posterior samples retained after burn-in. Then:

$$\hat{\pi}_{ij} = \hat{\mathbb{P}}(z_i = z_j \mid \mathbf{X}, \mathbf{W}) = \frac{1}{M} \sum_{s=1}^{M} \mathbb{1}_{z_i^{(s)}}(z_j^{(s)})$$

Although the number of clusters K is fixed across all samples, the label switching problem (Jasra et al., 2005) can still occur due to the invariance of the likelihood under permutations of the cluster labels. However, the posterior similarity matrix is invariant to such permutations and therefore remains robust to label switching. This makes it preferable to simpler alternatives such as selecting the posterior mode of $\mathbf{z}$ directly. Following Medvedovic et al. (2004), we apply hierarchical clustering with average linkage to the dissimilarity matrix $1 - \hat{\pi}_{ij}$, treating it as a pairwise distance measure.

Hierarchical clustering proceeds by iteratively merging the closest pairs of clusters based on a chosen linkage criterion. Using an average linkage, the distance between two clusters is defined as the average of all pairwise distances between points across the two clusters. This process produces a tree-like structure that encodes the nested grouping of nodes at different levels of granularity. We cut this tree at the level corresponding to K clusters to obtain the final partition.

3.2.5 Reweighting Embedding and Text Contributions

In the collapsed Gibbs sampler described above, the likelihood of node i under cluster k is computed by jointly considering both the spectral embedding $\mathbf{x}_i$ and the abstract text $\mathbf{w}_i$, following (3.1).

In practice, the embedding and text modalities may vary in their informativeness or reliability. To account for this, we introduce two non-negative weighting parameters, $\alpha_{\mathbf{X}} \geq 0$ and $\alpha_{\mathbf{W}} \geq 0$, which control the relative contribution of each modality to the posterior assignment probabilities. This flexible formulation allows the model to adapt to different conditions by amplifying or downweighting individual likelihoods based on domain knowledge or observed performance.

$$\begin{aligned}\log p(z_i = k|\mathbf{z}_{-i}, \mathbf{X}, \mathbf{W}) \propto\ & \alpha_{\mathbf{X}} \cdot \log p(\mathbf{x}_i|z_i = k, \mathbf{z}_{-i}, \mathbf{x}_{-i}) \\ & + \alpha_{\mathbf{W}} \cdot \log p(\mathbf{w}_i|z_i = k, \mathbf{z}_{-i}, \mathbf{w}_{-i}) \\ & + \log p(z_i = k|\mathbf{z}_{-i})\end{aligned}$$

Clearly, this formulation recovers equal weighting of modalities, as in (3.1), when $\alpha_{\mathbf{X}} = \alpha_{\mathbf{W}} = 1$. By tuning $\alpha_{\mathbf{X}}$ and $\alpha_{\mathbf{W}}$, the model can modulate the influence of each modality based on prior knowledge or empirical performance.

In Chapter 4, we explore this flexibility by fixing one coefficient and varying the other in the interval $[0, 1]$. Notably, the model reduces to a Dirichlet–Multinomial mixture when $(\alpha_{\mathbf{X}}, \alpha_{\mathbf{W}}) = (0, 1)$, and to a finite GMM when $(\alpha_{\mathbf{X}}, \alpha_{\mathbf{W}}) = (1, 0)$

Chapter 4

Results

4.1 Simulation Study

We begin by testing the model in a simple 2-class setting to test its robustness to various datasets. We will quantify the clustering performance of our using the Adjusted Rand Index (ARI) metric which is introduced below:

Adjusted Rand Index (ARI). The Adjusted Rand Index (Hubert and Arabie, 1985) is a clustering evaluation metric that quantifies the agreement between two partitions, correcting for random chance. It compares the set of pairwise decisions (i.e., whether two elements are in the same or different clusters) made by the model against the ground truth. The ARI takes a value of 1 when the clusterings are identical, 0 when the agreement is what would be expected by random chance, and can be negative if the agreement is worse than random.

4.1.1 Effect of embedding separation on clustering performance

To assess the model's ability to recover true cluster structure from embedding data, we begin with a controlled simulation varying the Euclidean separation between two clusters. Each cluster consists of multivariate embeddings sampled from spherical Gaussian distributions with identity covariance. The means of the clusters are placed symmetrically along the line $y = x$, with the Euclidean distance between them ranging from 0 to 2.5. For each separation value, we generate $n/2 = 100$ samples per cluster and hold the associated word vectors constant, allowing us to isolate the impact of embedding separation on clustering performance.

As expected, clustering performance improves with greater separation between cluster centres. When separation is low and the clusters are embeddings overlap, the embeddings alone are insufficient to distinguish them. In this regime, models with higher $\alpha_{\mathbf{W}}$, which place greater weight on the text likelihood, outperform those relying primarily on the embedding signal. As shown in Figure 4.1, each line corresponds to a different value of $\alpha_{\mathbf{W}}$, illustrating this trade-off. Performance converges across all models as separation increases, indicating that the embedding information alone becomes sufficient at higher separations.

Overall, this experiment highlights the benefit of incorporating textual information when the graph structure is ambiguous. In this controlled setting, where the word vectors are fixed and highly informative, increasing $\alpha_{\mathbf{W}}$ leads to improved clustering performance. This reflects the fact that assigning equal weight to the text and embedding likelihoods ($\alpha_{\mathbf{W}} = \alpha_{\mathbf{X}} = 1$) allows the model to fully exploit the available semantic signal. However, as we will see in the

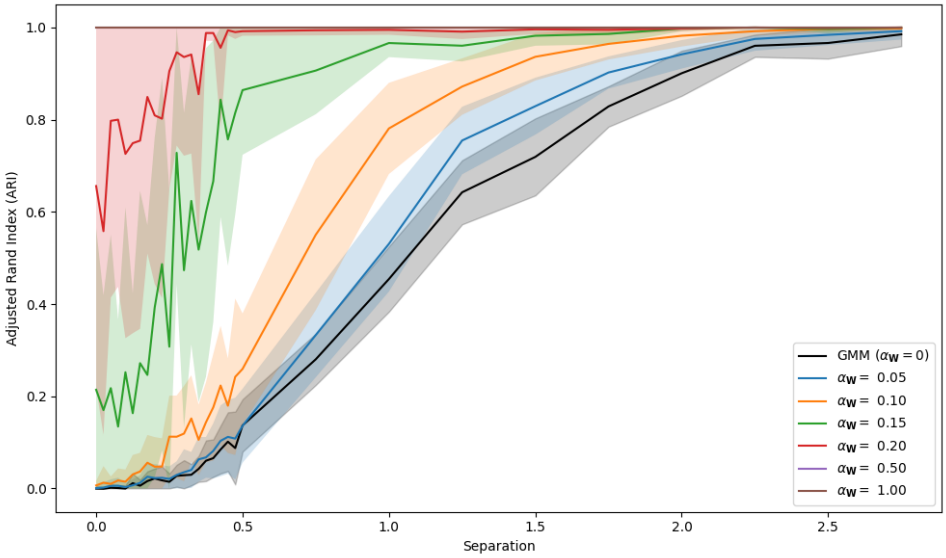


Figure 4.1 Clustering performance (measured by ARI) as a function of Euclidean separation between cluster centres in the embedding space. Each line corresponds to a different value of $\alpha_{\mathbf{W}}$, holding $\alpha_{\mathbf{X}} = 1$. For each separation level and $\alpha_{\mathbf{W}}$ setting, results are averaged over 20 independently simulated datasets. Shaded regions denote ± 1 standard deviation across these simulations. The black line shows the baseline GMM performance using embeddings alone ($\alpha_{\mathbf{W}} = 0$).

real-data setting, textual features may be noisier or less discriminative, in which case assigning higher weight to the embedding likelihood can yield better results.

4.1.2 Effect of textual distinctiveness on cluster performance

In this simulation, we investigate how the semantic distinctness of clusters, influences the model's ability to correctly recover them. In particular, we focus on the effect of the Dirichlet concentration parameter η on the separability of topic distributions across clusters.

Each cluster $k \in \{0, 1\}$ is assigned a word distribution $\boldsymbol{\phi}_k \sim \text{Dirichlet}_{|\mathcal{V}|}\left(\frac{\eta}{|\mathcal{V}|}, \ldots, \frac{\eta}{|\mathcal{V}|}\right)$, where $|\mathcal{V}|$ is the vocabulary size. Here the data-generating distribution follows the same form as the multinomial likelihood structure assumed in our model, ensuring consistency between simulation and inference. The parameter η controls the concentration of the word distribution:

- Low values of η yield sparse, peaked distributions (spiky), leading to distinctive vocabularies for each cluster.
- High values of η produce flatter, higher entropy word distributions, increasing the semantic overlap between clusters and making them harder to distinguish.

To isolate the effect of text, we fix the embeddings to be identical across clusters, and vary η across several orders of magnitude. For each value of η, we generate synthetic documents from the corresponding $\boldsymbol{\phi}_k$ and evaluate clustering performance of the model using ARI.

As illustrated in Figure 4.2, lower values of η produce sparse word distributions, making clusters highly distinguishable and enabling accurate recovery of the ground truth. In contrast, as η increases, the word distributions flatten and begin to overlap, impairing the model's ability to correctly assign nodes.

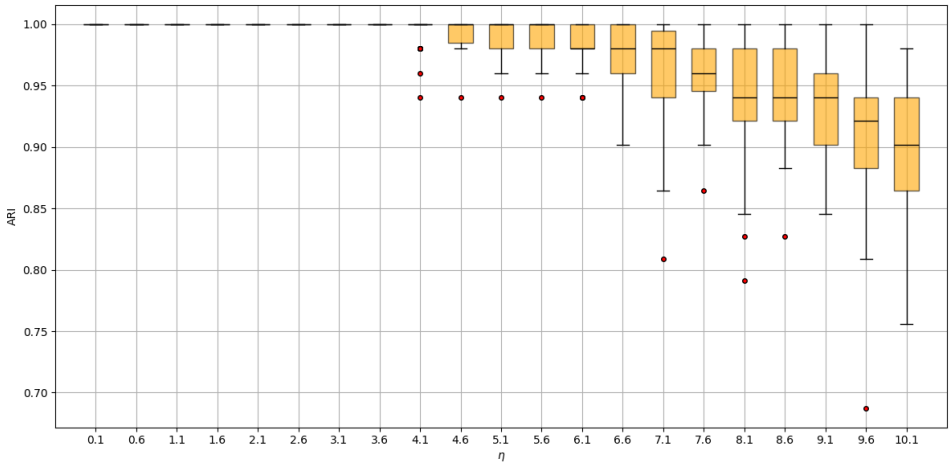


Figure 4.2 Boxplots showing the ARI between the ground truth cluster labels and the cluster assignments inferred by the model across 50 simulations per value of η.

We can also use the Jensen-Shannon (JS) Divergence as a proxy measure for the textual distinctiveness. We define the average pairwise JS divergence (just the JS divergence between ϕ_1 and ϕ_2 with two clusters) below:

For distributions ϕ_i, ϕ_j, let $\delta_{ij} = \text{JS}(\phi_i \mid\mid \phi_j)$. The average pairwise JS is defined as

$$\bar{\delta} = \frac{2}{k(k-1)} \sum_{i<j} \delta_{ij}$$

Figure 4.3 shows that ARI scores increase with higher JS divergence, indicating that more distinct word distributions facilitate better cluster recovery. The step-like appearance of ARI values stems from the small sample size. Here, minor node reassignments cause large discrete jumps in evaluation metrics.

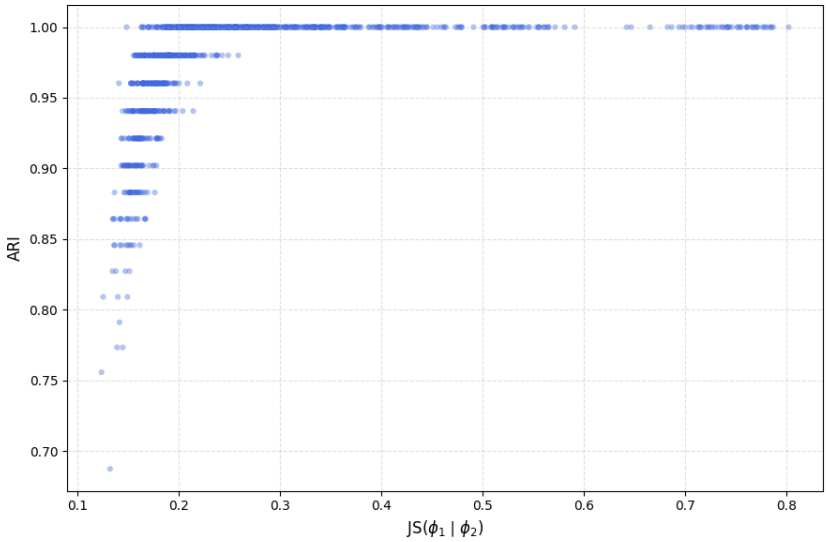


Figure 4.3 Scatter plot of ARI between the true and model-inferred cluster labels against the Jensen-Shannon divergence between the cluster-specific word distributions ϕ_1 and ϕ_2. Each point represents one simulation.

4.1.3 Cluster imbalance

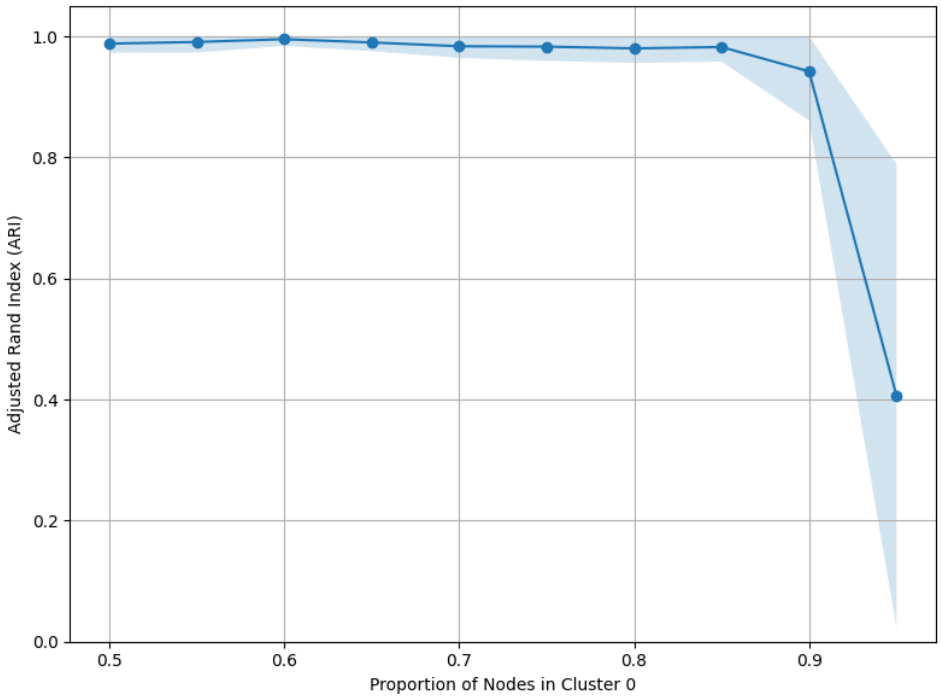


Figure 4.4 Clustering performance measured by ARI under increasing class imbalance. The x-axis denotes the proportion of nodes assigned to cluster 0, with the remaining nodes assigned to cluster 1. Shaded region indicates ± 1 standard deviation over 30 simulations.

To examine the impact of class imbalance on clustering accuracy, we conduct a simulation in which the number of nodes in each cluster is no longer equal. Specifically, we fix the total number of nodes and vary the proportion assigned to cluster 0 from 0.5 (balanced) to 0.95 (highly imbalanced). All other components of the simulation - including embeddings and textual features - are held constant across conditions.

In this experiment, the prior over cluster assignments is given by a symmetric Dirichlet distribution with concentration parameter $\gamma = 1$, i.e. $\psi \sim \text{Dirichlet}\left(\frac{1}{K}, \ldots, \frac{1}{K}\right)$. This is a commonly used prior that imposes minimal structure on the cluster proportions, treating all partitions with equal concentration as equally plausible.

As shown in Figure 4.4, the model exhibits robust performance across a wide range of cluster proportions, achieving near-perfect recovery up to a cluster 0 proportion of approximately 0.9. However, beyond this point, performance degrades rapidly, with both the mean ARI and its stability dropping off significantly.

This sharp decline highlights a key limitation of using a symmetric Dirichlet prior when the data-generating process involves extreme imbalance. Under such conditions, the model's implicit bias toward balanced partitions becomes increasingly mismatched with the observed data, leading to poor posterior inference. The under-represented cluster becomes difficult to distinguish from background noise, particularly when its membership falls below the threshold where it can contribute sufficient evidence to the likelihood.

These findings are particularly relevant for real-world datasets such as Cora, where class imbalance naturally arises. In such contexts, models that assume roughly equal cluster sizes, as induced by standard symmetric Dirichlet priors, may underperform when faced with highly skewed partitions. This highlights the importance of understanding how prior choices interact

with structural properties of the data, particularly in imbalanced settings.

4.2 Cora dataset

4.2.1 Overview

The Cora dataset is a widely-used benchmark in graph-based machine learning. It consists of a citation network of 2708 academic papers in machine learning and related fields. Each node represents a paper, and each edge denotes a citation between two papers. Nodes are labelled with one of seven research areas, which form the ground truth clusters.

- **Nodes:** 2708 (papers)
- **Edges:** 5429 (citations)
- **Clusters:** 7 ground truth categories (*Case-Based, Genetic Algorithms, Neural Networks, Probabilistic Methods, Reinforcement Learning, Rule Learning, Theory*)

Each node is also associated with a high-dimensional ($|\mathcal{V}| = 1433$) bag-of-words representation derived from the paper's abstract. While this provides rich semantic information, it also introduces challenges due to vocabulary sparsity and overlapping terminology across fields.

The citation graph is not fully connected; it comprises 78 connected components. Most nodes, however, belong to a single large connected component (LCC) containing 2485 nodes, 92% of the network. This subset captures the majority of the network's structure and is the focus later when performing modelling and evaluation to ensure meaningful spectral embeddings.

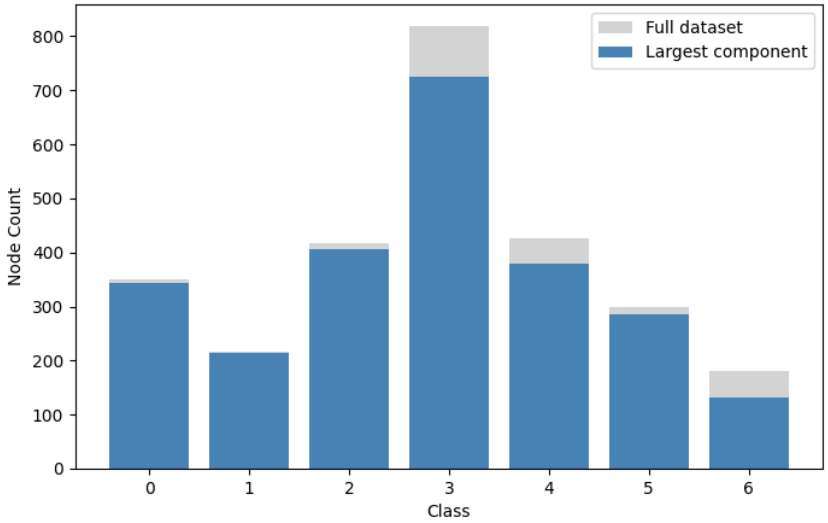


Figure 4.5 Class distribution of the full Cora dataset (grey) and of the largest connected component (blue). While the overall structure is preserved, some classes are slightly underrepresented in the LCC.

The class distribution is uneven, with class 3 being the most dominant. This imbalance, combined with graph fragmentation, further complicates clustering, especially when relying solely on connectivity or text.

Disconnected components introduce isolated subgraphs whose eigenvectors correspond to independent structural modes. While these modes carry valid local structure, their inclusion can distort the global spectral embedding, especially when dominated by small or singleton components. To mitigate this and produce more coherent embeddings, we restrict our analysis to the largest connected component (LCC). This results in a cleaner spectral profile and embeddings that more faithfully capture the community structure of the main graph.

4.2.2 Embedding choice

Adjacency spectral embedding

We begin by computing the ASE of the network. We use the scree plot to identify the optimal choice of d, which captures most of the latent structure (Zhu and Ghodsi, 2006).

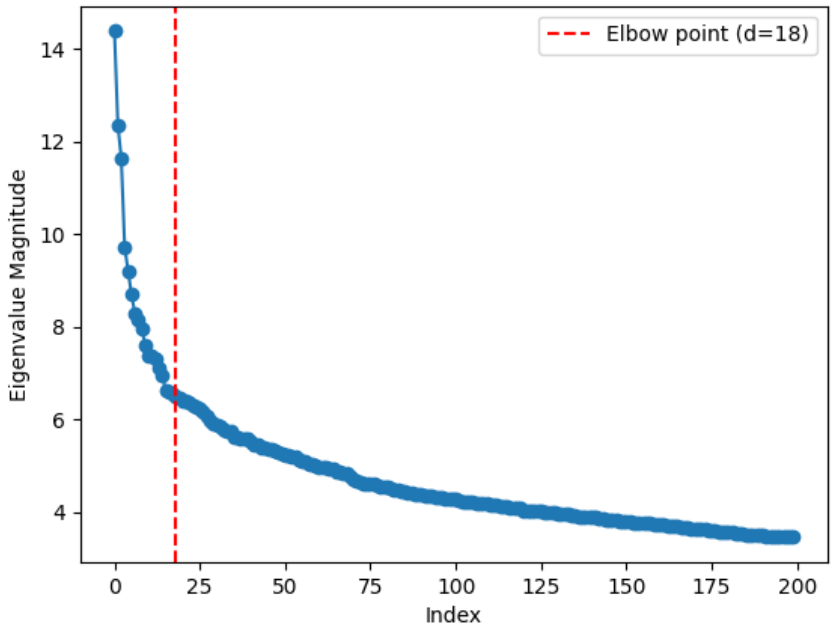


Figure 4.6 ASE Scree plot, with elbow identified using the Kneedle algorithm, $d = 18$, shown in red.

To determine an appropriate embedding dimension d, we apply the Kneedle algorithm proposed by Satopaa et al. (2011), which identifies the point of maximum curvature (the "elbow") in a scree plot of ordered eigenvalues. This method is especially useful in spectral clustering, where choosing the number of leading eigenvectors is important but often arbitrary. Kneedle provides an automated, data-driven way to detect where additional dimensions yield diminishing returns, helping to retain informative structure without unnecessary complexity. By removing the need for manually defined thresholds, it offers a more objective basis for selecting dimensionality in noisy or ambiguous graph data.

Despite the scree plot exhibiting a clear elbow, clustering based solely on the resulting ASE yields poor results. We attribute this to the sparsity of the graph - even when analysis is restricted to the largest connected component, which has a density of only 0.0016, calculated as $\frac{|E|}{n(n-1)}$. This sparsity degrades the quality of the spectral embedding, as many nodes are effectively indistinguishable based on their local connectivity patterns alone. The resulting clusters

prove less effective in capturing meaningful community structure, as evidenced by consistently lower ARI when applying GMMs to the ASE compared to the LSE.

This aligns with findings from Athreya et al. (2018) and Tang and Priebe (2018), who argue that Laplacian-based embeddings often outperform the ASE in clustering sparse networks. Accordingly, we proceed with the LSE, which benefits from an analogous central limit theorem (2.2) and improved empirical performance in such settings.

Laplacian spectral embedding

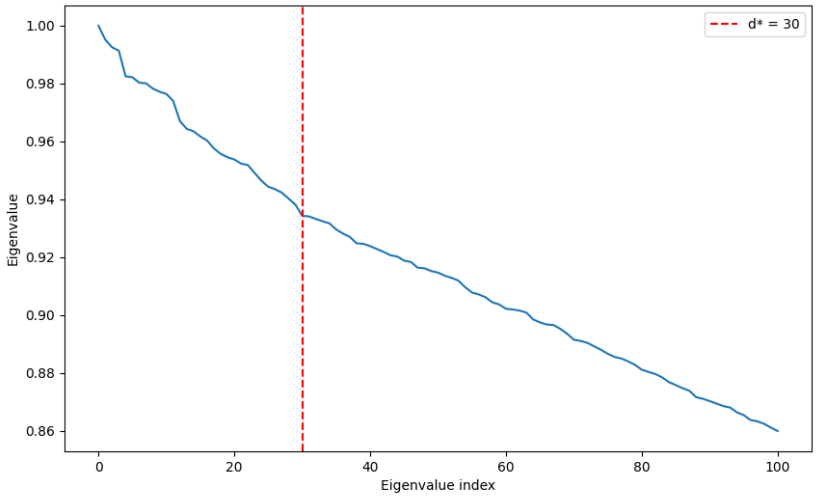


Figure 4.7 LSE Scree plot, with elbow identified using the Kneedle algorithm, $d = 30$, shown in red.

The scree plot for the Laplacian embedding of the largest connected component (Figure 4.7) displays a gradual spectral decay without a sharp elbow, suggesting the absence of strongly separated low-dimensional structure.

We follow Algorithm 2 of Ng et al. (2002), which prescribes normalising each row of the eigenvector matrix to have unit Euclidean norm prior to clustering. Empirically, this step significantly improves clustering performance: omitting it severely degrades results, with some clusters entirely unrepresented in the predicted labels. This appears to stem from scale differences in the raw embeddings adversely affecting the distance-based clustering algorithm. While such row normalisation is not standard in the Random Dot Product Graph (RDPG) embedding pipeline (see, for instance Athreya et al., 2018), we include it here as a pragmatic intervention to enhance alignment with ground truth labels.

That said, this normalisation step is mathematically questionable: by projecting points onto the unit sphere in $\mathbb{R}^d$, we are effectively fitting Gaussian clusters in $\mathbb{R}^d$ to data constrained to a $(d-1)$-dimensional hypersphere. This mismatch between the geometry of the data and the assumptions of the clustering model GMMs in Euclidean space may introduce distortion, though in our setting the empirical benefits appear to outweigh these concerns.

t-SNE visualisation. t-distributed Stochastic Neighbour Embedding (t-SNE) (van der Maaten and Hinton, 2008) is a non-linear dimensionality reduction technique commonly used to visualise high-dimensional data in two or three dimensions. It begins by converting pairwise

similarities between points in the high-dimensional space into conditional probabilities that reflect neighbourhood affinities. A similar probability distribution is then defined in the low-dimensional space, and t-SNE minimises the Kullback–Leibler (KL) divergence between these two distributions. This ensures that points that are close together in the original space remain close in the visualisation, preserving local structure. However, global distances and scales are not preserved, so t-SNE is best interpreted as a qualitative tool for visualising cluster formation or local groupings rather than global geometry.

Figure 4.8 presents a t-SNE projection of the LSE embeddings, with points coloured by their true labels and marker shapes denoting the clusters assigned by the GMM. Since clustering methods assign arbitrary labels to each cluster (e.g., cluster 0 in GMM is not guaranteed to correspond to class 0 in the ground truth), we use the Hungarian matching algorithm (Kuhn, 1955) to permute the predicted labels for optimal alignment with the true labels. This alignment aids visual and quantitative comparisons by ensuring that cluster labels correspond more intuitively to the true labels, making patterns in agreement or disagreement easier to interpret.

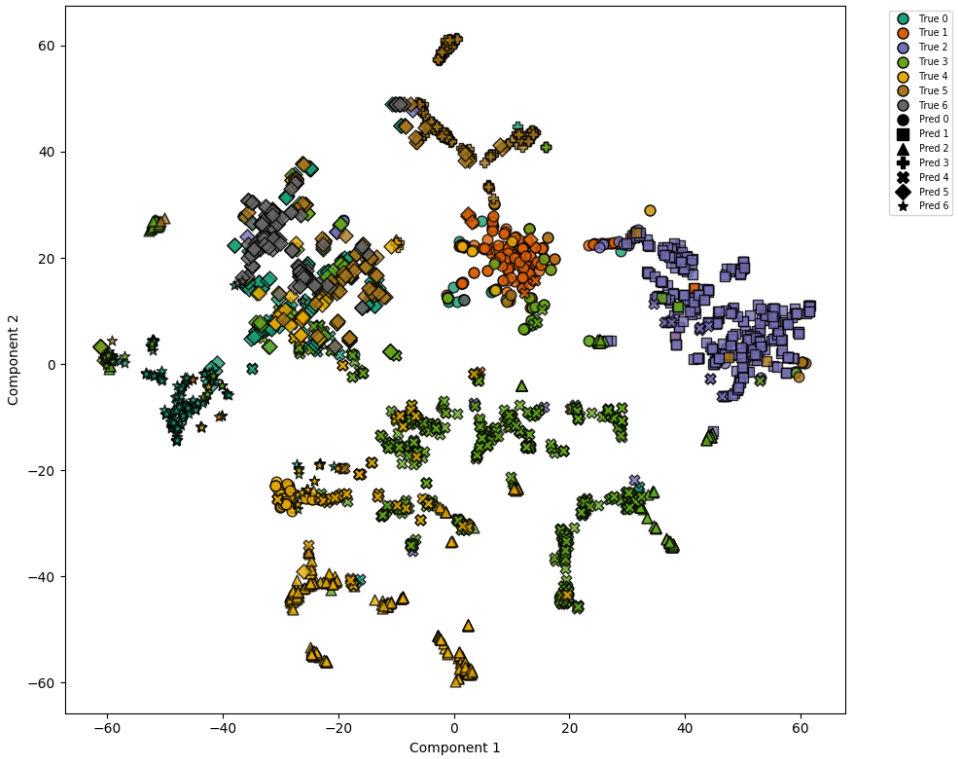


Figure 4.8 t-SNE visualisation of clustering via GMM of LSE. Colours denote the true labels, whilst marker shape shows the GMM clustering predictions.

Figure 4.9 further quantifies the agreement between the GMM cluster assignments and the ground truth using a confusion matrix. Diagonal dominance in most rows suggests that the LSE provides a clustering structure broadly aligned with the true labels. However, several classes exhibit substantial overlap. Notably, labels 3 and 4-corresponding to 'Probabilistic Methods' and 'Reinforcement Learning', respectively-are frequently confused with one another, likely reflecting their methodological similarity within the field of machine learning.

Class 6 (*Theory*), stands out as the most inaccurately classified cluster. It absorbs misclassified points from multiple other categories, particularly from class 0 (*Case Based*) and class 5 (*Rule Learning*). This misalignment is further illustrated in Figure 4.8, where the marker shapes, denoting GMM assignments, are highly heterogeneous within the colour cluster for class 6, suggesting that the embedding space does not clearly separate theoretical work from more applied

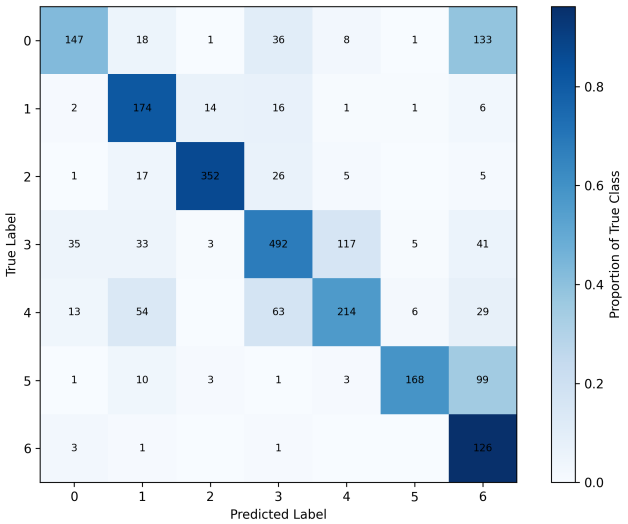


Figure 4.9 Confusion matrix showing counts and proportions of true labels and GMM cluster labels.

subfields. These findings highlight the limitations of graph structure-only embeddings in resolving semantically adjacent or conceptually diffuse categories, motivating the integration of textual features in downstream models.

To quantify the performance of different initialisation strategies, we computed the ARI across 50 runs of both KMeans and GMM, each with random initialisation. The median ARI and standard deviation are shown in Table 4.1. GMM slightly outperforms KMeans on average, justifying its use for initialising the cluster assignments $\mathbf{z}$ in our Bayesian inference procedure.

Table 4.1 Median ARI and standard deviation over 50 random initialisations for KMeans and GMM clustering on LSE.

Method	Median ARI	Std. Dev.
KMeans	0.356	0.025
GMM	0.395	0.035

4.2.3 Clustering with Combined Embedding and Text

Having evaluated clustering performance using spectral embeddings alone, we now apply the full Bayesian model, which integrates structural information from the graph (via LSE) with textual node features (via bag-of-words representations). This section outlines our prior specification and inference setup, highlighting the rationale for initialisation and hyper-parameter choices.

Rather than randomly initialising cluster assignments $\mathbf{z}$, we use the output of a GMM fitted to the LSE embedding. This ensures that the initial partition reflects meaningful structure in the graph, improving both the efficiency and convergence of the collapsed Gibbs sampler. We also use this GMM fit to calibrate the prior covariance parameter $\mathbf{S}_0$, which plays a critical role in stabilising inference as posterior behaviour is known to be sensitive to the scale of the covariance prior. Specifically, we compute the average within-cluster variance from the GMM assignments and set $\mathbf{S}_0$ as a diagonal matrix with these values, which aligns the prior scale with the observed variability while preserving independence across dimensions (see, for example

Sanna Passino et al., 2025).

Unless stated otherwise, we adopt the following default hyper-parameter values throughout: $\kappa_0 = \nu_0 = \gamma = \eta = 1$. We set $\mathbf{m}_0 = \mathbf{0}$, and initialise cluster assignments $\mathbf{z}^{(0)}$ using GMM as described above. These choices represent standard weakly-informative priors, which yield robust performance across both synthetic and real datasets.

The model is run for $M = 1000$ samples following a burn-in period of 200 iterations, and final cluster assignments are obtained using the posterior similarity matrix averaged over these post-burn-in samples.

Table 4.2 ARI comparison between embedding-only clustering methods and the joint model.

Method	Median ARI
KMeans (LSE)	0.356
GMM (LSE)	0.395
Joint clustering model	0.427

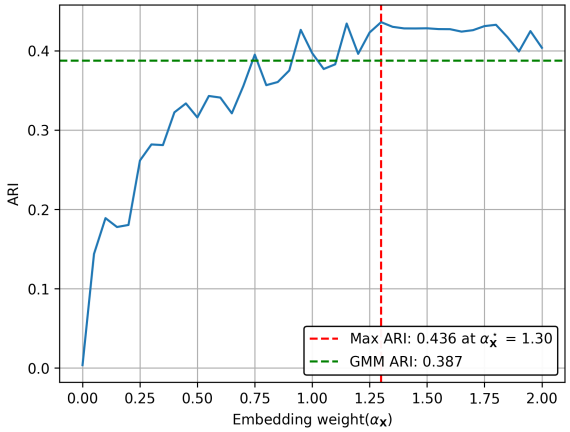
These results indicate a modest improvement in clustering performance when incorporating textual node features alongside spectral embeddings. While the joint model outperforms embedding-only baselines, the gain is relatively limited, suggesting that the additional signal provided by the text may be minor. In the following section, we investigate whether modulating the relative contribution of each modality can lead to more effective integration. Specifically, we explore how varying the weight assigned to the embedding and text likelihoods affects clustering accuracy.

Reweighting Textual Contribution

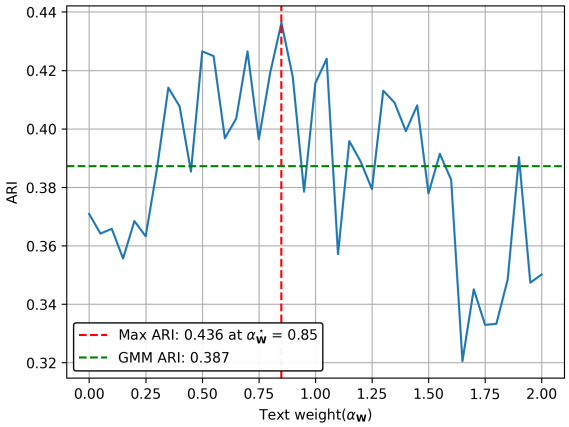
The relative informativeness of the textual and embedding modalities is highly variable. In the Cora dataset, for example, the bag-of-words representation alone does not sufficiently distinguish between certain clusters, due to shared vocabulary across topics.

The results shown in Figure 4.10 demonstrate that underweighting the embedding likelihood (i.e., setting $\alpha_{\mathbf{X}} < 1$) leads to consistently worse performance than the GMM baseline, suggesting that the structural information captured by the spectral embedding is critical for accurate clustering. As $\alpha_{\mathbf{X}}$ increases from 0 to 1, increasing the relative influence of the embedding, the ARI improves sharply, indicating that the model benefits from a stronger emphasis on graph structure. Beyond $\alpha_{\mathbf{X}} = 1$, performance plateaus, with clustering accuracy remaining consistently above the GMM baseline, but showing limited further gains. This suggests that while it is important not to underweight the embedding, aggressively upweighting it offers diminishing returns.

As an alternative to independently weighting the text and embedding likelihoods, we also consider a reparameterisation in which these two components form a convex combination controlled by a single parameter $\alpha \in [0, 1]$, setting $\alpha_W = 2\alpha$ and $\alpha_X = 2(1-\alpha)$. This ensures a fixed total contribution from the data likelihood terms, while varying the balance between modalities. As shown in Figure 4.11, performance peaks at $\alpha^\star = 0.36$, again favouring the embedding modality. However, 'best case' ARI scores are lower than in the previous setup with independent weights. This is likely due to the fact that for all $\alpha \neq 0.5$, the combined weight on the text and embedding likelihood terms exceeds that on the cluster assignment prior $\log p(z_i \mid z_{-i})$, effectively downweighting its influence. As a result, the model may be less sensitive to prior regularisation from the global clustering structure, leading to noisier or overfit assignments.



(a) Varying embedding weight $\alpha_{\mathbf{X}}$ with fixed text weight $\alpha_{\mathbf{W}} = 1$.



(b) Varying text weight $\alpha_{\mathbf{W}}$ with fixed embedding weight $\alpha_{\mathbf{X}} = 1$.

Figure 4.10 Effect of reweighting the text and embedding likelihoods on clustering performance (measured by ARI). Weights were tested in increments of 0.05 over the range $[0, 2]$. For each value, the model was run for $M = 500$ samples following a burn-in period of 200 iterations.

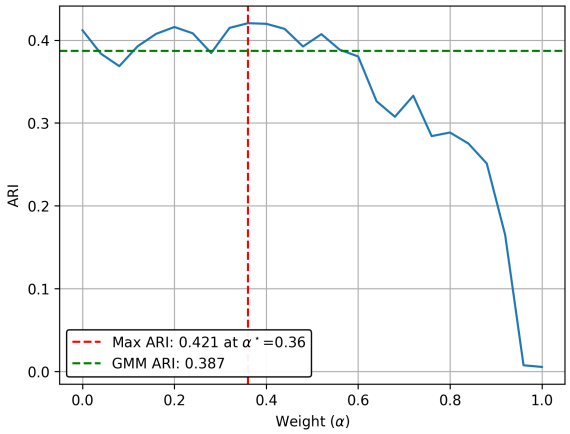


Figure 4.11 ARI as a function of weighting parameter α, with reference lines for the maximum ARI and the GMM baseline.

While this formulation offers a more interpretable trade-off between text and graph information, we find it less effective in practice and do not explore it further.

Chapter 5

Conclusion

This report has presented a novel Bayesian model for joint clustering in text-attributed networks, integrating structural information from graph embeddings with node-level textual covariates. The proposed model combines a finite GMM over spectral embeddings with Dirichlet-Multinomial topic models for the associated text, and performs inference via a collapsed Gibbs sampling scheme. The model is grounded in the asymptotic theory of spectral graph embeddings and RDPGs, which justifies the use of Gaussian likelihoods in the embedding space. A modality reweighting parameter was introduced to allow flexible integration of structure and content, adapting the model to varying signal quality across modalities.

The model was evaluated on both synthetic datasets-designed to test sensitivity to embedding separation, vocabulary distinctiveness, and class imbalance-and on the Cora citation network, a benchmark example of a text-attributed graph. The results demonstrate that the joint model improves clustering accuracy compared to text-only or structure-only baselines, particularly when neither modality alone provides sufficient discriminative power. To aid convergence and improve stability, cluster assignments were initialised using a GMM fit to the embeddings, while the prior covariance in the NIW distribution was set based on the average within-cluster variance across the GMM components.

Limitations and Future Work. Several directions could extend the proposed model. A limitation of the current model is its assumption of a fixed vocabulary size $|\mathcal{V}|$, with each cluster-specific word distribution ϕ_k drawn from a symmetric Dirichlet prior $\text{Dir}(\eta/|\mathcal{V}|, \ldots, \eta/|\mathcal{V}|)$. In many practical settings, particularly when clustering held-out data or modelling online network structures, new words may appear that were not observed in the original training corpus. To address this, the Dirichlet prior can be replaced with a Griffiths-Engen-McCloskey (GEM) distribution (Pitman, 2006), which arises as the limit of $\text{Dir}(\eta/V, \ldots, \eta/V)$ as $V \to \infty$. The GEM defines a distribution over an infinite vocabulary via the stick-breaking construction (Sethuraman, 1994), where $V_k \sim \text{Beta}(1, \eta)$ and $\pi_k = V_k \prod_{j=1}^{k-1}(1 - V_j)$. This construction allows the model to place positive probability on previously unseen words. See, for example, Sanna Passino et al. (2025), who adopt this approach in the context of session-level and command-level topic modelling for cyberattack data.

Similarly, placing a Dirichlet Process prior over cluster assignments would enable the number of clusters K to grow with the data, eliminating the need to fix it a priori. This is especially valuable when the true number of latent groups is unknown or expected to vary over time. These ideas fall naturally within the framework of hierarchical Bayesian non-parametrics, and are explored in related work such as Sanna Passino and Heard (2020). From an implementation perspective, such models typically involve auxiliary Gibbs sampling moves that merge, split, or instantiate new clusters during inference. Adapting these techniques to the current model would improve both its flexibility and applicability in more open-ended or evolving data

settings.

A second limitation concerns the computational scalability of the inference procedure. While collapsed Gibbs sampling improves mixing by integrating out nuisance parameters, its serial nature limits parallelisation and scalability to large networks. To address this, future work could use uncollapsed Gibbs sampling, which can easily be parallelised to allow for efficient sampling on GPUs. One promising strategy is the mean-for-mode approximation proposed by Tristan et al. (2015), which replaces the sampling of nuisance parameters with their posterior means. This approach retains much of the computational simplicity of collapsed methods, while enabling efficient, vectorised updates and parallelism. Alternatively, variational inference offers a deterministic and scalable alternative to sampling-based methods, and could be formulated for the current model to enable faster inference on large graphs with high-dimensional textual features (see, for example Teh et al., 2006)

Finally, the current model assumes that each cluster's word distribution ϕ_k is independently drawn from a symmetric Dirichlet prior, effectively treating topic vocabularies as disjoint and uncorrelated. This ignores potential lexical structure that may be shared across clusters, particularly in domains like scientific text, where related topics often use overlapping terminology. A natural extension would be to introduce a hierarchical Bayesian prior over these distributions. For instance, nested Dirichlet processes could be used to model shared structure across clusters, allowing cluster-specific vocabularies to borrow statistical strength from a global base distribution. Additionally, mixtures over shared secondary topics, such as those used in Sanna Passino et al. (2025) to capture global or high-frequency tokens, could be incorporated to improve topic interpretability and reduce redundancy in the estimated word distributions. These extensions would enable more coherent modelling in settings where topics are semantically related, and help avoid overfitting in situations with sparse text.

In summary, this work presents a Bayesian joint clustering model for text-attributed networks that integrates structural and textual information in a principled manner. The approach improves interpretability and adaptability in settings where neither modality alone is sufficient for reliable community detection.

Code. All code used in this project, including the model implementation, inference routines, and experiment notebooks, is available at the following repository: `https://github.com/bpellow/m4r-jcm`

Appendix A

Derivations for Collapsed Gibbs Sampling

To infer cluster assignments given the node embeddings and associated abstracts, we employ collapsed Gibbs sampling, which integrates out latent nuisance parameters to improve sampling efficiency. Specifically, we derive the collapsed conditional for z_i given all other cluster assignments $\mathbf{z}_{-i}$ and the observed data $\mathcal{D} = (\mathbf{X}, \mathbf{W})$.

$$\begin{aligned} p(z_i = k \mid \mathbf{z}_{-i}, \mathbf{X}, \mathbf{W}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma}) &\propto p(z_i = k \mid \mathbf{z}_{-i}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma}) p(\mathbf{X} \mid z_i = k, \mathbf{z}_{-i}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma}) \\ &\quad p(\mathbf{W} \mid z_i = k, \mathbf{z}_{-i}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma}) \\ &\propto p(z_i = k \mid \mathbf{z}_{-i}, \boldsymbol{\gamma}) p(\mathbf{x}_i \mid \mathbf{x}_{-i}, z_i = k, \mathbf{z}_{-i}, \boldsymbol{\beta}) \\ &\quad p(\mathbf{w}_i \mid \mathbf{w}_{-i}, z_i = k, \mathbf{z}_{-i}, \boldsymbol{\eta}) \end{aligned}$$

Each component is derived by first computing the marginal distribution obtained by integrating out its associated nuisance parameter. We then form a ratio of marginal likelihoods, corresponding to the inclusion or exclusion of node i in the relevant sufficient statistics. This yields the collapsed conditional terms used in the Gibbs sampler. Final expressions for the cluster prior, embedding likelihood, and text likelihood are given in Equations (A.3), (A.6), and (A.9), respectively.

A.1 Collapsed Cluster Assignment Probability

Let $N_{k,-i}$ be the number of data points assigned to cluster k, excluding node i.

The collapsed conditional distribution for z_i integrates over ψ:

$$p(z_i = k \mid \mathbf{z}_{-i}, \boldsymbol{\gamma}) = \frac{p(z_i = k, \mathbf{z}_{-i} \mid \boldsymbol{\gamma})}{p(\mathbf{z}_{-i} \mid \boldsymbol{\gamma})} = \frac{p(\mathbf{z} \mid \boldsymbol{\gamma})}{p(\mathbf{z}_{-i} \mid \boldsymbol{\gamma})} \tag{A.1}$$

Both the numerator and denominator are just marginals over the same Dirichlet-Multinomial distribution, differing only in whether they include or exclude the index i, and can therefore be evaluated in the same way on two subsets of assignments.

$$
\begin{aligned}
p(\mathbf{z} \mid \boldsymbol{\gamma}) &= \int_{\boldsymbol{\psi}} p(\mathbf{z} \mid \boldsymbol{\psi})\, p(\boldsymbol{\psi} \mid \boldsymbol{\gamma})\, d\boldsymbol{\psi} \\
&= \int_{\boldsymbol{\psi}} \left(\prod_{k=1}^{K} \psi_k^{N_k} \right) \cdot \frac{1}{\mathcal{B}(\boldsymbol{\gamma})} \prod_{k=1}^{K} \psi_k^{\gamma_k - 1}\, d\boldsymbol{\psi} \\
&= \frac{1}{\mathcal{B}(\boldsymbol{\gamma})} \int_{\boldsymbol{\psi}} \prod_{k=1}^{K} \psi_k^{N_k + \gamma_k - 1}\, d\boldsymbol{\psi} \\
&= \frac{\mathcal{B}(\mathbf{N} + \boldsymbol{\gamma})}{\mathcal{B}(\boldsymbol{\gamma})} \\
&= \frac{\prod_{k=1}^{K} \Gamma(N_k + \gamma_k)}{\Gamma\left(n + \sum_{k=1}^{K} \gamma_k\right)} \cdot \frac{\Gamma\left(\sum_{k=1}^{K} \gamma_k\right)}{\prod_{k=1}^{K} \Gamma(\gamma_k)} \\
&= \frac{\Gamma(\gamma)}{\Gamma(n+\gamma)\Gamma\left(\frac{\gamma}{K}\right)^K} \cdot \prod_{k=1}^{K} \Gamma\left(N_k + \frac{\gamma}{K}\right)
\end{aligned} \tag{A.2}
$$

Hence we can compute the ratio in Equation (A.1), up to a normalising constant:

$$
p(z_i = k \mid \mathbf{z}_{-i}) \propto \frac{N_{k,-i} + \gamma/K}{n + \gamma - 1}. \tag{A.3}
$$

A.2 Collapsed Embedding Likelihood

For notational clarity, we first derive the posterior of the parameters $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ given $\mathbf{X}^{(k)} = \{\mathbf{x}_i : z_i = k\}$, the set of nodes assigned to cluster k.

First the NIW prior is of the following form:

$$
\begin{aligned}
p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \mid \boldsymbol{\beta}) = &\frac{1}{Z_{\text{NIW}}(d, \kappa_0, \nu_0 + d - 1, \mathbf{S}_0)} |\boldsymbol{\Sigma}_k|^{-(\nu_0 + 2d + 1)/2} \\
&\times \exp\left(-\frac{\kappa_0}{2} (\boldsymbol{\mu}_k - \mathbf{m}_0)^\top \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{\mu}_k - \mathbf{m}_0) - \frac{1}{2} \operatorname{tr}(\boldsymbol{\Sigma}_k^{-1} \mathbf{S}_0) \right)
\end{aligned} \tag{A.4}
$$

with

$$
Z_{\text{NIW}}(d, \kappa_0, \nu_0 + d - 1, \mathbf{S}_0) = 2^{\frac{(\nu_0+d)d}{2}} \pi^{\frac{d(d+1)}{4}} \kappa_0^{-d/2} |\mathbf{S}_0|^{-(\nu_0+d-1)/2} \prod_{i=1}^{d} \Gamma\left(\frac{\nu_0 + d - i}{2} \right)
$$

Combining with the likelihood, we compute the posterior as follows.

$$
\begin{aligned}
p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \mid \mathbf{X}^{(k)}) &\propto p(\mathbf{X}^{(k)} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \mid \mathbf{m}_0, \kappa_0, \nu_0 + d - 1, \mathbf{S}_0) \\
&= \frac{(2\pi)^{-N_k d/2}}{Z_{\text{NIW}}(d, \kappa_0, \nu_0 + d - 1, \mathbf{S}_0)} |\boldsymbol{\Sigma}_k|^{-(\nu_0 + N_k + 2d + 1)/2} \\
&\quad \times \exp\left(-\frac{\kappa_0 + N_k}{2} \left(\boldsymbol{\mu}_k - \frac{\kappa_0 \mathbf{m}_0 + N_k \bar{\mathbf{x}}^{(k)}}{\kappa_k} \right)^\top \boldsymbol{\Sigma}_k^{-1} \left(\boldsymbol{\mu}_k - \frac{\kappa_0 \mathbf{m}_0 + N_k \bar{\mathbf{x}}^{(k)}}{\kappa_k} \right) \right. \\
&\quad \left. - \frac{1}{2} \operatorname{tr}\left(\boldsymbol{\Sigma}_k^{-1} \left(\mathbf{S}_0 + \mathbf{S}_{\bar{\mathbf{x}}^{(k)}} + \frac{\kappa_0 N_k}{\kappa_0 + N_k} (\bar{\mathbf{x}}^{(k)} - \mathbf{m}_0)(\bar{\mathbf{x}}^{(k)} - \mathbf{m}_0)^\top \right) \right) \right)
\end{aligned}
$$

By comparing this to the expression for the NIW prior, Equation (A.4), we recognise this as a NIW density with updated parameters $(\mathbf{m}_k, \kappa_k, \nu_k + d - 1, \mathbf{S}_k)$ corresponding to cluster k, where each depends implicitly on the cluster size N_k and other sufficient statistics.

$$p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \mid \mathbf{X}^{(k)}) = \text{NIW}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \mid \mathbf{m}_k, \kappa_k, \nu_k + d - 1, \mathbf{S}_k)$$

where:

$$\begin{aligned}
\mathbf{m}_k &= \frac{\kappa_0 \mathbf{m}_0 + N_k \bar{\mathbf{x}}^{(k)}}{\kappa_k} = \frac{\kappa_0}{\kappa_0 + N_k}\mathbf{m}_0 + \frac{N_k}{\kappa_0 + N_k}\bar{\mathbf{x}}^{(k)} \\
\kappa_k &= \kappa_0 + N_k \\
\nu_k &= \nu_0 + N_k \\
\mathbf{S}_k &= \mathbf{S}_0 + \mathbf{S}_{\bar{\mathbf{x}}^{(k)}} + \frac{\kappa_0 N_k}{\kappa_k}(\bar{\mathbf{x}}^{(k)} - m_0)(\bar{\mathbf{x}}^{(k)} - m_0)^\top \\
&= \mathbf{S}_0 + \mathbf{S}^{(k)} + \kappa_0 \mathbf{m}_0 \mathbf{m}_0^\top - \kappa_k \mathbf{m}_k \mathbf{m}_k^\top
\end{aligned}$$

where the cluster-specific sufficient statistics are defined as:

$$\begin{aligned}
\bar{\mathbf{x}}^{(k)} &= \frac{1}{N_k} \sum_{i:\, z_i = k} \mathbf{x}_i \\
\mathbf{S}_{\bar{\mathbf{x}}^{(k)}} &= \sum_{i:\, z_i = k} (\mathbf{x}_i - \bar{\mathbf{x}}^{(k)})(\mathbf{x}_i - \bar{\mathbf{x}}^{(k)})^\top \\
\mathbf{S}^{(k)} &= \sum_{i:\, z_i = k} \mathbf{x}_i \mathbf{x}_i^\top
\end{aligned}$$

With this notation we can now integrate out the nuisance parameters $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ by using the full joint from before:

$$\begin{aligned}
p(\mathbf{X}^{(k)} \mid \boldsymbol{\beta}) &= \int_{\boldsymbol{\mu}_k} \int_{\boldsymbol{\Sigma}_k} p(\mathbf{X}^{(k)}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \mid \boldsymbol{\beta})\, d\boldsymbol{\mu}_k\, d\boldsymbol{\Sigma}_k \\
&= \int_{\boldsymbol{\mu}_k} \int_{\boldsymbol{\Sigma}_k} p(\mathbf{X}^{(k)} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \mid \boldsymbol{\beta})\, d\boldsymbol{\mu}_k\, d\boldsymbol{\Sigma}_k \\
&= \frac{(2\pi)^{-N_k d/2}}{Z_{\text{NIW}}(d, \kappa_0, \nu_0 + d - 1, \mathbf{S}_0)} \int_{\boldsymbol{\mu}_k} \int_{\boldsymbol{\Sigma}_k} |\boldsymbol{\Sigma}_k|^{-(\nu_k + 2d + 1)/2} \\
&\quad \times \exp\left(-\frac{\kappa_k}{2} (\boldsymbol{\mu}_k - \mathbf{m}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{\mu}_k - \mathbf{m}_k) - \frac{1}{2} \operatorname{tr}(\boldsymbol{\Sigma}_k^{-1} \mathbf{S}_k) \right) d\boldsymbol{\mu}_k\, d\boldsymbol{\Sigma}_k \\
&= (2\pi)^{-N_k d/2} \frac{Z_{\text{NIW}}(d, \kappa_k, \nu_k + d - 1, \mathbf{S}_k)}{Z_{\text{NIW}}(d, \kappa_0, \nu_0 + d - 1, \mathbf{S}_0)} \\
&= \pi^{-N_k d/2} \frac{\kappa_0^{d/2} |\mathbf{S}_0|^{(\nu_0 + d - 1)/2} \prod_{i=1}^{d} \Gamma\left(\frac{\nu_0 + d - i}{2}\right)}{\kappa_k^{d/2} |\mathbf{S}_k|^{(\nu_k + d - 1)/2} \prod_{i=1}^{d} \Gamma\left(\frac{\nu_k + d - i}{2}\right)}
\end{aligned} \tag{A.5}$$

Using this marginal likelihood, we can compute the collapsed conditional embedding likelihood for node i by evaluating the ratio of marginal densities with and without $\mathbf{x}_i$ included in cluster k, where $\mathbf{X}_{-i}^{(k)} = \{\mathbf{x}_j : z_j = k,\, j \neq i\}$ denotes the set of embeddings in cluster k excluding node i.

$$
\begin{aligned}
p(\mathbf{x}_i \mid z_i = k, \mathbf{x}_{-i}, \mathbf{z}_{-i}, \boldsymbol{\beta}) &= p(\mathbf{x}_i \mid \mathbf{X}_{-i}^{(k)}, \boldsymbol{\beta}) \\
&= \frac{p(\mathbf{X}^{(k)} \mid \boldsymbol{\beta})}{p(\mathbf{X}_{-i}^{(k)} \mid \boldsymbol{\beta})} \\
&= \frac{(2\pi)^{-N_k d/2}}{(2\pi)^{-N_{k,-i} d/2}} \cdot \frac{Z_{\text{NIW}}(d, \kappa_k, \nu_k + d - 1, \mathbf{S}_k)}{Z_{\text{NIW}}(d, \kappa_{k,-i}, \nu_{k,-i} + d - 1, \mathbf{S}_{k,-i})}
\end{aligned}
$$

It can be shown that this collapsed predictive distribution takes the form of a multivariate Student's t-distribution with $\nu_{k,-i}$ degrees of freedom. (Murphy, 2012 - 2012, Chapter 4):

$$
p(\mathbf{x}_i \mid z_i = k, \mathbf{x}_{-i}, \mathbf{z}_{-i}, \boldsymbol{\beta}) = t_{\nu_{k,-i}} \left(\mathbf{x}_i \mid \mathbf{m}_{k,-i}, \frac{\kappa_{k,-i} + 1}{\kappa_{k,-i}} \frac{\mathbf{S}_{k,-i}}{\nu_{k,-i}} \right) \tag{A.6}
$$

where $\mathbf{m}_{k,-i}, \kappa_{k,-i}, \nu_{k,-i}, \mathbf{S}_{k,-i}$ denote the NIW parameters for cluster k defined above *excluding* node i.

A.3 Collapsed Text Likelihood

Similarly to the embedding case, we aim to compute the collapsed conditional for the text component by taking the ratio of marginal distributions with and without node i included in cluster k, where $\mathbf{W}^{(k)} = \{\mathbf{w}_i : z_i = k\}$ is the set of word count vectors for nodes assigned to cluster k, and $\mathbf{W}_{-i}^{(k)} = \{\mathbf{w}_j : z_j = k,\, j \neq i\}$ excludes node i.

$$
p(\mathbf{w}_i \mid z_i = k, \mathbf{w}_{-i}, \mathbf{z}_{-i}, \boldsymbol{\eta}) = p(\mathbf{w}_i \mid \mathbf{W}_{-i}^{(k)}, \boldsymbol{\eta}) = \frac{p(\mathbf{W}^{(k)} \mid \boldsymbol{\eta})}{p(\mathbf{W}_{-i}^{(k)} \mid \boldsymbol{\eta})} \tag{A.7}
$$

Integrating out the nuisance parameter $\boldsymbol{\phi}_k$ as follows:

$$
\begin{aligned}
p(\mathbf{W}^{(k)} \mid \boldsymbol{\eta}) &= \int_{\boldsymbol{\phi}_k} p(\mathbf{W}^{(k)}, \boldsymbol{\phi}_k \mid \boldsymbol{\eta})\, d\boldsymbol{\phi}_k \\
&= \int_{\boldsymbol{\phi}_k} p(\mathbf{W}^{(k)} \mid \boldsymbol{\phi}_k) p(\boldsymbol{\phi}_k \mid \boldsymbol{\eta})\, d\boldsymbol{\phi}_k \\
&= \int_{\boldsymbol{\phi}_k} \prod_{i:z_i=k} \text{Mult}(\mathbf{w}_i) \frac{1}{\mathcal{B}(\boldsymbol{\eta})} \prod_{v=1}^{|\mathcal{V}|} \phi_{kv}^{\frac{\eta}{|\mathcal{V}|}-1}\, d\boldsymbol{\phi}_k \\
&\propto \int_{\boldsymbol{\phi}_k} \prod_{i:z_i=k} \prod_{v=1}^{|\mathcal{V}|} \phi_{kv}^{w_{iv}} \frac{1}{\mathcal{B}(\boldsymbol{\eta})} \prod_{v=1}^{|\mathcal{V}|} \phi_{kv}^{\frac{\eta}{|\mathcal{V}|}-1}\, d\boldsymbol{\phi}_k \\
&= \frac{1}{\mathcal{B}(\boldsymbol{\eta})} \int_{\boldsymbol{\phi}_k} \prod_{v=1}^{|\mathcal{V}|} \phi_{kv}^{c_v^{(k)} + \frac{\eta}{|\mathcal{V}|}-1}\, d\boldsymbol{\phi}_k \\
&= \frac{\mathcal{B}(\mathbf{c}^{(k)})}{\mathcal{B}(\boldsymbol{\eta})} \\
&= \frac{\Gamma(\eta)}{\Gamma\left(\eta + C^{(k)}\right) \Gamma\left(\eta/|\mathcal{V}|\right)^{|\mathcal{V}|}} \prod_{v=1}^{|\mathcal{V}|} \Gamma\left(c_v^{(k)} + \eta/|\mathcal{V}|\right)
\end{aligned} \tag{A.8}
$$

Hence, the expression in Equation (A.7) follows directly by taking the ratio of the marginal likelihoods $p(\mathbf{W}^{(k)} \mid \boldsymbol{\eta})$ and $p(\mathbf{W}_{-i}^{(k)} \mid \boldsymbol{\eta})$, where the latter is computed as above but using only the subset of documents in cluster k excluding node i. Substituting the corresponding word count vectors into the ratio yields:

$$\begin{aligned} p(\mathbf{w}_i \mid z_i = k, \mathbf{w}_{-i}, \mathbf{z}_{-i}, \boldsymbol{\eta}) &= \frac{\Gamma\left(\eta + C_{-i}^{(k)}\right)}{\Gamma\left(\eta + C^{(k)}\right)} \prod_{v=1}^{|\mathcal{V}|} \frac{\Gamma\left(c_v^{(k)} + \eta/|\mathcal{V}|\right)}{\Gamma\left(c_{v,-i}^{(k)} + \eta/|\mathcal{V}|\right)} \\ &= \frac{\Gamma\left(\eta + C_{-i}^{(k)}\right)}{\Gamma\left(\eta + C_{-i}^{(k)} + M_i\right)} \prod_{v=1}^{|\mathcal{V}|} \frac{\Gamma\left(c_{v,-i}^{(k)} + w_{iv} + \eta/|\mathcal{V}|\right)}{\Gamma\left(c_{v,-i}^{(k)} + \eta/|\mathcal{V}|\right)} \\ &= \frac{1}{\prod_{r=0}^{M_i-1}\left(\eta + C_{-i}^{(k)} + r\right)} \prod_{v:\, w_{iv}>0} \frac{\Gamma\left(c_{v,-i}^{(k)} + w_{iv} + \eta/|\mathcal{V}|\right)}{\Gamma\left(c_{v,-i}^{(k)} + \eta/|\mathcal{V}|\right)} \end{aligned} \tag{A.9}$$

where:

- $\mathcal{B}$ is the multivariate Beta function,
- $\mathbf{c}_{-i}^{(k)} = \sum_{j \neq i, z_j = k} \mathbf{w}_j = \left(c_{1,-i}^{(k)}, \ldots, c_{|V|,-i}^{(k)}\right)$ is the vector of word counts of all nodes in cluster k, excluding node i,
- $C_{-i}^{(k)} = \sum_{v=1}^{|V|} c_{v,-i}^{(k)}$.

The final line follows by repeatedly applying the identity $\Gamma(x+1) = x\Gamma(x)$, and observing that terms in the product cancel whenever $w_{iv} = 0$. Repeated use of this identity yields the general form $\frac{\Gamma(x+n)}{\Gamma(x)} = \prod_{r=0}^{n-1}(x+r)$, which is applied here to simplify the ratio of gamma functions.

A.4 Joint marginal likelihood

The joint marginal likelihood under the model admits a natural factorisation into independent contributions from the embeddings, the textual features, and the cluster assignments (Kamper, 2015):

$$\begin{aligned} p(\mathbf{X}, \mathbf{W}, \mathbf{z} \mid \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\beta}) &= p(\mathbf{X} \mid \mathbf{z}, \boldsymbol{\beta}) p(\mathbf{W} \mid \mathbf{z}, \boldsymbol{\eta}) p(\mathbf{z} \mid \boldsymbol{\gamma}) \\ &= \left(\prod_{k=1}^{K} p(\mathbf{X}^{(k)} \mid \boldsymbol{\beta}) p(\mathbf{W}^{(k)} \mid \boldsymbol{\eta}) \right) p(\mathbf{z} \mid \boldsymbol{\gamma}) \end{aligned}$$

Closed-form expressions for each of the three components, embedding likelihood, text likelihood, and prior over assignments, have been derived previously in Equations (A.2), (A.5), and (A.8).

A.5 Incremental Update Formulae

We provide closed-form updates to posterior hyper-parameters when adding/removing a data point i to/from a particular cluster, with parameters $(\kappa, \nu, \mathbf{m}, \mathbf{S})$, the updated parameters (denoted $(\kappa', \nu', \mathbf{m}', \mathbf{S}')$ can be computed efficiently as follows:

A.5.1 Addition of node

$$
\begin{aligned}
\kappa' &= \kappa + 1 \\
\nu' &= \nu + 1 \\
\mathbf{m}' &= \frac{\kappa\mathbf{m} + \mathbf{x}_i}{\kappa'} \\
\mathbf{S}' &= \mathbf{S} + \mathbf{x}_i\mathbf{x}_i^\top - \kappa\mathbf{m}\mathbf{m}^\top + \kappa'\mathbf{m}'(\mathbf{m}')^\top
\end{aligned}
$$

A.5.2 Removal of node

$$
\begin{aligned}
\kappa' &= \kappa - 1 \\
\nu' &= \nu - 1 \\
\mathbf{m}' &= \frac{\kappa\mathbf{m} - \mathbf{x}_i}{\kappa'} \\
\mathbf{S}' &= \mathbf{S} - \left(\mathbf{x}_i\mathbf{x}_i^\top - \kappa\mathbf{m}\mathbf{m}^\top + \kappa'\mathbf{m}'(\mathbf{m}')^\top\right)
\end{aligned}
$$

Bibliography

N. Whiteley, A. Gray, and P. Rubin-Delanchy. Statistical exploration of the manifold hypothesis, 2025.

S. Young and E. R. Scheinerman. Random dot product graph models for social networks. In *Proceedings of the 5th International Workshop on Algorithms and Models for the Web-Graph (WAW)*, volume 4863 of *Lecture Notes in Computer Science*, pages 138–149. Springer, 2007.

P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

A. Athreya, D. E. Fishkind, M. Tang, C. E. Priebe, Y. Park, J. T. Vogelstein, K. Levin, and V. Lyzinski. Statistical inference on random dot product graphs: a survey. *Journal of Machine Learning Research*, 18(226):1–92, 2018.

P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 1983.

P. H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

A. Grover and J. Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 855–864. ACM, 2016.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 2003.

F. Sanna Passino, A. Mantziou, D. Ghani, P. Thiede, R. Bevington, and N. A. Heard. Nested dirichlet models for unsupervised attack pattern detection in honeypot data. *The Annals of Applied Statistics*, 19(1):586–613, 2025.

F. Sanna Passino and N. A. Heard. Bayesian inference for spectral graph clustering. *Journal of Machine Learning Research*, 2020.

J. S. Liu. The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427):958–966, 1994.

G. O. Roberts and S. K. Sahu. Updating schemes, correlation structure, blocking and parameterization for the gibbs sampler. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2):291–317, 1997.

A. Jasra, C. C. Holmes, and D. A. Stephens. Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, 20(1):50–67, 2005.

M. Medvedovic, K. Y. Yeung, and R. E. Bumgarner. Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, 20(8):1222–1232, 2004.

L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

M. Zhu and A. Ghodsi. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, 51(2):918–930, 2006.

V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st International Conference on Distributed Computing Systems Workshops*. IEEE, 2011.

M. Tang and C. E. Priebe. Limit theorems for eigenvectors of the normalized laplacian for random graphs. *The Annals of Statistics*, 46(5):2360–2415, 2018.

A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2002.

L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov), 2008.

H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2), 1955.

J. Pitman. *Combinatorial stochastic processes*, volume 1875 of *Lecture Notes in Mathematics*. Springer, 2006.

J. Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, 4(2):639–650, 1994.

J.-B. Tristan, J. Tassarotti, and G. Steele. Efficient training of lda on a gpu by mean-for-mode estimation. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 59–68, Lille, France, 07–09 Jul 2015. PMLR.

Y. W. Teh, D. Newman, and M. Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems 19*, pages 1353–1360. MIT Press, 2006.

K. P. Murphy. *Machine learning : a probabilistic perspective*. Adaptive computation and machine learning. The MIT Press, Cambridge, Massachusetts, 2012 - 2012.

H. Kamper. Gibbs sampling for fitting finite and infinite gaussian mixture models, 2015.