Imperial College London

MEng Individual Project

Imperial College London

Department of Computing

Node-Level Hypothesis Testing in P-Value Weighted Networks

Author:
Archie Licudi

Supervisor:
Prof. Nick Heard

Second Marker:
Dr. Francesco Sanna Passino

June 20, 2024

Abstract

Analysis of network-structured data is a mature and well-studied field of inter-disciplinary statistical research that endures in the modern age of machine learning. We present a comprehensive introduction to the field of approximate marginal inference in two-community weighted networks, considered through the lens of edge-to-node inference based on p-value weightings. We develop a novel *spectral pipeline* for community testing and p-value combination, drawing from the state-of-the-art techniques in information theory, statistical physics, and Bayesian inference. We provide theoretical guarantees as a product of carefully-chosen pipeline components and perform systematic simulation studies of the pipeline's effectiveness, demonstrating an advantage in hypothesis testing performance over existing proposals.

Acknowledgements

Thank you to my supervisor Prof Nick Heard for his guidance and my group at Arm research for helping me fall in love with statistical learning! Finally, thank you to my friends and family who supported me always as I worked on this paper.

Contents

List of Figures

Chapter 1

Introduction

Statistical networks are one of the most popular and versatile ways of representing real-world phenomena, in particular those that hinge on *pairwise interactions*: this includes *spin glasses* in quantum physics [BRU67], gene co-expression data in biology [DKMZ11a], and of course internet communications in computer science. As a result, much attention has been paid to network-structured data in both classical statistics and the emerging field of deep learning. A common question that arises in these analyses is how we can move from information about *connections between objects* to information about the objects themselves, and a slew of methods exist in the machine learning literature (for example, [WSL+19]), which we term **edge-to-node inference**.

A major roadblock faced by deep learning methods in applications to demanding problems that can be well-described by a traditional statistical framing is the lack of *principled* guarantees: whilst it may be possible to evaluate the empirical performance of deep learning on complex "easy-for-humans" tasks such as image recognition, the opaque and inherently approximative nature of deep models makes it difficult or even ineffective to transition these promising results into strong guarantees of correctness when distinguishing distributions that can be exactly described mathematically, even when these occur in the real world [GJB+23].

This leaves traditional statistical methodologies as an invaluable tool in modern network analysis, with approaches from both Bayesian and frequentist statistics being applied to and often outperforming deep models on many problems which admit a simple distributional description. One of the foundational tools of statistics is the *hypothesis test* and associated notion of *p-value*, and this naturally suggests that a rigorous framework for considering p-values in networks can be an invaluable tool for statistical analyses of these structures. In particular, we consider the edge-to-node inference problem in the context of *p-value combination*: given a network whose edges are weighted by p-values, how can we derive p-values on nodes? Many approaches have been proposed for p-value combination, and the necessity for it arises commonly, but the study of theoretical guarantees endowed by these combiners is comparatively sparse within the statistics landscape. Furthermore, to our knowledge there is no comprehensive analysis of p-value combination in the network setting currently extant in the literature, with single-node combinations occurring on a single sequence of independent p-values with no regard for pairwise interaction.

Building the natural model of hypothesis testing networks as having two communities of nodes (H_0^{node}, H_1^{node}) and two types of edge (H_0^{edge}, H_1^{edge}) immediately leads to deep and well-studied questions on inference in much more general classes of network once the question of likelihood ratios is considered, indeed subsuming many models that have been seminal in the past century of statistical research. This suggests the potential for exciting inter-disciplinary work in which methods developed to analyse p-values arising in genetics can be directly applied to models of quantum interaction in physics, and it is this lens through which we approach the question.

1.1 Contributions

We assume that the reader has familiarity with the basics of probability and statistics, along with undergraduate linear algebra and a passing acquaintance with the theories of optimization and combinatorial graphs. From this baseline, we introduce in Chapter 2 the theory of hypothesis testing and current state-of-the-art in general p-value combination methods, before describing our formalism for p-value networks (PVNs) in terms of the *weighted stochastic block model*. We also provide a brief introduction to information-theoretic measures of community separation and a pre-existing method which yields p-value networks.

Chapter 3 begins with formulating the likelihood-ratio test on PVN node hypotheses and situating this problem within the Ising model of statistical physics. From here, we tour the state-of-the-art in approximate inference in stochastic block/Ising models, providing the reader with a survey of techniques in Markov Chain Monte-Carlo, Variational Inference, Expectation-Maximisation, and Belief Propagation; at each stage, we derive novel versions of these algorithms for our setting. Section 3.5 then explores spectral methods on stochastic block models, focusing on the central limit theorem of [GJB+23] from which we provide a novel analysis of the SIMPLE algorithm.

Finally, we survey the modern landscape of information-theoretically optimal estimators for the stochastic block model, and use this to present our main contribution: a novel **spectral pipeline** framework for marginalisation and hence hypothesis testing in p-value networks that generalises each approach seen so far and is immediately applicable to *any* two-community weighted block model - in particular, the least well-studied regime of inference in *heterogenous* community structure. Each component of this approximate Bayesian inference with spectral prior framework is then comprehensively analysed through simulations in Chapter 4, including a new connection between phase transitions in detection and the Chernoff information in the network under spectral embedding. We provide numerical evidence that this method outperforms on medium to large networks prior proposals for both p-value combination and approximate inference in our setting, culminating in the proposal of the following combination method, applicable to any p-value network studied:

1. Transform each p-value in the adjacency matrix by Φ^{-1}, the inverse CDF of $\mathcal{N}(0,1)$
2. Embed each node in the graph by its components in the eigenvectors corresponding to the two highest eigenvalues in absolute value of $\Phi^{-1}(\mathbf{A})$
3. Fix a two-component Gaussian Mixture Model on these embeddings
4. Use this GMM as a prior for expectation-maximisation on any unknown parameters of the alternative distribution, with M-step given either by Belief Propagation or an MCMC method
5. After convergence of E-M, we have a marginal distribution on hypotheses of each individual nodes, from which a likelihood ratio is formed as the node-level test statistic

Chapter 2

Background

We begin with a brief survey of preliminaries from frequentist statistics, before introducing the theory of stochastic graphs and their lower-dimensional embeddings and context within the landscape of machine learning.

2.0.1 Statistical preliminaries

Let $X : \Omega \to \mathcal{X}$ be a random variable and p_θ a family of distributions parametric in $\theta \in \Theta$ such that $X \sim p_\theta$ for unknown parametrization. We make a **hypothesis** consisting of a partition of the **parameter space** Θ into two disjoint sets $\Theta = \Theta_0 \sqcup \Theta_1$. We refer to $H_0 : \theta \in \Theta_0$ as the **null hypothesis** and $H_1 : \theta \in \Theta_1$ (or equivalently, $\theta \in \Theta_0^c$ where complement is taken over Θ) as the **alternative hypothesis**.

- When $\Theta_0 = \{\theta_0\}$ (or $\Theta_1 = \{\theta_1\}$) this is known as a **simple** or **point** hypothesis
- Otherwise, it is a **composite** hypothesis.

Hypothesis testing is then the act of observing X to make a judgement on the probability of each hypothesis.

Definition (Hypothesis test)**.** A **test function** is a thus a mapping $\varphi : \mathcal{X} \to \{0, 1\}$ - or equivalently a subset $R \subseteq \Theta$ such that $\varphi(x) := \mathbf{1}_R$. This support set R is the **rejection region**, and a well-selected hypothesis test is one which most closely satisfies:

$$X \in R \iff \theta \in \theta_1$$

This immediately implies two types of error we are interested in, one in each direction of the equivalence. To support the case of composite hypotheses, we consider:

Definition (Power functions)**.** The **power function** of a hypothesis test φ is a mapping $\pi_\varphi : \Theta \to [0, 1]$ quantifying the likelihood of rejection given a particular $\theta \in \Theta$

$$\pi_\varphi(\theta) = \mathbf{E}[\varphi(X) \mid \theta] = \mathbf{P}(\varphi(X) = 1 \mid \theta) \tag{2.1}$$

Note that these functions are uniquely determined by p_θ, without any (Bayesian-like) consideration of the probability distribution on Θ.

Definition (Type 1 error)**.** Otherwise known as the **size**, this is the probability of *falsely rejecting the null hypothesis.* More formally, it is given by:

$$\alpha = \sup_{\theta \in \Theta_0} \pi_\theta(\varphi) = \sup_{\theta \in \Theta_0} \mathbf{P}(\varphi(X) = 1 \mid \theta) \tag{2.2}$$

In the important case of anomaly detection or binary classification more generally, this is analogous to the **false positive rate**. We say that a test with size $\leqslant \alpha$ has **level** α.

Conversely, **type 2 error** is the probability of failing to reject a false null hypothesis. In many cases, we will be interested in testing a simple null hypothesis against a composite alternative, notably in anomaly detection. It makes less sense then to reason about the supremum of type 2 errors as a single scalar "power" of the test, so we instead consider a hierarchy on power functions:

Definition (Uniformly Most Powerful test)**.** A level-α test φ is **Uniformly Most Powerful (UMP)** if, for any other level-α test φ':

$$\forall \theta \in \Theta_1, \quad \pi_{\varphi'}(\theta) \leqslant \pi_{\varphi}(\theta)$$

Remark. Note that for a simple null hypothesis, the power of the test can be written as a simple scalar value, denoted $1 - \beta$. UMP tests can then be decided simply by comparing values of the type 2 error β.

Assuming our hypotheses are such that we can estimate well $\mathbf{P}(X \mid \theta)$, a natural candidate for a test function would be to find the ratio of the probability that X occurs under H_1 to the probability that X occurs under H_0. Formally:

Definition (Likelihood Ratio Test)**.** Consider the random variable $X \sim p_\theta$ and hypotheses $H_0 : \Theta_0, H_1 : \Theta_1$. Then the **likelihood ratio test** is defined by:

$$\varphi_\Lambda(x) = \mathbf{1}_{\{\Lambda(x)>k\}}(x) \qquad \text{where} \qquad \Lambda(x) := \frac{\sup_{\theta\in\Theta_1} \mathbf{P}(X = x \mid \theta)}{\sup_{\theta\in\Theta_0} \mathbf{P}(X = x \mid \theta)}$$

k is thus a threshold parameter that uniquely decides the test's size α.

We note that this threshold parameter lets us choose a likelihood test of any level. The following is a foundational result in the theory of hypothesis testing and provides an unambiguous justification for the likelihood ratio test in the case of point hypotheses. We omit the proof here, but interested readers should consult an introductory text on hypothesis testing such as [CB02].

Lemma (Neyman-Pearson)**.** Let $X \sim p_\theta$ and consider the point hypotheses $H_0 : \theta = \theta_0$, $H_1 : \theta = \theta_1$. Then for any level α, the likelihood ratio test φ_Λ which has size equal α is UMP among level-α tests.

This lemma is exceptionally useful and gives a solid foundation for subsequent optimality results in the case of known distributions, but the point hypothesis restriction can severely limit its power for testing of models with unknown parameters. A generalisation to the case of *one-tailed unknown parameters* is originally due to Karlin and Rubin, but we will give without proof a slight modification of the version seen in [LB19].

Definition (Monotone Likelihood Ratio)**.** Let $\{p_\theta \mid \theta \in \Theta\}$ be a parametrised family of density or mass functions such that $\Theta \subset \mathbb{R}$. This family has **Monotone Likelihood Ratio (MLR)** if there exists a statistic $T(x)$ such that $\theta_1 < \theta_2$ implies the likelihood ratio $\Lambda(x) = \frac{p_{\theta_1}(x)}{p_{\theta_2}(x)}$ is monotone (non-increasing or non-decreasing) in $T(x)$ on the set for which $\Lambda(x)$ is defined - i.e. where $p_{\theta_1}(x), p_{\theta_2}(x) > 0$. More specifically, it has MLR if there exists some function T such that for any θ_1, θ_2 there exists some monotone function f_{θ_1,θ_2} defined on the same set as $\Lambda_{\theta_1,\theta_2}$ such that:

$$\theta_1 < \theta_2 \quad \implies \quad \forall x, \;\; f_{\theta_1,\theta_2}(T(x)) = \frac{p_{\theta_2}(x)}{p_{\theta_1}(x)} \tag{2.3}$$

Note that in the simplest case, $f_{\theta_1,\theta_2} = id$, hence the test statistic used is simply the likelihood ratio itself. Although not directly used in this version of the result, it is informative to consider choosing this T within the framework of *sufficient statistics*

Definition (Sufficient statistic)**.** Let $X \sim p_\theta$ with codomain $\mathcal{X}$. A statistic $T : \mathcal{X} \to \mathbb{R}$ is **sufficient** if the probability distribution of X, conditioned on $T(X)$, is independent of θ. Borrowing tools from information theory, T is sufficient if it encapsulates all information about θ:

$$I(\theta; T(X)) = I(\theta; X)$$

Where I denotes the ***mutual information***

Returning to the result:

Lemma (Karlin-Rubin)**.** Let $X \sim p_\theta$ where p_θ has non-decreasing MLR in $T(x)$. We define the two threshold tests

$$\varphi_+(x) = \begin{cases} 1 & T(x) > k \\ 0 & T(x) \leqslant k \end{cases} \qquad\qquad \varphi_-(x) = \begin{cases} 1 & T(x) < k' \\ 0 & T(x) \geqslant k' \end{cases}$$

Then:

$$\begin{aligned} &\varphi_+ \text{ is UMP for} && H_0 : \theta \leqslant \theta_0, \;\; H_1 : \theta > \theta_0 \\ &\varphi_- \text{ is UMP for} && H_0 : \theta \geqslant \theta_0, \;\; H_1 : \theta < \theta_0 \end{aligned}$$

These lemmas represent the two foundations from which we will build much of our analysis of hypothesis testing for combined statistics.

We recall some elementary facts about working with conditional probabilities of mixed conditional/discrete variables which will be used later for constructing likelihood ratios:

Definition (Mixed joint density)**.** Letting $A \sim p_A$ be a continuous random variable, $B \sim P_B$ a discrete random variable, we define the **mixed joint density**:

$$p_{A,B}(A, B) := p_A(A) P_B(B \mid A)$$

And hence the **mixed conditional density**:

$$p_A(A \mid B) := \frac{p_{A,B}(A, B)}{P_B(B)}$$

We will make a minor abuse of notation: if $A : \Omega \to \mathcal{A}$ is a random variable, then we may write $P(A)$ where A is shorthand for the event $A = \alpha$ for an arbitrary $\alpha \in \mathcal{A}$. Foundational tools of discrete probability hold in the mixed case:

Lemma (Mixed Bayes' theorem)**.** Let $A \sim p$ be a continuous random variable, $B \sim P$ a discrete random variable:

$$P(B \mid A) = \frac{p(A \mid B)P(B)}{p(A)} \qquad\qquad p(A \mid B) = \frac{P(B \mid A)p(A)}{P(B)}$$

Lemma (Mixed law of total probability)**.** Let $A \sim p$ be a continuous RV and $B \sim P$ be discrete with alphabet $\mathcal{B}$. Letting X denote any arbitrary event:

$$p(A \mid X) = \sum_{b \in \mathcal{B}} p(A \mid X, B = b) P(B = b \mid X) \qquad P(B \mid X) = \int_{\mathcal{A}} P(B \mid X, A = a) p(a \mid X) \, da$$

2.0.2 Combining p-values

We follow the approach of [Bir54] to motivate this question, who also gave a seminal result on its lack of a universal solution. We first make rigorous the notion of a p-value:

Definition (p-value)**.** Consider a hypothesis test based on statistic thresholding, as those seen in the Karlin-Rubin lemma. The **p-value** of an observation $t = T(x)$ is exactly:

$$\mathbf{P}(T(X) \text{ is } \textit{at least as extreme} \text{ as } t \mid H_0)$$

Where here “extreme" simply means either greater or less than - we can then see the p-value exactly as the type 1 error of the test if t is taken to be the threshold. The test of statistic T with threshold k is then equivalent to testing p-values as a statistic with threshold equal to the p-value of k.

The following proposition is well-known, but a proof can be illuminating:

Proposition 2.0.1. Let $X \sim p$ be continuous and T a test statistic with $T(X)$ the random variable of its value. Then, under H_0, the p-value of $T(X)$ is uniformly distributed on [0,1]

Proof. Wlog, we assume that this is a test for which the null hypothesis is rejected when $T(x) > k$. Letting $u(t)$ be the p-value of observation t, we note that:

$$u(t) = 1 - F_T(t \mid H_0) = \int_t^\infty p_T(t' \mid H_0) \, dt'$$

Where F_T, p_T are the CDF and PDF of $T(X)$ respectively (in the case where the PDF exists). Accordingly, we can find the CDF of the p-value distribution:

$$\begin{aligned} F_p(u \mid H_0) &= \mathbf{P}[1 - F_T(T(X) \mid H_0) < u \mid H_0] \\ &= 1 - \mathbf{P}\left[T(X) < F_T^{-1}(1 - u \mid H_0) \mid H_0\right] \\ &= 1 - F_T(F_T^{-1}(1 - u \mid H_0) \mid H_0) \\ &= u \end{aligned}$$

This is exactly the CDF of the uniform distribution, and since CDFs uniquely characterise distributions, we can conclude. □

Now suppose that we are able to devise a series of n independent statistics for the same hypothesis, and we test each one individually to yield a corresponding sequence of p-values. It is natural to now ask the optimal way to *combine* these p-values into a single test statistic on the hypothesis. The primary motivating example we consider in this paper is that of producing *node-level* statistics in a graph from *edge-level* tests: a p-value is assigned to each edge of a network corresponding to a hypothesis on its endpoints, and we seek a p-value for each node by combining the values on its incident edges. Denote the p-values arising from n tests as $u_1, ..., u_n$. By our previous result, we can thus give an equivalent test on p-values:

$$H_0 : \text{ each } u_i \sim \mathcal{U}[0,1] \qquad H_1 : \text{ one or more } u_i \sim g_i \text{ for } g_i \text{ non-uniform} \tag{2.4}$$

For the general case, we consider the alternative g_is to be unknown. When looking to develop a statistic for this test, we impose a “reasonableness" condition on the possible combiners:

Condition. If T is a most powerful combiner, then if H_0 is rejected for a given set of u_i, it will also be rejected for any p-value set u' where, for all i, $u'_i \leqslant u_i$. Equivalently, $T(u_1, ..., u_n)$ is monotone non-increasing in every component.

This is, in fact, the only restriction which we can place on optimal combiners in the most general setting. Indeed, following from [Bir55], this result is given in [Bir54]:

Theorem 2.0.1. For any combiner statistic $T(u_1, ..., u_n)$ of level α, there exists a set of alternative PDFs $g_1, ..., g_n$ such that it is UMP for the test given in (2.4).

This immediately implies that we cannot hope to find globally optimal combiners, but it is possible to reason about them for important classes of distribution. In particular, [HRD18] applies the results introduced in the previous section to derive optimal combiners for two common distributions

Proposition 2.0.2 (Heard, 2018)**.** Let $H_0 : u_i \sim \mathcal{U}[0,1]$

- In the case that $H_1 : u_i \sim \text{Beta}(\alpha, \beta)$, the UMP combiner is given by:

$$w \sum_i \log u_i - (1-w) \sum_i \log(1-u_i) \qquad w := \frac{1-\alpha}{\beta-\alpha}$$

  We encounter this situation of $u_i \sim$ Beta in the case where our underlying test is on a Poisson distribution:

$$H_0 : \lambda = \lambda_0 \qquad H_1 : \lambda > \lambda_0$$

- In the case that $H_1 : u_i \sim \Gamma_{[0,1]}(\alpha, \beta)$, the UMP combiner is given by:

$$w \sum_i \log u_i + (1-w) \sum_i u_i \qquad w := \frac{1-\alpha}{1+\beta-\alpha}$$

  Here $\Gamma_{[0,1]}$ denotes the gamma distribution truncated to the interval $[0,1]$.

Name	**Definition**
Fisher’s method	$S_F = \sum_i \log u_i$
Pearson’s method	$S_P = -\sum_i \log(1-u_i)$
George’s method	$S_G = S_F + S_P = \sum_i \log \frac{u_i}{1-u_i}$
Edgington’s method	$S_E = \sum_i u_i$
Stouffer’s method	$S_S = \sum_i \Phi^{-1}(u_i)$
Tippet’s method	$S_T = \min\{u_1, ..., u_n\}$

Table 2.1: Simple combiners

The same paper gives a survey of simple combiner proposals encountered in the literature, and both of the above are weighted combinations of these methods. We present them in Table 2.1 (where Φ is the CDF of $\mathcal{N}(0,1)$).

Sparse combiners

We now turn our attention to the motivating examples from high-throughput data analytics. Suppose we have a network in which some nodes are "anomalous" (satisfy the node-level hypothesis H_1) and each edge is anomalous (satisfies the edge-level hypothesis H_1) if an only if both its endpoints are. In this case, the edge values of each anomalous node represent a *sparse signal* for the alternative hypothesis - if the probability of a node being anomalous is $\pi = 0.1$, we can expect only 10% of the edge p-values at a given anomalous node to be non-uniform. To account for this sparsity, we follow the approach of [Hea22] and consider a **sparse mixture density** hypothesis:

$$H_1 : u_1, ..., u_n \sim \prod_i (1-\epsilon_n)\mathbf{1}_{[0,1]}(u_i) + \epsilon_n f_{1,n}(u_i) \tag{2.5}$$

Here $\epsilon_n \in [0,1]$ is the sparsity parameter, and $f_{1,n}$ an unknown alternative distribution. We impose a reasonableness condition on $f_{1,n}$ that it should be *stochastically smaller* than $\mathcal{U}[0,1]$:

Definition (Stochastic ordering)**.** We say A is **stochastically smaller** than B, denoted $A \preccurlyeq B$ if, for all x:

$$\mathbf{P}[A > x] \leqslant \mathbf{P}[B > x]$$

Given an observation of n independent p-values $u_1, ..., u_n$, we note that those with the highest probability of representing draws from the alternative distribution are the smallest (corresponding to the most extreme t_i observations under H_0) - conversely, the draws from the alternative distribution should be stochastically smaller than uniform. From this, it is natural to define **primitive statistics** on the lowest k values of u, inspired by S_F, S_P, S_E respectively:

$$s^F_{n,k} = -\sum_{i=1}^k \log u_i \qquad s^P_{n,k} = -\sum_{i=1}^k \log(1-u_i) \qquad s^E_{n,k} = \sum_{i=1}^k u_i$$

Name	**Definition**
Partial products	$\mathrm{PP}_n = \frac{\min_{1\leqslant k\leqslant n}\{\mathbf{E}(s^F_{n,k})-S^F_{nk}\}}{\sqrt{\mathbf{V}(s^F_{n,k})}}$
Complementary products	$\mathrm{PCP}_n = \frac{\min_{1\leqslant k\leqslant n}\{S^P_{nk}-\mathbf{E}(s^P_{n,k})\}}{\sqrt{\mathbf{V}(s^P_{n,k})}}$
Partial sums	$\mathrm{PS}_n = \frac{\min_{1\leqslant k\leqslant n}\{S^E_{nk}-\mathbf{E}(s^E_{n,k})\}}{\sqrt{\mathbf{V}(s^E_{n,k})}}$

Table 2.2: Normalised sparse combiners

Standardising these and optimising over k (since this is unknown) yields three statistics, presented in Table 2.2. Note, however, that these methods are not designed for the network setting, but rather where we observe *entirely independent* sequences of p-values for each node. In a network, we have external information: returning to our previous example, the weight that we place on each component of our signal (observed edge p-value) depends on both the value itself *and* the probability that the other endpoint is anomalous; we cannot therefore analyse the sparse signal at each node entirely independently if we seek an optimal combiner, as this does not allow us to take advantage of any other information available to us about its neighbourhood.

2.0.3 Statistical Graph Models

With a background in test statistics established, we turn our attention to the statistical modelling of network structured data. In particular:

- We consider an undirected graph $G = (X, E)$, where X is the *node set* of objects to be studied, and E is an *edge set* of random variables attached to each pair (i, j). Since we restrict to undirected graphs, this obeys the symmetry property $E_{i,j} = E_{j,i}$ for all $i, j \in X$.
- Equivalently, we consider a symmetric random matrix $\mathsf{A} : \Omega \to \mathbb{R}^{|X|\times|X|}$ corresponding to the weighted adjacency matrix of this graph.
- Given a single observation $\mathbf{A}$ of this random network, we are interested in learning some property of the nodes from the pairwise observations $\mathbf{A}_{ij}$

Much modern study of this problem in machine learning has focused on learning complex feature vectors associated with each node. This motivates consideration of the foundational **Latent Position Model (LPM)** - this was originally proposed in [HRH02] and interested readers can see [KRFR23] for a survey of the state-of-the-art.

Definition (Latent Position Model)**.** Let $\mathbf{A}$ be the adjacency matrix of an undirected, unweighted graph, and let $\mathbf{x}_i \in \mathbb{R}^d$ be a *position vector* associated to each node. This graph $(\mathbf{A}, \mathbf{X})$ is distributed according to the **latent position model** if, $\mathbf{x}_1, ..., \mathbf{x}_n \overset{iid}{\sim} F$ and, for some function κ:

$$\mathbf{A}_{ij} \mid \mathbf{x}_i, \mathbf{x}_j \overset{ind.}{\sim} \text{Bernoulli}(\kappa(\mathbf{x}_i, \mathbf{x}_j))$$

This is an exceptionally general model to derive properties of nodes from a network topology, but suffers from limited applicability to the hypothesis testing domain; testing a fixed hypothesis on a node corresponds to learning a very specific property: whether it is a member of the alternative or null set in the 2-community disjoint partition of the vertex set. Accordingly, whilst fitting an LPM to data might produce feature vectors which can be used in a variety of downstream tests, they introduce an assumed level of dimensionality which is at best tangential to the question of p-value combination. In the setting we are interested in, the connectivity distributions are decided exclusively by truth of the hypotheses on the endpoints. This motivates the following model with which we build p-value networks, as described in [GJB+23]:

Definition (Weighted stochastic block model)**.** Let $\mathbf{A}$ be the adjacency matrix of an n-node undirected graph, and let $z_i \in [k]$ be a *community label* associated to each node. This graph $(\mathbf{A}, \mathbf{z})$ is distributed according to the **weighted stochastic block model** if $z_1, ..., z_n \overset{iid}{\sim} F$ and there exists a symmetric family of distributions $\{H(z_1, z_2) \mid z_1, z_2 \in [k]\}$ such that, for all i, j:

$$\mathbf{A}_{ij} \mid z_i, z_j \overset{ind.}{\sim} H(\mathbf{z}_i, \mathbf{z}_j)$$

We say that $(\mathbf{A}, \mathbf{z}) \sim \text{WSBM}(n, k, H, F)$

If the set of community labels were arbitrary, WSBMs could be seen as a direct generalisation of LPMs, but the conventions these two definitions introduce for the heterogeneity amongst nodes has a nuanced difference:

- **Position models** assign vector-valued positions to each node which are used to decide the edge distribution.
- **Block models** assign labels from a set of finitely-many *communities*, and nodes are identically distributed within communities.

The block model in the unweighted case (corresponding to the LPM) has a long history and is exceptionally well-studied, especially in the $k = 2$ case [Abb23]:

Definition (Stochastic Block Model)**.** The graph $(\mathbf{A}, \mathbf{z})$ is distributed according to the **stochastic block model** if $z_1, ..., z_n \overset{iid}{\sim} F$ and there exists a function $\kappa : [k] \times [k] \to [0, 1]$ symmetric in its arguments such that:

$$\mathbf{A}_{ij} \mid z_i, z_j \overset{ind.}{\sim} \text{Bernoulli}(\kappa(z_i, z_j))$$

We say that $(\mathbf{A}, \mathbf{z}) \sim \text{SBM}(n, k, \kappa, F)$

Finally, we may combine previous definitions and these model frameworks to yield a formal notion of p-value networks:

Definition (p-value weighted SBM)**.** Let $(\mathbf{A}, \mathbf{z}) \sim \text{WSBM}(n, k, H, F)$. We say it is a **p-value weighted SBM** if:

- $k = 2$, hence $\mathcal{Z} = \{0, 1\}$, where zero denotes a *regular* node and one an *anomalous* node
- Each $H(b, b') \in \{\mathcal{U}[0, 1], p\}$, where p is a continuous distribution with range $[0, 1]$, stochastically smaller than uniform and with monotone non-increasing density on this range.

We consider two particular classes:

- The network is a **symmetric p-value weighted SBM** if:

$$H(b,b') = \begin{cases} p & b = b' \\ \mathcal{U}[0,1] & b \neq b' \end{cases}$$

  Here, we write $(\mathbf{A},\mathbf{z}) \sim \mathrm{SPVN}(n,p,\pi)$, where $\pi = F(H_1)$.

- It is **asymmetric** if

$$H(b,b') = \begin{cases} p & b = b' = 1 \\ \mathcal{U}[0,1] & \text{otherwise} \end{cases}$$

  In this case, we write $(\mathbf{A},\mathbf{z}) \sim \mathrm{APVN}(n,p,\pi)$

Note that the symmetricity of the PVN refers only to whether intra-community edges are distributed the same amongst the null and alternative communities, the underlying adjacency matrix is always symmetric.

2.1 Metrics of community separation

Both p-values and estimated partitions can be seen as one-dimensional embeddings for nodes, onto $[0,1]$ and $\{0,1\}$ respectively, and in this section we will consider the more general setting of embeddings into $\mathbb{R}$. Power-level analysis of a (one-tailed/likelihood ratio) hypothesis test then corresponds to one method of analysing the separation of embedded distributions under H_0 or H_1; in particular, it asks for the integral (β) of the embedded alternative density over the range which contains the (without loss of generality) highest α-quantile of the embedded null. To condense the information of this function β mapping α to a false negative rate, we consider a metric from the theory of estimation:

Definition (Bayes risk)**.** Given a particular prior π on parameters $\theta \in \Theta$ and loss function $\ell : \Theta \times \Theta \to \mathbb{R}$, the **Bayes risk** for estimators of θ is given by:

$$R^*_\pi = \inf_{\hat{\theta}} \mathbb{E}_{\theta\sim\pi} \ell(\theta,\hat{\theta})$$

Where the infimum is taken over all possible estimators $\hat{\theta}$. An estimator achieving this infimum is called a **Bayes solution**.

Following [AFL+17], we may specialise this to the case of one-tailed tests by considering $\ell(\theta,\hat{\theta}) = 1-\delta_{\theta\hat{\theta}}$. For a particular test statistic, the possible estimators are exactly the possible p-value thresholds α, hence:

$$\begin{aligned} R^*_\pi = \inf_{\alpha\in(0,1)} 1 - \mathbb{E}_{\theta\sim\pi}\delta_{\theta\hat{\theta}} &= \inf_{\alpha\in(0,1)} 1 - \mathbf{P}_\pi(\theta = \hat{\theta}) \\ &= \inf_{\alpha\in(0,1)} \pi_0\alpha + (1-\pi_0)\beta(\alpha) \end{aligned}$$

Note that when the test statistic is UMP, as the likelihood ratio, we consider the infimum ranging over α without loss of generality since this test must achieve the maximal TP rate amongst all estimators at a given level for the underlying distributions. This gives an informative metric for the quality of community separation in the 1d embedding, and a Bayes solution α^* gives an optimal achievable threshold to minimize the expected misclassification rate.

When testing nodes in a graph, each test statistic will be a one-dimensional embedding of an n-element row in an $n \times n$ matrix, so this more accurately defines a *family* of estimators θ_n for an inputs of a given size. We know by the law of large numbers that, where $\mathbb{E}p \neq \frac{1}{2}$ and the graph sparsity proportion are both constant in the size of the graph, community separation will become arbitrarily easy as $n \to \infty$ by a trivial one-tailed test on means - indeed, $\beta_n(\alpha_n) \to 0$ for any sequence $(\alpha_n \in (0,1))$. It is thus informative to consider the *efficiency rate* of a test in the limit, which both allows us to reason about when exact recovery is asymptotically possible, as well as the speed of convergence. An influential result from Chernoff [Che52] characterises this rate in terms of the underlying distributions.

Definition (Chernoff information)**.** Let f_0 and f_1 be density functions over $\mathbb{R}^d$. The **Chernoff information** quantifies their difference by:

$$C(f_0,f_1) := \sup_{t\in(0,1)} \left[-\log \int_{\mathbb{R}^d} f_0^t(\mathbf{x}) f_1^{1-t}(\mathbf{x})\, d\mathbf{x} \right] =: \sup_{t\in(0,1)} C_t(f_0,f_1)$$

For a fixed t, C_t is known as the **Chernoff divergence**.

Theorem 2.1.1 (Chernoff, 1952)**.** Let f_0, f_1 be the density functions of test statistics under H_0, H_1 respectively and let π be a prior on community assignments. Then, under the likelihood ratio test, we have:

$$\lim_{n\to\infty} \frac{1}{n} \inf_{\alpha_n \in (0,1)} \log(\pi_0 \alpha_n + (1-\pi_0)\beta_n(\alpha_n)) = -C(f_0, f_1)$$

In other words, the Chernoff information asymptotically characterises the optimal exponential decay rate of misclassification error achievable by testing this statistic.

This allows for an information-theoretic perspective on test statistics by considering if consistent recovery of communities is asymptotically possible under some optimal estimator. Much literature has been devoted to studying this problem on the unweighted SBM, and we follow [Abb23] in defining four notions of asymptotic estimator consistency.

Definition (Agreement)**.** Let x, y be community assig nments in $\{0,1\}^n$. Then the **agreement** between them is the maximum probability of overlap between x and either y or its complement:

$$A(x,y) := \max_{\pi \in S_2} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(x_i = \pi(y_i)) \qquad \text{(Agreement)}$$

$$\tilde{A}(x,y) := \max_{\pi \in S_2} \frac{1}{2} \sum_{k \in \{0,1\}} \frac{\sum_{i \leqslant n} \mathbf{1}(x_i = k)\mathbf{1}(x_i = \pi(y_i))}{\sum_{i \leqslant n} \mathbf{1}(x_i = k)} \qquad \text{(Normalised Agreement)}$$

Notation. We use the standard little- and big-oh notation to denote asymptotics w.r.t. number of nodes in the network. $f = \Omega(g)$ denotes that $g = O(f)$ and $f = \omega(g) \equiv g = o(f)$. In particular, $o(1)$ denotes a vanishing quantity and $\Omega(1)$ denotes a non-vanishing quantity.

Definition (Recovery requirements)**.** Letting $(\mathbf{A}, \mathbf{x}) \sim \text{PVN}(n, p, \pi)$, we have the following notions of asymptotic community recovery. These problems are "solved" for a given PVN model if there exists an algorithm accepting $\mathbf{A}$ and outputting $\hat{\mathbf{x}}$ such that

(Strong consistency/Exact recovery) $\qquad \mathbf{P}[A(\mathbf{x}, \hat{\mathbf{x}}) = 1] = 1 - o(1)$

(Weak consistency/Almost exact recovery) $\qquad \mathbf{P}[A(\mathbf{x}, \hat{\mathbf{x}}) = 1 - o(1)] = 1 - o(1)$

(Partial recovery) $\qquad \mathbf{P}[\tilde{A}(\mathbf{x}, \hat{\mathbf{x}}) \geqslant \alpha] = 1 - o(1)$

(Weak recovery) $\qquad \mathbf{P}[\tilde{A}(\mathbf{x}, \hat{\mathbf{x}}) \geqslant 1/2 + \Omega(1)] = 1 - o(1)$

We make an immediate connection between Chernoff information and recovery in a simplified case:

Proposition 2.1.1. Consider the sparse signal case described previously on a square matrix: that is to say we have an $n \times n$ matrix where each row is either anomalous or null, and consists of n random variables i.i.d. according to either the null (uniform) or alternative (mixture) distributions, representing n independent observations of an n-dimensional sparse signal. The difference between this setting and the APVN is in the independence of rows. Then exact recovery is solvable if and only if the Chernoff information between these distributions is nonzero. Furthermore, Chernoff information between embedded test statistics is maximised by a UMP combiner.

Proof. This is a trivial application of Bernoulli's inequality and Theorem 2.1.1. Denote by E_n the optimal estimator error, and then

$$\mathbf{P}[A(\mathbf{x}, \hat{\mathbf{x}}) = 1] = (1 - E_n)^n = \left(1 - e^{-nC(f_0,f_1)}\right)^n \leqslant 1 - ne^{-nC(f_0,f_1)}$$

This is clearly $1 - o(1)$ when $C(f_0, f_1) > 0$. Conversely, if exact recovery is solvable, then separation must become arbitrarily easy as $n \to \infty$, so $nE_n \to 0$, hence $n = O(E_n^{-1})$ and $\frac{1}{n} \log E_n \not\to 0$. That UMP statistics maximise CI is immediate since E_n must in turn be minimized by the UMP combiner. □

In this way, we can build a connection between the metrics one uses for p-value combination and point estimation.

2.2 Building P-Value Networks: the SIMPLE test

A natural way in which SPVN networks are induced from arbitrary 2-community block models would be to test at each edge whether this edge is *inter-community* or *intra-community*. [FFHL21] proposes a method based on spectral theory - we will provide a more thorough analysis of this algorithm in §3.5.1 but briefly introduce it now as background.

Suppose we have an undirected adjacency matrix $\mathbf{A}$, and denote by $\mathbf{U}\Lambda\mathbf{U}^\top$ the eigenvalue decomposition (this always exists since $\mathbf{A}$ real symmetric). For each node i, we denote by $\hat{\mathbf{V}}^d(i)$ the vector formed by the i-th components of the eigenvectors corresponding to the top d eigenvalues in absolute value, each scaled by the square root of the corresponding absolute eigenvalue. In other words, it is the i-th row of $\mathbf{U}^d\left|\Lambda^d\right|^{\frac{1}{2}}$. Let K be the number of communities in the model and we define:

$$T_{ij} := \left[\hat{\mathbf{V}}^K(i) - \hat{\mathbf{V}}^K(j)\right]^\top \Sigma^{-1} \left[\hat{\mathbf{V}}^K(i) - \hat{\mathbf{V}}^K(j)\right]$$

The asymptotic behaviour of this statistic on the unweighted SBM allows us to construct p-values:

Theorem 2.2.1 ([FFHL21])**.** Under certain regularity assumptions on a K-community unweighted SBM, the SIMPLE test statistic comparing nodes i,j with the same community assignment will follow:

$$T_{ij} \xrightarrow{\mathrm{d}} \chi^2_K$$

Similarly, if i and j do not share community assignment, then for some $\mu > 0$:

$$T_{ij} \xrightarrow{\mathrm{d}} \chi^2_K(\mu)$$

Where $\chi^2_K(\mu)$ denotes the *non-central chi-squared distribution* at μ.

Chapter 3

Combining p-values in networks

We follow the approach introduced in [HRD18] to construct a likelihood ratio from the observed p-values. Let H_0^i be the hypothesis that a given node is in the regular set and H_1^i be the hypothesis that the node is anomalous. Connections are either regular and distributed according to q_0, else they are anomalous and distributed according to q_1, determined by whether we are working in the symmetric or asymmetric model. We observe a matrix A of p-values testing the hypothesis $H_1^{i,j}$ that an edge is anomalous against $H_0^{i,j}$ that an edge is normal. For each configuration, we can characterise these hypotheses:

- **(Asymmetric)** $H_1^{i,j}$ is that both nodes are anomalous, $H_0^{i,j}$ that either is normal
- **(Symmetric)** $H_1^{i,j}$ is that the nodes belong to the same class, $H_0^{i,j}$ that they do not

Our objective is to combine these network-structured p-values for $H^{i,j}$ into vector-structured p-values for H^i. In particular, assuming the p-value for each $H^{i,j}$ is obtained from a UMP test, we are interested in obtaining a UMP statistic for H^i. Applying the Neyman-Pearson lemma, we have a direct route to this goal by constructing a likelihood ratio from density functions.

Remark. A hypothesis H_0^i can be seen as a Bernoulli(0,1) random variable, where $P(H_0^i 0 := P(H_0^i = 1)$ is the probability that the hypothesis holds. We have for any given pair of hypotheses $H_1 = 1 - H_0$, hence $P(H_1) = 1 - P(H_0)$. We may also write $H \sim \text{Bernoulli}(0, 1)$ to denote the random variable corresponding to which of the hypotheses H_0 or H_1 holds.

The exact approach to constructing a UMP test statistic requires the computation of the likelihood ratio for each node:

$$\frac{p_{\text{joint}}(A \mid H_0^i)}{p_{\text{joint}}(A \mid H_1^i)}$$

But these marginal distributions are intractable, even with perfect knowledge of the alternative density: the joint density of rows A_i and A_j depend not only on H^i, H^j, but also on H^k for every k in $N(i) \cap N(j)$ (Where $N(i)$ denotes the neighbourhood of node i). In the case that the PVN is complete, this yields a complexity exponential in the size of the nodes. In particular:

$$\begin{aligned}
\mathbf{P}(H_b^i \mid A) &\propto p_{\text{joint}}(A \mid H_b^i)\pi_b && (3.1)\\
&= \sum_{\mathbf{x} \in \{0,1\}^n,\, \mathbf{x}_i = b} p_{\text{joint}}(A \mid \mathbf{x})\pi(\mathbf{x}) && (3.2)\\
&= \sum_{\mathbf{x} \in \{0,1\}^n,\, \mathbf{x}_i = b} \left[\prod_{(j,k)} p(A_{jk} \mid \mathbf{x}_j, \mathbf{x}_k) \prod_j \pi(\mathbf{x}_j) \right] && (3.3)
\end{aligned}$$

This product decomposition is a consequence of pairwise independence of entries in A when conditioned on the community membership of endpoints.

Our best route forward, therefore, is to consider approximations to these intractable marginals. A trivial method would be to separate the network row-wise and apply our sparse combination statistics (2.2), but we will focus on direct approaches which actually seek to find this posterior marginal distribution and hence likelihood ratio.

3.1 Tools from statistical physics

With this form of posterior likelihood functions established, we may now introduce a connection to statistical physics, from which much of our literature will be drawn. The stochastic block model formulation is a favourite of statisticians and computer scientists, but it is closely linked to a model of pairwise interactions between particles that has been well-studied for decades among physicists. We follow [Mac03] for definitions.

Definition (Ising model)**. Ising models** seek to describe the interactions of *magnetic spin systems* in a lattice structure. They consist of an array of spins $\mathbf{x} \in \{-1, +1\}^n$, a symmetric affinity matrix $\mathbf{J} \in \mathbb{R}^{n \times n}$ and an external field vector $\mathbf{h} \in \mathbb{R}^n$. The total potential energy of the system is described by the **Hamiltonian**:

$$\mathcal{H}_I(\mathbf{x}; \mathbf{J}, \mathbf{h}) = -\sum_{i \leqslant j} \mathbf{J}_{ij}\mathbf{x}_i\mathbf{x}_j - \sum_i \mathbf{h}_i\mathbf{x}_i$$

Where $\mathbf{J}_{ij}$ is positive, we call the interaction **ferromagnetic**, otherwise it is **antiferromagnetic**.

Much literature within statistical physics has been devoted to analysing this model (see [BRU67] for a historical survey), and the Hamiltonian's similarity to the log-likelihood of a community assignment in the (W)SBM is immediate. Indeed, a generalisation of this model subsumes the weighted stochastic block model:

Notation. We denote by $[k]$ the set $\{1, 2, ..., k\}$.

Definition (Generalised Potts Model)**.** A q-label **generalised Potts model** has spins drawn instead from $\mathbf{x} \in [q]^n$ and a matrix of affinity functions $J_{ij} : [q] \times [q] \to \mathbb{R}$. The Hamiltonian of this system is then given by:

$$\mathcal{H}_P(\mathbf{x}; J, \mathbf{h}) = \sum_{i \leqslant j} J_{ij}(\mathbf{x}_i, \mathbf{x}_j) - \sum_i \mathbf{h}_i\mathbf{x}_i$$

For $q = 2$, we recover the Ising model where $J_{ij}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{J}_{ij}\mathbf{x}_i\mathbf{x}_j$.

Any weighted SBM trivially has log-likelihood proportional to the Hamiltonian of some generalised Potts model, and this is the comparison usually drawn in literature. The connection between log-likelihood and Hamiltonian is formalised by the Boltzmann distribution:

Definition (Boltzmann distribution)**.** For a Hamiltonian system with energy H, the **Boltzmann distribution** on states is given by:

$$\mu(\mathbf{x} \mid \mathbf{A}; H) := \frac{1}{Z} e^{-\beta H(\mathbf{x};\mathbf{A})}$$

Where Z is the usual normalising constant, and β is a free parameter. In the traditional mechanics setting, $\beta = (kT)^{-1}$ where k is the physical *Boltzmann constant* and T is the temperature of the system, but we will normally ignore k. This general form is sometimes known in statistics literature as the **Gibbs distribution**.

We thus recover the SBM posterior distribution as the Boltzmann distribution induced by the corresponding Potts model at unit temperature (hence $\beta = 1$) [DKMZ11a]. This generalisation is not, however, entirely necessary for the p-value network, since the uniform densities at the null distribution admit a simpler form of log-likelihood:

Claim. Any observation **(A,x)** $\sim \mathrm{SPVN}(n, \pi, p_1)$ is represented by an Ising model. Let $\mathbf{J}_{ij} = -\frac{1}{2} \log p_1(A_{ij})$ and $h = -\frac{1}{2} \log \frac{1-\pi}{\pi}$. Now, letting $\mathbf{x}_i = -1$ if node i is anomalous and $\mathbf{x}_i = 1$ otherwise:

$$\begin{aligned}
\ell(\mathbf{x}, \mathbf{A}) &= -\sum_{i \leqslant j} \log p_1(A_{ij})\delta_{\mathbf{x}_i\mathbf{x}_j} - \sum_i \log \pi_{\mathbf{x}_i} \\
&= -\frac{1}{2}\sum_{i \leqslant j} \log p_1(A_{ij})(\mathbf{x}_i\mathbf{x}_j + 1) - \frac{1}{2}\left[\sum_i (\mathbf{x}_i + 1)\log \pi - (\mathbf{x}_i - 1)\log(1 - \pi)\right] \\
&= -\sum_{i \leqslant j} \mathbf{J}_{ij}\mathbf{x}_i\mathbf{x}_j - \sum_i h\mathbf{x}_i + \frac{1}{2}\left[\sum_{i \leqslant j} \log p_1(A_{ij}) + \sum_i \log \pi(1 - \pi)\right] \\
&= \mathcal{H}_I(\mathbf{x}; \mathbf{J}, \mathbf{h}) + \text{const.}
\end{aligned}$$

Thus:

$$p(\mathbf{x} \mid \mathbf{A}) = \frac{e^{\ell(\mathbf{x},\mathbf{A})}}{\sum_{\mathbf{x}} e^{\ell(\mathbf{x},\mathbf{A})}} = \mu(\mathbf{x} \mid \mathbf{A}; \mathcal{H}_I)$$

The asymmetric PVN cannot be expressed in the form of a traditional Ising model with labels in $\{1, -1\}$ as this provides no way to differentiate between null-null and alternative-alternative connections, but a similar argument shows that it does correspond to an Ising model with labels in $\{0, 1\}$. Indeed:

Proposition 3.1.1. There is an equivalence between unit-temperature Boltzmann distributions on Ising models with $\mathbf{J}_{ij} > 0$ and posterior distributions on community assignments for SPVN models with a given alternative distribution. A similar equivalence holds between APVN models and Boltzmann distributions of $\{0, 1\}$-labelled Ising models.

Note that "equivalence" here does not mean bijection - rather that it is possible to construct an observed adjacency matrix $\mathbf{A} \sim$ PVN for which the posterior distribution on community assignments coincides with the Ising model Boltzmann distribution for any parameters $\mathbf{J}$,$\mathbf{h}$ and vice versa.

Remark. Since adding a constant value to the Hamiltonian does not affect the induced Boltzmann distribution, it is often formally convenient to add an $M \log N$ term (where M denotes the number of edges in the network) to the log-likelihood function so the system energy is *extensive in* (i.e. proportional to) N [DKMZ11a].

An important result concerning the intractability of our problem can now be brought over from the physics context:

Theorem 3.1.1 ([Luc14])**.** Maximum Likelihood Estimation in the Ising model is NP-Hard.

Corollary 1. *Maximum Likelihood Estimation in p-value networks is NP-Hard.*

Whilst this does not immediately imply that the formulation of a provably UMP test is also NP-Hard, we conjecture that it is - hence unless P = NP, optimal p-value combination in a general network of our model is intractable, so we may focus only on approximation.

Remark. For the remainder of this paper we will work with the PVN as our primary model and phrase our analysis in these terms. The reader should note, however, that much of our work applies directly to *any* two-community weighted SBM. This is novel in the APVN case since comparatively little literature has been devoted to weighted *heterogenous* SBMs, where edge distributions depend not only on whether two edges are in the same community, but also which communities they are in. This lack of understanding can be partially explained by their inherent computational complexity and the comparative lack of immediate parallels in well-studied areas of statistical physics.

3.1.1 The genie-aided case

As a warm-up, we consider the "genie-aided" hypothesis test - that is to say, testing $p(\mathbf{z}_i \mid \mathbf{A}, \mathbf{z}_{-i})$, where $\mathbf{z}_{-i}$ denotes all but the i-th index of $\mathbf{z}$:

$$\begin{aligned} P(\mathbf{z}_i = b \mid \mathbf{A}, \mathbf{z}_{-i}) &\propto \pi_b \prod_{j \leqslant k} p(\mathbf{A}_{jk} \mid \mathbf{z}_k, \mathbf{z}_j) \\ &\propto \pi_b \prod_{j} p(\mathbf{A}_{ij} \mid \mathbf{z}_i = b, \mathbf{z}_j) \end{aligned}$$

So we may completely characterise the posterior Bernoulli distribution on $\mathbf{z}_i$ by

$$P(\mathbf{z}_i = 1 \mid \mathbf{A}, \mathbf{z}_{-i}) = \frac{\pi \prod_j p(\mathbf{A}_{ij} \mid \mathbf{z}_i = b, \mathbf{z}_j)}{\sum_{b \in \{0,1\}} \pi_b \prod_j p(\mathbf{A}_{ij} \mid \mathbf{z}_i = b, \mathbf{z}_j)}$$

Alternatively, for a small subset of unknown indices I:

$$P(\mathbf{z}_I \mid \mathbf{A}, \mathbf{z}_{-I}) = \frac{\prod_{i \in I} \pi_{\mathbf{z}_i} \prod_{i \in I, j} p(\mathbf{A}_{ij} \mid \mathbf{z}_i, \mathbf{z}_j)}{\sum_{\mathbf{z}_I} \prod_{i \in I} \pi_{\mathbf{z}_i} \prod_{i \in I, j} p(\mathbf{A}_{ij} \mid \mathbf{z}_i, \mathbf{z}_j)} \tag{3.4}$$

3.2 Markov Chain Monte Carlo

This combination of intractable posterior marginal but easy-to-compute conditional immediately suggests an estimation routine based on **Gibbs sampling**:

1. Let $\mathbf{z}^{(0)} \sim \rho$ be a binary vector drawn from some prior distribution on the community assignments.
2. At each time step, choose some set of coordinates I. For each $i \in I$, sample:

$$\mathbf{z}_i^{(t+1)} \sim p(\cdot \mid \mathbf{A}, \mathbf{z}_{-i}^{(t)})$$

   Where $\mathbf{z}_{-i}^{(t)}$ denotes the time-t sample with only the i-th element excluded.

3. Continue for N iterations.

The simplest case is to take $I = [|\mathbf{z}|]$, but a stochastic variant which subsamples the vector may be more appropriate in larger networks. Despite its simplicity, the Gibbs sampler is often appealing due to two characteristic properties [GG84]:

Theorem 3.2.1 (Gibbs sampler)**.** Letting $\mathbf{z}^0, \mathbf{z}^{(1)}, ...$ be a Markov chain generated by Gibbs sampling $p(\mathbf{z} \mid \mathbf{A})$. Then assuming that each coordinate appears infinitely-often in I:

- **(Convergence)** As $t \to \infty$,

$$\left(\mathbf{z}_1^{(t)}, ..., \mathbf{z}_n^{(t)}\right) \xrightarrow{d} \mathbf{z} \sim p(\cdot \mid \mathbf{A})$$

- **(Ergodicity)** For any measurable function f, as $T \to \infty$,

$$\frac{1}{T} \sum_{t \leqslant T} f(\mathbf{z}^{(t)}) \xrightarrow{\text{a.s.}} \mathbb{E}_{p(\cdot|\mathbf{A})}[f(\mathbf{z})]$$

We note that $p(\mathbf{z} \mid \mathbf{A})$ characterises a multivariate Bernoulli distribution where $p(\mathbf{z}_i = 1 \mid \mathbf{A}) = \mathbb{E}_{p(\cdot|\mathbf{A})}[\mathbf{z}_i]$; we may thus use the ergodic property to approximate the MLE of marginal anomaly scores as the sample mean of a sufficiently long Gibbs chain, and our work in the previous section provides a characterisation of $p(\mathbf{z}_i \mid \mathbf{A}, \mathbf{z}_{-i}^{(t)})$.

These results do not, however, give any guarantees of the rate of convergence, and indeed the only bound that exists in the literature for the general case is that convergence occurs in time *at worst exponential* in the number of nodes in the graph. We can improve the convergence rate somewhat by grouping m nodes together and sampling from their joint distributions $p(\mathbf{z}_{i:i+m} \mid \mathbf{A}, \mathbf{z}_{-[i,i+m]})$ since this will have a tractable 2^m complexity, yielding a **blocked** Gibbs sampler, but at a corresponding increased computational cost per-step.

Algorithm 1: Gibbs Marginalization

```
Input: $\rho$ a prior on community assignments, $\mathbf{A}$ an observation, $\{\pi, p_1\}$ the PVN parameters, $k$
       a block size, $\omega$ a warm-up time, and $T$ the maximum iterations
Output: $\hat{\mathbf{z}}$ estimated anomaly probabilities
Initialize $\mathbf{z}^{(0)} \sim \rho$;
for $t \leqslant T$ do
    Divide the indices of $\mathbf{z}$ into $k$-sized blocks and, for each block, find $\rho^{(t+1)}$ the probability
      distribution on the block given the assignments $\mathbf{z}^{(t)}$ according to equation 3.4;
    For each block of indices $I$, sample $\mathbf{z}_I^{(t+1)} \sim \rho_I^{(t+1)}$;
    if $t \geqslant \omega$ then
        Initialize $\hat{\mathbf{z}} \leftarrow \mathbf{z}^{(t+1)}$ if $t = \omega$, otherwise update $\hat{\mathbf{z}}$ as the sample mean of each $\mathbf{z}^{(t)}$ for
          $t > \omega$;
return $\hat{\mathbf{z}}$
```

Alternative stopping criteria may also be used, such as convergence of $\hat{\mathbf{z}}$. When computing the mean, we will often let the chain run for some number (in our later simulations, 10) of iterations between each update in order to reduce the correlation between successive samples.

Remark (Extension to multigraphs)**.** We note that in this section and indeed all our iterative approximation methods, we work only with likelihoods of particular edge weights rather than the edge weights themselves. This implies a trivial extension to (independent) multigraphs where we simply consider the joint likelihood of every observed weight instead of the scalar weight on each individual edge.

Definition (Mixing time)**.** For a finite Markov chain with transition matrix P, state space Ω, and stationary distribution ρ, we define the **mixing time**:

$$\tau_\epsilon(P) := \min\left\{t \mid \max_{x\in\Omega} \left\|P^t(x,\cdot) - \pi\right\| \leqslant \epsilon\right\}$$

Here, $\|\cdot\|$ denotes total variation distance. In other words, τ_ϵ is the number of steps required for the chain distribution to be within ϵ of the stationary distribution, regardless of starting point.

3.2.1 The Swendsen–Wang Algorithm

The Gibbs sampler is a special case of **Markov Chain Monte Carlo (MCMC)** algorithms, which seek to perform inference of a distributional statistic by repeatedly drawing samples from a Markov chain and computing a sample estimate via the ergodic property. In particular, it is a *local* algorithm which peturbs variables one at a time, agnostic of the information we have about their dependence structure. We turn now to a seminal proposal of a *non-local* alternative, originally developed to sample from lattice Ising models [SW87]. The key principle of the **Swendsen-Wang** method is to introduce new latent variables $\mathbf{d} \in \{0,1\}^{n\times n}$ which represents the adjacency matrix of a second undirected, unweighted graph, known as the *bonding network*. The distributions $p(\mathbf{z} \mid \mathbf{d}, \mathbf{A})$, $p(\mathbf{d} \mid \mathbf{z}, \mathbf{A})$ should then be easy to sample from and we proceed by Gibbs sampling (alternating $\mathbf{z}$, $\mathbf{d}$) on this extended model. This allows us to make local changes to the bonding network which propagate to multi-node changes on the community assignments, hopefully accelerating the mixing process of the Markov chain. We derive a generalised form for our setting. First, recall that we may decompose the posterior distribution on $\mathbf{z}$ as a product of factors

$$P(\mathbf{z} \mid \mathbf{A}) = \frac{1}{Z}\prod_{i\leqslant j} f_{ij}(\mathbf{x}_i, \mathbf{x}_j)\prod_i \pi_{\mathbf{x}_i}$$

Where

$$f_{ij} = e^{J_{ij}\delta(\mathbf{x}_i,\mathbf{x}_j)} \qquad\qquad J_{ij} = \frac{1}{2}\log p_1(\mathbf{A}_{ij})$$

The comparison predicate $\delta : \{0,1\} \times \{0,1\} \to \{1,-1\}$ depends on whether we are in the symmetric or asymmetric case. Note that a common factor of $\prod_{i\leqslant j}\sqrt{p_1(A_{ij})}$ is cancelled in the normalising constant Z. We define the joint distribution on the bonding network:

$$P(\mathbf{z}, \mathbf{d} \mid \mathbf{A}) = \frac{1}{Z'}\prod_{i\leqslant j} g_{ij}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{d}_{ij})\prod_i \pi_{\mathbf{x}_i}$$

Where

$$g_{ij} = \begin{cases} g^+_{ij} & \log p_1(\mathbf{A}_{ij}) \geqslant 0 \\ g^-_{ij} & \log p_1(\mathbf{A}_{ij}) < 0 \end{cases}$$

$$g^+_{ij} = \begin{cases} e^{-J_{ij}} & \mathbf{d}_{ij} = 0 \\ e^{J_{ij}} - e^{-J_{ij}} & \mathbf{d}_{ij} = \delta(\mathbf{x}_i, \mathbf{x}_j) = 1 \\ 0 & \mathbf{d}_{ij} = 1, \delta(\mathbf{x}_i, \mathbf{x}_j) = -1 \end{cases} \qquad g^-_{ij} = \begin{cases} e^{J_{ij}} & \mathbf{d}_{ij} = 0 \\ e^{-J_{ij}} - e^{J_{ij}} & \mathbf{d}_{ij} = 1, \delta(\mathbf{x}_i, \mathbf{x}_j) = -1 \\ 0 & \mathbf{d}_{ij} = \delta(\mathbf{x}_i, \mathbf{x}_j) = 1 \end{cases}$$

Claim. This extension of the distribution to include latent variables is well-behaved:

- The partition functions Z and Z' coincide
- Marginalising over $\mathbf{d}$ recovers the original posterior: $P(\mathbf{z} \mid \mathbf{A}) = \sum_{\mathbf{d}} P(\mathbf{z}, \mathbf{d} \mid \mathbf{A})$

Proof. It is clear that both of these results follow immediately from the identity:

$$\sum_{\mathbf{d}}\left[\prod_{i\leqslant j} g_{ij}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{d}_{ij})\prod_i \pi_{\mathbf{x}_i}\right] = \prod_{i\leqslant j} f_{ij}(\mathbf{x}_i, \mathbf{x}_j)\prod_i \pi_{\mathbf{x}_i}$$

To see this, first note that the bias term does not depend on $\mathbf{d}$ so we factor it out. We may rewrite g as

$$g_{ij}^{\pm} = (1-\mathbf{d}_{ij})e^{\mp J_{ij}} + \mathbf{d}_{ij}\left(e^{\pm J_{ij}} - e^{-\delta(\mathbf{x}_i,\mathbf{x}_j)J_{ij}}\right)$$

Now notice that, during the summation, each $\mathbf{d}_{ij}$ will take the value 1 or 0 exactly $2^{|\mathbf{d}|-1}$ times. A standard result from combinatorics then gives that in this case we may exchange sums and products, i.e.

$$\sum_{\mathbf{d}}\prod_{i\leqslant j}(1-\mathbf{d}_{ij})e^{\mp J_{ij}} + \mathbf{d}_{ij}\left(e^{\pm J_{ij}} - e^{-\delta(\mathbf{x}_i,\mathbf{x}_j)J_{ij}}\right) = \prod_{i\leqslant j} e^{\mp J_{ij}} + \left(e^{\pm J_{ij}} - e^{-\delta(\mathbf{x}_i,\mathbf{x}_j)J_{ij}}\right)$$

Which is exactly $\prod_{i\leqslant j} f_{ij}$ as required. □

Due to the normalisation term, any constant multiple of g_{ij} (that is, depending only on J_{ij} and not $\mathbf{z}$ or $\mathbf{d}$) will yield an identical distribution. We rescale by $e^{J_{ij}}$ such that the two nonzero terms of g are a valid probability distribution:

$$\tilde{g}_{ij}^{+} := \begin{cases} 1-\beta_{ij}^{+} & \mathbf{d}_{ij}=0 \\ \beta_{ij}^{+} & \mathbf{d}_{ij}=\delta(\mathbf{x}_i,\mathbf{x}_j)=1 \\ 0 & \mathbf{d}_{ij}=1, \delta(\mathbf{x}_i,\mathbf{x}_j)=-1 \end{cases} \qquad \beta_{ij}^{+} := 1-e^{-2J_{ij}} = 1-p_1(A_{ij})^{-1}$$

$$\tilde{g}_{ij}^{-} := \begin{cases} 1-\beta_{ij}^{-} & \mathbf{d}_{ij}=0 \\ \beta_{ij}^{-} & \mathbf{d}_{ij}=1, \delta(\mathbf{x}_i,\mathbf{x}_j)=-1 \\ 0 & \mathbf{d}_{ij}=\delta(\mathbf{x}_i,\mathbf{x}_j)=1 \end{cases} \qquad \beta_{ij}^{-} := 1-e^{2J_{ij}} = 1-p_1(A_{ij})$$

The question now turns to sampling from these conditional distributions. For this, we must first introduce the **Metropolis-Hastings** algorithm and its general template for ensuring convergence of MCMC methods to the true probability distribution [THAS20]. For each node v at each iteration:

1. **(Proposal step)** Propose a value for $v^{(t+1)}$, sampled according to a distribution $q(\cdot \mid v^{t})$.

2. **(Rejection step)** Let $f(\mathbf{z})$ be any function proportional to the true density $p(\mathbf{z})$. This is especially useful in our setting, since it lets us choose the easy to compute $f(\mathbf{z} \mid \mathbf{A}) = p(\mathbf{A} \mid \mathbf{z})p(\mathbf{z})$. Let $\mathbf{z}'^{(t+1)}$ be $\mathbf{z}^{(t)}$ with node v updated to the proposal from above, and set

$$\alpha = \min\left(1, \frac{f(\mathbf{z}^{(t+1)})}{f(\mathbf{z}^{(t)})} \cdot \frac{q(v^{(t)} \mid v^{(t+1)})}{q(v^{(t+1)} \mid v^{(t)})}\right)$$

   With probability α, we set $\mathbf{z}^{(t+1)} = \mathbf{z}'^{(t+1)}$, else the proposal is rejected and $\mathbf{z}^{(t+1)} = \mathbf{z}^{(t)}$. Notice that if q is **symmetric**: $q(v' \mid v) = q(v \mid v')$, we need only consider f when computing α.

We have the following foundational theorem:

Theorem 3.2.2 ([THAS20])**.** The Metropolis-Hastings chain is both ergodic and satisfies *detailed balance*. This implies that the sampling distribution of the Metropolis-Hastings algorithm converges to the true distribution ergodically.

Notice that Gibbs sampling is a special case of the Metropolis algorithm - it is an interesting exercise to define the q, f functions used by Gibbs sampling and show that $\alpha = 1$ in all cases.

Metropolis-Hastings is a building block for more sophisticated MCMC methods by providing a template that f, q distributions can be fitted into with a reward of guaranteed convergence (but no guarantee of speed!)

- $P(\mathbf{d} \mid \mathbf{z})$ can be found trivially - note that the elements of $\mathbf{d}$ are independent conditioned on $\mathbf{z}$, so we may look at their componentwise formulae. If $\delta(\mathbf{z}_i, \mathbf{z}_j) = 1$ and $p_1(\mathbf{A}_{ij}) < 1$ (or $\delta(\mathbf{z}_i, \mathbf{z}_j) = -1$ and $p_1(\mathbf{A}_{ij}) \geqslant 1$), then $\mathbf{d}_{ij} = 0$, otherwise we set $\mathbf{d}_{ij} \sim \text{Bernoulli}(\beta_{ij})$.

- $P(\mathbf{z} \mid \mathbf{d})$ requires a little more care in the PVN setting than in the ferromagnetic Ising model. Bonded nodes place restrictions on δ:

- **(Symmetric case)** One node in each connected component is chosen and the labels of this representative set are sampled independently at random according to Bernoulli(π). The component is then traversed - when following an antiferromagnetic edge, the community assignment is flipped, and when following a ferromagnetic edge, the community assignments remain the same (note that connected components can only arise as a subgraph with this form of 2-colouring so this assignment is always possible).
- **(Asymmetric case)** Alternative nodes may not be bonded together along repelling edges but only alternative nodes may be bonded by attracting edges. To deal with this asymmetry, we make explicit a Metropolis-style proposal-rejection step. For each connected component, we traverse the nodes depth-first from an arbitrary starting point:
    1. Define $f_i(b) = ZP(\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_{i-1}, b, \mathbf{z}_{i+1}, ..., \mathbf{z}_n \mid \mathbf{d}, \mathbf{A})$ (i.e. the unnormalised probability) where $\mathbf{d}$ and $\mathbf{z}$ are each the current estimates (including any changes that have been made traversing the component). Supposing that $\mathbf{z}_i = b$ in the current estimate, flip this coordinate with probability $\min\{1, f_i(\neg b)/f_i(b)\}$. If the node is an endpoint of an attracting edge, then it must be set to anomalous since $f_i(0) = 0$.
    2. If the coordinate is flipped to anomalous, set each unvisited neighbour along attracting edges to alternative and along repelling edges to null - note that this step is technically unnecessary since $f(1) = 0$ on any neighbours, but it saves computation time and a potential division by zero.

We present the full Algorithm (2), noting its similarity to the Gibbs sampler.

```
Algorithm 2: PVN Swendsen-Wang
  Input: $\rho$ a prior on community assignments, $\mathbf{A}$ an observation, $\{\pi, p_1\}$ the PVN parameters, $\omega$
         a warm-up time, and $T$ the maximum iterations
  Output: $\hat{\mathbf{z}}$ estimated anomaly probabilities
  Initialize $\mathbf{z}^{(0)} \sim \rho$ and $\mathbf{d}^{(0)} \leftarrow 0$;
  for $t \leqslant T$ do
      foreach (unordered) pair of nodes $(i, j)$ such that $\delta(\boldsymbol{z}_i^{(t)}, \boldsymbol{z}_j^{(t)})$ do
          Sample $\mathbf{d}_{ij}^{(t+1)} \sim$ Bernoulli($\beta_{ij}$);
      foreach Connected component $I$ of the $\mathbf{d}^{(t+1)}$ graph do
          if $|I| > 1$ then
              Perform a Metropolis-style traversal as described above;
          else
              Sample a community assignment $z \sim$ Bernoulli($\pi$);
      if $t \geqslant \omega$ then
          Initialize $\hat{\mathbf{z}} \leftarrow \mathbf{z}^{(t+1)}$ if $t = \omega$, otherwise update $\hat{\mathbf{z}}$ as the sample mean of each $\mathbf{z}^{(t)}$ for
            $t > \omega$;
  return $\hat{\boldsymbol{z}}$
```

Ergodicity of this Markov chain is immediate since there is a nonzero probability of any partition transitioning to any other, and that it satisfies detailed balance in the Ising model has been well-studied (see, for example, [ES88]). Since the update dynamics of the community assignments are Metropolis style in both the symmetric and asymmetric case, we can extend the same proof to the APVN (asymmetric Ising model).

The mixing rate of this algorithm in the case of arbitrary network topologies has long been conjectured to be far superior to Gibbs sampling, and this is often backed up with empirical evidence. Recent advances in the study of *Random Cluster Networks* has led to the following result on *homogenous*, *ferromagnetic* Ising models:

Theorem 3.2.3 ([GJ18])**.** Suppose P_{SW} is a Markov transition matrix for the Swendsen-Wang process on a homogenous ferromagnetic Ising model, that is to say each that bonding graph edges must exist *intra*- rather than *inter*community, and each edge is set to zero with constant probability β. Then:

$$\tau_\epsilon(P_{SW}) \leqslant 8n^4m^2\left[(\log\epsilon)^{-1} - m\log\beta\right]$$

With n the number of nodes and m the number of edges.

Whilst this is still represents an $O(n^8)$ complexity in the complete graph setting, it is crucially a *polynomial* bound - we conjecture that a similar polynomial bound extends to the asymmetric *ferromagnetic* case. Our general PVN model, however, exhibits significant **geometric frustration**. The network topology is, in the general case, densely-connected and complex. Furthermore, we have a mixture of ferromagnetic and antiferromagnetic interactions in the same network since the alternative density takes values both above and below one on its domain. These general *spin glass* models are "notoriously" ([ZOK15]) difficult to extract information from, either numerically or analytically, and we have no such polynomial bounds on their simulation by Markov chains.

In the next section, we will explore another approach from the theory of spin glasses which improves Swendsen-Wang in high-frustration networks, but it is worth considering a ferromagnetic approximation of our Ising model resembling a mean field approach. In particular, suppose we define

$$J_{ij} = \frac{1}{2}\log Kp_1(A_{ij})$$

For some $K > 1$. This parameter controls the level of frustration, with $K = \min\{p_1(A_{ij})\}^{-1}$ ensuring that the network is entirely ferromagnetic. This leads to

$$e^J \propto p_1(A_{ij}) \qquad\qquad e^{-J} \propto K^{-1}$$

In particular, this will add a factor of K^{-m^-} to the unnormalised probability of $\mathbf{z}$, where m^- is the number of uniform connections induced by $\mathbf{z}$. The partition function, on the other hand, will be multiplied by a factor of $K^{-\mathbb{E}m^-}$, hence:

$$\frac{1}{Z}\prod_{i\leqslant j} f_{ij}(\mathbf{x}_i,\mathbf{x}_j)\prod_i \pi_{\mathbf{x}_i} = P(\mathbf{z}\mid\mathbf{A})\cdot K^{\mathbb{E}[m^-]-m^-(\mathbf{z})}$$

This therefore biases the probabilities in favour of assignments which maximise the number of non-uniform edges induced. In the symmetric case, this introduces a bias towards mostly-null or mostly-alternative assignments, and in the asymmetric case it simply biases mostly-alternative assignments. As with many questions in machine learning, this represents a *bias/efficiency* trade-off, where introducing a bias allows us to take advantage of more powerful estimation tools.

3.2.2 Replica Clustering and Population Annealing

As many known-hard optimisation problems can be phrased in terms of disordered, frustrated Ising models with arbitrary topology [ZFK16], efficient Monte Carlo simulations of them are often considered a "holy grail" of the statistical physics community. In 1986, Swendsen and Wang proposed the **Replica Monte Carlo** method [SW86], but this often fails to mix rapidly in the frustrated case. Following from this, Houdayer proposed in the early 2000s the **Houdayer Cluster Algorithm** [Hou01] which extends the RMC method and results in a large speedup for disordered frustration in two-dimensional lattice networks, but, as observed in the original paper, stagnates on arbitrary topologies (i.e. higher-dimensional lattices). Finally, in 2015, the **Isoenergetic Cluster Move (ICM)** method [ZOK15] was proposed, a small modification to HCA which can offer a similar speedup on higher-dimensional topologies. Both HCA and ICM are situated within the *parallel tempering* framework, which involves simulating multiple configuration chains but where the Boltzmann temperature β of the sampling distribution is heterogenous among them. In this vein, we will present both the ICM algorithm adapted to our PVN setting and the **Population Annealing (PA)** [WMK15] algorithm, which also seeks to improve Monte Carlo performance on Ising Models by varying the temperature - this is closely related to the canonical heuristic optimisation technique of *simulated annealing*, which was first developed to simulate Boltzmann-distributed systems [AK90]. These two methods represent the classical state-of-the-art in Monte Carlo simulations of our general model, although quantum computing shows promise to drastically improve performance by directly embedding the Ising model as entangled quantum particles [RME$^+$21].

ICM

ICM simulates chains for N_T different Boltzmann temperatures evenly spaced in $[\beta_{\min}, \beta_{\max}]$ and at each temperature 2 replica chains are simulated. A full update step of all the chains proceeds as follows:

- **(Metropolis Sweep)** First, every node in every replica chain is updated using the single-spin flip Metropolis algorithm or Gibbs sampler. This ensures ergodicity of each chain.
- **(Houdayer Cluster Move)** For each temperature $\beta \leqslant \mathcal{B}$, the overlap between the current states of the two replicas is computed:

  $$\mathbf{o}_i^{(t,T)} = \begin{cases} 1 & \mathbf{z}_i^{(t,T,0)} = \mathbf{z}_i^{(t,T,1)} \\ -1 & \mathbf{z}_i^{(t,T,0)} \neq \mathbf{z}_i^{(t,T,1)} \end{cases}$$

  We consider the subgraph such that two nodes i, j are connected if and only if they are connected in the underlying network and $\mathbf{o}_i^{(t,T)} = \mathbf{o}_j^{(t,T)}$. An index i such that $\mathbf{o}_i^{(t,T)} = -1$ is chosen uniformly at random and we find the connected component C of the overlap subgraph containing node i. For each $j \in C$, $\mathbf{z}_j^{(t,T,0)}$ *and* $\mathbf{z}_j^{(t,T,1)}$ are flipped.
- **(Parallel Tempering Exchange)** For each pair of neighbouring temperatures (β_1, β_2), pair one replica from β_1 with one replica from β_2. The states are then swapped between temperatures with (Metropolis-style) probability

  $$\min\left(1, e^{(\beta_2-\beta_2)(\mathcal{H}_1-\mathcal{H}_2)}\right)$$

  Where $\mathcal{H}_1, \mathcal{H}_2$ are the Hamiltonians evaluated at the corresponding states.

$\mathcal{B}$ is a tunable hyperparameter, but the original paper proposes $\mathcal{B} = \sqrt{\operatorname{Var} \mathbf{A}}$. Each step other than the HCM is Metropolis-style, so satisfies detailed balance, and we have:

Proposition 3.2.1 ([Hou01]). HCM is *isoenergetic*: the sum of the Hamiltonians of the two replicas is unchanged by a cluster move. Accordingly, the joint probability of two states before and after the move is unchanged and the forward transition probability is equal the reverse transition probability, so detailed balance is satisfied.

Proof. The original argument provided by Houdayer does not explicitly consider the asymmetric case, so it is worthwhile to show that it is still isoenergetic regardless of choice of δ. Recall that the Hamiltonian is given (up to an ignored constant) by:

$$\mathcal{H} = -\sum_{i \leqslant j \in E} \log p_1(\mathbf{A}_{ij})\delta(\mathbf{z}_i, \mathbf{z}_j) - \sum_i \log \pi_{\mathbf{x}_i}$$

Where δ here is any Boolean operator, not exclusively the Kronecker delta. The contribution of each node to the Hamiltonian is thus decomposed into the sum of its interaction with the external field and its pairwise interactions with each other node. For a node to be flipped by the HCM, it must be the case that the assignment is swapped between replicas, so the external field contribution to $\mathcal{H}^0 + \mathcal{H}^1$ doesn't change as one will change by $\pi_0 - \pi_1$ and the other by $\pi_1 - \pi_0$. Now suppose that node i is flipped and we consider the change in the pairwise interaction (i, j). For the flip to be isoenergetic, it suffices then to show:

$$\delta(\mathbf{z}_i^0, \mathbf{z}_j^0) + \delta(\mathbf{z}_i^1, \mathbf{z}_j^1) = \delta(\hat{\mathbf{z}}_i^0, \hat{\mathbf{z}}_j^0) + \delta(\hat{\mathbf{z}}_i^1, \hat{\mathbf{z}}_j^1) \tag{3.5}$$

Where $\hat{\mathbf{z}}$ denotes the community assignments after the flip. For i to be flipped, it must be the case that $\mathbf{z}_i^0 = \neg\mathbf{z}_i^1$, and trivially $\hat{\mathbf{z}}_i^0 = \neg\mathbf{z}_i^0$, $\hat{\mathbf{z}}_i^1 = \neg\mathbf{z}_i^1 = \mathbf{z}_i^0$. Accordingly, it suffices to show:

$$\delta(\mathbf{z}_i^0, \mathbf{z}_j^0) + \delta(\neg\mathbf{z}_i^0, \mathbf{z}_j^1) = \delta(\neg\mathbf{z}_i^0, \hat{\mathbf{z}}_j^0) + \delta(\mathbf{z}_i^0, \hat{\mathbf{z}}_j^1)$$

There are two cases:

- If $\mathbf{z}_j^0 = \mathbf{z}_j^1$, then j is outside of the connected component so not flipped, so (3.5) becomes:

  $$\delta(\mathbf{z}_i^0, \mathbf{z}_j^0) + \delta(\neg\mathbf{z}_i^0, \mathbf{z}_j^0) = \delta(\neg\mathbf{z}_i^0, \mathbf{z}_j^0) + \delta(\mathbf{z}_i^0, \mathbf{z}_j^0)$$

- If $\mathbf{z}_j^0 = \neg\mathbf{z}_j^1$ (and the pairwise interaction is nonzero), then j is inside the connected component and flipped, so (3.5) becomes:

  $$\delta(\mathbf{z}_i^0, \mathbf{z}_j^0) + \delta(\neg\mathbf{z}_i^0, \neg\mathbf{z}_j^0) = \delta(\neg\mathbf{z}_i^0, \neg\mathbf{z}_j^0) + \delta(\mathbf{z}_i^0, \mathbf{z}_j^0)$$

So $\mathcal{H}^0 + \mathcal{H}^1$ remains constant as required. □

$\mathcal{B}$ is chosen to prevent *percolation*, where the connected component to be flipped spans the whole graph hence no meaningful update occurs. The likelihood of a particular node to diverge between its two replicas is correlated with temperature so lower temperatures are far more likely to yield disconnected overlap subgraphs - essential when the underlying topology is dense.

The final marginals are computed as usual by averaging over the replicas at $T = 1$.

Population Annealing

Population Annealing is, like parallel tempering, a wrapper around other MCMC approaches involving replicas at different temperatures. Instead of keeping each chain at the same temperature and occasionally swapping states, we run the entire population of replicas at one temperature before resampling the population and reducing the temperature of the entire ensemble. The name is derived from the mechanical process of *annealing*, where material is heated to a malleable temperature and very slowly cooled to maximise the stability of the final equilibrium (compare to *tempering*, where the material is periodically reheated during the cooling process). In PA:

1. A population of R replica chains are initialized at temperature $T_{\max}$ and run for N_S MCMC steps
2. The population is now going to transition from β to β', but it is first resampled. Define:

   $$\tau_j(\beta, \beta') := \frac{1}{Z(\beta, \beta')} \cdot Re^{-(\beta'-\beta)\mathcal{H}_j}$$

   Where $\mathcal{H}_j$ is the Hamiltonian of replica j and $Z(\beta, \beta')$ the usual normalising term to ensure $\sum \tau_j = R$. Now:

   $$n_j := \begin{cases} \lceil \tau_j \rceil & \text{with probability } \tau_j - \lfloor \tau_j \rfloor \\ \lfloor \tau_j \rfloor & \text{otherwise} \end{cases}$$

   The population for temperature β'^{-1} is then the ensemble of n_j copies of each replica j. Small variance will be exhibited in population size from temperature-to-temperature, but the mean will be R.
3. Repeat until $T_{\min}$ is finished, and estimate the marginals as the mean state of the population.

Note that we may trivially use PA over PT in an ICM scheme, where instead we use only the Metropolis algorithm for the sweeps in the higher temperatures, and begin making Houdayer moves only once annealed to a low-enough temperature. Since we are interested in the final state only at $T = 1$ which may be too high for ICM to be effective, we may perform hybrid annealing/tempering where a final state is obtained by annealing to $T < 1$ and this is used to initialise a final run at $T = 1$.

This is no longer a strictly MCMC method, rather it is *sequential Monte Carlo*, since there is zero probability of a reverse transition between temperatures as in parallel tempering. Accordingly, the population average is a very slightly biased sample from the distribution for any given finite number of MCMC sweeps performed. Instead, it is asymptotically exact either as the population size $R \to \infty$ or the number of full PA runs which are then averaged over $\to \infty$. The large-population single-run case can often be faster in practice, but this does not provide the same theoretical guarantees as weighted averages over many runs. The reader is directed to [WMK15] for a discussion of how such an asymptotically exact weighted average scheme might be implemented.

3.3 Variational marginalisation via parameter estimation

Up until this point, we have considered the case where the alternative distribution of the p-value network is known. This may not, however, be appropriate for real-world settings, so it is natural to consider the question of estimating community membership knowing only the general properties of an alternative distribution of p-values. A canonical method for approximate maximum likelihood estimation when the marginal latent distributions are intractable is **expectation-maximisation** [DLR77]. This has been applied to the unweighted stochastic block model by [ZKRZ12], and we extend their work to weighted networks.

Formulated information-theoretically (as [Bis07]), Expectation-Maximisation seeks to estimate θ based on observed variables $\mathbf{X}$, which depend further on the unobserved latent variables $\mathbf{Z}$. Let p_θ denote the true distribution on $\{\mathbf{X}, \mathbf{z}\}$ and q_φ be some parametric family of distributions. The algorithm proceeds as *coordinate ascent* iteratively in two steps:

Definition (Variational Free Energy)**.** Define the **variational free energy** of this system to be

$$\mathcal{F}(q_\varphi, \theta) := \sum_{\mathbf{z}} q_\varphi(\mathbf{z}) \log p_\theta(\mathbf{X}, \mathbf{z}) - \sum_{\mathbf{z}} q_\varphi(\mathbf{z}) \log p_\theta(\mathbf{z})$$

This is also drawn from statistical physics

- **(E step)** Let $\varphi^{(t)}, \theta^{(t)}$ denote the estimates at time step t of φ, θ respectively. In the first step, we update the variational estimate:

$$\varphi^{(t+1)} \leftarrow \arg\max_{\varphi} \; \mathcal{F}(q_\varphi, \theta^{(t)})$$

- **(M step)** Now updating the parameter estimate:

$$\theta^{(t+1)} \leftarrow \arg\max_{\theta} \; \mathcal{F}(q_{\varphi^{(t+1)}}, \theta)$$

It is clear by unfolding definitions that the VFE is equivalent to

$$\begin{aligned} \mathcal{F}(q_\varphi, \theta) &= \mathbb{E}_{q_\varphi}[\log p_\theta(\mathbf{X}, \mathbf{z}) - \log q_\varphi(\mathbf{z} \mid \mathbf{X})] && (3.6) \\ &= -D_{KL}(q_\varphi(\mathbf{z}) \mid\mid p_\theta(\mathbf{x} \mid \mathbf{A})) + \ell(\theta) && (3.7) \end{aligned}$$

Where $\ell(\theta)$ denotes the log-likelihood of θ, i.e. the log-evidence of $\mathbf{A}$ w.r.t. θ. Since this does not depend on the approximation q_φ, we can read off immediately that the KL-divergence between q_φ and $p_\theta(\cdot \mid \mathbf{A})$ is minimized whenever q_φ maximises $\mathcal{F}(\cdot, \theta)$, and vice versa that $\ell(\theta)$ is maximised where $\mathcal{F}(q_\varphi, \cdot)$ is [ZKRZ12]. It is well-known [Wu83] that the successive approximations monotonically improve as $t \to \infty$ and converge to a stationary point w.r.t. the parameters, but there is no guarantee of convergence to the global MLE.

Remark. In the setting where the alternative distribution is known, we omit the M-step and compute $\mathcal{F}$ using the ground truth. In this case, we recover the traditional variational Bayes method and the free energy is often known as the ELBO score. When θ is unknown, this algorithm will produce a estimated probability distribution on the community assignments (hence approximate the marginals), but only a point estimate of the θ parameter. In the standard setting where we are interested only in community assignments of nodes and not knowledge of the underlying distributions, this method offers an advantage over a fully Bayesian variational estimator of $\{\varphi, \theta\}$ since using a point estimate for θ simplifies computation.

3.3.1 Mean field E-step

The simplest case corresponds to variationally fitting a mean field approximation to the marginals when the underlying distribution p is known. In particular:

$$q_\varphi(\mathbf{x}) = \prod_i q_\varphi^i(\mathbf{x}_i) \qquad\qquad q_\varphi^i(x_i) = \begin{cases} \varphi_i & x_i = 1 \\ 1 - \varphi_i & x_i = 0 \end{cases} \tag{3.8}$$

Letting $f : \mathbb{R}^2 \to \mathbb{R}$ and $g : \mathbb{R}^4 \to \mathbb{R}$ be arbitrary, we note the following:

$$\sum_{\mathbf{t} \in \mathcal{Z}^n} \left[q_\varphi(\mathbf{t}) \sum_{i<j\leqslant n} g(\mathbf{t}_i, \mathbf{t}_j, i, j) \right] = \sum_{i<j\leqslant n} \sum_{r,s \in \mathcal{Z}^2} \sum_{\mathbf{t}' \in \mathcal{Z}^{n-2}} q_\varphi^i(r) q_\varphi^j(s) q_\varphi(\mathbf{t}') g(r,s,i,j) \tag{3.9}$$

$$= \sum_{i<j\leqslant N} \sum_{r,s \in \mathcal{Z}^2} q_\varphi^i(r) q_\varphi^j(s) g(r,s,i,j) \tag{3.10}$$

$$\sum_{\mathbf{t} \in \mathcal{Z}^n} \left[q_\varphi(\mathbf{t}) \sum_{i\leqslant n} f(\mathbf{t}_i, i) \right] = \sum_{i\leqslant n} \sum_{r \in \mathcal{Z}} \sum_{\mathbf{t}' \in \mathcal{Z}^{n-1}} q_\varphi^i(r) q_\varphi(\mathbf{t}') f(r,i) \tag{3.11}$$

$$= \sum_{i\leqslant N} \sum_{r \in \mathcal{Z}} q_\varphi^i(r) f(r,i) \tag{3.12}$$

Where we are able to factor out $\sum_{\mathbf{t}'} q_\varphi(\mathbf{t}') = 1$. We can leverage these identities to reduce the exponential cost of computing $\mathcal{F}$ to a quadratic:

$$\begin{aligned} \mathcal{F}(q_\varphi) &= \sum_{\mathbf{z}} \left[q_\varphi(\mathbf{z}) \left(\sum_i \log p(\mathbf{z}_i) + \sum_{i\leqslant j} \log p(A_{ij} \mid \mathbf{z}_i, \mathbf{z}_j) \right) \right] - \sum_{\mathbf{z}} \left[q_\varphi(\mathbf{z}) \sum_i \log q_\varphi^i(\mathbf{z}_i) \right] \\ &= \sum_{i\leqslant j, r, s} \left[q_\varphi^i(r) q_\varphi^j(s) \log p(A_{ij} \mid r, s) \right] + \sum_{i,r} \left[q_\varphi^i(r) \log \frac{p(r)}{q_\varphi^i(r)} \right] \end{aligned}$$

We denote the estimated marginal $\psi_r^i := q_\varphi^i(r)$. Specialising to the PVN setting:

$$\begin{aligned} (\mathbf{A}, \mathbf{z}) \sim \mathrm{APVN}(n, p_1, \pi) \quad \mathcal{F}(q_\varphi) &= \sum_{i\leqslant j} \left[\psi_1^i \psi_1^j \log p_1(A_{ij}) \right] + \sum_i \left[\psi_0^i \log \frac{1-\pi}{\psi_0^i} + \psi_1^i \log \frac{\pi}{\psi_1^i} \right] \\ (\mathbf{A}, \mathbf{z}) \sim \mathrm{SPVN}(n, p_1, \pi) \quad \mathcal{F}(q_\varphi) &= \sum_{i\leqslant j} \left[(\psi_1^i \psi_1^j + \psi_0^i \psi_0^j) \log p_1(A_{ij}) \right] + \sum_i \left[\psi_0^i \log \frac{1-\pi}{\psi_0^i} + \psi_1^i \log \frac{\pi}{\psi_1^i} \right] \end{aligned}$$

Following [ZKRZ12], we wish to set $\nabla_\psi \mathcal{F} = 0$ and hence derive a set of self-consistent equations any maximum must obey.

$$\begin{aligned} \frac{\partial \mathcal{F}}{\partial \psi_1^i} &= \sum_j \psi_1^j \log p_1(A_{ij}) + \log \frac{\pi}{\psi_1^i} - 1 \\ \frac{\partial \mathcal{F}}{\partial \psi_0^i} &= \log \frac{1-\pi}{\psi_0^i} - 1 + \begin{cases} 0 & (\mathbf{A,z}) \sim \mathrm{APVN} \\ \sum_j \psi_0^j \log p_1(A_{ij}) & (\mathbf{A,z}) \sim \mathrm{SPVN} \end{cases} \end{aligned}$$

Giving a simple "self-referential" form of the ψ^i and hence trivial iterative estimation algorithm:

$$\psi_b^i \leftarrow \frac{\pi_b e^{h_b^i}}{\sum_r \pi_r e^{h_r^i}} \qquad h_1^i = \sum_j \psi_1^j \log p_1(A_{ij}) \qquad h_0^i = \begin{cases} 0 & (\mathbf{A,z}) \sim \mathrm{APVN} \\ \sum_j \psi_0^j \log p_1(A_{ij}) & (\mathbf{A,z}) \sim \mathrm{SPVN} \end{cases}$$

Note that we are able to exclude a -1 term from both h_b^i expressions as common factors cancel after normalisation. One may follow the same approach as [ZZ17] (specifically, the proof of Theorem 2.1) to see that this is exactly **Coordinate Ascent Variational Inference** (with known parameters). That is to say, the marginals are updated according to:

$$\psi_b^i \propto \exp \left[\mathbb{E}_{\psi^{-i}} \log p(\mathbf{z}_i = b \mid \mathbf{z}_{-i}, \mathbf{A}) \right] \tag{CAVI}$$

Here $\mathbf{z}_{-i}$ denotes the $\mathbf{z}$ vector excluding the i-th element, and $\mathbb{E}_{\psi^{-i}}$ denotes that the expectation is taken over $\mathbf{z}_{-i}$ w.r.t. the current estimate ψ.

3.3.2 M-step via discretization

In the previous section, we assumed knowledge of $\theta = \{\pi, p_1\}$, but here we consider the setting where either is unknown. One may know that p_1 is (well-approximated by) a member of a parametric family, in which case $\theta \setminus \{\pi\}$ becomes simply the set of parameters for this distribution and one may proceed by standard estimation methods for the family in question. However, we consider the most general case in which no information is assumed about p_1. In general, this problem fits an infinite-dimensional parameter and is intractable, so we consider a discretization of the edge weights into L bins by a partition of $[0, 1]$: $(\sigma_1, ..., \sigma_{L-1})$. We now have that $p_1 = \text{Categorical}(\mathbf{p})$, where $\mathbf{p}$ is a length-L vector of discrete probabilities to be learned. The null distribution in this case is then $\text{Categorical}(1/\sigma_1, 1/(\sigma_2 - \sigma_1)..., 1/(1 - \sigma_{L-1}))$, and we will focus our analysis on the case where the bins are evenly-sized (hence the categorial distribution is uniform), but this analysis may be extended to the non-uniform case quite easily.

Focusing for now on the APVN case, we seek:

$$\arg\max_{\mathbf{p}} \sum_{i \leqslant j} \psi_1^i \psi_1^j \log p_1(A_{ij} \mid \mathbf{p}) \quad \text{subject to } \sum_i \mathbf{p}_i = 1$$

The traditional Bayesian approach would be to consider a Dirichlet conjugate prior and take the mode of the posterior, but this does not allow us to take into account the weighted sum of logs and we are interested only in a point estimate for $\mathbf{p}$, so we approach it directly by the method of Lagrange multipliers. We seek to maximise:

$$f(\mathbf{p}) := \sum_{i \leqslant j} \psi_1^i \psi_1^j \log \mathbf{p}_{A_{ij}} + \lambda \left(1 - \sum_i \mathbf{p}_i \right)$$

Setting $\nabla_{\mathbf{p}} f = 0$ we find, for all k:

$$\lambda = \frac{1}{\mathbf{p}_k} \sum_{i \leqslant j} \psi_1^i \psi_1^j \mathbf{1}(A_{ij} = k)$$

By the sum constraint, it is clear that $\lambda = \sum_{i \leqslant j} \psi_1^i \psi_1^j$ hence we have the following M-step updates:

$$\text{(APVN)} \quad \mathbf{p}_k \leftarrow \frac{\sum_{i \leqslant j} \psi_1^i \psi_1^j \mathbf{1}(A_{ij} = k)}{\sum_{i \leqslant j} \psi_1^i \psi_1^j} \qquad \text{(SPVN)} \quad \mathbf{p}_k \leftarrow \frac{\sum_{i \leqslant j} (\psi_1^i \psi_1^j + \psi_0^i \psi_0^j) \mathbf{1}(A_{ij} = k)}{\sum_{i \leqslant j} (\psi_1^i \psi_1^j + \psi_0^i \psi_0^j)}$$

Note that this weighted proportion is exactly the mode of a Dirichlet distribution with concentration parameters given by the sum of estimated pairwise marginals over edges with a particular label. This suggests that we may trivially extend this E-M process to a fully Bayesian one by taking this as the posterior distribution on $\mathbf{p}$, but this is beyond the scope of the present discussion.

Finally, we may trivially fit π again by Lagrange multipliers to yield:

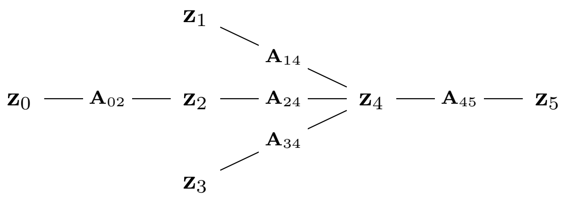
$$\pi \leftarrow \frac{1}{N} \sum_i \psi_1^i$$

We have formally derived these update rules, but they coincide with the natural notion that we should set these discrete probabilities equal to the weighted proportion of their coefficients.

Remark. Although we have derived these M-step update rules from the CAVI mean-field E-step, any approach for estimating the marginals may be used for the E-step and the same M-step updates performed. In particular, we may perform the E-step via an MCMC approach and make a mean-field approximation to update the parameter estimates with the same rules.

3.4 Belief Propagation

Whilst the previous methods, often originating from the study of lattices in statistical physics, have provided iterative methods for approximating intractable marginals on general networks, we may turn to traditional computer science for an iterative method which can solve exactly the marginalisation problem on tree-like networks [Pea82]. To understand tractability in tree-like networks, we need to consider the relationships which cause intractability in general. Consider a simplified network:



Now $\mathbf{A}_{02}$ and hence $\mathbf{z}_0$ is not independent of $\mathbf{A}_{45}$, despite the fact that they are independent conditioned on the community membership of their endpoints. This is due to a propagation of information along the dependence chain $\mathbf{A}_{45} \to \mathbf{z}_4 \to \mathbf{A}_{24} \to \mathbf{z}_2 \to \mathbf{A}_{02}$. This leads to a combinatorial explosion in complexity in the general case since, for a fully-connected network, the marginal probability of any node being alternative is dependent on every other node's community membership due to these propagation effects. But there is a clear bottleneck in this example network: if we suppose that $\mathbf{z}_4 = 1$ is revealed by a genie, then this endpoint independence blocks information propagation through this node. Accordingly, each marginal probability can be computed without marginalisation, except for $\mathbf{z}_0, \mathbf{z}_2$ which require marginalisation only over each other.

More generally, if we condition on the nodes $\mathbf{z}_I$ for some indexing set I, then we may remove each of these nodes I from the network, and each connected component of this fractured graph is independent. This is not particularly powerful in lattices (especially those of a higher dimension) since even a large set of conditioned variables will not reduce the size of connected components meaningfully, but it immediately implies an efficient recursive algorithm when our network is acyclic. Choose an arbitrary node $\mathbf{z}_r$ to be the root and interpret the network as a rooted tree:

- Suppose $\mathbf{z}_r$ has the children $\{\mathbf{z}_1, ..., \mathbf{z}_m\}$. Then the marginal of each child's subtree is independent conditioned on $\mathbf{z}_r$, hence

$$P(\mathbf{z}_r = b) = \frac{1}{Z}\pi_b \prod_{k\in[m]} \sum_{b'\in\{0,1\}} p(\mathbf{A}_{rk} \mid \mathbf{z}_r = b, \mathbf{z}_k = b')P(\mathbf{z}_k = b')$$

- If $\mathbf{z}_k$ is not a leaf, to find its marginal recurse down and consider it as the new root (forgetting the connection to the parent, since this has already been taken into account). If it is a leaf (the base case), we simply take $P(\mathbf{z}_m = b) = \pi_b$. Again, this step is valid by the independence on subtrees when conditioned on their common root.

To make this more amenable to the general graph topology, we interpret the child probabilities as *messages* passed from a child to parent, describing the "belief propagated" up from the leaf nodes. We write $\psi_b^{i\to j}$ to denote the message passed from node i to node j about the marginal b, hence defining:

$$\psi_b^{i\to j} := \frac{1}{Z_{ij}}\pi_b \prod_{(i,k)\in E, k\neq j} \sum_{b'\in\{0,1\}} \psi_{b'}^{k\to i} p_{bb'}(\mathbf{A}_{ik}) \tag{3.13}$$

Where the product is taken over each node other than j adjacent to i and Z_{ij} denotes the usual normalising constant. The marginal probability is then retrieved by taking every edge into account (including the "parent") and normalising:

$$\psi_b^{i} := \frac{1}{Z_i}\pi_b \prod_{(i,k)\in E} \sum_{b'\in\{0,1\}} \psi_{b'}^{k\to i} p_{bb'}(\mathbf{A}_{ik}) \tag{3.14}$$

The final step to make the algorithm well-defined on cyclic graphs is to remove the non-terminating recursion and instead iteratively update each message using the messages in the previous time step, akin to Gibbs sampling. It is clear that this iterative-update method will yield the correct marginals on a tree in d steps, where d is the tree's **diameter**: the longest path between any two nodes.

The algorithm in this form is known as **Exact Belief Propagation (EBP)**, and when applied to a cyclic graph it is known as **Loopy Belief Propagation (LBP)**. On a cyclic graph, we no longer have the subtree independence properties so the algorithm will not be exact, but it is hoped that it is a good approximation. In particular, LBP is a Markov chain method but does not rely on Monte Carlo sampling, so it often has a significant advantage in convergence time [DKMZ11a]. Although we derived loopy belief propagation from first principles by analysing trees, it has a deep historical connection to the analysis of Ising models via variational Bayesian theory. In particular, H.A. Bethe was concerned in the 1930s with the free energy of Ising models in an acyclic lattice configuration (known in the modern day as a *Bethe lattice*) [BB35]. On these lattices, he derived the following form of variational free energy (specialised to our setting) for an approximating distribution q:

$$\mathcal{F}_{\text{bethe}} := \sum_{i,j}\left[\sum_{b,b'} q_{ij}(b,b')\log\frac{q_{ij}(b,b')}{p_{bb'}(\mathbf{A}_{ij})\pi_b\pi_{b'}}\right] + \sum_{i}\left[m_i\sum_{b} q_i(b)\log\frac{q_i(b)}{\pi_b}\right] \tag{3.15}$$

Here m_i denotes the number of neighbours of node i minus 1 and $q_{ij}(b,b')$ is shorthand for $q(\mathbf{z}_i = b, \mathbf{z}_j = b')$. Furthermore, he showed that on acyclic graphs, this coincides with the usual free energy, and conjectured that it was a good approximation on loopy graphs. From the set-up, a connection to belief propagation may seem inevitable as they seek to answer very similar questions; the following elegant result due to Yedidia confirms this:

Theorem 3.4.1 ([YFW03])**.** Regardless of network topology, an approximated marginal distribution (*"set of beliefs"*) is a fixed point of the belief propagation update equations (3.13),(3.14) if and only if it is a local stationary point of the Bethe free energy (3.15)

We write $\mathcal{F}_{\text{bethe}}$ in terms of BP messages in our setting in a similar form to [DKMZ11a]:

$$\mathcal{F}_{\text{bethe}} := \sum_{i\leqslant j}\log Z^{ij} - \sum_{i}\log Z^{i} \qquad Z_{ij} = \sum_{b,b'} p_{bb'}(\mathbf{A}_{ij})\psi_b^{i\to j}\psi_{b'}^{j\to i} \tag{3.16}$$

$$Z_i = \sum_{b}\pi_b\prod_{(i,j)\in E}\sum_{b'} p_{bb'}(\mathbf{A}_{ij})\psi_{b'}^{j\to i} \tag{3.17}$$

Note that this takes a simpler form than in the cited paper as we cannot make the same rearrangements in our setting. Recalling the M-step we derived for the mean field variational approach, we may apply the same Lagrange multiplier method to find update equations for the discretized parameters based on maximising $\mathcal{F}_{\text{bethe}}$. In statistical physics, these equations (which describe the stationary points of the $\mathcal{F}$ approximation) are often called the **Nishimori Conditions**. While we are able to express these in terms of marginals, they can be written directly using messages:

$$\begin{aligned}
&\text{(APVN)} \quad \mathbf{p}_k \leftarrow \frac{\mathbf{p}_k}{N\pi^2}\sum_{(i,j)\in E}\frac{\mathbf{1}(A_{ij}=k)}{Z_{ij}}\psi_1^{i\to j}\psi_1^{j\to i}\\
&\text{(SPVN)} \quad \mathbf{p}_k \leftarrow \frac{\mathbf{p}_k}{N\pi^2}\sum_{(i,j)\in E}\frac{\mathbf{1}(A_{ij}=k)}{Z_{ij}}(\psi_1^{i\to j}\psi_1^{j\to i} + \psi_0^{i\to j}\psi_0^{j\to i})\\
&\qquad\qquad \pi_b \leftarrow \frac{1}{N}\sum_{i}\psi_b^{i}
\end{aligned}$$

3.4.1 Circular Belief Propagation

We have discussed informally the conjecture that LBP provides good approximations to the true marginals, but this is conditional on a *locally tree-like* form of sparsity. While it is true that real-world networks are often sparse, they similarly are often tightly modular - there is a high degree of connectivity between neighbours of any particular node. These networks can suffer from low *girth*, with many short-range cycles in the neighbourhood of long-range connecting "hub" nodes. In general, vanilla belief propagation can struggle to deal with these densely-connected clusters, but even a locally tree-like network gives no guarantees that LBP will converge, let alone to the true marginals. We will present two methods which seek to overcome this: one which adds new parameters to the BP update equations which allows for guaranteed convergence in arbitrary topologies, and one which seeks to exploit modular structure to break the network into a tree of tractable-sized components such that BP is exact with high probability (although no guarantees are given in the general case).

Circular belief propagation [BJD24] is a remarkably simple generalisation of LBP:

$$\psi_{b'}^{i\to j} \propto \sum_b p_{b'b}(\mathbf{A}_{ij})^{\beta_{ij}} \left[\pi_b^{\gamma_i} \prod_{(i,k)\in E, k\neq j} \psi_b^{k\to i} \left(\psi_b^{j\to i}\right)^{1-\alpha_{ij}/\kappa_i} \right]^{\kappa_i} \tag{3.18}$$

Here, $(\alpha, \beta, \gamma, \kappa)$ are parameters to be decided. We note that setting $(\alpha, \beta, \gamma, \kappa) = (1, 1, 1, 1)$ yields:

$$\psi_{b'}^{i\to j} \propto \sum_b p_{b'b}(\mathbf{A}_{ij}) \pi_b \prod_{(i,k)\in E, k\neq j} \psi_b^{k\to i} \tag{3.19}$$

Which is exactly vanilla BP in an "unrolled" form: the summing over possible assignments to node k is performed in the message $\psi^{k\to i}$ as opposed to $\psi^{i\to j}$, so we have a *sum-product* formulation of the update rule as opposed to a *product-sum* version. Marginal probabilities are then estimated in CBP by:

$$\psi_b^i \propto \left[\pi_b^{\gamma_i} \prod_{(i,j)\in E} \psi_b^{j\to i} \right]^{\kappa_i} \tag{3.20}$$

Aside from exponentiating parameters to be chosen, the primary difference between LBP and CBP is the "reverse term" $\psi_b^{j\to i}$ in the computation $\psi_{b'}^{i\to j}$ along with the "damping factor" κ. Notice that the message $j \to i$ is queried using the current assignment to i (b) as opposed to the assignment to j (b'); in this way j acts as both parent and child of i, weighting the contribution of each other child in a manner controlled by the α_{ij} parameter. We have the following convergence theorem:

Theorem 3.4.2 ([BJD24])**.** For an Ising model and any choice of (β, γ), it is possible to find a pair (α, κ) such that circular BP converges with at least linear rate to a unique fixed point. In particular, there exists some $\Lambda \in \mathbb{R}_{>0}$ such that it suffices to choose $\alpha_{ij} = \kappa_i = \lambda$ for any $0 < \lambda < \Lambda$.

This guarantees convergence in the SPVN case with a simple hyperparameter search to find Λ, and we conjecture that a similar result extends to the APVN case (although the difference between a symmetric and asymmetric Ising model prevents the same proof from being readily applied, despite a similar spectral norm argument likely being possible). In particular, it tells us that we may ignore the parameters (α, β, γ) and simply tune the damping level to reduce loop reverberations until they do not effect convergence. However, notice that no guarantee is made that the fixed point of this CBP dynamics is indeed the true marginal distribution - the linear rate of convergence ensures that this can never be true in general unless P = NP. In practice then, the parameters are fitted by iterative update, akin to the expectation maximisation procedure. Introduce the following likelihood ratios:

$$\Psi_{j\to i} := \frac{1}{2}\log\frac{\psi_0^{i\to j}}{\psi_1^{i\to j}} \qquad \Psi_i := \frac{1}{2}\log\frac{\psi_0^i}{\psi_1^i} \qquad \Psi_{\text{ext}} := \frac{1}{2}\log\frac{1-\pi}{\pi}$$

The authors of CBP propose the following unsupervised learning rules:

$$\begin{cases} \Delta\alpha_{ij} = \eta_\alpha \left[\Psi_{j\to i}(\Psi_i - \alpha_{ij}\Psi_{j\to i}) + \Psi_{i\to j}(\Psi_j - \alpha_{ij}\Psi_{i\to j})\right] \\ \Delta\kappa_i = -\eta_\kappa \Psi_{\text{ext}}(\Psi_i - \Psi_{\text{ext}}) \end{cases}$$

Where η_κ, η_α are learning rate parameters. Intuitively:

- The rule for κ seeks to minimize the correlation between the external field and the information i has about itself with the external field removed, hence mitigating the impact of external field contributions to each message reverberating around loops. κ is chosen to minimize this as the more general damping parameter.

- The rule for α seeks to maximise consensus of node information, equivalently minimizing the mean squared error between the information node i has about j and the information j has about itself. When the $\alpha\Psi_{\bullet\to\bullet}$ terms dominate the $\Psi_\bullet$ terms, α is increased, and vice versa.

The reader is directed to [BJD24] for a more formal motivation of these rules.

3.4.2 Generalized Belief Propagation

Finally, we come to an algorithm that can indeed solve exact marginalisation on general graphs, but at the cost of, in the worst case, providing no benefit in computation time. In the best case, it can improve computation by orders of magnitude. Consider a toy version of the modular network mentioned in the previous section: there are n fully-connected communities of m nodes each, and one node from each community is connected to a central "hub node". Belief propagation on this network is likely to have a high level of local innaccuracy by reverberations through the dense components, but if each community is absorbed to a single node, this graph is now a tree. Accordingly we may perform marginalisation at a cost of 2^m for each community, and incorporate this with exact BP on the *region tree.* The total complexity of this procedure is now only $\mathcal{O}(n2^m) \ll \mathcal{O}(2^{nm})$ by decomposing it into *non-overlapping subproblems.*

We follow the approach of [YFW00] to define **generalised belief propagation**

1. Choose the *region set* $\mathcal{R} \subseteq 2^V$ and initialize a *region graph* with regions as vertices
2. Find the pairwise intersections of each region, and add these *direct subregions* as vertices in the region graph, adding an edge to each of the parent regions.
3. Recursively perform this procedure, adding the pairwise intersections between the subregions generated in the previous step until each intersection is empty. Write $S \sqsubseteq R$ if S is a (potentially indirect, potentially improper) subregion of R.
4. We now perform BP on this region graph. Each region R sends a message to its direct subregions S for each possible assignment to the nodes in S. We denote by $\mathbf{z}_S$ an assignment to S, and by $\mathbf{z}_{R\setminus S}$ an assignment to the nodes in $R \cap S^c$. Introducing three further pieces of notation:
   - $M(R)$ is the set of messages which start outside R and end inside it whilst preserving dependencies between R and the start point. In particular:
   $$M(R) := \left\{\psi_{\mathbf{z}_{S'}}^{R'\to S'} \mid S' \sqsubseteq R,\ R' \cap R \subseteq S'\right\}$$
   - $M(R, S)$ is the subset of $M(S)$ starting inside R:
   $$M(R, S) := \left\{\psi_{\mathbf{z}_{S'}}^{R'\to S'} \mid R' \sqsubseteq R, S' \sqsubseteq S, R' \cap S \subseteq S'\right\}$$
   - $\Pi_R(\mathbf{z}_R)$ denotes the likelihood of $\mathbf{z}_R$, considered locally on the subgraph induced by R:
   $$\Pi_R(\mathbf{z}_R) := \prod_{(i,j)\in E|_R} p_{\mathbf{z}_i\mathbf{z}_j}(\mathbf{A}_{ij}) \prod_{i\in R} \pi_{\mathbf{z}_i}$$

   We may finally provide an expression for the inter-region messages and predicted marginals:
   $$\psi_{\mathbf{z}_S}^{R\to S} \propto \frac{1}{\prod_{\mu\in M(R,S)} \mu}\left[\sum_{\mathbf{z}_{R\setminus S}} \Pi_{R\setminus S}(\mathbf{z}_{R\setminus S}) \prod_{\mu\in M(R)\setminus M(S)} \mu\right]$$
   $$\psi_{\mathbf{z}_R}^{R} \propto \Pi_R(\mathbf{z}_R) \prod_{\mu\in M(R)} \mu$$

The way in which this is directly generalises vanilla BP is immediate, but we note the care taken to ensure that dependencies are resolved correctly. A canonical approach to generate such a region graph is the **Hugin** (or **factor tree**) **algorithm** which builds the region graph from the maximal cliques in the network - the reader is directed to Chapter 4 of [Jen96] for a comprehensive introduction. We have the following exactness result:

Theorem 3.4.3 ([YFW00])**.** If the region graph is acyclic and has at most two layers (that is to say, sub-regions are formed as intersections of regions but each sub-region is disjoint), then GBP is exact.

Whilst theoretically interesting and useful in specific contexts, the computational cost of GBP is often prohibitively high. Furthermore, even in a topology amenable to it, region graph selection must be done carefully to ensure exactness (or a close approximation). To quote Yedidia, construction of the region graph is often "more art than science", so we will not consider GBP in our simulation studies for a general combination framework and it is rarely directly applicable.

3.5 Spectral methods

Up until now, we have been working entirely with methods that iteratively construct a sequence of successively better approximations and, while giving strong theoretical guarantees, this can often incur an unacceptable computational expense. A natural solution then might be to turn to the theory of **spectral analysis**, which gives a trivial method of finding equilibria points of *linear* iterative dynamics by considering the eigenspace of a matrix induced by a particular update rule. Can we extract information for approximate inference from the spectra of carefully-chosen linear operators?

The simplest possible matrix to analyse is the adjacency matrix, and we motivate a spectral approach here - following [Abb23] but applied more rigorously to the asymmetric case. Assume for now that $\mathbb{E}p_1 < 0.5$, so we may distinguish node communities based on the mean (weighted) adjacency matrix $\mathbb{E}\mathbf{A}$. An immediate advantage of spectral analysis is that eigenspaces are equivariant under permutation: eigenvalues are preserved and eigenvectors are permuted equivalently to the matrix. Accordingly, we may without loss of generality assume that $\mathbb{E}\mathbf{A}$ is a block matrix with antidiagonal μ_0-blocks and diagonal μ_1 blocks (or only the top-left element in the asymmetric case). What is the spectrum of this expectation matrix?

First, note that it will in both cases have rank 2. In particular, it is of the form $\mathbf{x}^T\mathbf{x} - \mathbf{y}^T\mathbf{y}$. In the asymmetric case, this gives (letting n_b denote the number of nodes in community b):

$$\mathbf{x}^T = (\sqrt{\mu_0}, ..., \sqrt{\mu_0}) \in \mathbb{R}^N \qquad \mathbf{y}^T = (\underbrace{\sqrt{\mu_0 - \mu_1}, ..., \sqrt{\mu_0 - \mu_1}}_{n_1 \text{ times}}, \underbrace{0, ..., 0}_{n_0 \text{ times}})$$

$\operatorname{Im}\mathbb{E}\mathbf{A}$ is exactly $\operatorname{span}(\mathbf{x}, \mathbf{y})$, so we may find a representation of $\mathbb{E}\mathbf{A}$ in terms of this basis:

$$\mathbb{E}\mathbf{A}|_{\langle \mathbf{x},\mathbf{y}\rangle} = \begin{pmatrix} N\mu_0 & n_1\sqrt{\mu_0(\mu_0 - \mu_1)} \\ -n_1\sqrt{\mu_0(\mu_0 - \mu_1)} & n_1(\mu_0 - \mu_1) \end{pmatrix}$$

Note that the eigenvalues of this matrix are exactly the eigenvalues of $\mathbb{E}\mathbf{A}$ and we find eigenvectors of $\mathbb{E}\mathbf{A}$ in $\mathbb{R}^N$ by finding eigenvectors of this reduced matrix and changing the basis from $\{(1,0),(0,1)\} \mapsto \{\mathbf{x}, \mathbf{y}\}$. Solving the characteristic equation, a simple form for eigenvalues is not admitted as in the simplified symmetric case analysed in [AFWZ19], but we find

$$\lambda = \frac{1}{2}\left[N\mu_0 + C \pm \sqrt{(C - N\mu_0)^2 + 4(CN\mu_0 - K^2)}\right] \quad \text{where} \quad C := n_1(\mu_0 - \mu_1), \;\; K := n_1\sqrt{\mu_0(\mu_0 - \mu_1)}$$

Denoting these values by λ_1, λ_2 and setting the first coordinate of the eigenvectors to 1, we find the second coordinate:

$$y_{1,2} = \frac{\lambda_{1,2} - N\mu_0}{C}$$

In $\mathbb{R}^N$, the eigenvectors of $\mathbb{E}\mathbf{A}$ are thus spanned by $\mathbf{x} + y_{1,2}\mathbf{y}$. Now recall that null indices in $\mathbf{y}$ are zero, so (up to constant multiple), each null node will have component $\sqrt{\mu_0}$ in each eigenvector. Now, on the assumption that $0 \leqslant \mu_1 \leqslant \mu_0$, it is clear that $\sqrt{\mu_0} + y_2\sqrt{\mu_0 - \mu_1} < 0$ (we omit the computation for brevity). Accordingly, the eigenvector corresponding to the lower of the nonzero eigenvalues of $\mathbb{E}\mathbf{A}$ will be negative at every anomalous position and positive at every null position. A similar argument (for which the reader is referred to [AFWZ19]) yields the same result for the symmetric case.

It becomes natural now to consider the observation $\mathbf{A}$ under a *signal-plus-noise* model, that is to say

$$\mathbf{A} = \mathbb{E}\mathbf{A} + \mathbf{Z}$$

Where $\mathbf{Z}$ is a noise matrix distributed according to the WSBM but with each distribution re-centered around zero. Our interest then lies in *perturbation analysis* of the eigenvectors - how well can we (entry-wise) estimate the eigenvectors of the expectation matrix via the eigenvectors of the observed matrix? Placing bounds on this error in terms of the noise matrix $\mathbf{Z}$ allows us to reason about the estimation accuracy of a spectral embedding. This requires a large volume of finite-dimensional spectral theory to do rigorously, so for comprehensive analysis of this form, the reader is directed to [Abb23] or [AFWZ19].

We briefly consider the question of spectral embedding dimension by examining the spectrum of $\mathbb{E}\mathbf{A}$. Whilst a naive method would simply look at the second eigenvector, it is clear that a two-coordinate embedding of both the nonzero eigenvectors will increase separation between null and alternative nodes since, although the sign may not flip in the first eigenvector, it is still of the form $(\overbrace{a,...,a}^{n_1},\overbrace{b,...,b}^{n_0})$ where $a \neq b$. Does separation also increase if we increase the embedding dimension above the rank of the expectation matrix? Once the nonzero eigenvectors are accounted for, each additional coordinate will correspond to an element of a basis in $\mathbb{R}^N$ of $\ker \mathbb{E}\mathbf{A}$. It is natural to ask therefore whether a basis for this kernel carries information about community assignments. Focusing on the asymmetric case:

$$(\mathbb{E}\mathbf{A})\mathbf{v} = \begin{pmatrix} \mu_1 \sum_{i \leqslant n_1} \mathbf{v}_i + \mu_0 \sum_{i > n_0} \mathbf{v}_i \\ \vdots \\ \mu_0 \sum_{i \leqslant N} \mathbf{v}_i \\ \vdots \end{pmatrix}$$

A basis for this kernel would be given by the set of zero vectors with two neighbouring indices set to $\{1, -1\}$ - excluding the case where the pair straddles the boundary between null and alternative indices. This gives $N-2$ linearly independent vectors, so we conclude by rank-nullity that this is a basis. Notice that if $\mu_1 = \mu_2$, the expectation matrix will have rank 1, and the boundary-straddling vector is admitted as an element of the kernel to give an $(N-1)$-length basis. There are, of course, infinitely many possible choices for the basis, but it shows that, in the general case, no extra information about community assignments is contained within the vector. The vector we described does carry a signal as community assignments may be recovered as the (often unique) nontrivial partition of indices such that the sum of the elements in each partition is zero in each kernel basis vector, and a similar signal can be guaranteed on any choice of *orthogonal* kernel eigenbasis, it carries no more information than is present in the two nonzero eigenvectors.

Whilst this is useful in the case where eigenvalue order may not be preserved under peturbation, including these “redundant” zero eigenvectors adds no further information in the zero-variance case. It is conjectured in [CYM21] that, at least in the case of the unweighted SBM, this redundancy in the $\mathbb{E}\mathbf{A}$ spectrum carries over to $\mathbf{A}$, hence it is never advantageous to embed the network into a dimension higher than 2 - we reassert this conjecture in the weighted case, supported by numerical experiments.

3.5.1 Spectral embedding CLT

The following result provides a framework for a rigorous study of separation under spectral embedding, but we first require some technical preparations. Let $\mathbf{A}$ be an observed matrix distributed according to the K-community stochastic block model such that each edge distribution has exponential tails and finite expectation. Now let $\mathbf{B}$ and $\mathbf{C}$ denote the $K \times K$ block mean and variance matrices respectively, and let $\mathbf{X}_M$ denote the spectral embedding of matrix M into dimension $d = \text{rank}(\mathbf{B})$,

$$\mathbf{X}_M = \mathbf{U}_M \left|\Lambda_M\right|^{\frac{1}{2}}$$

With $\mathbf{U}_M$ an orthonormal eigenbasis of M and Λ_M the corresponding diagonal matrix of eigenvalues. It is clear that $\mathbf{B} = \mathbf{X_B}\mathbf{I}_{p,q}\mathbf{X}_\mathbf{B}^\top$, where $\mathbf{I}_{p,q}$ is the diagonal matrix of p 1s followed by q -1s. The *signature* of a particular embedding is this pair (p, q) (since signs may be flipped in the embedding). Finally, we denote by Δ the second moment matrix $\mathbb{E}[\mathbf{X}\mathbf{X}^\top]$ of $\mathbf{X}_i = (\mathbf{X_B})_{\mathbf{z}_i}$, with expectation taken over the community assignment vectors $\mathbf{z}$.

Theorem 3.5.1 (WSBM CLT [GJB+23])**.** For all $k \in [K]$ we write:

$$\Sigma_k := \mathbf{I}_{p,q}\Delta^{-1}\left[\sum_{\ell=1}^{K} \pi_\ell \mathbf{C}_{k\ell}(\mathbf{X_B})_\ell(\mathbf{X_B})_\ell^\top\right]\Delta^{-1}\mathbf{I}_{p,q}$$

There is then some sequence of *indefinite orthogonal transformations*

$$\mathbf{Q}_n \in \mathbb{O}(p,q) := \left\{M \in \mathbb{R}^{d\times d} \mid M\mathbf{I}_{p,q}M^\top = \mathbf{I}_{p,q}\right\}$$

such that, for all $\mathbf{x} \in \mathbb{R}^d$:

$$\mathbf{P}\left\{n^{\frac{1}{2}}(\mathbf{X_A}\mathbf{Q}_n - \mathbf{X})_i^\top \leqslant \mathbf{x} \mid \mathbf{z}_i = k\right\} \longrightarrow \Phi(\mathbf{x}, \Sigma_k)$$

Where $\Phi(\mathbf{x}, \Sigma_k)$ denotes the CDF applied to $\mathbf{x}$ of $\mathcal{N}(0, \Sigma_k)$. In other words, the (realigned) spectral embedding of community k converges in distribution to $\mathcal{N}((\mathbf{X_B})_k, \Sigma_k/n)$.

Remark. This is quite a unique form of "central limit theorem", since the result is asymptotic as $n \to \infty$, but this n does not correspond to repeated independent samples, rather the size of the network and associated sequence of *orthogonal realignment matrices* used to coerce the spectral embedding of the observed matrix into a form compatible with the embedding of the expectation matrix. This is an immediate consequence of our observations in the previous section that we may only reason about orthonormal eigenbases up to transformation by (indefinite) orthogonal matrices.

This implies an immediate p-value combination algorithm:

1. Embed the adjacency matrix into $d = 2$ space
2. Fit a two-component Gaussian mixture model to the data and find the likelihood ratio of each point

Notice that this is fully adaptive: we need no knowledge of $\mathbf{Q}_n$ or even the alternative distribution to fit such a GMM under this CLT since the variance and realigned $\mathbf{B}$ embedding may be fitted by expectation-maximisation or variationally. As shown in [GJB$^+$23], an oracle with access to the true embeddings may compute Σ_k and $\mathbf{Q}_n$, but this is irrelevant to our setting. Assuming we have reached convergence in the CLT, we are now able to find the asymptotic efficiency of such a likelihood ratio test on the embedded GMM:

Corollary 2 ([GJB$^+$23])**.** *The size-adjusted Chernoff information between the two communities embedded as before is given by*

$$\lim_{n\to\infty} n^{-1}\mathcal{C} = \sup_{t\in(0,1)} \left[\frac{t(1-t)}{t}(\boldsymbol{e}_0 - \boldsymbol{e}_1)\boldsymbol{B}\Pi\boldsymbol{S}(t)^{-1}\boldsymbol{B}(\boldsymbol{e}_0 - \boldsymbol{e}_1)\right]$$

Where $\boldsymbol{S}(t) := (1-t)\,\mathrm{diag}(\boldsymbol{C}_0) + t\,\mathrm{diag}(\boldsymbol{C}_1)$ *is a convex combination of the diagonal matrices induced by rows of the block variance matrix.*

Note that this corollary uses the fact that Chernoff information is invariant under indefinite orthogonal transformation.

Remark (Extension to multigraphs)**.** Although we do not cover it here, [JRD21] provides a simple extension of the spectral embedding to multigraphs known as the *Unfolded Adjacency Spectral Embedding* and accompanying central limit theorem. In this case, the CLT is proven on unweighted graphs with richer community structure, but we conjecture that a similar limit theorem applies in the case of multigraphs. An alternative and potentially more robust approach is to simply combine the p-values with a trivial combiner as seen in Chapter 2 and use this as the single-edge p-value going forward since pairwise interactions no longer need to be considered.

SIMPLE for community recovery

This strong theoretical result yields a new lens through which we may analyse SIMPLE applied to the weighted SBM. We recall the SIMPLE statistic:

$$T_{ij} := \left[\hat{\mathbf{V}}(i) - \hat{\mathbf{V}}(j)\right]^\top \Sigma^{-1} \left[\hat{\mathbf{V}}(i) - \hat{\mathbf{V}}(j)\right]$$

Noting that this is simply the squared Mahalanobis distance between the observed $\hat{\mathbf{V}}(i) - \hat{\mathbf{V}}(j)$ and its distribution under the hypothesis that the nodes are colocated. The main result of [FFHL21] that allows for p-value threshold tests using the SIMPLE test statistic on WSBMs can now be reduced to a simpler conjecture which extends the CLT:

Conjecture. Under the previous regularity assumptions for a K-community weighted SBM, the SIMPLE test statistic comparing nodes i,j with the same community membership will follow

$$T_{ij} \xrightarrow{\mathrm{d}} \chi^2_K$$

Proof. Under the hypothesis that i and j belong to the same community, their embedding distributions will be centered around the same point, thus $\mathbb{E}[\hat{\mathbf{V}}(i) - \hat{\mathbf{V}}(j)] = 0$. On the assumption that the difference in the embedded representations is also distributed according to a Gaussian (implied by a more general conjecture of asymptotic independence of the representations), we can conclude by a well-known result: the squared Mahalanobis distance of a sample from its multidimensional Gaussian distribution is distributed under χ^2_d, and since we take $d = K$, the result is immediate since only convergence in distribution is required. □

We note that the original paper provided this result on a model which allows for a richer probabilistic community structure, but does not permit weighted edges. Different regularity conditions are also imposed as their proof proceeds differently, so it is difficult to assert that this theorem is a particular generalisation or specialisation, rather a sister result that provides a theoretical backing for using SIMPLE on a model it was not initially developed for.

In a similar vein, we may consider a statistic for testing membership in a specific community by finding the Mahalanobis distance from the distribution under the hypothesis that the node is distributed according to this community. The CLT gives this result trivially by finding

$$T_b(i) := [(\mathbf{X_A}\mathbf{Q}_n)_i - (\mathbf{X_B})_b]^\top \Sigma_b^{-1} [(\mathbf{X_A}\mathbf{Q}_n)_i - (\mathbf{X_B})_b]$$

Or where the mean and variance are fitted from the observations. Indeed, the squared Mahalanobis distance on Gaussians is, up to addition of a constant term, exactly the negative log-likelihood of an observation. From this we may conclude that the likelihood ratio test described in the previous section coincides almost exactly with a SIMPLE-style test of community membership, differing only in the addition of the terms $\log \det(2\pi\Sigma_b)$ on the numerator and denominator of the computed ratio. By the Neyman-Pearson lemma therefore, a SIMPLE-style test cannot asymptotically outperform the previously described method, and a simulation study suggests that it provides no advantage in the pre-convergence regime.

3.5.2 Spectral Partition and Optimal Estimators

Until this point, methods we have drawn from the literature have been focused on marginalisation, but we turn now to proposals for asymptotically exact estimation, adapt them to our setting, and modify them to emit confidence levels as opposed to a single label. These confidence levels may not be “p-values” in the traditional sense with a tractable distribution, rather test statistics that enable a level-power flexibility in false-positive/false-negative rates. Much attention has been focused on community recovery in the unweighted SBM and information-theoretic upper bounds on accuracy, along with algorithms that attain them, have been developed - see [Abb23] for a modern survey. We note that an estimator that is *optimal* in that it asymptotically solves exact recovery whenever information-theoretically possible is not necessarily *rate-optimal* since this places no bounds on the speed of convergence to exactness. The theory is less well-developed for the weighted setting, but we have two important results to work from:

- Yun and Proutiere [YP16] establish an information-theoretic bound on the expected asymptotic error in the Labelled SBM and devise a *spectral partition* algorithm which attains it.
- Xu et al. [XJL18] establish a similar risk bound and spectral algorithm which achieves it in the *homogenous* (i.e. symmetric) weighted SBM case. In particular, this risk lower bound is proven on the *balanced* case, where the two communities are of approximately equal size. The attaining algorithm is derived largely as an extension of the previous paper by first discretizing the model to convert it to a labelled network (albeit with a coarseness level that is dependent on network size), and they conjecture that a similar discretization argument can be applied to extend Yun and Proutiere's result to more general cases. We will not focus on the details of this algorithm, but use their result as justification for the claim that preserving continuity of the edge weights is uneccessary.

Notably, no such result exists in the literature that we know of for the *weighted*, *unbalanced* and potentially *heterogenous* model that arises in our setting. In this paper, we do not seek to close this question, rather use these algorithms and associated theoretical guarantees as a template for our p-value combination methods in the hope that statistically-principled proofs of optimal accuracy might be within future reach.

We provide an outline of the spectral partition algorithm of Yun and Proutiere, a structure mimicked by Xu et al.

1. **Initialisation** adds noise to the network and trims (a vanishing proportion of) the highest-degree nodes which have an outsized impact on estimation error. In Xu et al. this notably also includes uniformly binning the weights
2. **Spectral clustering** finds a rough initial assignment on the nodes

3. **Parameter estimation** based on the rough clustering is included as the algorithm's optimality does not depend on knowledge of the label distributions. The estimates are computed in the simplest possible way, analogous to the CAVI M-step, as the empirical probabilities given the rough clustering.

4. **Refinement** is the most novel aspect of the algorithm. For $\log(n)$ iterations, the rough clustering is *refined* by choosing the clustering in time step $t+1$ as the MLE w.r.t. the clustering found in time step t

The proof of optimal risk then proceeds by first showing that the spectral clustering finds a first approximation with asymptotically bounded error, and given these error bounds the refinement step will ensure exact recovery.

Steps 3 and 4 of this algorithm bear a striking resemblance to expectation-maximisation, and indeed represent a modified version of the CAVI algorithm described in section 3.2. In particular:

- Step (2) corresponds to the first E-step, where initial estimates for the latent community variables are proposed
- Step (3) corresponds to a single M-step to estimate unknown parameters
- Step (4) corresponds to CAVI where the variational distribution fitted is a Dirac delta around the proposed parameters - hence $\mathbb{E}_{\psi_{-i}^{(t)}}[\log p(\mathbf{z}_i \mid \mathbf{z}_{-i}, \mathbf{A})] = \log p(\mathbf{z}_i \mid \mathbf{z}_{-i}^{(t)}, \mathbf{A})$. The lack of an expectation term also implies a resemblance to Gibbs sampling - indeed, this is a Gibbs sampler with a thresholded distribution having mass of 1 at the MLE. We refer to this method going forward as **SP-refine**.

Stripped away from its information-theoretic context then, this algorithm is a special case of ones we have seen before - it is non-Bayesian variational inference with a spectral prior. This simplified version of the process is, furthermore, sufficient to ensure estimator optimality on the LSBM.

The simplicity of the algorithm allows us to trivially extend it to the Bayesian case by considering any other MCMC/VI methods which will fit an estimated distribution on community assignments rather than point estimates, and we conjecture that a similar style of optimality result will hold thanks to the strong theoretical guarantees of these methods to converge eventually to the true marginal distribution. The spectral prior seeks then to shortcut the warm-up period experienced by these algorithms before they converge to the typical set of the sampling distribution. The refinement step of [XJL18] takes a non-iterative approach:

1. A rough spectral clustering σ_u is produced at each node u by excluding it from the graph and clustering the remainder
2. $\sigma_u(u)$ is then set as the MLE w.r.t. the σ_u clustering on $\mathbf{A} \setminus \{u\}$
3. A consensus step then occurs where a distinguished node r is chosen arbitrarily and the assignment b of the node u is chosen to maximise the overlap between the nodes which σ_r assigns to b and the nodes colocated with u according to σ_u.

Whilst formally interesting, we will not consider this consensus method in our simulation studies, and instead use SP-refine as the point estimate benchmark due to the intense computational cost of this procedure on any reasonably-sized graph (although optimisations are certainly possible).

Remark (Noising)**.** The procedure of Yun and Proutiere does not add noise explicitly, but rather chooses a value $\omega_\ell \sim \mathcal{U}[0,1]$ to represent each label in the network. Whilst the nondeterminism this introduces is formally useful, it is a consequence of the discrete edge labels having no intrinsic numerical meaning. On the other hand, in our setting, we already have each edge weight taking a value in $[0,1]$ and this value directly carries information about the community of the endpoints, so this step is unnecessary. Similarly, Xu et al. introduce noise on the discretized network by replacing each edge label with one drawn uniformly at random with probability $\frac{2(L+1)}{N}$ (where L is the number of labels). This is done for technical reasons in the proof to ensure each label occurs with probability at least $2/N$ and numerical studies do not support any evidence that this improves performance. Accordingly, we will not consider the addition of noise in our spectral prediction pipeline.

Spectral clustering details

The proposals for spectral clustering in these papers are markedly different from the algorithms we have presented so far - in particular, they operate on the low-rank approximation matrix $\mathbf{X}_\mathbf{A}\mathbf{X}_\mathbf{A}^\top$ rather than the eigenvector matrix directly, and nonparametrically seek clusters in a non-iterative method which resembles a single-step approximation of k-medians. The Yun and Proutiere algorithm includes steps for estimating the number of communities where this is unknown and other differences are minimal, so we present only the Xu et al. version, termed **SP-cluster**, for simplicity:

1. First, embed each node i as $\hat{\mathbf{A}}_i$ where $\hat{\mathbf{A}}$ is the rank-d approximation of $\mathbf{A}$ obtained by SVD. As usual, we will specialise to our $d = K = 2$ setting.

2. For each node u, order the network by embedding distance. In particular, we define an ordering ν_u on the other nodes such that, for all $i \leqslant n$

$$\left\|\hat{\mathbf{A}}_u - \hat{\mathbf{A}}_{\nu_u(i)}\right\|_2 \leqslant \left\|\hat{\mathbf{A}}_u - \hat{\mathbf{A}}_{\nu_u(i+1)}\right\|_2$$

We define a notion of embedding centrality:

$$D_b(u) := \left\|\hat{\mathbf{A}}_u - \hat{\mathbf{A}}_{\nu_u(\lceil \mu_b N \rceil)}\right\|_2$$

In the original formulation, $\mu_b = \mu$ is a tuning parameter initially set to (in the $K = 2$ case) half the size of the smallest community. Since we are assuming unevenly-sized clusters in our setting, we add a community-dependence and propose $\mu_b = \frac{\pi_b}{2}$

3. Choose $u_0 := \arg\min_u D_0(u)$ and

$$u_1 = \arg\max_{\tilde{u}} \left\|\hat{\mathbf{A}}_{\tilde{u}} - \hat{\mathbf{A}}_{u_0}\right\|_2$$

Where $\tilde{u}$ ranges over the $(1-\mu)$-quantile of $D_1(\tilde{u})$.

4. Finally, assign

$$\mathbf{z}_v = \arg\min_i \left\|\hat{\mathbf{A}}_v - \hat{\mathbf{A}}_{u_i}\right\|_2$$

If the mode of $\mathbf{z}$ is 1 but $\pi_1 < \pi_0$, the community assignments in $\mathbf{z}$ are flipped in order to be consistent with the influence of the external field

It is clear that this algorithm is chosen to be efficient and more effective clusterings could be found with greater computational cost, but it is *good enough* to achieve the optimality results when combined with a refinement step. In Chapter 4 we show that the GMM procedure is usually more effective at prior generation, whilst also giving explicit likelihood ratios which may be used as Bayesian priors. This is supported by the observation from [GJB+23] that embedded clusters are often non-spherical but any method based on Euclidean distance minimisation cannot account for this.

An important aspect of the optimality proofs in both papers is that the method by which an initial clustering was found does not affect the results as long as it satisfies the same asymptotic accuracy bounds as the above method, so we are able to substitute any initial guess method we can show asymptotically improves this procedure without losing the theoretical guarantees of the overall algorithm.

3.6 The spectral pipeline

At last, we reach the proposed **spectral pipeline** for p-value combination in networks, unifying every approach discussed so far:

- **(Operator selection)** First, a linear operator $\mathbf{G}$ is chosen to represent the network. In our version, this should be an adjacency matrix of a WSBM to take advantage of the central limit theorem, but the edge weights may be transformed to maximise the CI between embedded communities:

  - A first choice is $\log(\mathbf{A})$. Intuitively, the spectral embedding is a linear transformation but probabilities are combined multiplicatively so this provides a more concrete link between the weighted geometric average suitable for probabilities and the weighted arithmetic average obtained by a linear operator. This is precisely the adaptation of *Fisher's Method* for p-value combination
  - We recall that the alternative edge distribution is, by construction, monotone nonincreasing and stochastically smaller than uniform. Accordingly, the greatest divergence between the two communities will occur in the extremes, so to maximise the difference in block mean and variance matrices and hence embedded CI, we wish to choose a transformation that reflects this. Both *George's Method* - $\log \frac{\mathbf{A}}{1-\mathbf{A}}$ - and *Stouffer's Method* - $\Phi^{-1}(\mathbf{A})$ - are clear candidates, both having the correct shape (where computations are done elementwise).
  - Finally, if the alternative density p_1 is known, the log-likelihood $\log p_1(\mathbf{A})$ is a natural choice since, by the Neyman-Pearson lemma, $p_1(\mathbf{A})^{-1}$ maximises the order separation between uniformly- and alternatively-distributed elements of $\mathbf{A}$, and this is a monotone transformation of that test statistic which places more weight on the extreme values.

- **(Spectral prior)** We find an orthonormal eigenvalue decomposition of $\mathbf{G}$ and truncate it to rank d, yielding $\mathbf{X_G}$, and embed each node as its corresponding row in either $\mathbf{X_G}$ or $\mathbf{X_G X_G^\top}$. This embedding is turned into a prior:

  - Following the analysis of the spectrum of $\mathbb{E}\mathbf{A}$ from the beginning of the section, we may find the prior directly by looking at the eigenvector corresponding to the second-highest eigenvalue. [CRV15] proposes an algorithm achieving weak consistency on the unweighted homogenous SBM which first projects the all-ones vector onto the rank-2 eigenspace and finds the unit vector orthogonal to this projection, using this as a surrogate for the second eigenvector of the expectation matrix via a robustness result from the Davis-Kahan theorem. Adapting to our (unbalanced) setting, the smallest πN elements from this vector are then chosen and their corresponding nodes marked as anomalous. The robustness however comes from the fact that the all-ones vector is guaranteed to be the first eigenvector of the expectation matrix in the homogenous setting, but this is not the case for the heterogenous case we are interested in. If the alternative expectation is known, we can simply compute the first eigenvector of the expectation matrix and proceed in the same way, however this method cannot be adapted to the heterogenous case if the alternative mean is unknown.
  - If we are interested only in an initial point estimate we may cluster the low rank approximation by either k-medians or the SP scheme
  - If we are using an elementwise transformation of the adjacency matrix, the CLT is available and we perform the aforementioned GMM clustering. If a point estimate is needed, a likelihood ratio test is used, otherwise the likelihood ratio is converted into a prior distribution on the marginals

- **(Refinement)** Once a prior σ is obtained, we may proceed by any Markov chain or variational method. Since our primary goal is “p-value combination", we do not consider producing point estimates, but the SP or consensus approach may be used unmodified if this is desired.

  - The predicted initial marginals may be used directly to initialize mean field variational inference or (circular) belief propagation. This can be used as part of an E-M procedure if the alternative distribution parameters are unknown. GBP methods may also be used, but these are excluded from our simulation studies due to their high computational cost and it is unlikely that the accuracy advantage they give is relevant to the refinement step.
  - With known alternative distribution, the point estimate obtained from the spectral prior may be used as an initial state for any MCMC method: M-H, Gibbs S-W, or ICM. Whilst population annealing can be very powerful for cold-start inference, we focus on ICM+PT as it is better adjusted to refinement from a prior since a high-temperature “exploration" stage can damage an otherwise good prior.

This framework may seem very general with multiple components to choose from at each step, but we emphasise that the NP-hard nature of our problem and generality of the PVN model makes it impossible to declare a single algorithm the “best” in the general case, especially when different use-cases have different success criteria. The shape of the alternative distribution and network topology will both influence the optimality of different approaches, so we encourage application-specific experimentation. The next section will provide rigorous numerical comparisons of various configurations applied to synthetic and real-world data and develop a suite of suggested “canonical” pipelines for different general classes of network.

Chapter 4

Simulation Studies

We recall the proposed components of the spectral pipeline:

Initial Transforms	**Clusterings**	**Refinement**
Log-likelihood	Direct eigenvector	Gibbs sampler
Stouffer's method	Gaussian Mixture	Swendsen-Wang
George's method	K-Means	SP-refine
Fisher's method	SP-cluster	ICM+PT
		CAVI
		(Circular) BP

Table 4.1: Spectral pipeline components

The multi-stage nature of the process suggests that we may evaluate each of the three steps individually, in particular:

- The effectiveness of adjacency matrix transforms may be evaluated analytically using Corollary 2. Since Chernoff Information measures the asymptotic performance of *any* combiner, we are justified in our assertion that the most effective transform for a particular model will be the one attaining the highest embedded CI, regardless of latter steps in the pipeline.

- Similarly, the performance of a rough clustering algorithm is a strong proxy for the distance between generated marginals (resp. point estimates) and the true values (resp. true ground states of the Hamiltonian), and this divergence from the typical set is the most important metric for prior selection at the refinement stage. We may accordingly choose a "best" clustering for a particular model and then apply this same clustering method for evaluating each of the refinement steps.

Along with heterogeneity structure, and the choice of alternative distribution, we are interested in evaluating the performance of different methods on different graph topologies. To this end, we consider three important graph statistics:

- **(Sparsity)** This is simply d/N, where d is the average degree of the nodes and N the size of the network. We may generate networks with a given sparsity and no other structure by the Erdős-Rèyni uniform random sampling process.

- **(Modularity)** Many real-world networks are strongly *modular*, with densely-connected communities and comparatively few inter-community connections [BP16] (recall the example given when describing generalised belief propagation). Suppose we have a partition of the network's nodes into $\{V_1, ..., V_m\}$. The **modularity** score is defined as the proportion of intra-community edges in the network, minus the proportion we would expect in an Erdős-Rèyni graph of the same parameters. Note that this notion of "community" is purely in relation to the graph topology and does not necessarily have any relation to the SBM communities.

$$Q_{\{V_i\}} := \frac{1}{2|E|} \sum_{i,j \in V} \delta_{c_i c_j} \left(A_{ij} - \frac{d_i d_j}{2|E|} \right)$$

Where d_i is the degree of node i and c_i its community. With no *a priori* partitioning, the modularity is defined as the maximum over all possible partitions, but we will generate modular graphs from disconnected community sets with the following procedure:

1. Let N be the intended number of nodes, M the number of edges, and c the number of communities. Let $p \in [0, 1]$ be a *rewiring probability* that controls the modularity level
2. Generate an $(\frac{N}{c}, \frac{M}{c})$ Erdős-Rèyni graph G_i for each of the c communities and let $G = \bigcup_i G_i$
3. For each edge (i, j) in these initial partitions, with probability p, choose a node h uniformly at random which does not share a community with i or j and replace the edge with (i, h).
4. Return G and compute Q with the ground truth partitions used for generation.

- **(Small-Worldness)** Intuitively, a *small-world* network is one which is sparse overall, but the neighbourhood of each node is densely-connected, and there is a relatively short path between any two nodes. This is measured by the **small-world index**, which is informally the ratio (clustering level)/(mean path length), divided by the expectation of the same ratio in an Erdős-Rèyni graph. Formally, we define the **local clustering coefficient** at a node (where $\mathbf{N}(i)$ denotes the neighbourhood of i):

$$\gamma_i := \frac{|\{(j, k) \in E \mid j, k \in \mathbf{N}(i)\}|}{\frac{1}{2} d_i(d_i - 1)}$$

We may take the mean as the **global clustering coefficient** $\gamma = \frac{1}{N} \sum_i \gamma_i$. We also define λ to be the mean length of the shortest path between any two sets of nodes. The small-world index is then defined:

$$\sigma := \frac{\gamma}{\lambda} \cdot \frac{N \log N}{d \log d}$$

Where elementary results from graph theory show that $\frac{d \log d}{N \log N}$ is the expected ratio in an E-R graph [WS98]. The canonical way of generating small-world networks is the **Watts-Strogatz** algorithm:

1. Create a ring lattice graph with N nodes and constant degree d (where d is even). This is simply the graph where nodes are arranged in a circle and connected to their d nearest-neighbours.
2. For each edge (i, j) in the ring lattice, with probability p, remove it and replace it with (i, h), where $h \notin \{i, j\}$ is chosen uniformly at random.

Generating graphs with these algorithms and varying the key statistics will thus allow us to gain insight into which methods perform the strongest in a wide range of real-world contexts.

4.1 Choice of linear operator

We first analyse adjacency matrix transforms in terms of their Chernoff information via Corollary 2. This result greatly simplifies the comparison process: since this is entirely decided by properties of the distribution, the only topology statistic which has an effect is the sparsity, so without loss of generality we consider only Erdős-Rèyni graphs. Following [HRD18], we will focus on alternative distributions of the form (all truncated to $[0, 1]$):

- $\text{Beta}_{[0,1]}(a, b)$ for $a \in (0, 1], b \in [1, \infty)$
- $\Gamma_{[0,1]}(a)$ for $a \in (0, 1], b \in [0, \infty)$. Note that this includes χ^2_2 (appearing in the SIMPLE test) as a special case where $a = 1$.
- An important special case is $\text{ChiPVal}(k, \mu)$, the alternative distribution of p-values when testing $H_0 \sim \chi^2_k$ against $H_1 \sim \chi^2_k(\mu)$, the non-central chi-squared distribution with mean μ - this arises when p-value testing with SIMPLE. To our knowledge, no closed form of this distribution exists since $\chi^2_2(\mu)$ has a CDF defined as an infinite series of gamma functions, but we conjecture it is (approximately if not exactly) a beta distribution so do not test it separately. Figure 4.1 shows histograms for a representative sample of μ parameters, along with beta distributions which closely match the shape.

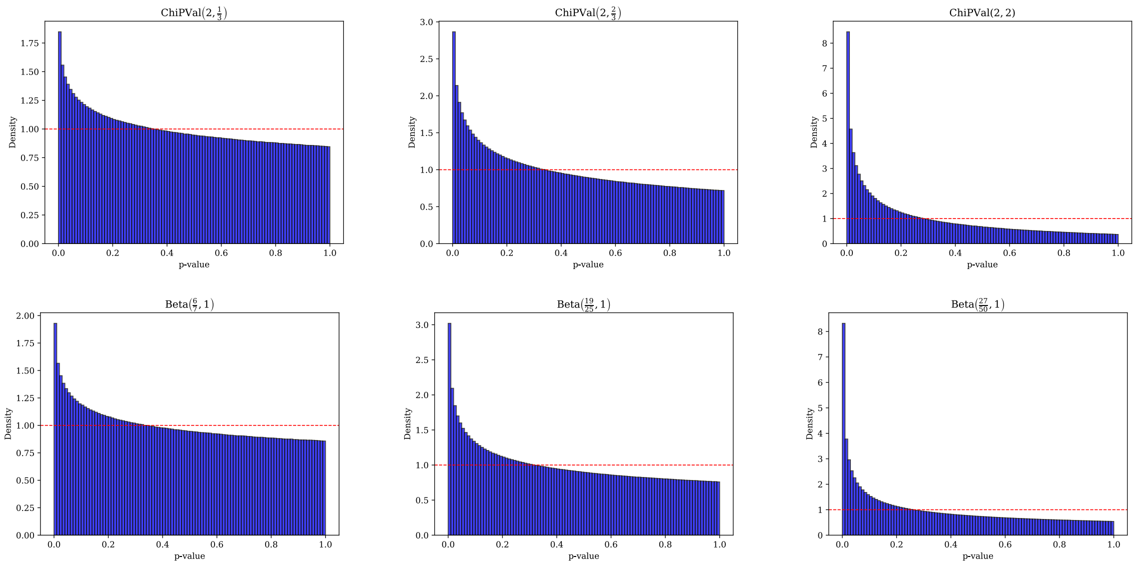


Figure 4.1: Histograms of ChiPVal$(2, \mu)$ (top row) and Beta$(a, 1)$ (bottom). Each was generated from 10^9 samples.

Sparsity manifests itself simply by introducing a Dirac delta to the density functions. Let ρ be the probability of an edge existing and p the unmodified weight density - then the in-network density is the zero-inflated distribution:

$$(1-\rho)\delta_0 + \rho f$$

In the case where average degree is constant among communities, an unobserved edge carries no useful information, so we minimize its contribution by setting any unobserved edge to zero in $T(\mathbf{A})$, *after* performing the transformation T. It is important to note that each of our refinement methods automatically compensate for unobserved edges by construction since they are just excluded from any computations, so it is only the spectral embedding step where the representation of an unobserved edge has an effect.

Figure (4.4) illustrates the optimal transforms in terms of embedded Chernoff information for representative parameter choices of $\Gamma_{[0,1]}(a, b)$ alternative distributions and figure (4.5) shows the same for the Beta$_{[0,1]}(a, b)$ alternative. While our regions do coincide to some extent with the same plots found in [GJB$^+$23] for the Beta alternative, we notice that the majority of the plots are dominated by either the log-likelihood or Stouffer transforms, neither of which were considered by the authors of that paper. Similarly, for Γ, we see our intuition that the transform should be selected which maximises the impact of extreme values (whether close to 0 or 1) is correct in most regimes. The plots found in [GJB$^+$23] may suggest that thresholding schemes are often the optimal choice, but these plots considering a wider variety of transforms show this to be misleading. As expected, there is a critical sparsity level before which thresholding schemes perform poorly, since this can lead to nodes being disconnected or otherwise lacking enough information for accurate inference. We can see that this critical sparsity is lower in balanced, symmetric networks - to be expected since this regime intuitively maximises the combined cohesion of the two communities, so removing a large proportion of edges is most likely to yield two dense intra-community clusters. Finally, we see that thresholding schemes are optimal only in the $a \leqslant 0.1$ band. When we consider the means of the two distributions (Figure 4.2), this becomes intuitively clear - “all-or-nothing" thresholding schemes are only effective when the alternative distribution has very high divergence from uniform. Especially where $\mathbb{E}[p_1] \approx \frac{1}{2}$, they are of little use, and a scheme such as Stouffer/George’s method which can more effectively differentiate the shape of the distributions is more effective.

We omit further phase diagrams for brevity, but note that if log-likelihood is excluded (as in the case of unknown alternative distribution), the beta distribution landscapes closely resemble the ones for gamma and are dominated by Stouffer’s/George’s method. Furthermore, the difference between these methods is rarely large. Two simple line plots of ρ against size-adjusted CI can be seen in Figure (4.3) which illustrate these observations.

In conclusion:

- Where the alternative distribution is unknown, one should use the Stouffer or George transforms
- Where the alternative distribution is known and not a case covered by these plots:
  - If the graph is complete or nearly complete ($\rho > 0.95$), the log-likelihood transform should be considered.
  - If $\mu_1 \ll 0.5$ ($\mu_1 \leqslant 0.05$), a threshold scheme at $\tau = 0.01$ should be considered.
  - In all other cases, the Stouffer or George transforms should also be used.

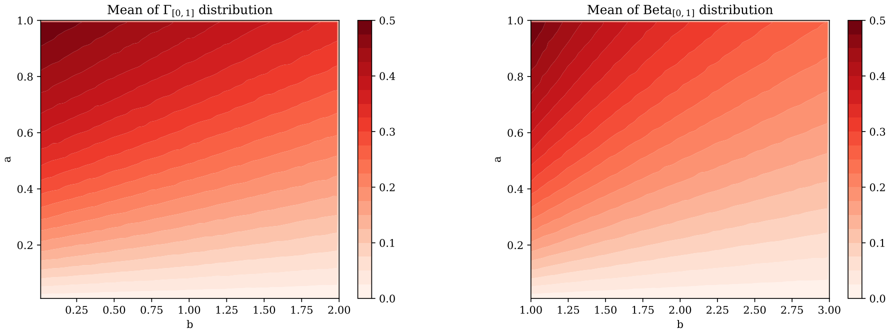


Figure 4.2: Means of the truncated alternative distributions over parameter range considered.

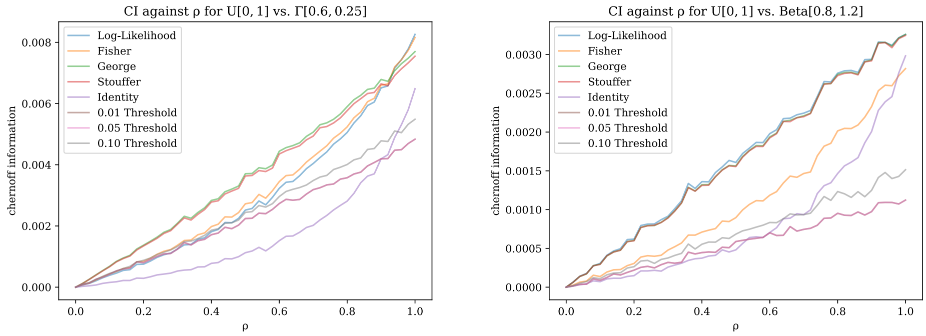


Figure 4.3: Line plots showing size-adjusted CI against ρ for Gamma (left) and Beta (right) alternative distributions. Notice on the Beta plot that the George and Stouffer method lines coincide almost exactly, along with the 0.01 and 0.05 threshold lines.

4.2 Choice of clustering algorithm

Moving from matrix operations (applied in origin space) to clustering algorithms (applied in the embedded space), the sensitivity of our choices to specific parameters of the distribution is reduced. Instead, we focus on evaluating the accuracy of the algorithms as more general statistics are varied:

- The mean and variance of the underlying alternative distributions - in particular the difference between μ_1 and μ_0.
- The Chernoff information of the embedded points to cluster
- Size and sparsity of the underlying network

The performance of the different clustering choices are illustrated for a representative selection of parameters. We consider two performance metrics: accuracy and component-wise sum of **Hellinger distance** from the proposed marginal distribution to the ground truth. Hellinger distance is defined, for discrete probability distributions P, Q on $\Omega = \{\omega_1, ..., \omega_n\}$:

$$D_H(P,Q) := \frac{1}{\sqrt{2}}\left[\sum_{i=1}^{n}\left(\sqrt{P(\omega_i)} - \sqrt{Q(\omega_i)}\right)^2\right]^{\frac{1}{2}}$$

To produce an accuracy value from methods which provide a likelihood ratio or marginal, we simply order the elements of the prior and choose the greatest πN of them as the predicted anomalies - note that this bounds the error rate of random guessing at $2\pi(1-\pi)$ since the threshold is not fixed. In each of these experiments, the adjacency remains un-transformed.

- 10 samples of $N = 1000$ node networks were drawn with $\pi = 0.2, \rho = 1$ from $p_1 = \Gamma_{[0,1]}(1, b)$, where b is varied along $[0.2, 2]$ to change the mean. Comparative plots of the accuracy and Hellinger distance from truth of each algorithm are shown in Figure (4.7).
- 10 samples of $N = 1000$ node networks were then drawn with $\pi = 0.2$ from $p_1 = \Gamma_{[0,1]}(1, 3/2)$, where ρ is varied along $(0, 1]$. This selection of Γ parameters was chosen as it yields $\mu_1 \approx 0.42$, representative of a nontrivial but manageable regime.Comparative plots are shown in Figure (4.8).
- 10 samples of $N = 1000$ node networks were then drawn with $\rho = 0.5$ from $p_1 = \Gamma_{[0,1]}(1, 3/2)$, where π is varied along $[0.1, 0.5]$. This selection of Γ parameters was chosen as it yields $\mu_1 \approx 0.42$. Comparative plots are shown in Figure (4.9).
- The plot in Figure (4.6) combines the above three experiments. For each graph generated according to each choice of parameters, the Chernoff information of the embedding was computed and the combined performance plotted as a function of these CIs.

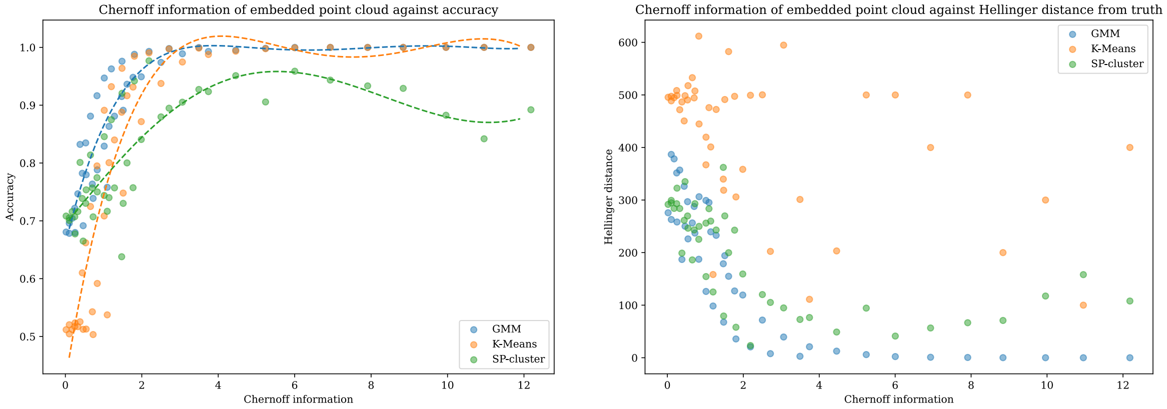


Figure 4.6: Accuracy and Hellinger distance line plots for clustering methods on the APVN($\Gamma_{[0,1]}(1, b)$) embedded space, with μ_1 varying and $N = 1000, \rho = 1, \pi = 0.2$

We observe a sharp phase transition in the plot of accuracy against mean (4.7) in the region $\mu_1 \in [0.4, 0.44]$ where the GMM, K-Means, and eigenvector methods move from perfect accuracy to indistinguishability (note that K-means is the only method which does not have the $2\pi(1-\pi) \approx 0.32$ error rate bound). This aligns with the similar phase transitions observed and well-studied in the unweighted SBM which occur as the probability of in-community and out-community edges grow close together ([DKMZ11b], [KMM$^+$13]). The plot (4.6) provides an explanation for this observed phase transition as a corollary of the phase transition from detectable to undetectable that occurs when the Chernoff information of the embedded point cloud drops below ≈ 2 - since each of the individually plotted parameters contribute to CI to some extent. Although the CI is not a perfect proxy for the rough cluster accuracy, we can see from the splines-of-best-fit that it approximates the phase transition very closely (the eigenvector method is excluded from this plot since it is more sensitive to specific parameters of the network). For the GMM and SP-cluster methods, a similar phase transition can be seen in the plot for Hellinger distance, where the trajectory curves sharply upwards at CI ≈ 2. This justifies why we do not transform the adjacency matrix for these experiments: the evidence strongly suggests that the accuracy of clustering methods (other than direct eigenvector) are largely sensitive only to the embedding CI (along with network size), whether this comes from network structure or parameters of the alternative distribution.

In conclusion:

- We may interpret the plots in figure (4.6) as suggesting that GMM is in most cases the superior choice for the rough cluster step, providing in most cases the best performance in terms of both accuracy and Hellinger distance (although the direct eigenvector method is excluded from this figure, the other plots suggest it is rarely a beneficial choice due to lack of robustness and inability to discern communities when $\mu_1 = \frac{1}{2}$). This is supported by the fact that it perfectly characterises the distribution as $N \to \infty$, inhibited only by the performance of the fitting algorithm (in this case, Expectation-Maximisation using the `GaussianMixture` class from `scikit-learn`).

- Figure (4.10) suggests that SP-cluster may be preferred where the network is small ($N \leqslant 750$), corresponding to a weaker central limit theorem. Similarly, suggested by figure (4.8) and confirmed by smaller experiments omitted for brevity, where the network is sparse ($\rho \leqslant 0.1$) but not too large ($N < 5,000$), SP-cluster should be preferred - although the difference is usually small. GMM should, however, be preferred for more balanced networks ($\pi \geqslant 0.3$).

4.3 Choice of refinement procedure

It is in this section that we will pay close attention to the topology of the graph as these algorithms take it explicitly into account. Our general procedure for testing will be the following:

- Sample an unweighted graph topology according to one of our generation methods and assign anomalous nodes

- For a representative sample of alternative distributions, weight the edges of this same topology

- Generate a rough clustering with the optimal method and run each of the refiners on the same clustering

Whilst Swendsen-Wang provides an important theoretical backdrop and optimized implementations can likely be fast, the algorithm we have derived for application to the APVN setting requires a large volume of non-trivial graph operations at each iteration, and is hence prohibitively slow for large-scale simulations. We exclude it from our analysis here, but preliminary small-scale experiments suggest it is significantly outperformed by ICM and even Gibbs sampling in the refinement setting. We note that the primary advantage of ICM/S-W is that the flips of large connected components allows for rapidly "tunnelling through" the energy landscape away from local minima, hence increasing the convergence rate of the Markov chain. In the refinement setting, however, this is mostly only important when the prior is of poor quality, so large-cluster moves occurring at lower temperatures and mixing up to $\beta = 1$ as in ICM is generally more effective than the S-W approach.

We begin by analysing the accuracy of each algorithm on networks with varying sparsity and no further structure. As usual, we generate 10 samples of $N = 1000$ APVN networks with $\pi = 0.2$. Figure (4.11) shows plots of accuracy and ROC-AUC score when sampled against $\Gamma_{[0,1]}(1, 0.4)$ - this is a fairly challenging regime with $\mu_1 \approx 0.47$. We see in this plot a sharp phase transition in accuracy from an undetectable baseline ≈ 0.68 to ≈ 0.78 at $\rho = 0.7$. Interestingly, such a phase transition does not occur for the ROC-AUC plot, where the evolution is much more gradual. That SP-refine lags behind in ROC-AUC is to be expected as it generates point estimates, but it is further important to note that it also performs worse on accuracy than MCMC-based methods. An interesting feature is exhibited after the phase transition, where the trajectory of each algorithm overlaps almost completely. This phenomenon is seen more clearly in Figure (4.12), sampled against $\Gamma_{[0,1]}(1, 0.9)$ - an easier regime with $\mu_1 \approx 0.42$, on the other side of the phase transition from Figure (4.7). Remarkably, there is little to no differentiation between the methods: SP-refine is the only outlier, falling slightly behind the other algorithms. Whilst this seems shocking, we draw comparisons to [KMM+13], [DKMZ11a], and [ZKRZ12], each of which studied the unweighted SBM and observed a similar phenomenon of different algorithms converging to the exact same accuracy trajectory as the inter- and intra-community edge probabilities vary. We conjecture, therefore, that this represents an information-theoretic rate-optimal limit in detection accuracy associated with ground states of the free energy, which each method has achieved independently.

An artefact of our graph generation is that, in order to increase run-to-run consistency of our metrics, *exactly* πN nodes will be marked as anomalous in each sample. Accordingly, marking the highest πN elements as anomalous will introduce a degree of uniformity between algorithms when computing accuracy statistics; this does not explain the convergent behaviour we see, however, since it would not carry over to the ROC-AUC scores if this was the case. An interesting feature we remark upon is the performance of (vanilla, loopy) BP being seemingly unaffected by graph sparsity. In our implementation, BP is run until the ELBO score (free energy) converges rather than a fixed number of iterations, so it may be the case that more challenging regimes will prevent BP from converging, but we have not observed this. For completeness, we fix $\rho = 0.5$ and plot in figure (4.13) the performance of each algorithm as μ_1 varies. We note that there is no pronounced phase transition in this case as may be predicted from the rough clustering performance.

Finally, the experiments on modular (4.14) and Wattz-Strogatz (4.15) small-world networks show less variation than expected, so this is unfortunately quite uninteresting. As may be expected, BP consistently outperforms other methods on both styles of modular network.
In conclusion:

- There is little differentiation amongst the iterative refinement methods. The fastest convergence overall is seen by BP in almost every circumstance, although especially in modular networks.
- One may choose an MCMC method if guaranteed convergence is a necessity, but otherwise one should err towards belief propagation regardless of parameter choice.

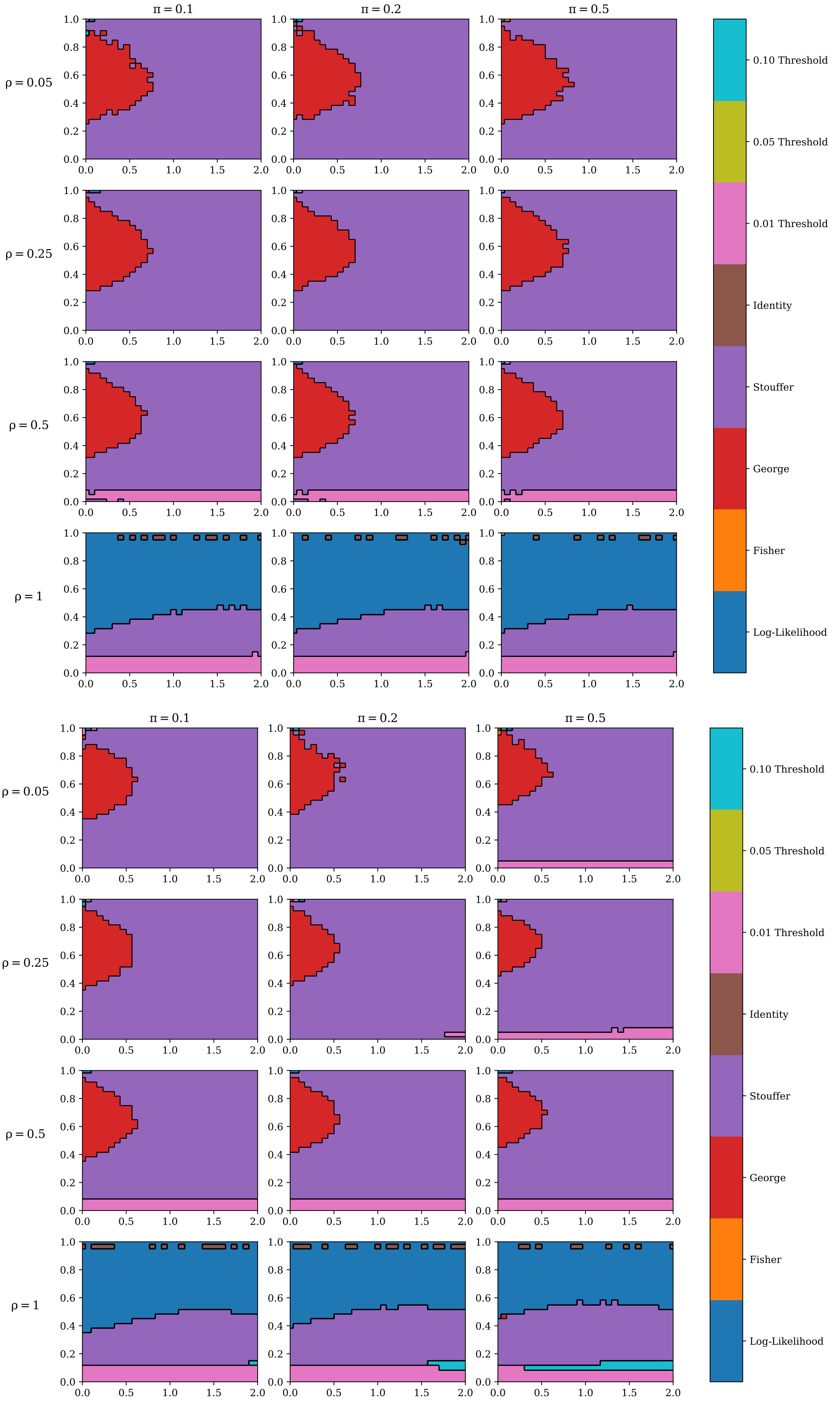


Figure 4.4: Phase diagrams of optimal adjacency transforms for asymmetric (top) and symmetric (bottom) p-value networks with $\Gamma_{[0,1]}(a, b)$ alternative distribution. The x-axis runs along b and the y-axis along a.

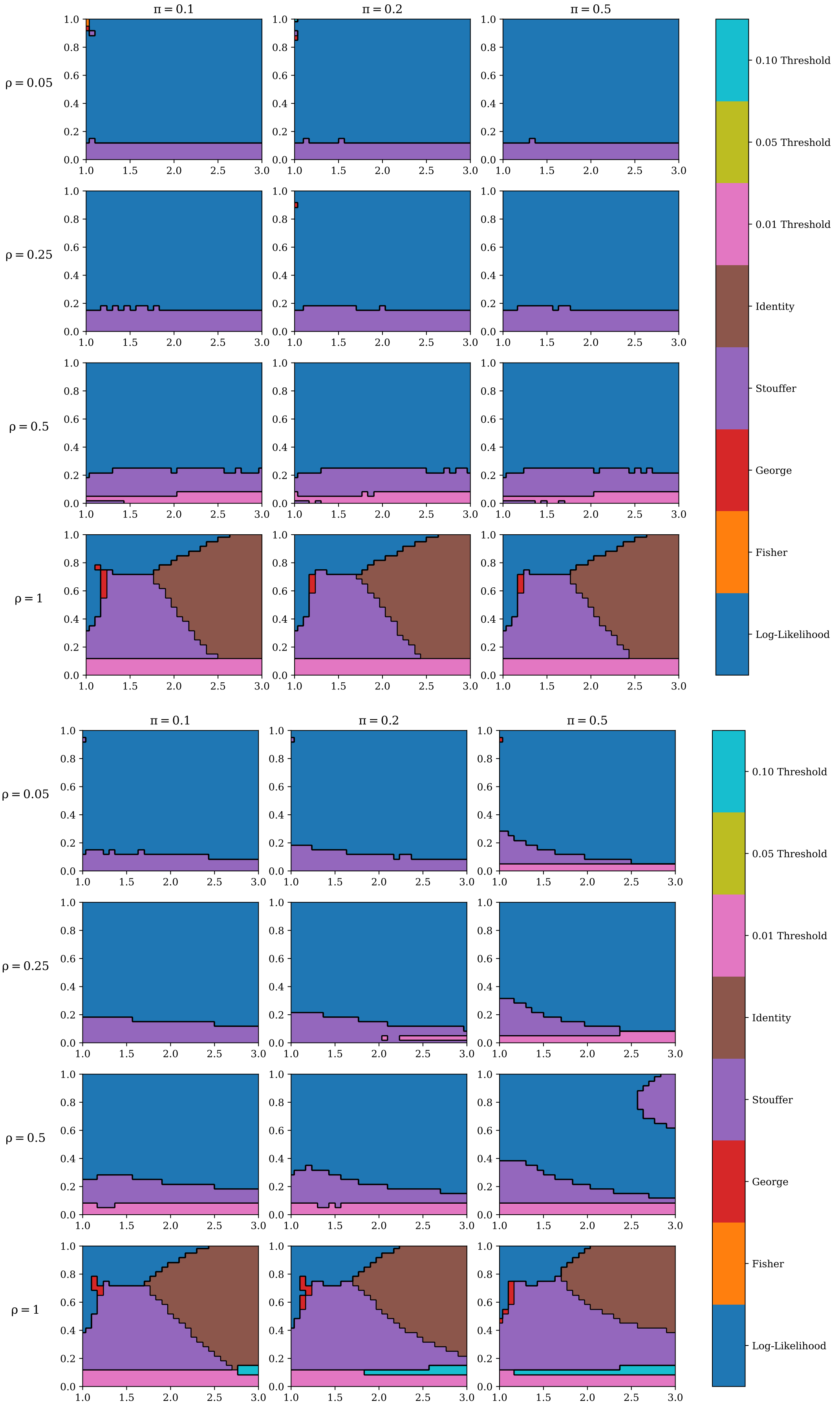


Figure 4.5: Phase diagrams of optimal adjacency transforms for asymmetric (top) and symmetric (bottom) p-value networks with $\text{Beta}_{[0,1]}(a, b)$ alternative distribution. The x-axis runs along b and the y-axis along a.

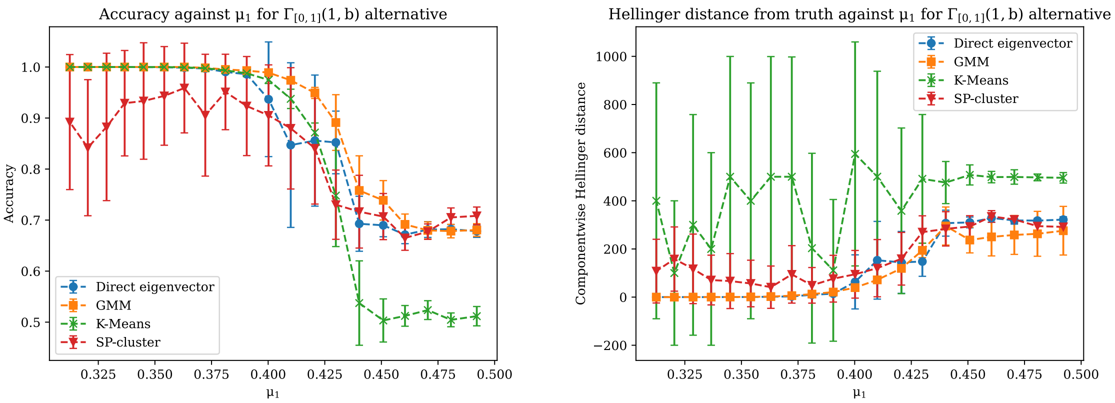


Figure 4.7: Accuracy and Hellinger distance line plots for clustering methods on the APVN($\Gamma_{[0,1]}(1, b)$) embedded space, with μ_1 varying and $N = 1000, \rho = 1, \pi = 0.2$

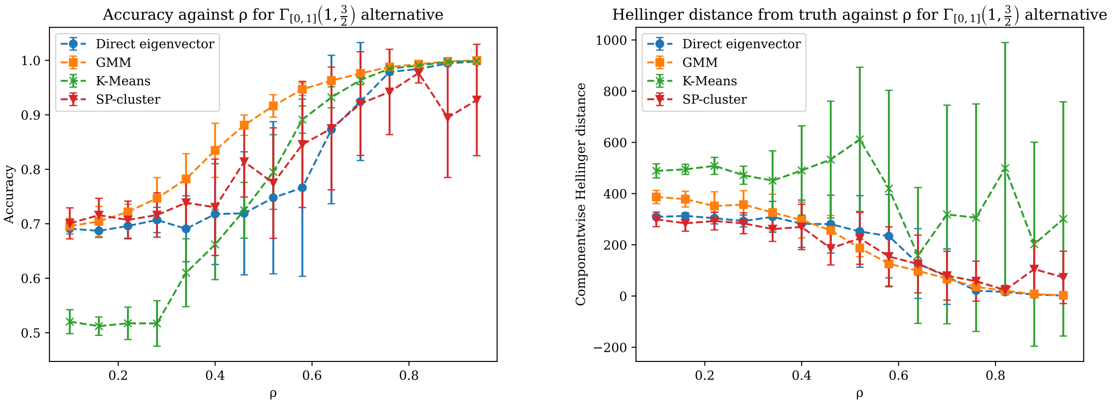


Figure 4.8: Accuracy and Hellinger distance line plots for clustering methods on the APVN($\Gamma_{[0,1]}(1, 3/2)$) embedded space, with ρ varying and $N = 1000, \pi = 0.2$

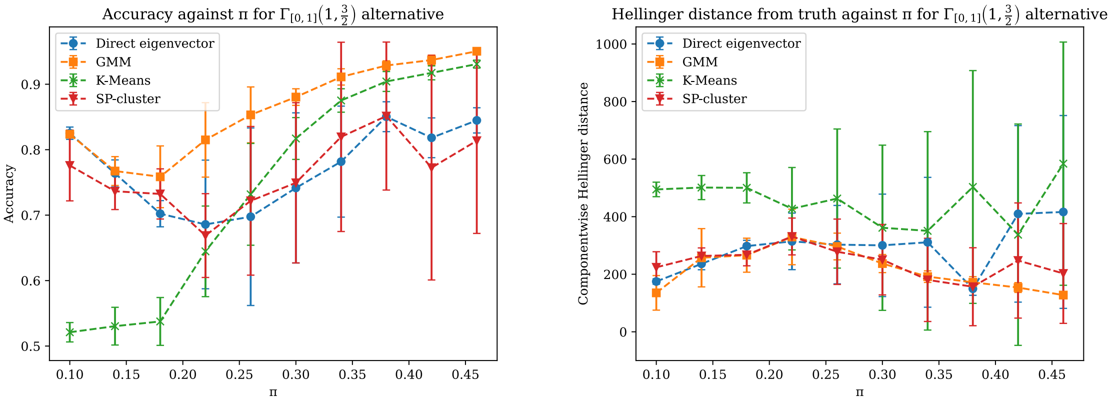


Figure 4.9: Accuracy and Hellinger distance line plots for clustering methods on the APVN($\Gamma_{[0,1]}(1, 3/2)$) embedded space, with π varying and $N = 1000, \rho = 0.5$

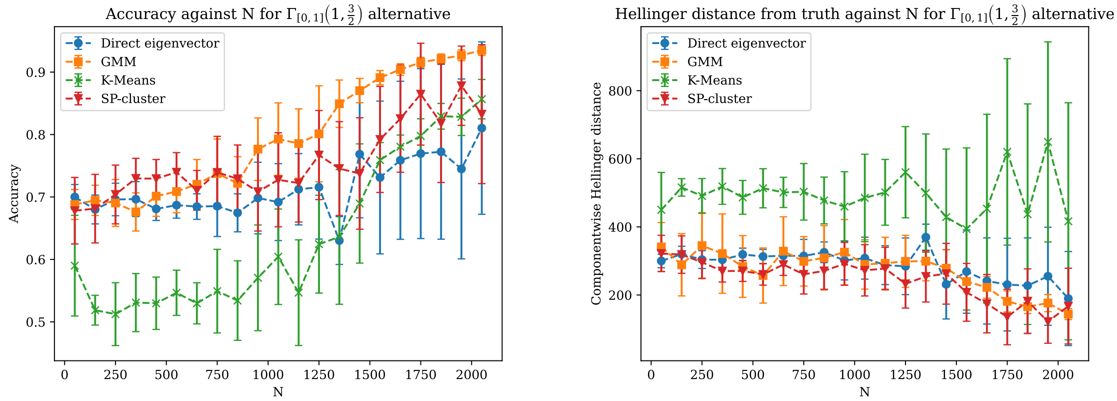


Figure 4.10: Accuracy and Hellinger distance line plots for clustering methods on the APVN($\Gamma_{[0,1]}(1, 3/2)$) embedded space, with N varying and $\rho = 0.5, \pi = 0.2$

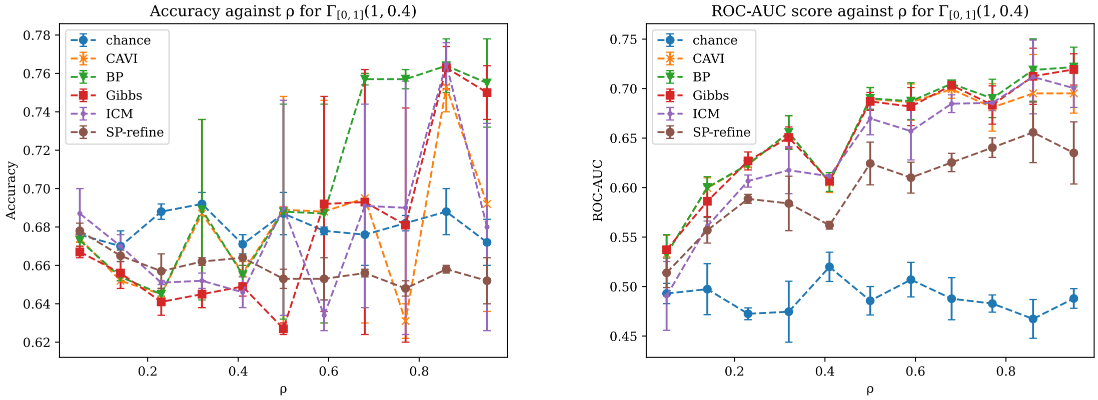


Figure 4.11: Accuracy and ROC-AUC line plots for refinement steps after optimal prior choice for APVN($\Gamma_{[0,1]}(1, 0.4)$), with ρ varying and $N = 1000, \pi = 0.2$

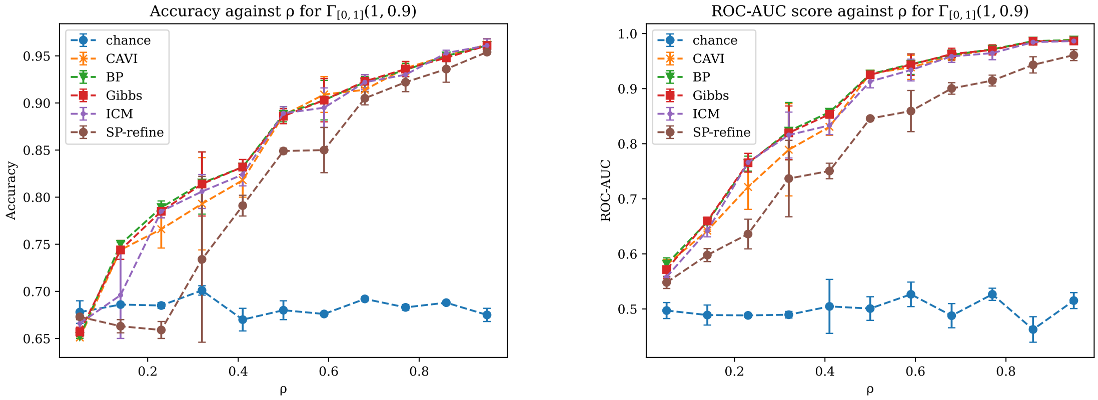


Figure 4.12: Accuracy and ROC-AUC line plots for refinement steps after optimal prior choice for APVN($\Gamma_{[0,1]}(1, 0.9)$), with ρ varying and $N = 1000, \pi = 0.2$

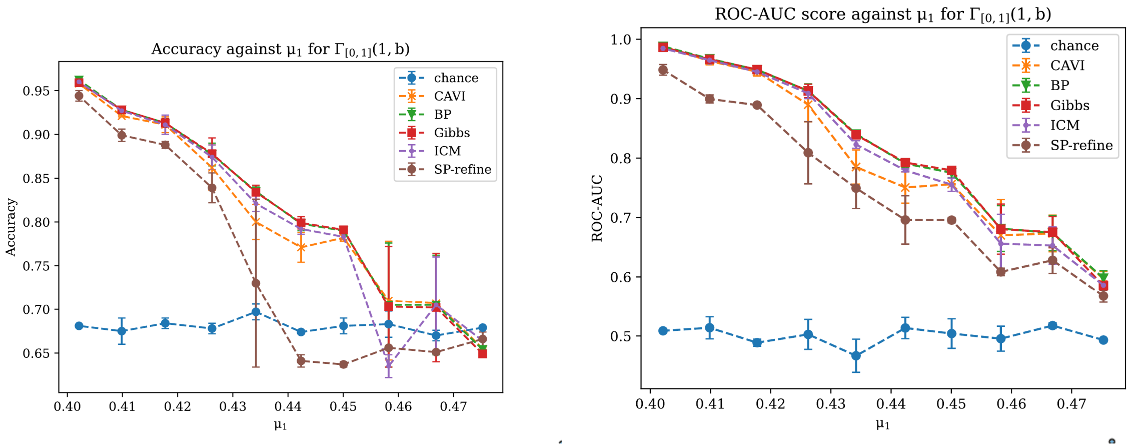


Figure 4.13: Accuracy and ROC-AUC line plots for refinement steps after optimal prior choice for APVN($\Gamma_{[0,1]}(1, b)$), with μ_1 varying and $\rho = 0.5, N = 1000, \pi = 0.2$

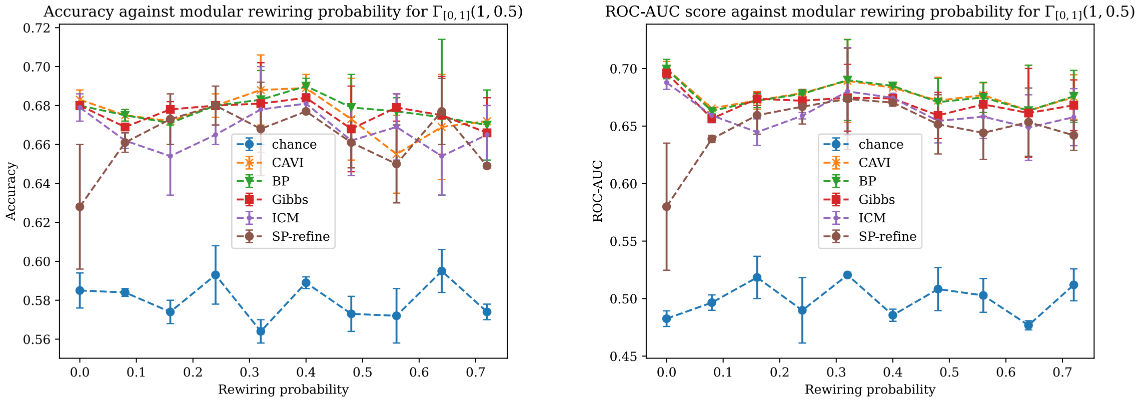


Figure 4.14: Accuracy and ROC-AUC line plots for refinement steps APVN($\Gamma_{[0,1]}(1, 0.5) = \chi_2^2$) on a 5-community modular network with $\rho \approx 0.2$ and $N = 1000, \pi = 0.3$

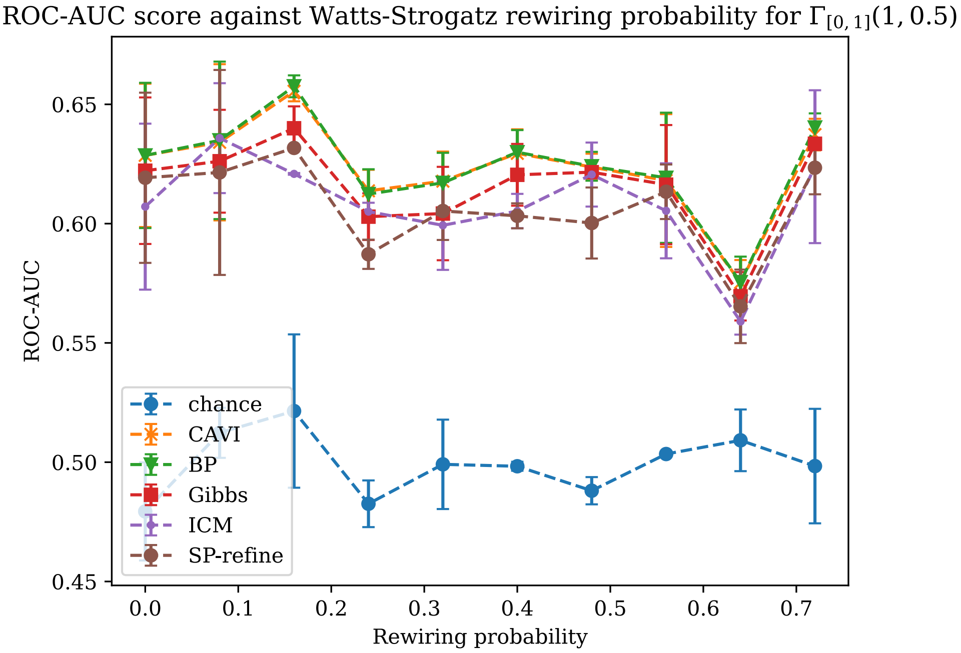


Figure 4.15: ROC-AUC line plots for refinement steps APVN($\Gamma_{[0,1]}(1, 0.5) = \chi_2^2$) on a Watts-Strogatz small-world network, $\rho \approx 0.2$ and $N = 1000, \pi = 0.3$

Chapter 5

Conclusion

In this project, we have given a thorough survey of the fields of both approximate inference in networks and p-value combination, with detours through spectral theory and statistical physics. We have introduced the generalisable and highly-performant spectral pipeline for community marginalisation and p-value combination in two-community weighted block models with arbitrary topology, achieving our initial goal of developing a flexible framework into which more research in this field can fit.

This is not to say we have answered every question, however, and indeed this work is more of a starting point that concluding remark. Particular directions for future work include:

- **Applications of the framework**
  As of yet, we have not applied our framework to real-world data, instead focusing on comprehensive simulation studies to account for a wide variety of possible datasets. Immediate applications may be found in cybersecurity and other anomaly detection procedures in networks, as evidenced by [HRD16], [PWTH18].

- **Strengthening theoretical guarantees**
  Despite guaranteed eventual convergence of any MCMC-based refinement, we do not yet have any guarantees of optimality of our method or analysis of asymptotic recovery potential in an estimation regime. We conjecture, due to the derivation of our framework from methodologies in which these optimality proofs do exist, that finding them (or slight modifications to the method which enables them) is within reach, and believe that this is a fruitful area of future research.

- **Generalisations to richer structures**
  We have been confined in the present analysis to p-value weighted *graphs*. Similar spectral approaches exist in literature for hypergraphs ([ALS18]), and we believe our framework can likely be adapted to provide p-value combination and community inference in these settings too. Furthermore, much interest recently has been generated in *topological data analysis* ([MHM20]) which focuses on simplicial complexes. Since networks are simply 1-dimensional simplicial complexes and community inference is closely linked to the cycle structure of the graph (the object of study in *(simplicial) homology theory*), we believe similar approaches to be adaptable to more general topological settings.

5.0.1 Ethical considerations

We note that, although we do not explicitly work with any end-user (or even real-world) data, we have developed a methodology for the analysis of networks which allows inference of node (potentially *user*) properties based on pairwise interactions (potentially *communication* or other *internet activity*). As such, we run into the same ethical considerations of privacy as any other research into machine learning - consideration *must* be made when applying any methods we have developed to real-world data for the privacy implications of the inference being done.

On the other hand, a more robust understanding of when inference is possible may lead to advances in applications of privacy (especially *differential privacy*); so, as with much in life, ethical considerations can “go either way” and the main takeaway should always be to exercise caution and careful consideration.

Bibliography

[Abb23] Emmanuel Abbe. Community detection and stochastic block models, 2023.

[AFL^{+}17] Avanti Athreya, Donniell E. Fishkind, Keith Levin, Vince Lyzinski, Youngser Park, Yichen Qin, Daniel L. Sussman, Minh Tang, Joshua T. Vogelstein, and Carey E. Priebe. Statistical inference on random dot product graphs: a survey, 2017.

[AFWZ19] Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank, 2019.

[AK90] Emile H. L. Aarts and Jan H. M. Korst. Simulated annealing and boltzmann machines - a stochastic approach to combinatorial optimization and neural computing. In *Wiley-Interscience series in discrete mathematics and optimization*, 1990.

[ALS18] Kwangjun Ahn, Kangwook Lee, and Changho Suh. Hypergraph spectral clustering in the weighted stochastic block model. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):959–974, October 2018.

[BB35] H. A. Bethe and William Lawrence Bragg. Statistical theory of superlattices. *Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences*, 150(871):552–575, 1935.

[Bir54] Allan Birnbaum. Combining independent tests of significance*. *Journal of the American Statistical Association*, 49(267):559–574, 1954.

[Bir55] Allan Birnbaum. Characterizations of Complete Classes of Tests of Some Multiparametric Hypotheses, with Applications to Likelihood Ratio Tests. *The Annals of Mathematical Statistics*, 26(1):21 – 36, 1955.

[Bis07] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007.

[BJD24] Vincent Bouttier, Renaud Jardri, and Sophie Deneve. Circular belief propagation for approximate probabilistic inference, 2024.

[BP16] Albert-László Barabási and Márton Pósfai. *Network science.* Cambridge University Press, Cambridge, 2016.

[BRU67] STEPHEN G. BRUSH. History of the lenz-ising model. *Rev. Mod. Phys.*, 39:883–893, Oct 1967.

[CB02] George. Casella and Roger L. Berger. *Statistical inference.* Duxbury advanced series. Duxbury/Thomson Learning, Pacific Grove, Calif, 2nd ed., international student ed. edition, 2008 - 2002.

[Che52] Herman Chernoff. A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the sum of Observations. *The Annals of Mathematical Statistics*, 23(4):493 – 507, 1952.

[CRV15] Peter Chin, Anup Rao, and Van Vu. Stochastic block model and community detection in the sparse graphs: A spectral algorithm with optimal rate of recovery, 2015.

[CYM21] Youngser Park Congyuan Yang, Carey E. Priebe and David J. Marchette. Simultaneous dimensionality and complexity model selection for spectral graph clustering. *Journal of Computational and Graphical Statistics*, 30(2):422–441, 2021.

[DKMZ11a] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6), December 2011.

[DKMZ11b] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Inference and phase transitions in the detection of modules in sparse networks. *Physical Review Letters*, 107(6), August 2011.

[DLR77] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

[ES88] Robert G. Edwards and Alan D. Sokal. Generalization of the fortuin-kasteleyn-swendsen-wang representation and monte carlo algorithm. *Phys. Rev. D*, 38:2009–2012, Sep 1988.

[FFHL21] Jianqing Fan, Yingying Fan, Xiao Han, and Jinchi Lv. Simple: Statistical inference on membership profiles in large networks, 2021.

[GG84] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984.

[GJ18] Heng Guo and Mark Jerrum. Random cluster dynamics for the Ising model is rapidly mixing. *The Annals of Applied Probability*, 28(2):1292 – 1313, 2018.

[GJB$^+$23] Ian Gallagher, Andrew Jones, Anna Bertiger, Carey Priebe, and Patrick Rubin-Delanchy. Spectral embedding of weighted graphs, 2023.

[Hea22] N. A. Heard. Standardized partial sums and products of p-values. *Journal of Computational and Graphical Statistics*, 31(2):563–573, 2022.

[Hou01] J. Houdayer. A cluster monte carlo algorithm for 2-dimensional spin glasses. *The European Physical Journal B*, 22(4):479–484, August 2001.

[HRD16] Nick Heard and Patrick Rubin-Delanchy. Network-wide anomaly detection via the dirichlet process. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pages 220–224, 2016.

[HRD18] N A Heard and P Rubin-Delanchy. Choosing between methods of combining p-values. *Biometrika*, 105(1):239–246, January 2018.

[HRH02] Peter D. Hoff, Adrian E. Raftery, and Mark S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

[Jen96] Finn V. Jensen. *Introduction to Bayesian Networks*. Springer-Verlag, Berlin, Heidelberg, 1st edition, 1996.

[JRD21] Andrew Jones and Patrick Rubin-Delanchy. The multilayer random dot product graph, 2021.

[KMM$^+$13] Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940, November 2013.

[KRFR23] Hardeep Kaur, Riccardo Rastelli, Nial Friel, and Adrian E. Raftery. Latent position network models, 2023.

[LB19] B. Li and G.J. Babu. *A Graduate Course on Statistical Inference*. Springer Texts in Statistics. Springer New York, 2019.

[Luc14] Andrew Lucas. Ising formulations of many np problems. *Frontiers in Physics*, 2, 2014.

[Mac03] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Copyright Cambridge University Press, 2003.

[MHM20] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

[Pea82] Judea Pearl. Reverend bayes on inference engines: a distributed hierarchical approach. In *Proceedings of the Second AAAI Conference on Artificial Intelligence*, AAAI'82, page 133–136. AAAI Press, 1982.

[PWTH18] Matthew Price-Williams, Melissa Turcotte, and Nick Heard. Time of day anomaly detection. In *2018 European Intelligence and Security Informatics Conference (EISIC)*, pages 1–6, 2018.

[RME$^+$21] Marek M. Rams, Masoud Mohseni, Daniel Eppens, Konrad Jałowiecki, and Bartłomiej Gardas. Approximate optimization, sampling, and spin-glass droplet discovery with tensor networks. *Phys. Rev. E*, 104:025308, Aug 2021.

[SW86] Robert H. Swendsen and Jian-Sheng Wang. Replica monte carlo simulation of spin-glasses. *Phys. Rev. Lett.*, 57:2607–2609, Nov 1986.

[SW87] Robert H. Swendsen and Jian-Sheng Wang. Nonuniversal critical dynamics in monte carlo simulations. *Phys. Rev. Lett.*, 58:86–88, Jan 1987.

[THAS20] Dimiter Tsvetkov, Lyubomir Hristov, and Ralitsa Angelova-Slavova. On the convergence of the metropolis-hastings markov chains, 2020.

[WMK15] Wenlong Wang, Jonathan Machta, and Helmut G. Katzgraber. Population annealing: Theory and application in spin glasses. *Physical Review E*, 92(6), December 2015.

[WS98] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[WSL$^+$19] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph cnn for learning on point clouds, 2019.

[Wu83] C. F. Jeff Wu. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11(1):95 – 103, 1983.

[XJL18] Min Xu, Varun Jog, and Po-Ling Loh. Optimal rates for community estimation in the weighted stochastic block model, 2018.

[YFW00] Jonathan S Yedidia, William Freeman, and Yair Weiss. Generalized belief propagation. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000.

[YFW03] Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. *Understanding belief propagation and its generalizations*, page 239–269. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2003.

[YP16] Se-Young Yun and Alexandre Proutiere. Optimal cluster recovery in the labeled stochastic block model, 2016.

[ZFK16] Zheng Zhu, Chao Fang, and Helmut G. Katzgraber. borealis - a generalized global update algorithm for boolean optimization problems, 2016.

[ZKRZ12] Pan Zhang, Florent Krzakala, Jörg Reichardt, and Lenka Zdeborová. Comparative study for inference of hidden classes in stochastic block models. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(12):P12021, December 2012.

[ZOK15] Zheng Zhu, Andrew J. Ochoa, and Helmut G. Katzgraber. Efficient cluster algorithm for spin glasses in any space dimension. *Physical Review Letters*, 115(7), August 2015.

[ZZ17] Anderson Y. Zhang and Harrison H. Zhou. Theoretical and computational guarantees of mean field variational inference for community detection, 2017.