Some inference problems on networks with applications

By

Shuqi Yu

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

(Mathematics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2023

Date of final oral examination: January 6, 2023

The dissertation is approved by the following members of the Final Oral Committee:

Sébastien Roch, Professor, Mathematics

Karl Rohe, Professor, Statistics

Cécile Ané, Professor, Statistics and Botany

Hanbaek Lyu, Assistant Professor, Mathematics

Acknowledgements

At the very beginning, I want to thank my advisor, colleague, family, and friends, and it would be impossible to have this dissertation and my Ph.D. study at the University of Wisconsin-Madison without their help and support. After the pandemic, the lifestyle changed crazily for everyone who suffers from the world-spread virus and we have to stay at home and move all communication online. I feel lucky to have all these supportive people around and finish my study at UW under this uneasy global disease.

It was one of the most important choices for continuing studying in Mathematics in my 20s. I feel so fortunate that I joined UW Math, and I believe the choice was proven right. In UW Math, I learned the upright attitude from the outstanding Scientists and felt their passion for Math and their work-life balance attitude for life. They could work on the research until midnight and reply to emails in minutes. I was joking with my family that, I'm kind of working with a group of people who are more talented than me and also more hard-working than me.

There are many people I want to show my appreciation to. It was a long way in my study at UW Madison. I want to thank Professor Shi Jin, Qin Li, Lu Wang and Louis Fan, I could never forget the warm and encouraging conversations with them during my application for graduate school in 2016. As a quiet Asian girl, I was very uncertain about my ability to study math in graduate school and pursue a doctoral degree, they encourage me and taught me to be strong to face whatever challenges ahead. I also want to thank Professor Timo Seppalainen, my academic advisor, who showed me his respectful and elegant attitude toward research and jobs, I feel fortunate to have chance to work with him in my first year.

I could not be more grateful to have Sébastien Roch as my thesis advisor, who is super careful about every little detail when it comes to research. In the meantime, he always gives generous support to students, encourages us to explore the unknowns and brings up the core idea direct to solve the problem when we are stuck. Sébastien is also a very supportive advisor who guides us to collaborate with others and provides us opportunities to join other research groups like IFDS. He shows me the power of mathematics as a tool to solve application problems without the tedious algebraic computation. I really enjoyed the process of modeling and solving complicated biological problems with neat math tricks. It's not easy to have a career plan during the pandemic as the job market is staying cold nowadays and Sébastien is willing to share and give detailed advice. I am so glad that I had the opportunity to learn about writing, research, and much more from him.

I would like to thank Professor Cécile Ané, it was a great pleasure to be a member of Ané & Roch lab, and she is always warm and generous to offer her help not only on the research in phylogenetic but also the guidance toward a career plan. She is very supportive of the students in the lab and shares professional comments in every discussion.

I also would like to thank Professor Karl Rohe, who can always come up with so many fun research ideas that are neatly directed to the core. I could not forget the shock when we saw he had the baby with him in the course "Modern Multivariate Statistics". He is very encouraging of the students and has a lot of passion to play with the data with lots of cool ideas. I was also impressed with the friendly conversation style between fellow graduate students from Rohe lab.

I would like to thank several faculty members at UW, including Hanbaek Lyu, Gheorghe Craciun, Hao Shen, David Anderson, Benedek Valko, Stephen Wright, Nan Chen, Hung Tran, Jean-Luc Thiffeault, Betsy Stovall, Andreas Seeger, Jose Rodriguez,

Michael Ferris, Daniel Erman, Dima Arinkin, Uri Andrews. Thank you for being so generous in sharing and for kindly helping me navigate through so many questions.

I want to also thank my colleagues who work in or collaborate with me, including Yu Sun, Ben Teo, Max Bacharach, Sijia Fang, Fan Chen, John Fogg, Brandon Legied, Yun Li, Xiao Shen, Tianhong Huang and Tony Yuan. Thank you for the affable company along the journey and for all the deep and happy conversations that are related or unrelated to the research.

Finally, I am deeply thankful to my dear parents, Lei and Yuhua, and my fiance, Hongxu, without whom I could not possibly start this adventure. Thank you for your endless, enormous, and loving support and for continuously offering advice. I would also like to thank my lovely fluffy friend, Graygray, who always provides her endless emotional support during the journal of studying toward my degree.

Abstract

Networks are one of the basic structures to represent the relations between objects. There is a lot of application of random networks in daily life, including social networks, evolution history, navigation maps and protein structures. This dissertation will introduce this graphical tool to analyze applications in two fields: social data analysis and phylogenetics.

Along with the data explosion, an ever-increasing share of human communication and social interaction, making available vast quantities of data. How to extract the signal from the vast noise becomes a more and more important question for large-scale social media data analysis. We introduce a notion of cross-validated eigenvalues, which guide us to find the correct dimension of random graphs under a class of random graph models. We provide a simple estimation procedure, the central limit theorem that gives a p-value for the statistical significance of each sample eigenvector, and proof of consistency for estimating the number of communities in a network.

A phylogenetic network is a graphical tool to analyze the evolutionary history between species. We show the identifiability of phylogeny in the presence of gene flow caused by several sources including lateral gene transfer, hybridization and incomplete lineage sorting. We first show an algorithmic proof to identify the phylogeny tree in the presence of lateral gene transfer events with the transfer rate constantly large, and a notion reconstruction algorithm is also presented with the proof of consistency. Then we identify the phylogeny from multi-sequence data under multispecies coalescent model using the multiple independent sites per locus under a large class of substitution models. We extend the idnetifiability result considering hybridization, which identify the level-1 phylogenetic network as we can view the networks as a collection of displayed trees.

Contents

Chapter 1

Introduction

The network is one of the basic structures which can represent the relations between vertices and has plenty of applications in different fields such as social contact, neuronal networks, protein-protein interactions, and evolution relations. This dissertation introduces a set of applications on networks for social media data analysis and phylogeny inference.

1.1 Introduction to social network analysis

The development of new technologies has brought more and more human activity information available online, digital media like Twitter records an ever-increasing share of human communication and social interaction, making available vast quantities of big data, in the forms of text, audio, and video. This brings a big opportunity for the data scientist to study individuals and society at large unobtrusively, and the raw and large-scale data also brings the big challenge to extracting the signal from the vast noise.

In the application of social data analysis, this dissertation tries to answer the question that **how to select the correct number of communities using the spectral methods**. Spectral methods hold a central place in statistical data analysis. Spectral methods refer to a collection of algorithms built upon the eigenvectors(resp. singular vectors) and eigenvalues (resp. singular values) of some properly designed matrices of data. Classical spectral methods include principal components analysis

(PCA), in which a low-dimensional subspace that explains most of the variance in the data is sought [Pea01];[Hot33]; Fisher's discriminant analysis, which aims to determine a separating hyperplane for data classification [Fis36]; and multidimensional scaling, used to realize metric embeddings of the data [Kru64]. Recent developments in spectral methods have highlighted their strengths in handling large-scale, high-dimensional, and noisy data [BW09]; [Che+21b], including community detection in networks [McS01];[RCY11a], sampling [HC17];[Roh19]; [CZR19], clustering [Von07]; [RZ20], dimensionality reduction [BN03]; [CR20], low-rank matrix estimation [AM07]; [KMO09], among others.

In social network analysis, a large and widespread class of models supposes that each person has a set of k latent characteristics. The probability that a pair of people are friends depends only on that pair's k characteristics. Typically, for example, if two people have similar features, then they are more likely to become friends. One common diagnostic is the scree plot, which plots the largest sample eigenvalues in decreasing order; the user searches this plot for a "gap" or "elbow" in the decaying eigenvalues. This diagnostic has two key limitations. First, the eigenvalues often have multiple gaps and elbows. Second, in statistical models with k true dimensions, bias differentially affects the k and $k+1$ sample eigenvalues, and this bias blurs any gap or elbow between them. A more general problem is that a useful theory and methodology must confront the possibility that only some of the leading k population eigenvectors are estimable. In this situation, the "correct" choice of k is the number of statistically useful dimensions. To confront these problems, Chapter 2 introduces a notion of cross-validated eigenvalues. Under a large class of random graph models, we provide (1) a simple estimation procedure, (2) a central limit theorem that gives a p-value for the statistical significance of each sample eigenvector, and (3) a proof of consistency. This approach can be used to estimate the number of statistically useful sample eigenvectors, naturally adapting to the complexity of the estimation task. In simulations and a data

example, the proposed estimator compares favorably to alternative approaches in both computational and statistical performance.

1.2 Introduction to the phylogenetic network inference

A phylogeny is a proposal of how organisms are related by their evolutionary history, based on the evidence that all living things are related by common descent. Phylogenetic networks are a graphical tool used to perform analysis in mathematics and biology [HRS10]. The leaves represent the species at the current time, the internal vertices represent the ancestor species in evolutionary history and the edges represent evolutionary events between their incident vertices. Every species has observable traits that may be binary (e.g. 'horns' or 'no horns') or greater (e.g. DNA or amino acid sequences). Evolutionary events are comprised of mutations that influence these traits, and there are many ways to model them. Traditional sequence-based substitution models include the Jukes-Cantor [JC+69] and the generalized time-reversible model [Tav+86]. However, phylogenetic networks inferred from different genes often imply different, conflicting evolutionary histories from which they have been sampled [Pol+06; GD08; Cra+09; Nak13]. In addition to statistical errors in gene tree estimation, there are several well-recognized causes of gene tree incongruence, including incomplete lineage sorting (ILS), hybridization, lateral gene transfer (LGT), gene duplication and loss.

Another key question this dissertation tries to answer is whether **can we infer phylogenetic networks from the gene sequence in the presence of the gene flow?** Observing measured traits between taxa and using them to determine discrete relationships and the evolutionary distance between taxa, and the phylogenetic network topologies are recovered from these sub-networks or pairwise distances. Lots of prior works have been done in phylogenetic networks inference including probability-

or likelihood-based methods [Bou+13], quartet-based methods [Lar+10], [Mir+14], [RSM19], concatenation [RS15], sequence-based methods[TKF91], [TKF92],[DR13a].

In chapter 3, this dissertation first focuses on identifying a phylogeny tree in the presence of LGT, which has an important role in the evolution [SSJ03; McD+10; WC11; MC17; Hib+21]. A stochastic model of LGT was introduced by Roch and Snir [RS13], LGT events occur at random along the phylogeny according to a Poisson process, for each gene independently. The goal is to recover the species phylogeny from a collection of gene trees, each of which can be thought of as a randomly scrambled instance of the species phylogeny. A related model was also studied in [LRH07; Ste+13; SS13]. It was proved in [RS13] that under the assumptions in Section 3.2.2, a species phylogeny with n leaves can be recovered from a logarithmic number of genes when the LGT rate is at most $O(1/\log n)$ per unit time. [RS13] also showed that the species phylogeny cannot be distinguished with constant probability from the same number of genes when the LGT rate is of the order of $\Omega(\log\log n)$ per unit time. Under the same assumptions, the algorithm result in [DR16] improved the LGT rate to a small constant bound by a recursive approach which progressively builds the species phylogeny from the leaves up, using the information obtained from partially reconstructed subtrees to reach further into the past. A similar result was given in [Ste+13] by showing the subtree spanned by any three leaves is statistically consistent when the LGT rate is small, and then the species tree can be reconstructed from the gene trees by majority voting. To further close this gap for a constantly large transfer rate, Chapter 3 introduces an algorithmic idea to reconstruct the phylogeny bottom up. since the LGT events happen chronologically from the root and the transition matrix is invertible, the previous divergence time can be identified because it's the first time that two clusters merge in the distribution.

In chapter 4, we are motivated by the example given by [Ste+13] showing that the mixtures of two sets of branch lengths on a tree of one topology can exactly mimic

the different (expected) site pattern frequencies of a tree of a different topology under the two-state symmetric model. This dissertation introduces a new estimator which can overcome this issue by using two independent sites per gene. Under the DNA substitution model including the Jukes-Cantor model and the general time reversible model, chapter 4 introduces a notion of multiple sites estimator whose expectation satisfies the four points condition in the presence of incomplete lineage sorting. One can further infer the whole phylogeny tree using quartet-based algorithms like [Erd+99]. One can extend the identifiability result by considering different rate across loci or under the mixture of identical trees model. This dissertation further discuss the identifiability on level-1 phylogenetic networks as we can view the networks as a collection of displayed trees and the four points condition of the ideal version of the estimator is also held by linearity.

1.3 Collaboration note

Chapter 2 is based on the collaboration work in [Che+21a] with Fan Chen, Karl Rohe and Sébastien Roch from University of Wisconsin Madison. Chapter 3 is based on the work with Elchanan Mossel from Massachusetts Institute of Technolog, Allan Sly from Princeton University and Sébastien Roch. The section 4.2 in chapter 4 is based on the work with Sébastien Roch and the section 4.3.2 in chapter 4 is based on the work with Cécile Ané, Sébastien Roch,Yu Sun and Jingcheng Xu from University of Wisconsin Madison. My main role in all these projects are rigorous theory development and paper writing.

Chapter 2

Estimating Graph Dimension with Cross-validated Eigenvalues

2.1 Background and Notations

In social network analysis, a large and popular class of models supposes that each person has a set of k latent characteristics and the probability that a pair of people are friends depends only on that pair's k characteristics. Typically, for example, if two people have similar characteristics, then they are more likely to become friends. We aim to estimate the number of characteristics k using a class of models where every edge is statistically independent, conditionally on the characteristics. This includes the Latent Space Model, the Aldous-Hoover representation, and graphons [HRH02; Ald85; Hoo89; Lov12; JC14].

Denote the adjacency matrix $A \in \mathbb{N}^{n \times n}$ as recording the number of edges between i and j in element A_{ij}. We are particularly interested in the class of models where every person i is assigned a vector of characteristics $Z_i \in \mathbb{R}^k$ and

$$\mathbb{E}(A_{ij}) = \sum_{\ell=1}^{k} Z_{i\ell} Z_{j\ell} = \langle Z_i, Z_j \rangle. \tag{2.1}$$

This model is often called the random dot product model [Ath+13], which includes the Stochastic Blockmodel, along with its degree-corrected and mixed membership variants

[KN11; Air+08]. In the Stochastic Blockmodel, k is the number of blocks. Under mild identifiability conditions, $\mathbb{E}(A)$ and the normalized form of this matrix defined in Equation (2.2) below have k non-zero eigenvalues. We use this fact to estimate k.

The scree plot gives the sample eigenvalues. We look for an elbow or a gap.

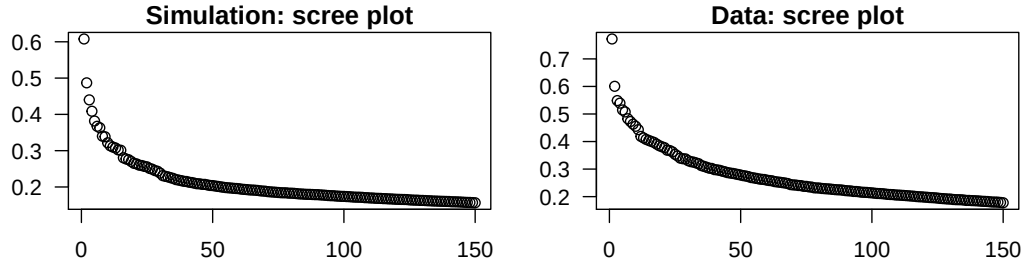

Figure 2.1: In these examples, it is difficult to detect a gap or elbow. In the left panel, the graph is simulated from a Degree-Corrected Stochastic Blockmodel with $n = 2560$. In the right panel, the graph is a citation graph among $n = 22,688$ academic journals. Displayed are the largest 150 eigenvalues of the normalized and regularized adjacency matrix, L, defined in Equation (2.2). Section 2.1.1 gives more details for these figures.

Unfortunately, the eigenvalues of A or of its normalized form (i.e., the "sample eigenvalues") in the scree plot often fail to provide a clear estimate of k. As an illustration, this is the case in Figure 2.1. We address this problem with a cross-validation technique. There are three basic pieces that, when put together, enable our cross-validation approach to estimating the latent dimension k.

The *first piece* provides two identically distributed adjacency matrices, $\tilde{A}$ and $\tilde{A}_{\text{test}}$, from a single adjacency matrix A. For each element i, j, define the elements of $\tilde{A}$ and $\tilde{A}_{\text{test}}$ as

$$\tilde{A}_{ij} \sim \text{Binomial}(A_{ij}, 1-\varepsilon) \text{ and } [\tilde{A}_{\text{test}}]_{ij} = A_{ij} - \tilde{A}_{ij}.$$

If $\varepsilon = 1/2$, then $\tilde{A}$ and $\tilde{A}_{\text{test}}$ are identically distributed. If A has independent elements with $\mathbb{E}(A_{ij}) = \langle Z_i, Z_j \rangle$, then $\tilde{A}$ also has independent elements with $\mathbb{E}(\tilde{A}_{ij}) = \langle Z_i, Z_j \rangle / (1-\varepsilon)$. Moreover, we will model A_{ij} as Poisson, which makes $\tilde{A}$ and $\tilde{A}_{\text{test}}$ statistically independent. It is common to model A_{ij} as Poisson because it is convenient

and, in sparse graphs specifically, the difference between the Poisson and Bernoulli models becomes negligible [KN11; FP+20; CD18; CCB16; ZA20]. See Section 2.5 for further comparison between Bernoulli and Poisson graphs.

The *second piece* is that $\mathbb{E}(A), \mathbb{E}(\tilde{A})$, and $\mathbb{E}(\tilde{A}_{\text{test}})$ have nearly identical spectral properties that reveal k. Because $\mathbb{E}(\tilde{A}) = \mathbb{E}(A)/(1-\varepsilon)$, dividing an eigenvalue of $\mathbb{E}(A)$ by $1-\varepsilon$ gives an eigenvalue of $\mathbb{E}(\tilde{A})$. As a result, they have the same number of non-zero eigenvalues, which is the value k that we aim to estimate. Importantly, these matrices all have identical eigenvectors.

While the technical parts of the chapter use the eigenvectors of an adjacency matrix, the proposed algorithm instead uses eigenvectors from the normalized and regularized adjacency matrices. To define this matrix for the full adjacency matrix A, define the node degrees $d_i = \sum_j A_{ij}$ and the regularization parameter $\tau = n^{-1} \sum_i d_i$. The normalized and regularized adjacency matrix is

$$L = DAD, \text{ where } D \text{ is a diagonal matrix with } D_{ii} = (d_i + \tau)^{-1/2}. \tag{2.2}$$

Importantly, the normalized and regularized form of $\mathbb{E}(A)$ also has k non-zero eigenvalues. For dense graphs, L has similar statistical properties to A. However, for sparse graphs, L has better statistical properties [LLV17; ZR18].

The *third and final piece* of the proposed approach is that for an eigenvector $\tilde{x}$ of $\tilde{A}$ (or its normalized and regularized version $\tilde{L}$), the "cross-validated eigenvalue,"

$$\lambda_{\text{test}}(\tilde{x}) = \tilde{x}^T \tilde{A}_{\text{test}} \tilde{x} = \sum_{ij} \tilde{x}_i \tilde{x}_j [\tilde{A}_{\text{test}}]_{ij} \tag{2.3}$$

is a weighted sum of independent random variables which converges to the normal distribution; this result is conditional on $\tilde{x}$, which is independent of $\tilde{A}_{\text{test}}$ in the Poisson

model. We test whether the expected value of $\lambda_{\text{test}}(\tilde{x})$ is zero,

$$\mathbb{E}(\lambda_{\text{test}}(\tilde{x})|\tilde{A}) = \tilde{x}^T\,\mathbb{E}(\tilde{A}_{\text{test}}|\tilde{A})\tilde{x} = \tilde{x}^T\,\mathbb{E}(\tilde{A}_{\text{test}})\tilde{x} = \sum_{\ell}(\sum_{i} Z_{i\ell}\tilde{x}_i)^2/2. \tag{2.4}$$

When this value is zero, it means that $\tilde{x}$ is orthogonal to the latent space; in this case, we say that $\tilde{x}$ is not statistically useful (although, perhaps, it is still useful for tasks other than estimating $Z_1, \ldots, Z_n$). Ideally, the first k eigenvectors will provide large values of λ_{test}, with large Z-scores, while the following eigenvectors have $\lambda_{\text{test}} \approx 0$ and Z-scores normally distributed, with mean zero and variance one. The main theoretical result, stated in Theorem 2.4.1 below, shows that the proposed technique is consistent.

2.1.1 Motivating examples

In Figure 2.1, the simulated graph comes from a Degree-Corrected Stochastic Blockmodel, with $k = 128$ hierarchically arranged blocks. Many of the 128 dimensions cannot be estimated from the data. As such, it is not surprising that no artifacts arise in the scree plot around 128.

In Figure 2.2 below, the simulated scree plot from Figure 2.1 is repeated as a black line. The first two eigenvalues have been removed to improve the display. The blue line gives $\tilde{x}^T\tilde{L}_{\text{test}}\tilde{x}$, where $\tilde{x}$ runs through the leading 150 eigenvectors of $\tilde{L}$. The red line gives the population version of this quantity, where the normalized and regularized matrix is constructed from the matrix $\mathbb{E}(A)$. The red and blue lines reveal that the eigenvectors computed from the data do not correlate with the underlying Z's after around $\hat{k} \approx 60$. Moreover, the Z-scores in the right panel start to cluster around the cutoff at $\hat{k} \approx 60$. The standard scree plot (black line in left panel) does not reveal anything around this value. The full details of this simulation model are provided in supplementary Section 2.8.1. This illustration splits the edges ten separate times, each with probability $\varepsilon = .1$, and averages the results over those ten folds.

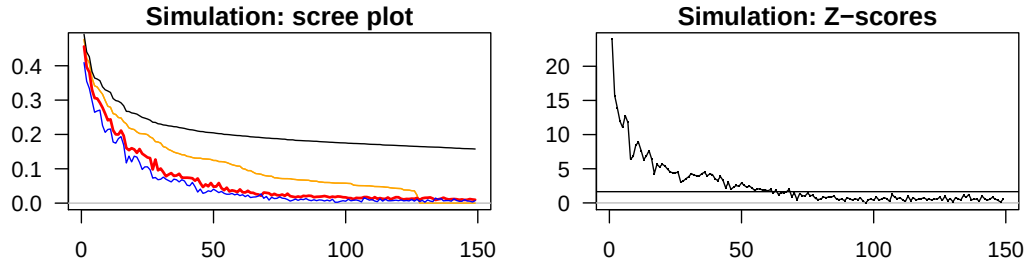

Figure 2.2: In the left panel, the black line gives the empirical eigenvalues (repeated from the left panel of Figure 2.1) and the orange line gives the $k = 128$ non-zero population eigenvalues. The blue line gives $\tilde{x}^T \tilde{L}_{\text{test}} \tilde{x}$ and red line gives the population version of this quantity. In the right panel, the Z-scores test the null hypothesis that the quantity in (2.4) is zero. The horizontal black line gives the cutoff for .05 significance. In this example, a good choice for $\hat{k}$ would be around 60.

The right panel of Figure 2.1 gives the scree plot for a citation graph on 22,688 academic journals. This graph was constructed from the Semantic Scholar database [Amm+18] of roughly 220 million academic papers. Citations from one paper to another were converted to citations between the journals that published the papers. If there were more than 5 citations from journal i to journal j using a 5% sample of all edges, then A_{ij} is set to one. Otherwise, A_{ij} is zero. This graph was originally constructed and studied in [RZ20]. For simplicity, the graph was symmetrized by setting $A_{ij} = 1$ if $A_{ji} = 1$. The average journal degree is 35. In the simulation in Figure 2.2 above, the red and orange lines give "population quantities," constructed with $\mathbb{E}(A)$. In Figure 2.3, we use the blue line as an estimate of the red line and the Z-scores to test the null hypothesis that Equation (2.4) is equal to zero.

One way of understanding the difficulty of using spectral approaches for estimating k is this: the sample eigenvectors can "overfit" to the noise in large-scale graphs and it is hard to tell when this overfitting happens. This is the same overfitting that makes the sample eigenvalues in the scree plot differentially biased; in Figure 2.2 in the left panel, the gap from the black line to the orange and red lines is smaller on the left

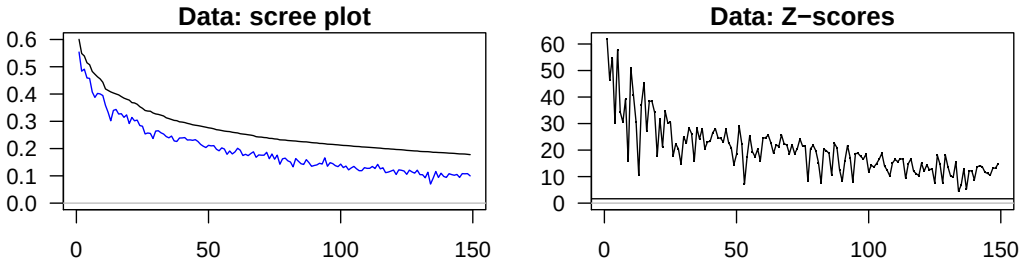

Figure 2.3: In this example, $\varepsilon = .05$ and for illustration, the data was divided only one time. All of the leading 150 dimensions are highly statistically significant. This is consistent with the results in [RZ20] that showed the leading 100 dimensions reveal groups of journals that form coherent academic areas. In this example, L has 789,980 non-zero elements, spread across 22,688 rows and columns. Despite the relatively large size of this graph, computing all 150 cross-validated eigenvalues and their Z-scores requires less than 10 seconds in `R` on a 2020 MacBook Pro. This speed is enabled by the sparse matrix packages `Matrix` and `RSpectra` [BM21; QM19]

and larger on the right. In this chapter, we exploit a notion of cross-validated eigenvalues as a new approach to estimating k. Here, the eigenvectors and cross-validated eigenvalues are computed on different graphs which is made possible by splitting the edges into two graphs [AS15; ABH16]. This removes the bias from overfitting.

Under a large class of random graph models, we provide a simple procedure to compute cross-validated eigenvalues.

A related holdout approach was previously explored in the econometrics literature [ADZ14; Lam16] for covariance estimation. In this chapter and in those prior papers, the eigenvectors are estimated with a portion of the data and the "signal strength" of those vectors is estimated with the remaining held-out data. There are three key differences with this previous work. First, the observed data is of a different nature; for a covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$, it is assumed in [Lam16] that we observe $y_i = \Sigma^{1/2} x_i$, where x_i are unobserved and contain independent, identically distributed (i.i.d.) random variables. Second, the notion of sample splitting is different; the approach in [ADZ14] constructs

a sample covariance matrix with a subsample of the observed vectors $y_1, \ldots, y_n$. Third, the target of estimation is different; we provide p-values to estimate k, while the prior work aims to estimate the eigenvalues of Σ, which is presumed to be full rank.

For cross-validated eigenvalues, we provide an intuitive central limit theorem, which leads to a p-value for the statistical significance of a sample eigenvector. This can be used to estimate the number of statistically useful sample eigenvectors, and thus k. We provide consistency results for the proposed estimator of k, allowing for weighted and sparse graphs. Finally, through simulations and real data applications, we show that this estimator compares favorably to alternative approaches in both computational and statistical performance.

2.1.2 Prior literature

Numerous methods have been proposed to estimate k under the Stochastic Blockmodel and its degree-corrected version [BLM15; BS16; Lei16; WB17; CL18; MSZ19; LL19a; Liu+19; Jin+20]. One previous technique has been proposed to estimate the dimension of the more general random dot product graph [LLZ20]. These methods roughly fall into one of three categories: spectral, cross-validation, and (penalized) likelihood-based approaches. Methods based on likelihood or cross-validation are actively researched, yet the majority of them are commonly restrained by the scale of networks. Spectral methods are highly scalable for estimating k in large networks, although their rigorous analyses require delicate, highly technical random matrix arguments [AEK17; BBK19; CCH20; DZ19; BBK20; HLY20].

Among the likelihood-based approaches, the authors of [WB17] proposed to estimate k by solving a Bayesian information criterion (BIC) type optimization problem, where the objective function is a sum of the log-likelihood and of the model complexity. The computation is often not feasible because the likelihood contains exponentially many

terms. In [MSZ19], a pseudo-likelihood ratio is used to compare the goodness-of-fit of models with differing ks that have been estimated using spectral clustering with regularization [RCY11b; QR13; JY16; SWZ19], speeding up the computation. However, the two methods allow little node degree heterogeneity. Related to the goodness-of-fit technique, the authors of [Jin+20] present a stepwise testing approach based on the number of quadrilaterals in the networks. Computing the statistic requires at least n^2 multiplication operations, regardless of the sparsity of the graph, thus is infeasible for large n. More recently, cross-validation [PC84; AC10] has also been adapted to the context of choosing k. For example, in [CL18], a block-wise node-pair splitting technique is introduced. In each fold, a block of rows of the adjacency matrix are held out from the Stochastic Blockmodel fitting (including the community memberships), then the left-out rows are used to calculate a predictive loss. In [LLZ20], the authors propose to hold out a random fraction of node-pairs, instead of nodes (thus all the incidental node-pairs). In addition, they suggest using a general low-rank matrix completion (e.g., a singular value thresholding approach [Cha15]) to calculate the loss on the left-out node-pairs. Theoretical conditions for not under-estimating k were established in both cross-validation based methods [CL18; LLZ20]. Calculating the loss on either held-out rows or on scattered values in the adjacency matrix requires $O(n^2)$ computations, regardless of sparsity. This limits the ability of these techniques to scale to large graphs.

In [BS16; Lei16], hypothesis tests using the top eigenvalue or singular value of a properly normalized adjacency matrix are proposed, based on edge universality and other related results for general Wigner ensembles [TW94; Sos99; Erd+12; Erd+13; Ale+14]. The analyses of these hypothesis tests assume dense graphs. In [Liu+19], a version of the "elbow in the scree plot" approach (see, e.g., [ZG06] for a discussion of this approach) is analyzed rigorously under the Degree-Corrected Stochastic Blockmodel, also in the dense case. For sparser graphs, the spectral properties of other matrices associated to graphs have been used to estimate k, including the non-backtracking

matrix [Krz+13; BLM15; LL19a] and the Bethe-Hessian matrix [LL19a]. However, their theoretical analysis currently allow little node degree heterogeneity in the sparse case.

There is also related work on bootstrapping [SB99; Tho+16; GS17; LL19b; LLS20a], jackknife resampling [LLS20b] and subsampling [BB15; LS19; Nau+21] in network analysis. In particular, in [LS19], subsampling schemes are applied to the nonzero eigenvalues of the adjacency matrix under low-rank graphon models. Weak convergence results are established under some technical conditions, including sufficient edge density (i.e., average degree growing asymptotically faster than $\sqrt{n}$); simulation results also indicate that sparsity leads to poor performance for the estimators considered, especially in the case of the eigenvalues closer to the bulk.

2.2 The statistical model, sample eigenvalues, and a new measure of signal strength

We consider a connected multigraph $G = (V, E)$ consisting of the set of nodes $V = \{1, \ldots, n\}$ and edges E, where we allow multiple edges and self-loops. The adjacency matrix $A \in \mathbb{N}^{n \times n}$ records the number of edges between i and j in element A_{ij}.

The introduction motivated the chapter by expressing $\mathbb{E}(A_{ij}) = \langle Z_i, Z_j \rangle$ in Equation (2.1). If the $Z_1, \ldots, Z_n \in \mathbb{R}^k$ span $\mathbb{R}^k$ and the elements A_{ij} are independent Poisson variables, then this model is included in Definition 1, which is the focus of this chapter.

Definition 1 (Poisson graph)**.** We consider random graph models where the elements

of A are independent Poisson random variables and $\mathbb{E}(A)$ has the eigendecomposition

$$\mathbb{E}(A) = U\Lambda U^{\mathrm{T}} \tag{2.5}$$

for $U \in \mathbb{R}^{n\times k}$ with orthonormal columns and diagonal matrix $\Lambda \in \mathbb{R}^{k\times k}$ with positive elements $\lambda_1, \ldots, \lambda_k \in \mathbb{R}$ down the diagonal in non-increasing order.

Define the population (or expectation) matrix $P = \mathbb{E}(A)$. The diagonal of Λ contains the leading k eigenvalues of P and their corresponding eigenvectors are in the columns of U. For $j > k$, the eigenvalues of P are $\lambda_j = 0$. In this chapter, we aim to estimate the number of nonzero eigenvalues.

This chapter makes two simplifying assumptions. The first simplifying assumption is that A is symmetric (i.e., edges are undirected). This assumption can be relaxed; Remark 2.4.1 discusses directed graphs, contingency tables, and rectangular incidence matrices. The second simplifying assumption is that the elements of A are Poisson. A simulation in Section 2.5 demonstrates that the proposed technique provides reliable p-values in Bernoulli graphs as well.

2.2.1 Sample eigenvalues: a poor diagnostic

A common approach to estimating the eigenvalues of P is to use a plug-in estimator, i.e., estimating the eigenvalues of P with the eigenvalues of A. The symmetric matrix $A \in \mathbb{N}^{n\times n}$ has eigenvectors $\hat{x}_1, \ldots, \hat{x}_n \in \mathbb{R}^n$ that are the solution to

$$\hat{x}_j \;=\; \underset{x\in\hat{S}_j}{\operatorname{argmax}} \;\; x^{\mathrm{T}}Ax, \tag{2.6}$$

where $\hat{S}_j = \{x \in R^n : \|x\|_2 = 1 \text{ and } x^{\mathrm{T}}\hat{x}_\ell = 0 \text{ for } \ell = 1, \ldots, j-1\}$. The eigenvalues $\hat{\lambda}_j$ for $j = 1, \ldots, n$ are defined as

$$\hat{\lambda}_j = \hat{x}_j^{\mathrm{T}} A \hat{x}_j. \tag{2.7}$$

Note that the quadratic form that defines the eigenvalues in Equation (2.7) is identical to the objective function that the eigenvectors optimize in equation (2.6). This leads to overfitting and bias.

The eigenvalues of A, i.e., $\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3, \ldots$, are often plotted against their index $1, 2, 3, \ldots$. This is called a scree plot and it is used as a diagnostic to estimate k. In this scree plot, there might be a "gap" or an "elbow" at the kth eigenvalue, which reveals k.

However, there is a fundamental problem with the plug-in estimator for the population eigenvalues which can make the "gap" or "elbow" in the scree plot more difficult to observe. The leading eigenvalue estimates $\hat{\lambda}_1, \ldots, \hat{\lambda}_k$ are asymptotically unbiased, so long as their corresponding population eigenvalues $\lambda_1, \ldots, \lambda_k$ are large enough (see, e.g., [CCH20] for related results). However, when n is much larger than k, $\hat{\lambda}_{k+1}$ is a biased estimate of $\lambda_{k+1} = 0$, with $\mathbb{E}(\hat{\lambda}_{k+1}) > \lambda_{k+1} = 0$ (see, e.g., [BBK20] for related results). So, if λ_k is not large enough, then this bias diminishes the appearance of a "gap" or "elbow" between $\hat{\lambda}_k$ and $\hat{\lambda}_{k+1}$ in the scree plot.

This can be seen in the left panel of Figure 2.2. In that figure, the gap between the sample eigenvalues (black line) and the population eigenvalues (orange line) is smaller on the left and larger on the right. Note that in this figure, it is not the eigenvalues of A and $\mathbb{E}(A)$ that are shown, but rather the eigenvalues of the normalized and regularized forms of A and $\mathbb{E}(A)$.

2.2.2 Measuring the signal strength of a sample eigenvector

Even if an oracle were to tell us that the population eigenvector x_j has population eigenvalue $\lambda_j \neq 0$, we should only use the sample eigenvector $\hat{x}_j$ for statistical inference if $\hat{x}_j$ is close to x_j and other leading population eigenvectors. Because of this requirement, the population eigenvalues λ_j do not measure the signal strength of $\hat{x}_j$.

In the realistic setting where the signal is not overwhelming for every sample eigenvector $\hat{x}_j$ for $j < k$, we will measure the signal strength of this vector in the following way. Define the population (or expected) cross-validated eigenvalue of a sample eigenvector $\hat{x}_j$ as

$$\lambda_P(\hat{x}_j) = \hat{x}_j^{\mathrm{T}} P \hat{x}_j. \tag{2.8}$$

Of course, this is not actually an eigenvalue in the traditional sense. However, there are three reasons to think of this quantity as an analogue. First, the quadratic form in Equation (2.8) mimics the form for the sample eigenvalue in Equation (2.7), which is also the objective function that the sample eigenvectors optimize in Equation (2.6). As such, for a population eigenvector x_j, $\lambda_P(x_j)$ is the corresponding population eigenvalue λ_j.

The second reason to think of λ_P as an analogue of an eigenvalue is that for a sample eigenvector $\hat{x}$, there always exists a vector $\hat{x}^\perp \in \mathbb{R}^n$ that is orthogonal to $\hat{x}$ and

$$P\hat{x} = \hat{\lambda}\hat{x} + \hat{x}^\perp$$

for some value $\hat{\lambda} \in \mathbb{R}$. If $\hat{x}^\perp = 0$, then $\hat{x}$ is an eigenvector of P with eigenvalue $\hat{\lambda}$. Even when $\hat{x}^\perp \neq 0$, if $\|\hat{x}\|_2 = 1$, then $\hat{\lambda} = \lambda_P(\hat{x})$.

$$\lambda_P(\hat{x}) = \hat{x}^{\mathrm{T}} P \hat{x} = \hat{x}^{\mathrm{T}}(\hat{\lambda}\hat{x} + \hat{x}^\perp) = \hat{\lambda}$$

The third reason also provides intuition for why λ_P is a measure of signal strength; indeed λ_P provides for the optimal reconstruction of P as conceptualized in the following proposition. See supplementary Section 2.7 for a proof.

Proposition 2.2.1 ([Lam16])**.** *The solution to*

$$\min_{\hat{\lambda}_1,\ldots,\hat{\lambda}_q} \left\| P - \sum_{j=1}^{q} \hat{\lambda}_j \hat{x}_j \hat{x}_j^{\mathrm{T}} \right\|_F$$

is $\hat{\lambda}_j = \lambda_P(\hat{x}_j)$ *for* $j = 1, \ldots, q$.

A key advantage of the population quantity $\lambda_P(\hat{x}_j)$ over λ_j is that it reveals information about the vector that we can compute, i.e., $\hat{x}_j$, not the eigenvector that we wish we had, i.e., x_j. If a sample eigenvector $\hat{x}_j$ is close to its population counterpart x_j, then for the reasons above, it is reasonable to presume that $\lambda_P(\hat{x}_j)$ is close to the population eigenvalue λ_j. However, if the estimation problem is too difficult and $\hat{x}_j$ is nearly orthogonal to the eigenvectors of P that have non-zero eigenvalues, then $\lambda_P(\hat{x}_j) \approx 0$. Notably, and most importantly, this can happen even if $\hat{x}_j$'s corresponding population eigenvector x_j has a non-zero eigenvalue. For example, this happens when the estimation problem is too difficult as happens for the 60th-128th sample eigenvectors in Figure 2.2.

The notion of $\lambda_P(\hat{x}_j)$ was originally proposed and studied in [ADZ14] and [Lam16] for optimal estimation of eigenvalue shrinkage under a different statistical model. While the signal strength λ_P is unknown, cross validation can provide an unbiased estimator that is asymptotically normal. The next three sections develop and study a statistical testing procedure for the null hypothesis

$$H_0 : \lambda_P(\tilde{x}_j) = 0 \tag{2.9}$$

conditionally on $\tilde{x}_j$, where $\tilde{x}_j$ are estimates of the eigenvectors of P (or its normalized form) constructed from a large subsample of the edges in the graph.

2.3 The three key pieces for cross-validated eigenvalues

This section details the three key pieces that enable the estimation of cross-validated eigenvalues. First, edge splitting a Poisson random graph creates two independent Poisson random graphs. Second, the expected adjacency matrices of the resulting graphs are "spectrally invariant" to this edge splitting. Third, the quadratic form for λ_{test} in Equation (2.3) satisfies a central limit theorem.

2.3.1 The first piece: edge splitting creates independent Poisson graphs

The `ES` procedure in Algorithm 1 splits the edges of a graph into two graphs and outputs the two adjacency matrices $\tilde{A}$ and $\tilde{A}_{\text{test}}$. Notice that under `ES` and conditionally on A_{ij}, $\tilde{A}_{ij} \sim \text{Binomial}(A_{ij}, \varepsilon)$ and $[\tilde{A}_{\text{test}}]_{ij} = A_{ij} - \tilde{A}_{ij}$. The independence of $\tilde{A}$ and $\tilde{A}_{\text{test}}$ follows from the next lemma, often referred to as thinning (see, e.g., [Dur19, Section 3.7.2]).

Lemma 2.3.1. Define $X \sim \text{Poisson}(\lambda)$ and conditionally on X, define $Y \sim \text{Binomial}(X, p)$ and $Z = X - Y$. Unconditionally on X, the random variables Y and Z are independent Poisson random variables and, further, $Y \sim \text{Poisson}(p\lambda)$ and $Z \sim \text{Poisson}((1-p)\lambda)$.

To apply the lemma, let X be A_{ij}, let λ be P_{ij}, let Y and Z be the (i,j)-th elements of $\tilde{A}$ and $\tilde{A}_{\text{test}}$ respectively, and let $p = \varepsilon$. Lemma 2.3.1 implies that $\tilde{A}$ and $\tilde{A}_{\text{test}}$ are independent Poisson random graphs.

Input: Adjacency matrix $A \in \mathbb{N}^{n \times n}$ and edge splitting probability $\varepsilon \in (0, 1)$.
Procedure `ES`(A, ε)**:**
1. Convert A into $G = (V, E)$, where $\{i, j\}$ is repeated in the edge set E potentially more than once if $A_{ij} > 1$.
2. Initiate $\tilde{E}_{\text{test}}$ and $\tilde{E}$, two empty edge sets on V.
3. **for** *each copy of edge* $\{i, j\} \in E$ **do**
   assign it to $\tilde{E}_{\text{test}}$ with probability ε. Otherwise, assign it to $\tilde{E}$.
4. Convert $(V, \tilde{E}_{\text{test}})$ into an adjacency matrix $\tilde{A}_{\text{test}} \in \mathbb{N}^{n \times n}$ and $(V, \tilde{E})$ into an adjacency matrix $\tilde{A} \in \mathbb{N}^{n \times n}$

Output: $\tilde{A}$ and $\tilde{A}_{\text{test}}$.

Algorithm 1: Edge splitting

2.3.2 The second piece: the population graphs are "spectrally invariant" to splitting

The next proposition shows that `ES` preserves the spectral properties of the population adjacency matrices $\mathbb{E}(\tilde{A})$ and $\mathbb{E}(\tilde{A}_{\text{test}})$. This result does not require any distributional assumptions on A, only that its elements are integers (so that $\tilde{A}$ and $\tilde{A}_{\text{test}}$ are defined).

Proposition 2.3.1. *If $\tilde{A}$ and $\tilde{A}_{\text{test}} \in \mathbb{N}^{n \times n}$ are generated by applying `ES` to $A \in \mathbb{N}^{n \times n}$ with splitting probability ε, then*

1. *The eigenvectors of $\mathbb{E}(A), \mathbb{E}(\tilde{A})$, and $\mathbb{E}(\tilde{A}_{\text{test}})$ are identical.*
2. *If λ_j is an eigenvalue of $\mathbb{E}(A)$, then $(1 - \varepsilon)\lambda_j$ is an eigenvalue of $\mathbb{E}(\tilde{A})$ and $\varepsilon\lambda_j$ is an eigenvalue of $\mathbb{E}(\tilde{A}_{\text{test}})$.*

Here, all expectations are unconditional on A.

Proof. Define $P = \mathbb{E}(A)$ and let $P = U\Lambda U^{\mathrm{T}}$ be its eigendecomposition; if A is not random, then $P = A$ and U, Λ potentially have n columns. It follows directly from the construction in `ES` that $\mathbb{E}(\tilde{A}) = (1 - \varepsilon)P$ and $\mathbb{E}(\tilde{A}_{\text{test}}) = \varepsilon P$. Rearranging terms reveals

the eigendecomposition of $\mathbb{E}(\tilde{A}_{\text{test}})$,

$$\mathbb{E}(\tilde{A}_{\text{test}}) = \varepsilon P = U(\varepsilon\Lambda)U^{\mathrm{T}}$$

and similarly for $\mathbb{E}(\tilde{A})$. This shows that they have the same eigenvectors and the simple relationship between their eigenvalues in the statement. □

2.3.3 The third piece: cross-validated eigenvalues are asymptotically Gaussian

To state the theorem formally, we consider a sequence of random adjacency matrices $B^{(n)} \in \mathbb{N}^{n\times n}$ from Poisson random graphs with $\mathbb{E}(B^{(n)}) = Q^{(n)} \in \mathbb{R}^{n\times n}$ satisfying $\max_{ij} Q_{ij}^{(n)} \leq 1$, and a sequence of unit vectors $x^{(n)} \in \mathbb{R}^n$. To simplify the notation, we suppress the explicit dependence on n. We will impose the following delocalization condition on x:

$$\|x\|_\infty^2 = o(\sigma), \tag{2.10}$$

where

$$\sigma^2 = 2(x^2)^{\mathrm{T}}Q(x^2) - (x^2)^{\mathrm{T}}\mathrm{diag}(Q)(x^2),$$

with x^2 being the vector x with entries squared and $\mathrm{diag}(Q)$ being the diagonal matrix containing the diagonal elements of Q. Similarly, we also define

$$\hat{\sigma}^2 = 2(x^2)^{\mathrm{T}}B(x^2) - (x^2)^{\mathrm{T}}\mathrm{diag}(B)(x^2).$$

In the next section, we will apply the theorem to $B := \tilde{A}_{\text{test}}$, $Q := \varepsilon P$ and x an eigenvector of $\tilde{A}$ or of its normalized form.

Theorem 2.3.1 (CLT for cross-validated eigenvalue)**.** Let B, Q, σ and $\hat{\sigma}$ be as above.

Assume that x satisfies Condition (2.10). Then,

$$\frac{\lambda_B(x) - \lambda_Q(x)}{\hat{\sigma}} \Rightarrow N(0,1). \tag{2.11}$$

The proof of Theorem 2.3.1 is in supplementary Section 2.7.

Remark. If the elements of B are Bernoulli instead of Poisson, then a similar central limit theorem holds with a new variance γ^2 that contains the sum $\sum_{ij} x_i^2 x_j^2 P_{ij}(1 - P_{ij})$. The key difference compared to the Poisson variance is the inclusion of $(1 - P_{ij})$ which is difficult to estimate. Because $(1 - P_{ij}) \leq 1$, The Poisson model formula for σ^2 provides an upper bound for γ^2. As such, σ^2 and $\hat{\sigma}^2$ can still be used to provide conservative inference. Moreover, when the graph is sparse, we have $(1 - P_{ij}) \to 1$ and σ^2, and $\hat{\sigma}^2$ become better approximations. The key problem with Bernoulli graphs is not the lack of normality, or estimating the variance. The key problem is that $\tilde{A}$ and $\tilde{A}_{\text{test}}$ are no longer independent. So the CLT in Theorem 2.3.1 cannot be applied to $\hat{\lambda}(\tilde{x})$ if A contains Bernoulli elements. This is further discussed and studied in Section 2.5.

Remark. Regarding the delocalization condition (2.10), when all entries of Q are of the same order $\rho = o(1)$, then $\sigma = \Theta(\rho^{1/2})$ and the condition boils down to $\|x\|_\infty = o(\rho^{1/4})$. In supplementary Section 2.7.2 (Corollary 2.7.1), we discuss a sufficient condition for $\|x\|_\infty^2 = o(\sigma)$ to hold in terms of the expected number of edges in B.

2.4 Cross-validated eigenvalue estimation

In this section, we use Theorem 2.3.1 to test the null hypothesis $H_0 : \lambda(\tilde{x}_j) = 0$, where $\tilde{x}_j$ is an eigenvector of $\tilde{A}$. Section 2.4.1 states the algorithm and Section 2.4.2 provides the main theoretical result, i.e., that it is consistent.

2.4.1 The algorithm

The algorithm reports a p-value for each eigenvector. These are then used to estimate k. In addition to the splitting probability ε, the algorithm takes two more parameters: (i) the maximum number $k_{\max}$ of eigenvectors to consider and (ii) the significance level α. We describe the algorithm for an undirected graph with the adjacency matrix $A \in \mathbb{N}^{n\times n}$ in Algorithm 2; see Remark 2.4.1 for rectangular or asymmetric A. After `EigCV`, a few remarks on the theory and the implementation are in order.

Input: Adjacency matrix $A \in \mathbb{N}^{n\times n}$, edge splitting probability $\varepsilon \in (0,1)$, and significance level $\alpha \in (0,1)$

Procedure `EigCV`$(A, \varepsilon, k_{\max}, \text{folds})$**:**

1. **for** $f = 1, \ldots, \text{folds}$ **do**
   i. $\tilde{A}$, $\tilde{A}_{\text{test}} \leftarrow$`ES`$(A, \varepsilon)$ `// Algorithm 1`
   ii. (Optional) Compute $\tilde{L}$ with $\tilde{A}$ as in Equation (2.2).
   iii. Compute the leading $k_{\max}$ eigenvectors of $\tilde{L}$ (or $\tilde{A}$), as $\tilde{x}_1, \ldots, \tilde{x}_{k_{\max}}$.
   iv. **for** $\ell = 2, \ldots, k_{\max}$ **do**
   compute the test statistic

$$T_{f,\ell} = \frac{\tilde{\lambda}_{\text{test}}(\tilde{x}_\ell)}{\tilde{\sigma}_\ell},$$

   where $\tilde{\lambda}_{\text{test}}(x) = x^{\mathrm{T}}\tilde{A}_{\text{test}}x$, and $\tilde{\sigma}_\ell = \sqrt{2\varepsilon(\tilde{x}_\ell^2)^{\mathrm{T}}A\tilde{x}_\ell^2 - \varepsilon(\tilde{x}_\ell^2)^{\mathrm{T}}\text{diag}(A)\tilde{x}_\ell^2}$ is the standard error evaluated using the full graph. Here, $\tilde{x}_\ell^2 \in \mathbb{R}^n$ is the vector $\tilde{x}_\ell$ with each element squared.
2. **for** $\ell = 2, \ldots, k_{\max}$ **do**
   Compute T_ℓ as the mean of the $T_{1,\ell}, \ldots, T_{\text{folds},\ell}$ and compute the one-sided p-value $p_\ell = 1 - \Phi(T_\ell)$, where Φ is the cumulative distribution function of the standard normal distribution.

Output: The graph dimensionality estimate: $\text{argmin}_{k\le k_{\max}}\{p_k \ge \alpha\} - 1$.

Algorithm 2: Eigenvalue cross-validation

Remark. Theorem 2.4.1 studies `EigCV` with folds = 1 and $\tilde{A}$ instead of $\tilde{L}$ in step ii. Moreover, there is an additional step needed in Theorem 2.4.1 to check for delocalization. This technical requirement is further discussed in Section 2.4.2. This step is not used

in `EigCV` or in our code. We allow for these modifications in `EigCV` because they are practically advantageous. Increasing folds > 1 helps to remove the randomness in the p-values generated from edge splitting `ES` (see supplementary Section 2.8.2 for further discussion of folds > 1). Using $\tilde{L}$ instead of $\tilde{A}$ helps reduces localization of eigenvectors [LLV17; ZR18]. Finally, we do not include a check for delocalization because we find in the simulation in Section 2.5 that when an eigenvector $\tilde{x}_\ell$ delocalizes, then $(\tilde{x}_\ell^2)^{\mathrm{T}} A \tilde{x}_\ell^2$ in the formula for $\tilde{\sigma}_\ell$ is very large, thus leading to conservative inferences.

Remark. If folds $= 1$ and the p-values p_k are used to select eigenvectors, then the eigenvectors $\tilde{x}_k$ should be used (not $\hat{x}_k$). This is because the p-value p_k is only associated with the eigenvector $\tilde{x}_k$. It is tempting to compute the eigenvectors of A or L with all of the edges and then give the k-th eigenvector $\hat{x}_k$ the p-values p_k. However, when the left-out edges are also used to compute the eigenvectors, this alters the eigenvectors. In addition to slightly changing the elements of the eigenvectors, it is common for the order of the eigenvectors to also change. Or, for the new eigenvectors to be a more general rotation of the subsampled eigenvectors. It is an area for future research to understand if and how the p-values can be extended. By making ε small we can ensure that the subsampled eigenvectors $\tilde{x}_k$ are nearly as good as $\hat{x}_k$.

`EigCV` easily extends to two other settings, rectangular incidence matrices and a test of independence for contingency tables.

Remark. Rectangular incidence matrices. If the matrix $A \in \mathbb{N}^{r \times c}$ is either rectangular or asymmetric (e.g., the adjacency matrix for a directed graph, the incidence matrix for a bipartite graph, a contingency table, etc.), then eigenvectors should be replaced by singular vectors. In step 3 of `EigCV`, compute the singular vector pairs $\tilde{u}_\ell$

and $\tilde{v}_\ell$. Then, the test statistic is

$$T_\ell = \frac{\tilde{u}_\ell^{\mathrm{T}} \tilde{A}_{\text{test}} \tilde{v}_\ell}{\sqrt{\varepsilon (\tilde{u}_\ell^2)^{\mathrm{T}} A \tilde{v}_\ell}}.$$

Theorem 2.3.1 extends under analogous conditions to this setting. Our `R` package (`https://github.com/RoheLab/gdim`) includes this extension.

Remark. Contingency tables. Suppose $A \in \mathbb{N}^{r \times c}$ is a contingency table with multinomial elements. Note that the χ^2 test of independence tests the null hypothesis that $\mathbb{E}(A)$ is rank 1, i.e., $k = 1$. To test if $\mathbb{E}(A)$ has rank greater than $k = 1$ and potentially reject independence, one can apply `EigCV` to A with $k_{max} = 2$, using the extension to rectangular matrices in Remark 2.4.1. The three key pieces for the cross-validation apply to this setting, thus enabling this approach. First, if the distribution of A is multinomial, then `ES` provides two independent matrices. Second, in expectation, those matrices have identical singular vectors and the same number of non-zero singular values. Third, the CLT in Theorem 2.3.1 extends to this data generating model with analogous conditions. `EigCV` is potentially powerful for alternative hypotheses where $\mathbb{E}(A)$ has a large second eigenvalue. In contrast to the traditional Pearson's χ^2 test for independence, `EigCV` handles a large number of rows and columns and a sparse A, where the vast majority of elements are zeros. Moreover, it has the added advantage that the singular vectors $\tilde{u}_2$ and $\tilde{v}_2$ estimate where the deviation from independence occurs, thus making the results more interpretable. This is an area of our ongoing research.

2.4.2 Statistical consistency

This section states a consistency result for a modified version of the algorithm stated in Algorithm 3 in supplementary Section 2.7.3. The main modification is the addition of a delocalization test. We use K for the true latent dimension and $\hat{K}$ for its estimate.

We will make some further assumptions. Let $P = \rho_n P^0$, where $0 < \rho_n < 1$ controls the sparsity of the network, and $P^0 = U\Lambda^0 U^{\mathrm{T}}$ is a matrix of rank k with $P^0_{ij} \leq 1$ for all i, j. Here, $\Lambda^0 = \mathrm{diag}(\lambda^0_1, \cdots, \lambda^0_K)$ is the diagonal matrix of its non-increasing eigenvalues, and $U = (u_1, \cdots, u_K)$ contains the corresponding eigenvectors. We first consider the signal strength in the population adjacency matrix. The magnitude of the leading eigenvalues characterize the useful signal in the data; only if they are sufficiently large is it possible to identify them from a finite graph sample. As such, the first assumption requires that the leading eigenvalues of the population graph are of sufficient and comparable magnitude. We also include necessary assumptions on the sparsity of the graph.

Assumption 1 (Signal strength and sparsity)**.** We assume that there exist positive constants ψ_1, ψ_1' such that

$$\kappa := \lambda^0_1/\lambda^0_K \in (0, \psi_1), \qquad \lambda^0_1 \geq \psi_1' n.$$

In addition, we assume that $P^0_{ij} \leq 1$ for all i, j and that the network sparsity satisfies $c_0 \frac{\log^{\xi_0} n}{n} \leq \rho_n \leq c_0' n^{-\xi_1}$, for some constants $\xi_0 > 2$, $\xi_1 \in (0, 1)$, $c_0, c_0' > 0$.

Observe that Assumption 1 implies in particular that $\psi_1^{-1}\psi_1' n\rho_n \leq \lambda_K \leq \lambda_1 \leq n\rho_n$ since $\lambda_1 \leq \mathrm{tr}(P) \leq n\rho_n$. Assumption 1 is less strict than the assumptions in [LLZ20]. This is because we do not require a minimum gap between distinct eigenvalues, which is hard to satisfy in practice.

Next, we consider a property of the population eigenvectors. The notion of coherence was previously introduced by [CR09]. Under the parametrization of Assumption 1, the coherence of U is defined as

$$\mu(U) = \max_{i \in [n]} \frac{n}{K} \|U^{\mathrm{T}} e_i\|^2 = \frac{n}{K} \|U\|^2_{2,\infty},$$

where e_i is the i-th standard basis vector. A lower coherence indicates that the population eigenvectors are more spread-out—that is, they are not concentrated on a few coordinates.

Assumption 2 (Coherence)**.** We assume $\mu(U) \leq \mu_0$, for some constant $\mu_0 > 1$.

Our main theoretical result asserts the consistency of our cross-validated eigenvalue estimator for estimating the latent dimension. The proof of Theorem 2.4.1 is in supplementary Section 2.7.

Theorem 2.4.1 (Consistency)**.** Suppose $A \in \mathbb{R}^{n \times n}$ is a Poisson graph satisfying Assumptions 1 and 2. Let K be the true latent space dimension, and let $\hat{K}$ be the output of Algorithm 3 (see supplementary Section 2.7.3) with edge splitting probability ε. Then,

$$\mathbb{P}\left(\hat{K} = K\right) \to 1 \quad \text{as} \quad n \to \infty.$$

2.5 Poisson vs Bernoulli

The Poisson model has previously been used to study statistical inference with random graphs [KN11; FP+20; CD18; CCB16; ZA20]). In the sparse graph setting, these models produce very similar graphs (e.g. Theorem 7 in [Roh+18]). Moreover, under the Poisson model, the edges are exchangeable [CCB16; CD18; Roh+18].

This section shows that `EigCV` continues to perform well in settings where the elements of A are Bernoulli random variables. The technical results and derivations in this chapter do not directly apply to this setting. The key difficulty comes from the edge splitting. Under the Bernoulli model, $A_{ij} \in \{0, 1\}$. So, it is not possible for both $\tilde{A}$ and $\tilde{A}_{\text{test}}$ to get the edge i, j. This creates *negative* dependence. Because the fitting and

test graph are dependent, a central limit theorem akin to Theorem 2.3.1 becomes more difficult. Nevertheless, a theorem akin to Theorem 2.4.1 still holds with minor changes to the conditions.

This simulation shows that the negative dependence created from Bernoulli edges makes the testing procedure more conservative. That is, when we are testing

$$H_0 : \mathbb{E}(\lambda_{\text{test}}(\tilde{x}_\ell)) = 0$$

for $\ell > k$, the $\alpha = .05$ test rejects with probability less than .05. This type of miscalibration is traditionally considered acceptable.

This section simulates from a $k = 2$ Stochastic Blockmodel and examines the distribution of T_3 and T_4, under both the Bernoulli and Poisson models for edges. In all of these simulations, there are $n = 2000$ nodes. Each node is randomly assigned to either block 1 or 2 with equal probabilities. Let i and j be any two nodes in the same block and u and v be any two nodes in different blocks. Across simulation settings, $P_{ij}/P_{u,v} = 2.5$. While keeping this ratio constant, the values in P increase to make the expected degree of the graph go from 3.5 to 105, by increments of 3.5. In the Poisson model, $A_{ij} \sim \text{Poisson}(P_{ij})$ and in the Bernoulli model, $A_{ij} \sim \text{Bernoulli}(P_{ij})$. We compute T_3 and T_4 in `EigCV` (Algorithm 2) with edge split probability $\varepsilon = .05$ and folds $= 1$. Supplementary Section 2.8.2 gives the identical simulation, but for folds $= 10$.

We use the one-sided rejection region $T_\ell > 1.65$. If $T_\ell \sim N(0, 1)$, then this has level $\alpha = .05$. We refer to the simulated probability that $T_\ell > 1.65$ as the rejection probability. Figure 2.4 estimates the rejection probability in two ways. First, each dot gives the proportion of 1000 replicates in which $T_\ell > 1.65$. Second, the line gives the values

$$\hat{\alpha} = 1 - \Phi((1.65 - \bar{T}_\ell)/\text{SD}(T_\ell)), \tag{2.12}$$

where Φ is the cumulative distribution function (CDF) of the standard normal, $\bar{T}_\ell$ is the average value of T_ℓ over the 1000 replicates, and $\text{SD}(T_\ell)$ is the standard deviation of these 1000 replicates. This is an estimate of the rejection probability, under the assumption that T_ℓ is normally distributed.

Across simulation settings, the $\alpha = .05$ test is conservative.

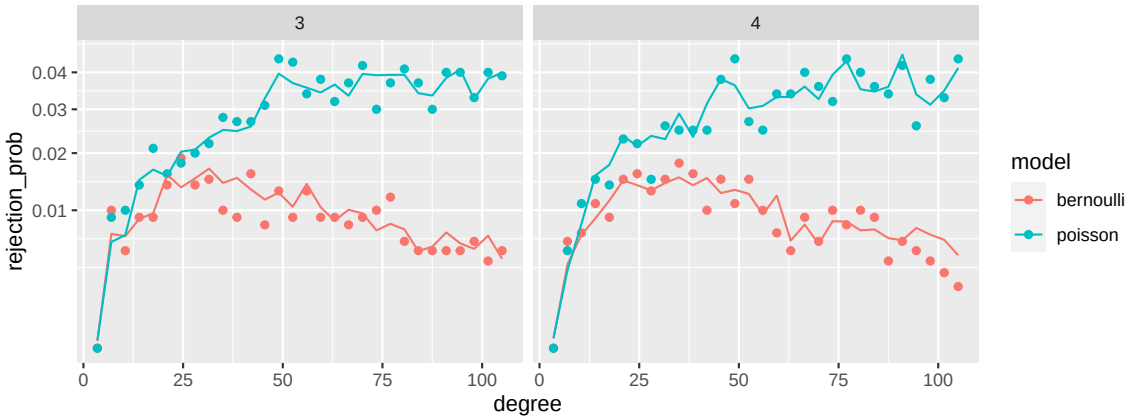

Figure 2.4: The left panel gives results for T_3 and the right panel gives results for T_4. Each point corresponds to the proportion of 1000 replicates that the test statistic was greater than 1.65. The line is a smoothed version of $\hat{\alpha}$.

Under the simulation settings described above, Figure 2.4 shows four things.

1. The points and lines are below .05. So, the proposed test is conservative.

2. On the left side of both panels, the points and lines are well below .05. So, the proposed test is particularly conservative for very sparse graphs. Note that the bottom left point in both figures is over-plotted with two points. This is because none of the 4000 tests in the lowest degree graphs were rejected. The corresponding $\hat{\alpha}$'s are on the order 1/20,000.

3. On the right side of both panels, the red line decreases. So, for Bernoulli graphs, the test is increasingly conservative for denser graphs.

4. The points are scattered around their respective lines, which suggests that the

normal distribution provides a reasonable approximation for the right tail of the distribution (but T_ℓ does not have expectation zero and variance one).

The test is increasingly conservative for dense Bernoulli graphs. In particular, the line giving the normal approximation $\hat{\alpha}$ is decreasing. So, either $\mathbb{E}(T_3) < 0$ or $Var(T_3) < 1$ or both (and similarly for T_4). The next figure, Figure 2.5, shows that the expectation of T_3 and T_4 decreases away from zero for dense Bernoulli graphs.

Under the Bernoulli model, for $\ell > k$, the negative dependence makes the expectation of T_ℓ decrease away from zero as the graph becomes more dense.

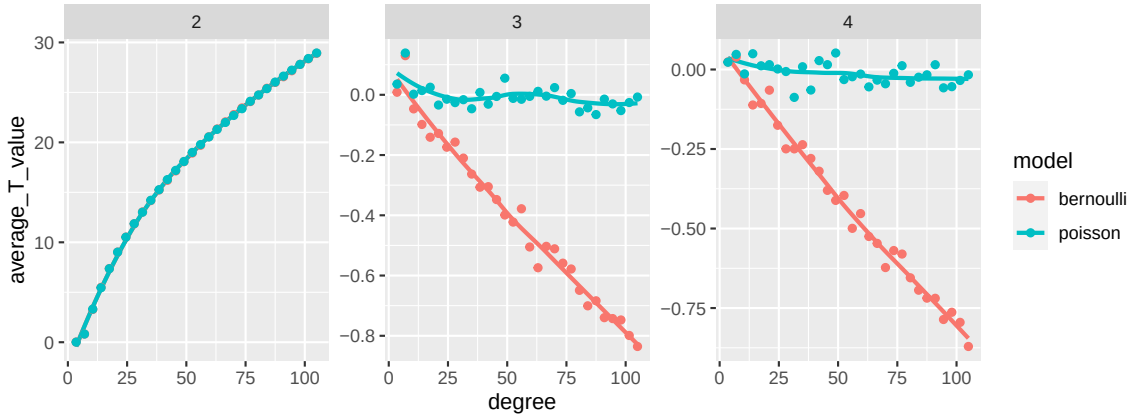

Figure 2.5: Each dot gives the average value of T_ℓ over 1000 replicates; T_2 on the left, T_3 in the middle, and T_4 on the right. The previous figure, Figure 2.4, shows that as the Bernoulli graph becomes denser, T_3 and T_4 are less likely to exceed 1.65. This figure shows that this is because their expectation decreases away from zero.

The expected values of T_3 and T_4 decrease due to the negative dependence between $\tilde{A}$ and $\tilde{A}_{\text{test}}$ under the Bernoulli model. That is, if an edge appears in $\tilde{A}$, then it cannot appear in $\tilde{A}_{\text{test}}$, and vice versa. Figure 2.5 shows that $\tilde{x}_\ell^{\text{T}} \tilde{A} \tilde{x}_\ell$ and $\tilde{x}_\ell^{\text{T}} \tilde{A}_{\text{test}} \tilde{x}_\ell$ are negatively correlated when (1) the edges are Bernoulli, (2) the graph is dense, (3) and $\ell > k$. In the two-block Stochastic Blockmodel, this negative dependence does not shift the expectation of T_2, even for dense Bernoulli graphs (see also Figure 2.8 in supplementary Section 2.8). The fact that $\tilde{x}_\ell$ displays negative correlation, but only for $\ell > k$, is

particularly interesting. Here is one interpretation: because these eigenvectors do not estimate signal, they only find noise in $\tilde{A}$. Then, that noise disappears in $\tilde{A}_{\text{test}}$ due to the negative dependence.

On the very left of both panels in Figure 2.4, the expected degree is 3.5. In this very sparse regime, which is below the weak recovery threshold [MNS15] (see also [Abb18]), the test are conservative for both Bernoulli and Poisson graphs. In this very sparse regime, the eigenvectors $\tilde{x}_\ell$ often localize on a few nodes that have a large degree simply due to chance. When this happens, $(\tilde{x}_\ell^2)^{\mathrm{T}} A \tilde{x}_\ell^2$ over-estimates the population quantity $(\tilde{x}_\ell^2)^{\mathrm{T}} \mathbb{E}(A) \tilde{x}_\ell^2$. This causes $\tilde{\sigma}_\ell$ in the denominator of T_ℓ to be large, thus making the standard deviation of T_ℓ small and the test statistic less likely to exceed 1.65. Figure 2.9 in supplementary Section 2.8 displays the standard deviation of the test statistics. It is comforting that the test is conservative in such scenarios because these localized eigenvectors $\tilde{x}$ are often particularly troubling artifacts of noise. For example, the consistency result in Section 2.4.2 requires an additional step to `EigCV` that discards localized eigenvectors. The current simulation suggests that the additional step is a technical requirement. For this reason, we do not include the additional step in `EigCV`.

Supplementary Section 2.8.2 repeats these figures with folds = 10. Increasing the number of folds makes the T_3 and T_4 even more conservative, while making T_2 more powerful. This happens because the variation in T_ℓ comes from both the original data A and the edge splitting. By increasing the number of folds, the second source of variation diminishes while not disrupting the expectation of T_ℓ.

2.6 Comparing to other approaches

This section compares the proposed method (`EigCV`) with some existing graph dimensionality estimators using both simulated and real graph data. Throughout, we

set the graph splitting probability ε to 0.05, set the significance level cut-off at $\alpha = 0.05$, and folds = 10.

We compare `EigCV` to (1) BHMC, a spectral method based on the Bethe-Hessian matrix with correction [LL19a]; (2) LR, a likelihood ratio method adapting a Bayesian information criterion [WB17]; (3) ECV, an edge cross-validation method with an area under the curve criterion [LLZ20]; (4) NCV, a node cross-validation using an binomial deviance criterion [CL18]; and (5) StGoF (with $\alpha = 0.05$), a stepwise goodness-of-fit estimate [Jin+20]. We performed all computations in `R`. For (1)-(4), we invoked the `R` package `randnet`, and for (5), we implemented the original Matlab code (shared by the authors) in `R`.[1]

2.6.1 Numerical experiments

This section presents several simulation studies that compare our method with other approaches to graph dimensionality. We sampled random graphs with $n = 2,000$ nodes and $k = 10$ blocks from the Degree-Corrected Stochastic Blockmodel (DCSBM). Specifically, for any $i, j = 1, 2, ..., n$,

$$A_{ij} \overset{\text{ind.}}{\sim} \text{Bernoulli}\left(\theta_i \theta_j B_{z(i)z(j)}\right),$$

where $z(i) \in \{1, 2, ..., k\}$ is the block membership of node i, and $B \in \mathbb{R}^{k \times k}$ is the block connectivity matrix, with $B_{ii} = 0.28$ and $B_{ij} = 0.08$ for $i, j = 1, 2, ..., k$, and $\theta_i > 0$ is the degree parameters of node i. We investigated the effects of degree heterogeneity by drawing θ_i's from three distributions (before scaling to unit sum): (i) a point mass distribution, (ii) an Exponential distribution with rate 5, (iii) a Pareto distribution with location parameter 0.5 and dispersion parameter 5. From (i) to (iii), the node degrees become more heterogeneous. Finally, to examine the effects of sparsity, we

[1] The `R` code is also available at `https://github.com/RoheLab/gdim`.

chose the expected average node degree in $\{25, 30, ..., 60\}$. For each simulation setting, we evaluated all methods 100 times.

Figure 2.6 displays the accuracy of all graph dimensionality methods. Here, the accuracy is the fraction of times an estimator successfully identified the true underlying graph dimensionality (which is 10).[2] From the results, both BHMC and ECV offered satisfactory estimation when the graph is degree-homogeneous and the average degree becomes sufficiently large, while they were affected drastically by the existence of degree heterogeneity. The LR estimate was affected by degree heterogeneity as well (although less than BHMC or ECV) and also required a relatively large average node degree to estimate the graph dimensionality. The NCV methods failed to estimate the graph dimensionality under most settings. The StGoF estimate worked better for degree-heterogeneous graphs but required a larger average node degree for accuracy. It is also worth pointing out that the LR and StGoF methods tended to over-estimate the graph dimensionality when the average degree is large, especially for the power-law graphs (see supplementary Figure 2.12). Finally, our method provided a much more accurate dimensionality estimate overall, requiring smaller average node degree and allowing degree heterogeneity. In addition, our testing approach also enjoys a strong advantage of reduced computational cost. To show this, Figure 2.7 depicts the average runtime for each method. It can be seen that the proposed method and BHMC are faster than competing methods by several orders of magnitude. The computational complexity of each StGoF iteration (or test) is at least $O(n^2)$, regardless of whether the graph is sparse or not. Consequently, StGoF requires the longest runtime.

[2]Besides comparison of accuracy, we also compared the deviation of the estimation by each method, for which similar results hold consistently (see supplementary Figure 2.12).

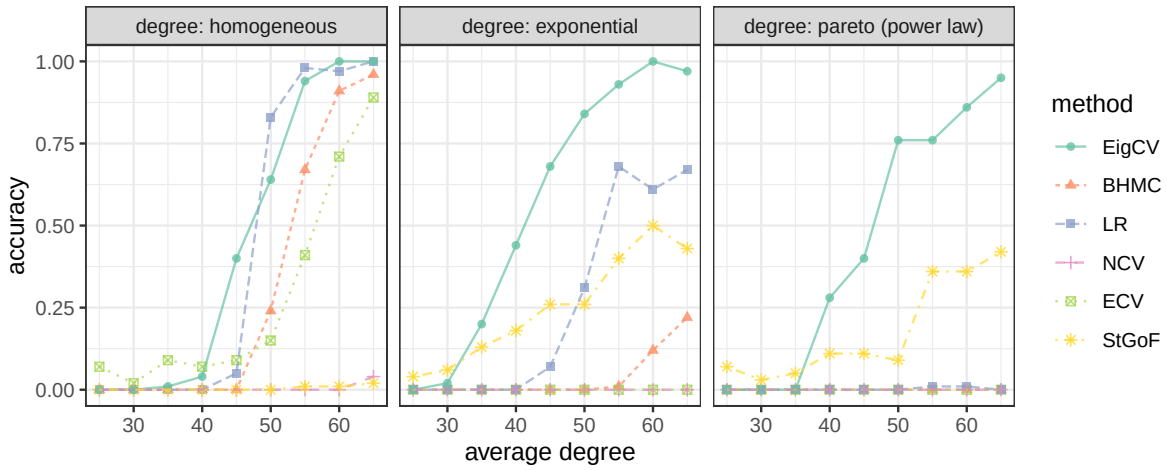

Figure 2.6: Comparison of accuracy for different graph dimensionality estimates under the DCSBM. The panel strips on the top indicate the node degree distribution used. Within each panel, each colored line depicts the relative error of each estimation method as the average node degree increases. Each point on the lines are averaged across 100 repeated experiments.

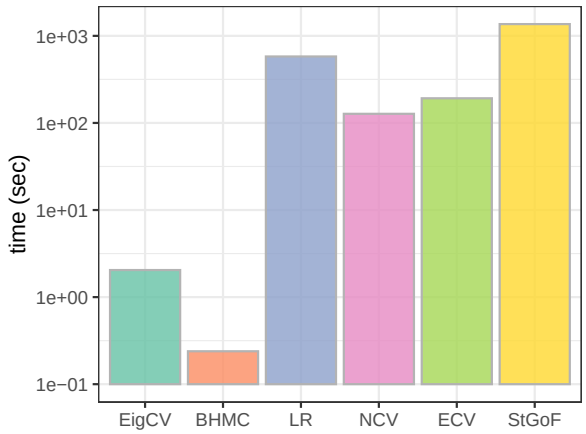

Figure 2.7: Comparison of runtime for the different graph dimensionality methods. Each colored bar indicates the runtime of applying each method on a DCSBM graph with 2000 nodes and 10 blocks. The maximum graph dimensionality is set to 15 for all methods. The runtime was averaged across 100 repeated experiments.

2.6.2 Email network

A real data network was generated using email data within a large European research institution, with each node representing one of the 1005 core members [LKF07]. There is an edge from node i to node j, if i sent at least one email to j. The dataset also contains 42 "ground-truth" community memberships of the nodes. That is, each individual belongs to exactly one of 42 departments at the research institute. For simplicity, we removed the 14 small departments that consist of less than 10 members (see supplementary Table 2.2 for similar results without the filter). This resulted in a directed and unweighted network with a total of 936 nodes from 28 communities.

We applied the graph dimensionality methods to estimate the number of clusters in the network. For the randomized methods (including ECV, NCV, and our proposed method), we ran them 25 times and report the mean and standard deviation of the estimates. For the methods that report a p-value (including StGoF and our proposed method), we use a significance level of $\alpha = 0.01$, followed by a multiplicity correction using the procedure of [BH95]. We set $k_{\max} = 50$. Finally, we chose the splitting probability to be 0.05, as the network is sparse with an average node degree of 23.5. Table 2.1 lists the inferences made by each method. As shown, our method provided an estimate that is close to the true number of departments within the institute. BHMC, LR, NCV, and ECV all estimated small numbers of clusters, while StGoF went significantly larger (≥ 50). These observations were consistent with the simulation results (see supplementary Figure 2.12). Among all the others, only the proposed method provided a close estimate (≈ 28) to the true number of departments. Similarly to the simulation results, the BHMC method and our method are more computationally efficient, with much shorter runtime than the others.

Table 2.1: Comparison of graph dimensionality estimates using the email network among members in a large European research institution. Each members belongs to one of 28 departments.

Method	Estimate (mean)	Runtime (second)
EigCV	28.3	0.68
BHMC	14	0.02
LR	17	85.19
NCV	6.5	204.97
ECV	16.5	41.07
StGoF	≥ 50	397.47

2.7 Technical proofs

2.7.1 Proof of Proposition 2.2.1

Proof. Let $\hat{X} \in \mathbb{R}^{n \times q}$ contain the leading q sample eigenvectors $\hat{x}_1, \ldots, \hat{x}_q$ in its columns. We aim to show that

$$\min_{\Gamma \text{ is diagonal}} \|P - \hat{X}\Gamma\hat{X}^{\mathrm{T}}\|_F$$

contains $\lambda_P(\hat{x}_1), \ldots, \lambda_P(\hat{x}_q)$ down the diagonal.

Using the symmetry of P and the cyclic property of the trace, we obtain

$$\begin{aligned} \|P - \hat{X}\Gamma\hat{X}^{\mathrm{T}}\|_F^2 &= \mathrm{tr}(P^2) + \mathrm{tr}(\hat{X}\Gamma^2\hat{X}^{\mathrm{T}}) - 2\mathrm{tr}(P\hat{X}\Gamma\hat{X}^{\mathrm{T}}) \\ &= \mathrm{tr}(P^2) + \mathrm{tr}(\Gamma^2) - 2\mathrm{tr}(\hat{X}^{\mathrm{T}}P\hat{X}\Gamma). \end{aligned}$$

Taking a derivative with respect to the diagonal of Γ and setting equal to zero gives

$$\Gamma = \mathrm{diag}(\hat{X}^{\mathrm{T}}P\hat{X})$$

which contains $\lambda_P(\hat{x}_1), \ldots, \lambda_P(\hat{x}_q)$ down the diagonal. □

2.7.2 Proof of Theorem 2.3.1

Proof. We will use Lyapunov's CLT for triangular arrays with fourth moment condition (see, e.g., [Dur19, Exercise 3.4.12]). Recall that B_{ij} is Poisson with mean Q_{ij}. Its mean and variance are Q_{ij}, while its central fourth moment is $Q_{ij}(1+3Q_{ij}) \leq 4Q_{ij}$ under the assumption $Q_{ij} \leq 1$. Note that $\sigma^2 = 2\sum_{ij}(x_ix_j)^2Q_{ij} - \sum_{i=j}(x_ix_j)^2Q_{ij} = \sum_{i\leq j}\left(2-\mathbf{1}\{i=j\}\right)^2(x_ix_j)^2Q_{ij}$. To use Lyapunov's CLT, taking into account the symmetry of B, we show that the following ratio converges to zero:

$$\begin{aligned}
&\frac{\sum_{i\leq j}\mathbb{E}\left|\left(2-\mathbf{1}\{i=j\}\right)x_ix_j(B_{ij}-Q_{ij})\right|^4}{\sigma^4} \\
&\leq \frac{\sum_{i\leq j}\left(2-\mathbf{1}\{i=j\}\right)^4(x_ix_j)^4(4Q_{ij})}{\sigma^4} \\
&\leq \frac{16\|x\|_\infty^4\sum_{i\leq j}\left(2-\mathbf{1}\{i=j\}\right)^2(x_ix_j)^2Q_{ij}}{\sigma^4} \\
&= \frac{16\|x\|_\infty^4}{\sigma^2} \\
&= o(1),
\end{aligned} \tag{2.13}$$

where we used the bound on the fourth moment in the first inequality and the delocalization condition on the last line. This shows that

$$\frac{\lambda_B(x)-\lambda_Q(x)}{\sigma^2} \Rightarrow N(0,1). \tag{2.14}$$

Via Slutsky's Lemma, we can multiply the ratio in Equation (2.14) by any sequence that converges to one in probability and the result still holds. The proof is then concluded by showing that $\sigma/\hat{\sigma}$ converges to one in probability. Indeed, we have

$$\begin{aligned}
\mathrm{Var}\left(\frac{\hat{\sigma}^2}{\sigma^2}\right) &= \frac{\mathrm{Var}[(x^2)^{\mathrm{T}}Bx^2-(x^2)^{\mathrm{T}}\mathrm{diag}(B)x^2]}{\sigma^4} \\
&= \frac{\mathrm{Var}[\sum_{i\leq j}\left(2-\mathbf{1}\{i=j\}\right)^2(x_ix_j)^2B_{ij}]}{\sigma^4}
\end{aligned}$$

$$= \frac{\sum_{i \leq j} \left(2 - \mathbf{1}\{i = j\}\right)^4 (x_i x_j)^4 Q_{ij}}{\sigma^4},$$

which is Equation (2.13) up to a factor of 4 and thus $o(1)$. So, by Chebyshev's inequality, $\hat{\sigma}^2/\sigma^2$ converges in probability to its expectation. Note that $\mathbb{E}(\hat{\sigma}^2/\sigma^2) = 1$ and that taking the inverse and the square root is continuous transformation. So, the ratio $\sigma/\hat{\sigma}$ converges in probability to one. □

Corollary 2.7.1

The following corollary gives a sufficient condition for $\|x\|_\infty^2 = o(\sigma)$ to hold in terms of m and the expected number of edges in B.

Corollary 2.7.1. Using the setting of Theorem 2.3.1, let $\pi \in \mathbb{R}^n$ be a probability distribution on the nodes with π_i proportional to a node's expected degree. Define $\langle \pi, x^2 \rangle$ be the expected value of x_I^2 for I drawn from π and define $m = 2^{-1} \sum_i d_i$ as the expected total number of edges. If Q is positive semi-definite and

$$\frac{\|x\|_\infty^2}{\langle \pi, x^2 \rangle} = o\left(\sqrt{m}\right),$$

then the CLT in Equation (2.11) holds.

Proof. The proof of Corollary 2.7.1 follows directly from the next lemma.

Lemma 2.7.1. Suppose $Q \in \mathbb{R}_+^{n \times n}$ is positive semi-definite. Define $d = Q\mathbf{1}_n \in \mathbb{R}^n$ to be the expected degrees of the nodes $1, \ldots, n$, where $\mathbf{1}_n \in \mathbb{R}^n$ is a vector of 1's. Then,

$$\sigma^2 = 2(x^2)^{\mathrm{T}} Q x^2 - (x^2)^{\mathrm{T}} \mathrm{diag}(Q) x^2 \geq \frac{\langle d, x^2 \rangle^2}{\sum_i d_i}.$$

Proof. Define $y = x^2, \theta = d^{1/2}, \Theta = \mathrm{diag}(\theta) \in \mathbb{R}^{n \times n}, y_\theta = \Theta y$, and $\mathscr{L} = \Theta^{-1} Q \Theta^{-1}$.

Because the elements of θ are non-negative, $\mathscr{L}$ is non-negative definite.

The first part of the proof is to show that $\mathscr{L}\theta = \theta$. This is because $\Theta^{-2}Q$ is a Markov transition matrix. So,

$$\Theta^{-2}Q\mathbf{1}_n = \mathbf{1}_n \implies \Theta^{-1}Q\Theta^{-1}\Theta\mathbf{1}_n = \Theta\mathbf{1}_n$$

and this implies that $\mathscr{L}\theta = \theta$. So, by the Perron-Frobenius Theorem, θ is the leading eigenvector of $\mathscr{L}$ with eigenvalue 1.

Let $\mathscr{L}$ have eigenvectors and eigenvalues $(\phi_1, \lambda_1), \ldots, (\phi_n, \lambda_n)$, where $\phi_1 = \theta/\|\theta\|_2$, $\lambda_1 = 1$ and $0 \leq \lambda_j \leq 1$ for $j \neq 1$. Then,

$$y^{\mathrm{T}}Qy = y_\theta^{\mathrm{T}}\mathscr{L}y_\theta = \sum_{\ell=1}^{n} \lambda_\ell \langle \phi_\ell, y_\theta \rangle^2.$$

Keeping only the first order term on the right-hand side, we have

$$y^{\mathrm{T}}Qy \geq \lambda_1 \langle \phi_1, y_\theta \rangle^2 = \frac{\langle d, x^2 \rangle^2}{\sum_i d_i}.$$

The desired result follows from the fact that $\sigma^2 = y^{\mathrm{T}}Qy + y^{\mathrm{T}}(Q - \mathrm{diag}(Q))y \geq y^{\mathrm{T}}Qy$, since y and Q have non-negative entries. □

Applying the bound in the lemma to the delocalization condition and rearranging gives the claim. □

2.7.3 Proof of consistency

This section details the proof of Theorem 2.4.1.

Notation We use the notation $[n]$ to refer to $\{1, 2, ..., n\}$. For any real numbers $a, b \in \mathbb{R}$, we denote $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. For non-negative a_n and b_n that depend on n, we write $a_n \lesssim b_n$ to mean $a_n \leq Cb_n$ for some constant $C > 0$, and similarly for $a_n \gtrsim b_n$. Also we write $a_n = O(b_n)$ to mean $a_n \leq Cb_n$ for some constant $C > 0$. The matrix spectral norm is $\|M\| = \max_{\|x\|_2=1} \|Mx\|_2$, the matrix max-norm is $\|M\|_{\max} = \max_{i,j} |M_{ij}|$, and the matrix $2 \to \infty$ norm is $\|M\|_{2,\infty} = \max_i \|M_{i,\cdot}\|_2$.

Modified algorithm

Algorithm 3 is used in the consistency result.

Input: Adjacency matrix $A \in \mathbb{N}^{n \times n}$, edge splitting probability $\varepsilon \in (0, 1)$
Procedure `EigCV'`$(A, \varepsilon, k_{\max})$**:**

1. Obtain $\tilde{A}$, $\tilde{A}_{\text{test}} \leftarrow$`ES`$(A, \varepsilon)$ from splitting A and set $S = \emptyset$.
   `// Algorithm 1`
2. **for** $k = 2, \ldots, k_{\max}$ **do**
   a - compute $\tilde{\lambda}_{\text{test}}(\tilde{x}_k) = \tilde{x}_k^{\mathrm{T}} \tilde{A}_{\text{test}} \tilde{x}_k$ and
   $\tilde{\sigma}_k = \sqrt{\frac{\varepsilon}{1-\varepsilon} (\tilde{x}_k^2)^{\mathrm{T}} \left(2\tilde{A} - \operatorname{diag}(\tilde{A})\right) \tilde{x}_k^2}$
   b - if
   $$\|\tilde{x}_k\|_\infty^2 \leq \min\left\{\frac{\tilde{\sigma}_k^2}{\log^2 n}, \frac{\log n}{n}\right\},$$
   add k to S and compute
   $$T_k = \frac{\tilde{\lambda}_{\text{test}}(\tilde{x}_k)}{\tilde{\sigma}_k}.$$

Output: The graph dimensionality estimate:
$\hat{K} = |\{T_k \geq \sqrt{n \log n} : k \in S\}|$.

Algorithm 3: Modified eigenvalue cross-validation

Some concentration bounds

We will need several concentration bounds for Poisson random variables. We derive them from standard results.

We begin with a simple moment growth bound.

Lemma 2.7.2 (Poisson moment growth)**.** Let Z be a Poisson random variable with mean $\mu \leq 1$. There exists a universal constant $C > 0$ such that, for all integers $p \geq 2$,

$$\mathbb{E}[|Z - \mu|^p] \leq C\mu \, \frac{p!}{2} \left(\frac{e}{2}\right)^{p-2}.$$

Proof. We show that

$$\mathbb{E}[|Z - \mu|^p] \leq C'\mu \, \left(\frac{p}{2}\right)^p, \tag{2.15}$$

for some constant $C' > 0$. The claim then follows from Stirling's formula in the form

$$\sqrt{2\pi} p^{p+1/2} e^{-p} \leq p!, \qquad \forall p \geq 1.$$

By the definition of the Poisson distribution and using the fact that $0 \leq \mu \leq 1$ by assumption, we have

$$\begin{aligned}
\mathbb{E}[|Z - \mu|^p] &= \sum_{z \geq 0} |z - \mu|^p e^{-\mu} \frac{\mu^z}{z!} \\
&= |\mu|^p e^{-\mu} + |1 - \mu|^p e^{-\mu} \mu + \sum_{z \geq 2} |z - \mu|^p e^{-\mu} \frac{\mu^z}{z!} \\
&\leq 2\mu + \mu^2 e \left\{ \sum_{z \geq 0} z^p \frac{e^{-1}}{z!} \right\}.
\end{aligned}$$

The term in curly brackets on the last line is the p-th moment of a Poisson random variable with mean 1, which is $\leq C'' \left(\frac{p}{2}\right)^p$ for some constant $C'' > 0$ by [Ahl21, Theorem 1]. Eq. (2.15) follows. □

The moment growth bound implies concentration for linear combinations of independent Poisson random variables.

Lemma 2.7.3 (General Bernstein for Poisson variables)**.** Let $Z_1, \ldots, Z_m$ be independent Poisson random variables with respective means $\mu_1, \ldots, \mu_m \leq 1$. For any $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m) \in \mathbb{R}^m$ and $t > 0$,

$$\mathbb{P}\left[\sum_{i=1}^{m} \alpha_i (Z_i - \mu_i) \geq t\right] \leq \exp\left(-\frac{t^2}{C'\mu_{\max}\|\boldsymbol{\alpha}\|_2^2 + C''\|\boldsymbol{\alpha}\|_\infty t}\right),$$

where $\mu_{\max} = \max_i \mu_i$ and $C', C'' > 0$ are universal constants.

Proof. We use [BLM13, Corollary 2.11]. Observe that

$$\sum_{i=1}^{m} \mathbb{E}[\alpha_i (Z_i - \mu_i)^2] = \sum_{i=1}^{m} \alpha_i^2 \mu_i \leq \mu_{\max}\|\boldsymbol{\alpha}\|_2^2.$$

Moreover, by Lemma 2.7.2 and Stirling's formula,

$$\begin{aligned}
\sum_{i=1}^{m} \mathbb{E}[\alpha_i^p (Z_i - \mu_i)_+^p] &\leq \sum_{i=1}^{m} \alpha_i^p C \mu_i \frac{p!}{2}\left(\frac{e}{2}\right)^{p-2} \\
&\leq C\mu_{\max}\|\boldsymbol{\alpha}\|_2^2 \frac{p!}{2}\left(\frac{e}{2}\|\boldsymbol{\alpha}\|_\infty\right)^{p-2} \\
&\leq \frac{p!}{2} v \left(\frac{e}{2}\|\boldsymbol{\alpha}\|_\infty\right)^{p-2},
\end{aligned}$$

where we define

$$v := \max\{1, C\}\, \mu_{\max}\|\boldsymbol{\alpha}\|_2^2.$$

The claim then follows from [BLM13, Corollary 2.11]. □

The moment growth bound also implies spectral norm concentration.

Lemma 2.7.4 (Spectral norm of Poisson graph)**.** Suppose $B \in \mathbb{R}^{n \times n}$ is the adjacency matrix of a Poisson graph with mean matrix Q satisfying $Q_{ij} \leq 1$ for all i, j. Let $q_{\max} = \max_{ij} Q_{ij}$ and assume that $nq_{\max} \geq c_0 \log^{\xi_0} n$ for some $\xi_0 > 2$. Then, for any

$\delta > 0$, there exists a constant $C''' > 0$ such that

$$\|B - Q\| \leq C''' \sqrt{n q_{\max} \log n},$$

with probability at least $1 - n^{-\delta}$.

Proof. We use [Tro12, Theorem 6.2]. We first rewrite the matrix as a finite sum of independent symmetric random matrices

$$B - Q = \sum_{i=1}^{n} \sum_{j=i}^{n} (B_{ij} - Q_{ij}) E^{i,j},$$

where $E^{i,j} \in \mathbb{R}^{n \times n}$ with $E^{i,j}_{ij} = E^{i,j}_{ji} = 1$ and 0 elsewhere.

Observe that, for $i \neq j$,

$$(E^{i,j})^p = \begin{cases} E^{i,i} + E^{j,j} & \text{if } p = 2, 4, \dots \\ E^{i,j} & \text{if } p = 3, 5, \dots \end{cases}$$

while, if $i = j$,

$$(E^{i,i})^p = E^{i,i}, \qquad p \geq 2.$$

Let $X^{i,j} := (B_{ij} - Q_{ij}) E^{i,j}$. Then $\mathbb{E} X^{i,j} = 0$. Moreover, for $i \neq j$ and $p = 2, 4, \dots$, we have

$$\mathbb{E}(X^{i,j})^p = \mathbb{E}(B_{ij} - Q_{ij})^p \left(E^{i,i} + E^{j,j}\right) \preceq C q_{\max} \frac{p!}{2} \left(\frac{e}{2}\right)^{p-2} (E^{i,i} + E^{j,j}),$$

by Lemma 2.7.2. Similarly, for $i \neq j$ and $p = 3, 5, \ldots$,

$$\mathbb{E}(X^{i,j})^p = \mathbb{E}(B_{ij} - Q_{ij})^p \, E^{i,j} \preceq C q_{\max} \frac{p!}{2} \left(\frac{e}{2}\right)^{p-2} (E^{i,i} + E^{j,j}),$$

where we used the fact that the matrix $\left(\begin{smallmatrix} 1 & \alpha \\ \alpha & 1 \end{smallmatrix}\right)$ has eigenvalues $1 + \alpha, 1 - \alpha \geq 0$ when $|\alpha| \leq 1$. When $i = j$,

$$\mathbb{E}(X^{i,i})^p = \mathbb{E}(B_{ii} - Q_{ii})^p \, E^{i,i} \preceq C q_{\max} \frac{p!}{2} \left(\frac{e}{2}\right)^{p-2} (2E^{i,i}).$$

Define

$$(\Sigma^2)^{i,j} := C q_{\max}(E^{i,i} + E^{j,j}).$$

and

$$\sigma^2 \;=\; \left\| \sum_{i=1}^{n} \sum_{j=i}^{n} (\Sigma^2)^{i,j} \right\| = \left\| C q_{\max} \sum_{i=1}^{n} \sum_{j=i}^{n} (E^{i,i} + E^{j,j}) \right\| \leq 2C q_{\max} n,$$

where the inequality holds since $\sum_{i=1}^{n} \sum_{j=i}^{n} (E^{i,i} + E^{j,j})$ is a diagonal matrix with maximum entry $2n$. Then, by [Tro12, Theorem 6.2],

$$\begin{aligned} \mathbb{P}\left[\|B - Q\| \geq t\right] &= \mathbb{P}\left[\left\| \sum_{i=1}^{n} \sum_{j=i}^{n} X^{i,j} \right\| \geq t \right] \\ &\leq n \exp\left(\frac{-t^2/2}{\sigma^2 + (e/2)t} \right) \\ &\leq n \exp\left(\frac{-t^2/2}{2C q_{\max} n + (e/2)t} \right). \end{aligned}$$

Taking $t = C''' \sqrt{n q_{\max} \log n}$ and using the fact that $n q_{\max} \geq c_0 \log^{\xi_0} n$, $\xi_0 > 2$, gives the result. $\square$

Key properties of sample eigenvectors

Consider the adjacency matrix A of a Poisson graph satisfying Assumptions 1 and 2. Fixing $\varepsilon \in (0,1)$, let $\tilde{A}$ and $\tilde{A}_{\text{test}}$ be as in Section 2.3. Let $P = \rho_n P^0 = \mathbb{E}A = \sum_{j=1}^{K} \lambda_j x_j^{\mathrm{T}} x_j$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_K > 0$. Let $\{\tilde{x}_l\}_{l=1}^{k_{\max}}$ be the collection of eigenvectors associated with eigenvalues $\{\tilde{\lambda}_l\}_{l=1}^{k_{\max}}$ of $\tilde{A}$. Without loss of generality, we assume $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \cdots \geq \tilde{\lambda}_{k_{\max}}$. Define

$$\hat{U} = (\tilde{x}_1, \cdots, \tilde{x}_K) \qquad \text{and} \qquad U = (x_1, \cdots, x_K) \in \mathbb{R}^{n \times K}. \tag{2.16}$$

We will need the following event:

$$\mathcal{E}^0 = \left\{ \left\| \tilde{A} - (1-\varepsilon)P \right\| \leq C''' \sqrt{n\rho_n \log n} \right\}.$$

Applying Lemma 2.7.4 with $B := \tilde{A}$ and $Q := (1-\varepsilon)P$ shows that $\mathcal{E}^0$ holds with high probability.

Concentration of signal eigenspace First, we use a version of the Davis-Kahan theorem to show that the signal sample eigenvectors are close to the signal population eigenspace.

Lemma 2.7.5 (Signal eigenspace)**.** Under event $\mathcal{E}^0$, there exists an orthonormal matrix $O \in \mathbb{R}^{K \times K}$ such that, for all $k \in [K]$,

$$\|\tilde{y}_k - x_k\|_2 = O\left(\sqrt{\frac{\log n}{n\rho_n}}\right), \qquad \|\tilde{x}_k - y_k\|_2 = O\left(\sqrt{\frac{\log n}{n\rho_n}}\right),$$

where

$$\tilde{y}_l = \left(\hat{U}O\right)_{\cdot l} = \left(\sum_{i=1}^{K} (\tilde{x}_i)_j O_{il}\right)_{j=1}^{n}, \quad y_l = \left(UO^{\mathrm{T}}\right)_{\cdot l} = \left(\sum_{i=1}^{K} (x_i)_j O_{li}\right)_{j=1}^{n}.$$

Moreover, for all $k \in [K]$, $s \in [K]$, and $t \in [k_{\max}] \setminus [K]$,

$$\langle x_s, y_k \rangle = O_{ks}, \qquad \langle \tilde{x}_t, \tilde{y}_k \rangle = 0.$$

Proof. We use the variant of the Davis-Kahan theorem in [YWS15, Theorem 2]. Under $\mathcal{E}^0$, $\left\| \tilde{A} - (1-\varepsilon)P \right\| = O(\sqrt{n\rho_n \log n})$. By [YWS15, Theorem 2], there exists an orthonormal matrix $O \in \mathbb{R}^{K \times K}$ such that, for all $l \in [K]$,

$$\|\tilde{y}_l - x_l\|_2 \leq \|\hat{U}O - U\|_{\mathrm{F}} = O\left(\frac{\|\tilde{A} - (1-\varepsilon)P\|}{\lambda_K} \right) = O\left(\sqrt{\frac{\log n}{n\rho_n}} \right),$$

and

$$\|\tilde{x}_l - y_l\|_2 \leq \|\hat{U} - UO^{\mathrm{T}}\|_{\mathrm{F}} = \|(\hat{U}O - U)O^{\mathrm{T}}\|_{\mathrm{F}} = \|\hat{U}O - U\|_{\mathrm{F}} = O\left(\sqrt{\frac{\log n}{n\rho_n}} \right),$$

where we used $\lambda_K \geq \psi_1^{-1}\psi_1' n\rho_n$, which holds under Assumption 1.

By the orthonormality of $\{x_l\}_l$ and $\{\tilde{x}_l\}_l$, we have for $s \in [K]$,

$$\langle x_s, y_k \rangle = \sum_{l=1}^{n} (x_s)_l \left(\sum_{i=1}^{K} (x_i)_l O_{ki} \right) = \sum_{i=1}^{K} \left(\sum_{l=1}^{n} (x_s)_l (x_i)_l \right) O_{ki} = O_{ks},$$

and for $t \in [k_{\max}] \setminus [K]$,

$$\langle \tilde{x}_t, \tilde{y}_k \rangle = \sum_{l=1}^{n} (\tilde{x}_t)_l \left(\sum_{i=1}^{K} (\tilde{x}_i)_l O_{ik} \right) = \sum_{i=1}^{K} O_{ik} \left(\sum_{l=1}^{n} (\tilde{x}_t)_l (\tilde{x}_i)_l \right) = \sum_{i=1}^{K} O_{ik} \mathbf{1}_{\{i=t\}} = 0.$$

□

Bounds on population quantities The previous lemma implies bounds on the population quantity of interest, $\lambda_P(\tilde{x}_l)$.

Lemma 2.7.6 (Bounding $\lambda_P(\tilde{x}_l)$)**.** Under event $\mathcal{E}^0$,

$$\begin{aligned} \tilde{x}_l^{\mathrm{T}} P \tilde{x}_l &\gtrsim n\rho_n, \qquad && \forall l \in [K], \\ \tilde{x}_l^{\mathrm{T}} P \tilde{x}_l &\lesssim \log n, \qquad && \forall l \in [k_{\max}] \setminus [K]. \end{aligned}$$

Proof. For $s \in [K]$, expanding $\tilde{x}_s$ over an orthonormal basis including $\{x_l\}_{l \in K}$, we get

$$\begin{aligned} \tilde{x}_s^{\mathrm{T}} P \tilde{x}_s &= \sum_{k=1}^{K} \lambda_k \langle \tilde{x}_s, x_k \rangle^2 \\ &= \sum_{k=1}^{K} \lambda_k \left[\langle x_k, y_s \rangle^2 - \langle x_k, y_s - \tilde{x}_s \rangle \langle x_k, \tilde{x}_s + y_s \rangle \right] \\ &\geq \sum_{k=1}^{K} \lambda_k O_{sk}^2 - \sum_{k=1}^{K} \lambda_k \|y_s - \tilde{x}_s\|_2 \|x_k\|_2^2 \left(\|\tilde{x}_s\|_2 + \|y_s\|_2 \right) \qquad (2.17) \\ &\geq \psi_1^{-1} \psi_1' n\rho_n - O\left(2Kn\rho_n \sqrt{\frac{\log n}{n\rho_n}} \right) \qquad (2.18) \\ &\gtrsim n\rho_n \end{aligned}$$

where inequality (2.17) follows from Cauchy–Schwarz, the triangle inequality and $\langle x_k, y_s \rangle^2 = O_{sk}$ by Lemma 2.7.5. Inequality (2.18) holds since $\sum_{k=1}^{K} O_{sk}^2 = 1$, $\psi_1^{-1} \psi_1' n\rho_n \leq \lambda_k \leq n\rho_n$ by Assumption 1, $\|\tilde{x}_s - y_s\|_2 = O\left(\sqrt{\frac{\log n}{n\rho_n}} \right)$ by Lemma 2.7.5 and $\|\tilde{x}_k\|_2 = \|x_k\|_2 = \|y_s\|_2 = 1$.

For $t \in [k_{\max}] \setminus [K]$,

$$\begin{aligned} \tilde{x}_t^{\mathrm{T}} P \tilde{x}_t &= \sum_{k=1}^{K} \lambda_k \langle \tilde{x}_t, x_k \rangle^2 \\ &= \sum_{k=1}^{K} \lambda_k \langle \tilde{x}_t, x_k - \tilde{y}_k + \tilde{y}_k \rangle^2 \\ &= \sum_{k=1}^{K} \lambda_k \left[\langle \tilde{x}_t, x_k - \tilde{y}_k \rangle + \langle \tilde{x}_t, \tilde{y}_k \rangle \right]^2 \end{aligned}$$

$$= \sum_{k=1}^{K} \lambda_k \langle \tilde{x}_t, x_k - \tilde{y}_k \rangle^2 \tag{2.19}$$

$$\leq K\lambda_1 \max_{k\in[K]} \|x_k - \tilde{y}_k\|_2^2 = O(\log n) \tag{2.20}$$

where equality (2.19) follows from $\langle \tilde{x}_t, \tilde{y}_k \rangle = 0$ by Lemma 2.7.5. Equation (2.20) holds since $\|\tilde{y}_k - x_k\|_2 = O\left(\sqrt{\log n / n\rho_n}\right)$ by Lemma 2.7.5 and $\lambda_k \leq \lambda_1 \leq n\rho_n$ by Assumption 1. □

Delocalization of signal eigenvectors To establish concentration of the estimate $\tilde{\lambda}_{\text{test}}(\tilde{x}_l)$ around $\varepsilon\lambda_P(\tilde{x}_l)$ for $l \in [K]$, we first need to show that $\tilde{x}_l$ is delocalized. That result essentially follows from an entrywise version of Lemma 2.7.5 based on a technical result of [Abb+20].

Lemma 2.7.7 (Delocalization of signal sample eigenvectors)**.** There exist constants $\delta_1 > 0$, $C_1 > 0$ such that the event

$$\mathcal{E}^1 = \left\{ \|\tilde{x}_l\|_\infty \leq C_1 \sqrt{\frac{\mu_0}{n}}, \forall\, l \in [K] \right\}$$

holds with probability at least $1 - 3n^{-\delta_1}$.

Proof. We use [Abb+20, Theorem 2.1] on $\tilde{A}$, which requires four conditions. We check these conditions next. First, let $\tilde{A}^* = (1-\varepsilon)P$, $\Delta^* = \lambda_K$,

$$\kappa = \frac{\lambda_1}{\lambda_K} \leq \psi_1, \tag{2.21}$$

where the inequality follows from Assumption 1,

$$\varphi(x) = \frac{1}{32\psi_1} \min\{\sqrt{n}x, 1\},$$

and

$$\gamma = C'''\psi_1(\psi_1')^{-1}\sqrt{\frac{\log n}{n\rho_n}} \gtrsim \sqrt{\frac{\log n}{n^{1-\xi_1}}}, \tag{2.22}$$

where C''' is the constant in Lemma 2.7.4 and $\psi_1, \psi_1' > 0$, $\xi_1 \in (0,1)$ are the constants in Assumption 1.

(A1) *(Incoherence)* By [Abb+20, Eq. (2.4)] and the remarks that follow it, the incoherence condition is satisfied provided

$$\mu(U) := \frac{n}{K}\|U\|_{2,\infty}^2 \leq \frac{n\gamma^2}{K\kappa^2}.$$

Under Assumption 2, $\mu(U) \leq \mu_0$ while (2.22) implies $n\gamma^2 = \Omega(\log n)$ and (2.21) implies $\kappa = O(1)$. Hence the condition is satisfied.

(A2) *(Row and columnwise independence)* By Lemma 2.3.1, $\tilde{A}$ is the adjacency matrix of a Poisson graph with independent entries. In particular, $\{\tilde{A}_{ij} : i = m \text{ or } j = m\}$ are independent of $\{\tilde{A}_{ij}; i \neq m, j \neq m\}$.

(A3) *(Spectral norm concentration)* As observed previously, applying Lemma 2.7.4 with $B := \tilde{A}$, $Q := (1-\varepsilon)P$ and $\delta > 0$ shows that the event

$$\mathcal{E}^0 = \left\{ \left\| \tilde{A} - (1-\varepsilon)P \right\| \leq C'''\sqrt{n\rho_n \log n} \right\},$$

holds with probability $1 - n^{-\delta}$. Moreover, by the remark after Assumption 1,

$$\gamma\Delta^* = C'''\psi_1(\psi_1')^{-1}\sqrt{\frac{\log n}{n\rho_n}}\lambda_K \geq C'''\sqrt{n\rho_n \log n}.$$

Hence,

$$\mathbb{P}\left[\left\| \tilde{A} - \tilde{A}^* \right\| \leq \gamma\Delta^* \right] \geq 1 - n^{-\delta}.$$

Note further that, under Assumption 1, $\gamma = o(1)$, which implies

$$32\kappa \max\{\gamma, \varphi(\gamma)\} \leq 32\kappa \max\left\{\gamma, \frac{1}{32\psi_1}\right\} \leq 1,$$

for n large enough, as required in [Abb+20, Assumption (A3)], where we used (2.21).

(A4) *(Row concentration)* As required in [Abb+20, Assumption (A4)], the function φ is continuous and non-decreasing on $\mathbb{R}_+$ with $\varphi(0) = 0$ and $\varphi(x)/x$ nonincreasing on $\mathbb{R}_+$. Let $W \in \mathbb{R}^{n\times K}$. By standard norm bounds

$$\frac{1}{\sqrt{n}} \leq \frac{\|W\|_F}{\sqrt{n}\|W\|_{2,\infty}} \leq 1.$$

As a result, by definition of φ,

$$\varphi\left(\frac{\|W\|_F}{\sqrt{n}\|W\|_{2,\infty}}\right) = \frac{1}{32\psi_1}.$$

Let

$$g = \Delta^*\|W\|_{2,\infty}\varphi\left(\frac{\|W\|_F}{\sqrt{n}\|W\|_{2,\infty}}\right) = \frac{1}{32\psi_1}\lambda_K\|W\|_{2,\infty}.$$

Fix $m \in [n]$ and $r \in [K]$. Applying Lemma 2.7.3 on $\tilde{A}_{m\cdot}$ with $\max_{ij}\mathbb{E}\tilde{A}_{ij} \leq (1-\varepsilon)\rho_n$, there exist $c_2 > 0$, $c_2' > 1$ such that

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{i\in[n]}(\tilde{A}_{mi} - \tilde{Q}_{mi})W_{ir}\right| \geq g/\sqrt{K}\right) \\
\leq\ & 2\exp\left(-\frac{g^2/K}{C'(1-\varepsilon)\rho_n\|W_{\cdot r}\|_2^2 + C''\|W_{\cdot r}\|_\infty g/\sqrt{K}}\right) \\
=\ & 2\exp\left(-\frac{\lambda_K^2\|W\|_{2,\infty}^2}{32^2\psi_1^2 KC'(1-\varepsilon)\rho_n\|W_{\cdot r}\|_2^2 + 32\psi_1\sqrt{K}C''\|W_{\cdot r}\|_\infty\lambda_K\|W\|_{2,\infty}}\right)
\end{aligned}$$

$$\begin{aligned} &\leq 2\exp\left(-\frac{\lambda_K^2}{32^2\psi_1^2 KC'(1-\varepsilon)n\rho_n + 32\psi_1\sqrt{K}C''\lambda_K}\right) \\ &\leq 2\exp(-c_2 n\rho_n) \\ &\leq n^{-c_2'}, \end{aligned}$$

where C' and C'' are the constants in Lemma 2.7.3 and we used again that, by the remark after Assumption 1, $\lambda_K \geq \psi_1^{-1}\psi_1' n\rho_n$. In the final inequality, we use that $n\rho_n \geq c_0 \log^{\xi_0} n$, $\xi_0 > 2$ under Assumption 1. Since

$$\left\|(\tilde{A}-\tilde{Q})_{m\cdot}W\right\|_2 \leq \sqrt{K}\sup_r \left|\sum_{i\in[n]}(\tilde{A}_{mi}-\tilde{Q}_{mi})W_{ir}\right|,$$

a union bound over r implies

$$\mathbb{P}\left[\left\|(\tilde{A}-\tilde{Q})_{m\cdot}W\right\|_2 \leq g\right] \geq 1 - Kn^{-c_2'}.$$

Recall the definition of $\hat{U}$ and U from (2.16). Applying [Abb+20, Theorem 2.1] and using [Abb+20, Eq. (2.4)] again, there exists $\tilde{C} > 0$ such that

$$\begin{aligned} \max_{l\in[K]}\|\tilde{x}_l\|_\infty &\leq \left\|\hat{U}\right\|_{2,\infty} \\ &\leq \tilde{C}(2\kappa+\varphi(1))\|U\|_{2,\infty} \\ &\leq \tilde{C}\left(2\psi_1 + \frac{1}{32\psi_1}\right)\sqrt{K}\sqrt{\frac{\mu_0}{n}}, \end{aligned}$$

with probability $1-n^{-\delta}-2n^{-(c_2'-1)}$, where we used Assumption 2 on the last line. Taking $C_1 = \tilde{C}(2\psi_1 + \frac{1}{32\psi_1})\sqrt{K}$ and $\delta_1 = \min\{\delta, c_2'-1\} > 0$ gives the claim. □

Concentration of quadratic forms Next, we show that $\tilde{\lambda}_{\text{test}}(\tilde{x}_l)$ is concentrated around $\varepsilon\lambda_P(\tilde{x}_l)$.

Lemma 2.7.8 (Concentration of $\tilde{\lambda}_{\text{test}}(x)$)**.** Let $x \in \mathbb{R}^n$ be a unit vector such that

$$\|x\|_\infty^2 \le \frac{\log n}{n}, \tag{2.23}$$

then there exists $\delta_2 > 1$ such that

$$\mathbb{P}\left[\left|\sum_{i,j} x_i x_j (\tilde{A}_{\text{test}} - \varepsilon P)_{ij}\right| \le \sqrt{\rho_n \log n}\right] \ge 1 - n^{-\delta_2}.$$

Proof. We use Lemma 2.7.3. From $\|x\|_2 = 1$, we get

$$\begin{aligned}
&\mathbb{P}\left[\left|\sum_{i,j} x_i x_j (\tilde{A}_{\text{test}} - \varepsilon P)_{ij}\right| \ge \sqrt{\rho_n \log n}\right] \\
&\le 2\exp\left(-\frac{(\sqrt{\rho_n \log n})^2/2}{C'\varepsilon\rho_n \sum_{i,j}(x_i x_j)^2 + C'' \max_{ij}|x_i x_j| \sqrt{\rho_n \log n}}\right) \\
&\le 2\exp\left(-\frac{\rho_n \log n/2}{C'\varepsilon\rho_n + C''\|x\|_\infty^2 \sqrt{\rho_n \log n}}\right).
\end{aligned}$$

By Assumption 1, $\rho_n \gg \frac{\log n}{n}$ while $\sqrt{\rho_n \log n} = o(1)$. By (2.23), the denominator on the last line is $\lesssim \rho_n$ and the claim follows. □

We also bound the variance estimate for the signal eigenvectors.

Lemma 2.7.9 (Bound on the variance estimate)**.** Under event $\mathcal{E}^0 \cap \mathcal{E}^1$, for all $l \in [K]$,

$$\tilde{\sigma}_l^2 := \frac{\varepsilon}{1-\varepsilon}(\tilde{x}_l^2)^{\mathrm{T}} \left(2\tilde{A} - \operatorname{diag}(\tilde{A})\right) \tilde{x}_l^2 = \Theta(\rho_n).$$

Proof. Let $\tilde{Q} = (1-\varepsilon)P$. We first show $(\tilde{x}_l^2)^{\mathrm{T}} \tilde{A} \tilde{x}_l^2$ can be controlled via $(\tilde{x}_l^2)^{\mathrm{T}} \tilde{Q} \tilde{x}_l^2$. Indeed observe that for each $l \in [K]$

$$\left|(\tilde{x}_l^2)^{\mathrm{T}} \tilde{A} \tilde{x}_l^2 - (\tilde{x}_l^2)^{\mathrm{T}} \tilde{Q} \tilde{x}_l^2\right| \;=\; |(\tilde{x}_l^2)^{\mathrm{T}} (\tilde{A} - \tilde{Q}) \tilde{x}_l^2|$$

$$\begin{aligned} &\leq \left\|\tilde{A}-\tilde{Q}\right\| \|\tilde{x}_l^2\|_2^2 \\ &\leq \left\|\tilde{A}-\tilde{Q}\right\| \|\tilde{x}_l\|_\infty^2 \|\tilde{x}_l\|_2^2 \\ &= O\left(\sqrt{n\rho_n \log n}\cdot\frac{1}{n}\right) \\ &= O\left(\sqrt{\frac{\rho_n \log n}{n}}\right) \end{aligned}$$

where we used that $\left\|\tilde{A}-\tilde{Q}\right\| = O(\sqrt{n\rho_n \log n})$ under event $\mathcal{E}^0$ and $\|\tilde{x}_l^2\|_\infty = \|\tilde{x}_l\|_\infty^2 = O(\frac{1}{n})$ under $\mathcal{E}^1$. Moreover, observe that $\sqrt{\rho_n \log n/n} \ll \rho_n$ since $n\rho_n \geq c_0 \log^{\xi_0} n$ under Assumption 1. So

$$\left|(\tilde{x}_l^2)^{\mathrm{T}}\tilde{A}\tilde{x}_l^2 - (\tilde{x}_l^2)^{\mathrm{T}}\tilde{Q}\tilde{x}_l^2\right| \ll \rho_n. \tag{2.24}$$

To get an upper bound on $\tilde{\sigma}_l^2$, note that

$$\begin{aligned} (\tilde{x}_l^2)^{\mathrm{T}} P \tilde{x}_l^2 &\leq \lambda_1 \|\tilde{x}_l^2\|_2^2 \\ &\leq \lambda_1 \cdot \|\tilde{x}_l\|_\infty^2 \cdot \|\tilde{x}_l\|_2^2 \\ &= O\left(n\rho_n \cdot \frac{1}{n}\cdot 1\right) \\ &= O(\rho_n), \end{aligned}$$

where we used $\lambda_1 \leq n\rho_n$ by Assumption 1. Hence, we get

$$\begin{aligned} \tilde{\sigma}_l^2 &= \frac{\varepsilon}{1-\varepsilon}\left[2(\tilde{x}_l^2)^{\mathrm{T}}\tilde{A}\tilde{x}_l^2 - (\tilde{x}_l^2)^{\mathrm{T}}\mathrm{diag}(\tilde{A})\tilde{x}_l^2\right] \\ &\leq \frac{2\varepsilon}{1-\varepsilon}(\tilde{x}_l^2)^{\mathrm{T}}\tilde{A}\tilde{x}_l^2 \\ &\leq \frac{2\varepsilon}{1-\varepsilon}|(\tilde{x}_l^2)^{\mathrm{T}}\tilde{A}\tilde{x}_l^2 - (\tilde{x}_l^2)^{\mathrm{T}}\tilde{Q}\tilde{x}_l^2| + \frac{2\varepsilon}{1-\varepsilon}(\tilde{x}_l^2)^{\mathrm{T}}\tilde{Q}\tilde{x}_l^2 \\ &\leq \frac{2\varepsilon}{1-\varepsilon}|(\tilde{x}_l^2)^{\mathrm{T}}\tilde{A}\tilde{x}_l^2 - (\tilde{x}_l^2)^{\mathrm{T}}\tilde{Q}\tilde{x}_l^2| + 2\varepsilon(\tilde{x}_l^2)^{\mathrm{T}}P\tilde{x}_l^2 \\ &= O(\rho_n), \end{aligned}$$

by (2.24).

In the other direction, by Cauchy-Schwarz,

$$(\tilde{x}_l^2)^{\mathrm{T}} P \tilde{x}_l^2 \geq \frac{\left(\tilde{x}_l^{\mathrm{T}} P \tilde{x}_l\right)^2}{\sum_{ij} P_{ij}} \gtrsim \frac{(n\rho_n)^2}{n^2 \rho_n} \gtrsim \rho_n,$$

where the middle inequality follows from Lemma 2.7.6. Combining with (2.24), we have

$$\begin{aligned}
\tilde{\sigma}_l^2 &= \frac{\varepsilon}{1-\varepsilon}\left[(\tilde{x}_l^2)^{\mathrm{T}} \tilde{A} \tilde{x}_l^2 + (\tilde{x}_l^2)^{\mathrm{T}} \left(\tilde{A} - \mathrm{diag}(\tilde{A})\right) \tilde{x}_l^2\right] \\
&\geq \frac{\varepsilon}{1-\varepsilon}(\tilde{x}_l^2)^{\mathrm{T}} \tilde{A} \tilde{x}_l^2 \\
&\geq \frac{\varepsilon}{1-\varepsilon}(\tilde{x}_l^2)^{\mathrm{T}} \tilde{Q} \tilde{x}_l^2 - \frac{\varepsilon}{1-\varepsilon}\left|(\tilde{x}_l^2)^{\mathrm{T}} \tilde{A} \tilde{x}_l^2 - (\tilde{x}_l^2)^{\mathrm{T}} \tilde{Q} \tilde{x}_l^2\right| \\
&= \varepsilon(\tilde{x}_l^2)^{\mathrm{T}} P \tilde{x}_l^2 - \frac{\varepsilon}{1-\varepsilon}\left|(\tilde{x}_l^2)^{\mathrm{T}} \tilde{A} \tilde{x}_l^2 - (\tilde{x}_l^2)^{\mathrm{T}} \tilde{Q} \tilde{x}_l^2\right| \\
&\gtrsim \rho_n.
\end{aligned}$$

That concludes the proof. □

Proof of Theorem 2.4.1

Now, we are ready to prove Theorem 2.4.1.

Proof of Theorem 2.4.1. By Lemmas 2.7.4 and 2.7.7, the event $\mathcal{E}^0 \cap \mathcal{E}^1$ holds with probability $1 - 4n^{-\delta_1}$. Under $\mathcal{E}^0 \cap \mathcal{E}^1$, which depends only on $\tilde{A}$, the claims in Lemmas 2.7.5, 2.7.6 and 2.7.9 also hold. For the rest of the proof, we condition on $\mathcal{E}^0 \cap \mathcal{E}^1$ and use the fact that $\tilde{A}_{\mathrm{test}}$ is independent of $\tilde{A}$ by Lemma 2.3.1.

Let $\tilde{x}_l$, $l \in [k_{\max}]$, be the top $k_{\max}$ unit eigenvectors of $\tilde{A}$ and let

$$\tilde{\sigma}_l^2 = \frac{\varepsilon}{1-\varepsilon}(\tilde{x}_l^2)^{\mathrm{T}} \left(2\tilde{A} - \mathrm{diag}(\tilde{A})\right) \tilde{x}_l^2.$$

Define

$$S = \left\{ l \in [k_{\max}] : \|\tilde{x}_l\|_\infty^2 \leq \min\left\{ \frac{\tilde{\sigma}_l^2}{\log^2 n}, \frac{\log n}{n} \right\} \right\},$$

to be the subset of $[k_{\max}]$ corresponding to sufficiently delocalized eigenvectors. Recall that the test statistic associated to $\tilde{x}_l$ is

$$T_l = \frac{\tilde{x}_l^{\mathrm{T}} \tilde{A}_{\mathrm{test}} \tilde{x}_l}{\tilde{\sigma}_l}.$$

We say that l is rejected if

$$l \in S \quad \text{and} \quad |T_l| \geq \sqrt{n \log n} =: \tau_n.$$

No under-estimation We show that the test statistic associated with the K leading eigenvectors of $\tilde{A}$ will reject the null hypothesis with high probability, that is,

- $[K] \subset S$; and
- $|T_l| \geq \tau_n$, $\forall l \in [K]$.

Fix $s \in [K]$. First, we check that $s \in S$. Under $\mathcal{E}^1$, $\|\tilde{x}_s^2\|_\infty = O(1/n) \ll \log n/n$. We need to check that $\left\|\tilde{\underline{x}}_s^2\right\|_\infty \leq \tilde{\sigma}_s^2/\log^2 n$, for n sufficiently large. This follows from the fact that $\tilde{\sigma}_s^2 = \Theta(\rho_n)$ by Lemma 2.7.9 and $\rho_n \geq c_0 n^{-1} \log^{\xi_0} n$ with $\xi_0 > 2$ under Assumption 1.

Next, we bound $|T_s|$ from below. We have, with probability $1 - n^{-\delta_2}$, where $\delta_2 > 1$ is the constant in Lemma 2.7.8,

$$|T_s| \quad = \quad \left| \frac{\tilde{x}_s^{\mathrm{T}} \tilde{A}_{\mathrm{test}} \tilde{x}_s}{\tilde{\sigma}_s} \right|$$

$$
\begin{aligned}
&\geq \frac{\varepsilon \left|\tilde{x}_s^{\mathrm{T}} P \tilde{x}_s\right| - \left|\tilde{x}_s^{\mathrm{T}} (\tilde{A}_{\mathrm{test}} - \varepsilon P) \tilde{x}_s\right|}{\tilde{\sigma}_s} \\
&\gtrsim \frac{\varepsilon n \rho_n - \sqrt{\rho_n \log n}}{\sqrt{\rho_n}} \\
&\gtrsim n\sqrt{\rho_n} \\
&\gg \sqrt{n \log n}, \qquad (2.25)
\end{aligned}
$$

where the dominating term is controlled through $|\tilde{x}_s^{\mathrm{T}} P \tilde{x}_s| \gtrsim n\rho_n \gg \log n$ by Lemma 2.7.6, the term $|\tilde{x}_s^{\mathrm{T}} (\tilde{A}_{\mathrm{test}} - \varepsilon P) \tilde{x}_s|$ is bounded above by $\sqrt{\rho_n \log n} \ll \log n$ from Lemma 2.7.8 and the denominator satisfies $\tilde{\sigma}_s^2 = \Theta(\rho_n)$ by Lemma 2.7.9. The final bound follows from Assumption 1. By a union bound, (2.25) holds simultaneously for $s \in [K]$ with probability $1 - Kn^{-\delta_2}$.

No over-estimation Then, we show that the noise eigenvectors of $\tilde{A}$ will either be too localized or the test statistic associated with them will fail to reject the null hypothesis. In other words, we show that for any $s \in S \setminus [K]$, it holds that $|T_s| < \tau_n$ with high probability.

Let $t \in S \setminus [K]$. We bound $|T_t|$ from above as follows

$$
\begin{aligned}
|T_t| &= \left| \frac{\tilde{x}_t^{\mathrm{T}} \tilde{A}_{\mathrm{test}} \tilde{x}_t}{\tilde{\sigma}_t} \right| \\
&\leq \frac{\varepsilon \left|\tilde{x}_t^{\mathrm{T}} P \tilde{x}_t\right| + \left|\tilde{x}_t^{\mathrm{T}} (\tilde{A}_{\mathrm{test}} - \varepsilon P) \tilde{x}_t\right|}{\tilde{\sigma}_t} \qquad (2.26) \\
&= O\left(\sqrt{\frac{n}{\log^2 n}} \cdot (\log n + \sqrt{\rho_n \log n}) \right) \\
&= O\left(\sqrt{n}\right). \qquad (2.27)
\end{aligned}
$$

The first term in the numerator of (2.26) satisfies $\left|x_t^{\mathrm{T}} P x_t\right| = O(\log n)$ by Lemma 2.7.6 while the term $|\tilde{x}_t^{\mathrm{T}} (\tilde{A}_{\mathrm{test}} - \varepsilon P) \tilde{x}_t|$ in (2.26) is bounded above by $\sqrt{\rho_n \log n} \ll \log n$ from

Lemma 2.7.8. For the denominator $\tilde{\sigma}_t$, $t \in S$ implies that $\|\tilde{x}_t^2\|_\infty \leq \tilde{\sigma}_t^2/\log^2 n$, thus

$$\tilde{\sigma}_t^2 \geq \log^2 n \cdot \|\tilde{x}_t^2\|_\infty \geq \frac{\log^2 n}{n} \cdot n\|\tilde{x}_t^2\|_\infty \geq \frac{\log^2 n}{n} \cdot \|\tilde{x}_t\|_2^2 = \frac{\log^2 n}{n}.$$

By a union bound, (2.27) holds simultaneously for $t \in S \setminus [K]$ with probability at least $1 - n^{-\delta_2+1}$.

Consistency Therefore, it follows that the algorithm outputs $\hat{K} = K$ with probability tending to 1. □

2.8 Supporting figures and tables

2.8.1 Details of Degree-Corrected Stochastic Blockmodel in the introduction, Figures 2.1 and 2.2

In the introduction, the simulated graph comes from a Degree-Corrected Stochastic Blockmodel (DCSBM). See Section 2.6.1 for a description of this model and its parameters. In Figures 2.1 and 2.2, the DCSBM has $k = 128$ blocks. The 2,560 nodes are randomly assigned to the 128 blocks with uniform probabilities. On average, each block contains 20 nodes. The smallest block has 10 nodes and the largest block has 32. The degree parameters θ_i are distributed as Exponential($\lambda = 1$) and the B matrix is hierarchically structured. In order to specify the elements of B, let $\mathbb{T}$ be a complete binary tree with 7 generations (i.e., $2^7 = 128$ leaves). Each leaf node is assigned to one of the $k = 128$ blocks. Define $u \wedge v \in \mathbb{T}$ as the most recent common ancestor of u and v (i.e., the node closest to the root along the shortest path between u and v). Define $g(u, v)$ as the distance in $\mathbb{T}$ from the root to $u \wedge v$. So, if the shortest path between u and v passes through the root, then $g(u, v) = 0$. Moreover, $g(u, u) = 7$ for all leaf nodes u. Set $B_{u,v} = p2^{g(u,v)}$, where $p = .0008$ is chosen so that the average expected degree of

the nodes is equal to 20.

2.8.2 Bernoulli vs. Poisson

Two clarifying figures for folds = 1

The average value of T_3 and T_4 are different under the Poisson and Bernoulli models (see Figure 2.5 in Section 2.5). To further illustrate this difference and to show that T_2 does not have the same property, Figure 2.8 subtracts the average value of the test statistic *under the Bernoulli model* from the corresponding average under the Poisson model. The difference appears for T_3 and T_4, but not for T_2. This shows that $\tilde{x}_\ell^{\mathrm{T}} \tilde{A} \tilde{x}_\ell$ and $\tilde{x}_\ell^{\mathrm{T}} \tilde{A}_{\mathrm{test}} \tilde{x}_\ell$ are negatively correlated when (1) the edges are Bernoulli, (2) the graph is dense, (3) and $\ell > k$. In the two-block Stochastic Blockmodel, this negative dependence does not shift the expectation of T_2, even for dense Bernoulli graphs.

Figure 2.9 displays the standard deviation for the test statistics T_2, T_3, T_4 over the 1000 replicates in the simulation. This figure is repeated with folds = 10 in Figure 2.11.

Increasing the number of folds in the Bernoulli vs. Poisson comparison

Figure 2.10 repeats Figure 2.4, but with folds = 10 instead of 1. It shows that increasing the number of folds makes T_3 and T_4 more conservative. While Theorem 2.3.1 shows that T_ℓ is asymptotically normal with folds = 1, increasing the number of folds induces unknown dependence between the test statistics. Figure 2.10 suggests that even with 10 folds, the right tail of the distribution of T_3 and T_4 are well approximated by the normal distribution; this is because the bumpy line that gives $\hat{\alpha}$ is close to the points. The tests are conservative because their expectation is not zero and their variance is not one.

Figure 2.11 shows that increasing the number of folds dramatically reduces the

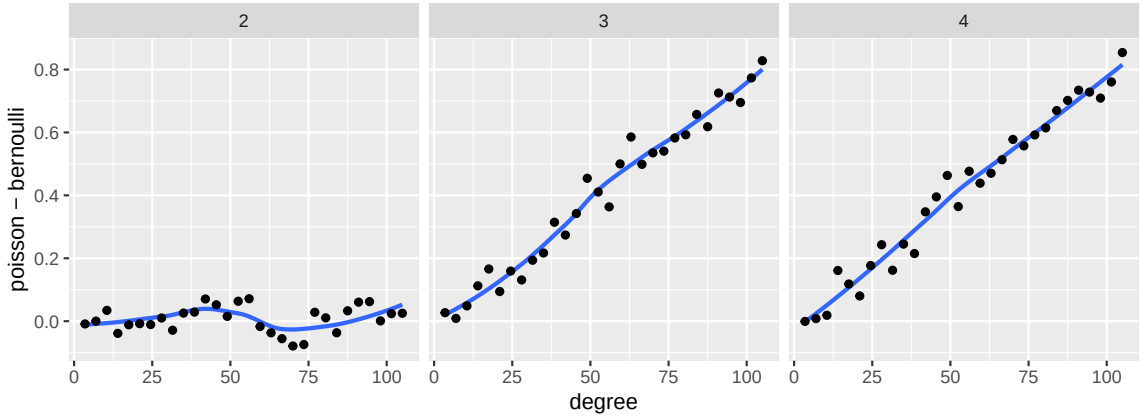

Figure 2.8: Each panel gives the difference between the Poisson and Bernoulli results from Figure 2.5. In this simulation model, there are $k = 2$ blocks in the Stochastic Blockmodel. Taken together, this suggests that the negative dependence in Bernoulli graphs between the fitting and testing adjacency matrices diminishes the expected value of T_ℓ when $\ell > k$ and the graph is more dense. However, the negative dependence does not appear to diminish the expected value of T_2. When a test is conservative under the null, one typically suffers a reduced power under the alternative. However, this result suggests that the negative dependence require us to pay this price. The Bernoulli model makes T_3 conservative, without making T_2 less powerful.

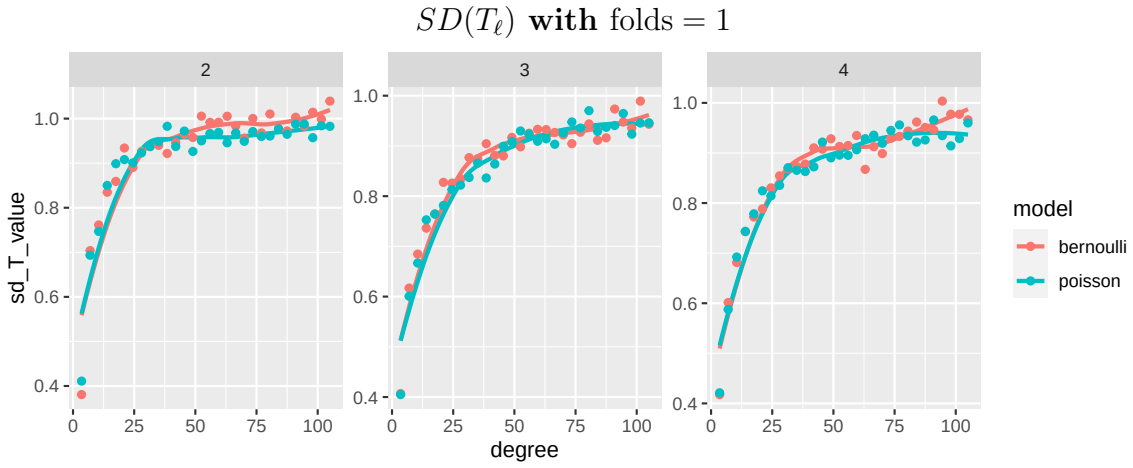

Figure 2.9: As discussed in Section 2.5, the localization of the eigenvectors on small degree graphs (on the left of each panel) makes $\tilde{\sigma}$ over estimate the standard error. This makes the standard deviation of the test statistics over the 1000 replicates smaller than 1.

variation in T_2, T_3, T_4. This makes T_2 more powerful and T_3, T_4 more conservative.

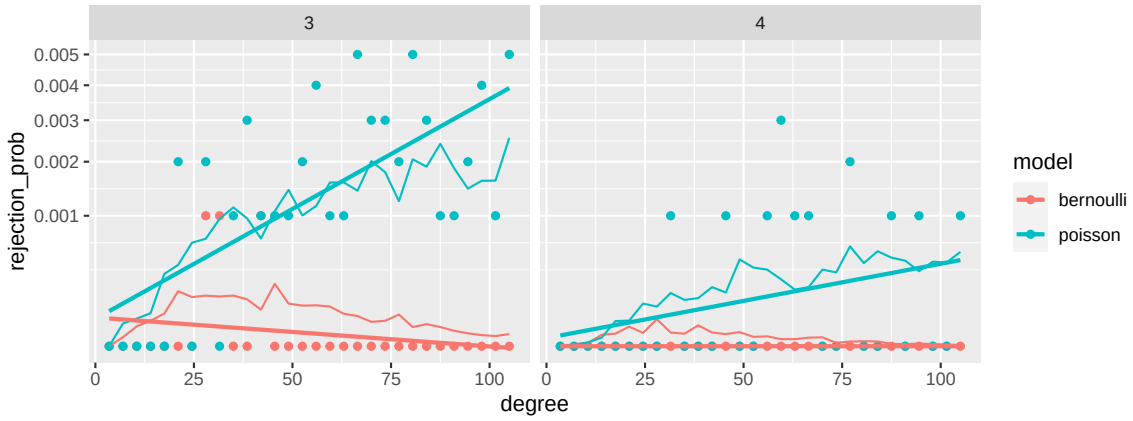

Figure 2.10: This is a repeat of Figure 2.4, but with folds = 10 instead of 1. The vertical axis gives estimates of the rejection probability for T_3 (left) and T_4 (right). Each dot give the proportion of 1000 simulations for which the test statistic exceeds 1.65. The bumpy line interpolates the values $\hat{\alpha}$, defined in Equation (2.12). Because there are so few rejections, particularly for T_4, it is difficult to see whether the bumpy line is close to the points. The straight line is the ordinary least squares fit to the points. It roughly aligns with the bumpy line. This suggests that the right tails of the distributions for T_3 and T_4 are approximately normally distributed.

2.8.3 Comparing to other techniques

In Figure 2.6, we evaluated the accuracy of each method when requiring the exact recovery of k. In order to illustrate how each method either under-estimates or over-estimates k, Figure 2.12 displays the results in Figure 2.6 by the relative error for each estimate $\hat{k}$, which is defined as

$$\text{relative error} = \frac{\hat{k} - k^*}{k^*},$$

where $k^* = 10$ is the true k. From the simulation results, we observed that most methods under-estimate k when the average degree of the graph is smaller (i.e., sparser), except for StGoF which over-estimates it. In addition, from the standard deviation of the relative

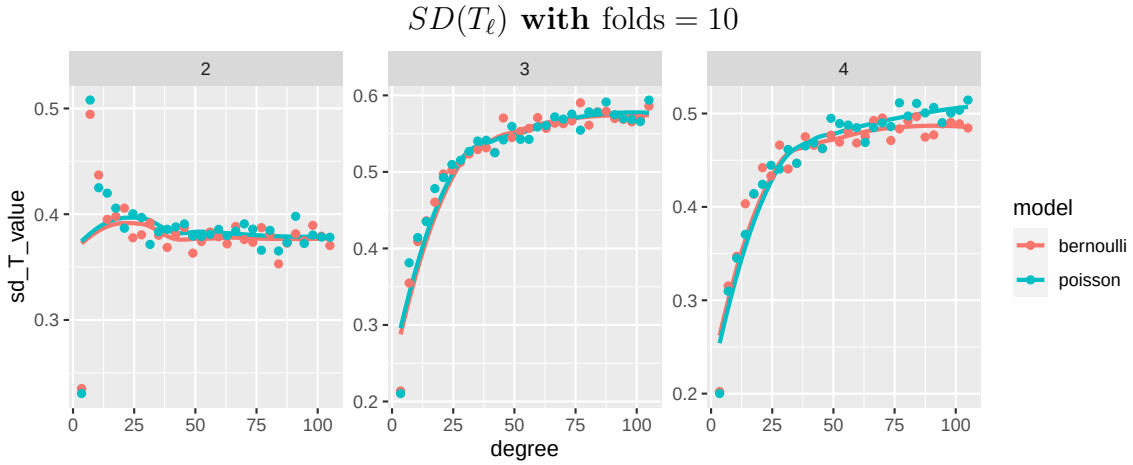

Figure 2.11: This is a repeat of Figure 2.9, but with folds = 10 instead of 1. The vertical axis gives the standard deviation of the test statistics T_2, T_3, T_4 over the 1000 replicates. Notice that they are all significantly less than one. This explains why the rejection probabilities in Figure 2.10 are significantly lower than .05.

error, we observe that `EigCV` provides a more accurate and less variable estimation of k as the graph sparsity varies.

In Section 2.6.2, we removed the 14 small departments that consist of less than 10 members. Among these, two departments have only one members, and eight departments have less than five members. Table 2.2 compares six methods using this email network without filtering. We observed similarly that `EigCV` provided a closer estimate of k than other methods.

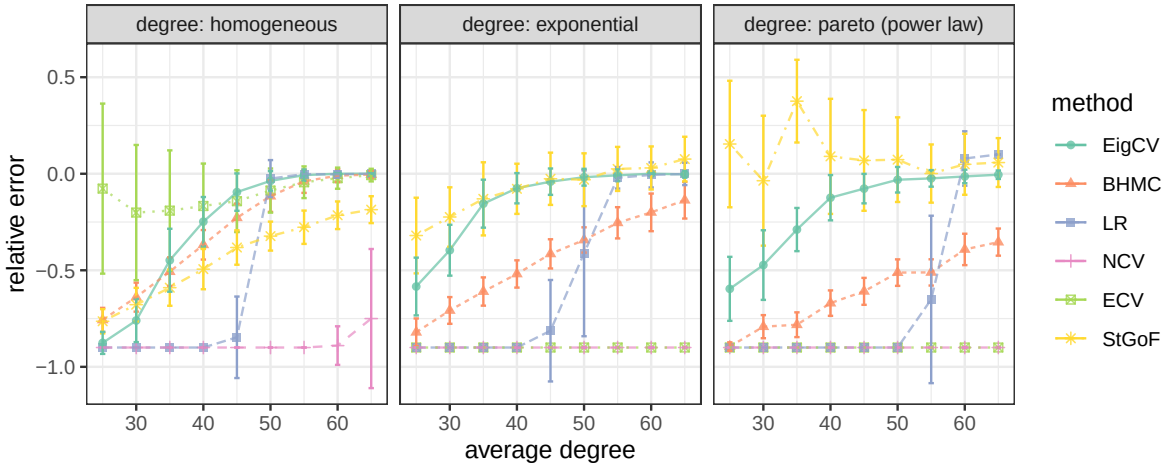

Figure 2.12: Comparison of relative error for different graph dimensionality estimators under the DCSBM. The panel strips on the top indicate the node degree distribution used. Within each panel, each colored line depicts the relative error of each estimation method as the average node degree increases. Each point on the lines are averaged across 100 repeated experiments. For each point, an error bar indicates the sample standard deviation of relative errors.

Table 2.2: Comparison of graph dimensionality estimates using the email network among members in a large European research institution. Each members belongs to one of 42 departments.

Method	Estimate (mean)	Runtime (second)
EigCV	30.56	0.81
BHMC	14.00	0.04
LR	13.00	128.17
NCV	6.96	271.15
ECV	20.08	60.13
StGoF	> 50	544.66

Chapter 3

Species tree inference in the presence of lateral gene transfer: identifiability and consistency

3.1 Background

Phylogenetic inference is one of the central problems in modern phylogenomics. However, phylogenetic networks inferred from different genes often imply different, conflicting evolutionary histories from which they have been sampled [Pol+06; GD08; Cra+09; Nak13]. In addition to statistical errors in gene tree estimation, there are several well-recognized causes of gene tree incongruence, including incomplete lineage sorting (ILS), hybridization, lateral gene transfer (LGT), gene duplication and loss.

Among these evolutionary processes, ILS and hybridization are the two most studied processes. Firstly, the ILS is a population genetic phenomenon, which describes the failure of ancestral gene copies to coalesce into a common ancestral copy until earlier than the previous speciation event. It has been proved that the topology of the phylogeny can be reconstructed from sufficiently many genes in the presence of the ILS under some mild assumptions [MK06; LP07; MR08; SR08]. Secondly, hybridization is also an important source of gene incongruence in bacteria, plants and animals [Arn97;

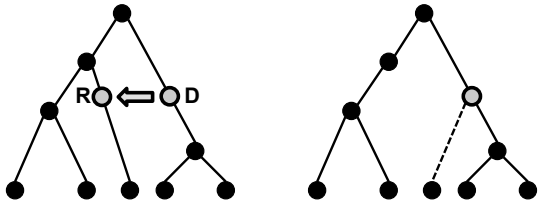

Figure 3.1: An LGT event. On the left, the species phylogeny is shown with the donor (D) and recipient (R) locations. On the right, the resulting (unweighted) gene tree is shown after the LGT event.

Rie97; Mal07; Lam+18]. One can identify the hybridization by observing morphological or molecular intermediacy and/or combination if exists. There are some works on hybridization detection in the presence of the ILS using joint analysis of hybrids and putative parental taxa [MK09; YDN12].

The LGT is another source of phylogenetic inference conflicting, and also the focus of this chapter, see the example in Figure 3.1 from [DR16]. Growing evidences show that LGT play important role in the evolution [SSJ03; McD+10; WC11; MC17; Hib+21], so there is a key problem in modern phylogenomics: *Is the species tree identifiable from the gene trees regardless of the LGT rate?*

A stochastic model of LGT was introduced by Roch and Snir [RS13], LGT events occur at random along the phylogeny according to a Poisson process (see Section 3.2 for details), for each gene independently. The goal is to recover the species phylogeny from a collection of gene trees, each of which can be thought of as randomly scrambled instance of the species phylogeny. A related model was also studied in [LRH07; Ste+13; SS13]. It was proved in [RS13] that under the assumptions in Section 3.2.2, a species phylogeny with n leaves can be recovered from a logarithmic number of genes when the LGT rate is at most $O(1/\log n)$ per unit time. [RS13] also showed that the species phylogenies cannot be distinguished with constant probability from the same number of genes when the LGT rate is of the order of $\Omega(\log\log n)$ per unit time. Under the same

assumptions, the algorithm result in [DR16] improved the LGT rate to a small constant bound by a recursive approach which progressively builds the species phylogeny from the leaves up, using the information obtained from partially reconstructed subtrees to reach further into the past. Similar result was given in [Ste+13] by showing the subtree spanned by any three leaves is statistically consistency when the LGT rate is small, and then the species tree can be reconstructed from the gene trees by majority voting.

In this chapter, we further close the gap in LGT rate, that is, we show a phylogeny is identifiable from the distribution of the gene trees when the LGT rate is arbitrarily constant per unit of time. When the LGT rate is constantly large, any subtree of the gene tree with leaves more than three mismatches the species phylogeny restricting to the same leaf-set with high probability. Because of this statistical inconsistency, the idea like partially recovering from the leaves in [DR16] and the majority voting in [Ste+13] both fail. Instead, we view the LGT process as a non-homogeneous continuous-time Markov process on graph space. Since the LGT events happen chronologically on the phylogeny from the root, we can reconstruct the phylogeny backward in time from the distribution of the gene tree from the leaves up by Markov property once we define the generator matrix Q_t correctly. Observing that the generator matrix Q_t is determined by the edges on the species phylogeny at time t and Q_t changes only when crossing a speciation time, so we can recover Q_t recursively from the leaf-edges on gene trees. This idea can be adapted to a consistent reconstruction algorithm. In the counter-example which has long leaf edges on the phylogeny, we show the limiting distribution only depends on the topology of the gene trees which implies the failure of majority voting. Under the same assumptions in [RS13; DR16] with the pairwise distance between leaves on gene trees, we improve the algorithm result in [DR16] in the sense of relaxing LGT rate, that is, we show a species phylogeny with n leaves can be reconstructed from an exponential number of genes even when the LGT rate is large in constant. We observe that the pairwise distance between two leaves depends on the last LGT event on the

path between these two leaves. By the property of the continuous-time Poisson process, an LGT event happens at a specific location on the phylogeny has probability zero. After considering all $\binom{n}{2}$ pairwise distances between the n leaves on all gene trees, we can identify the species' phylogeny from a collection of gene trees by majority voting.

3.2 Definitions and Results

In this section, we introduce the stochastic model of lateral genetic transfer (LGT), which is based on a related model of Galtier [Gal07; GD08]. We assume that LGT events occur at random along the species phylogeny and follow roughly the presentation in [RS13] and [DR16]. We first introduce the model in Section 3.2.1, and then we give the formal statement of our main results in Section 3.2.2.

3.2.1 Stochastic Model of LGT

A species phylogeny is a graphical representation of the speciation history of a group of organisms. The leaves correspond to extant species. Each branching indicates a speciation event. We associate each edge with a positive value corresponding to the time elapsed along that edge.

Definition 2 (Phylogeny)**.** A species phylogeny $T_s = (V_s, E_s; r, \tau, \mu)$ is a directed tree rooted at r with vertex set V_s, edge set E_s and n labelled leaves $L = [n] = \{1, 2, \cdots, n\}$ such that 1) the degree of all internal vertices $V_s - L$ is exactly 3 except the root r which has degree 2, and 2) the edges are assigned inter-speciation times $\tau : E_s \to (0, +\infty)$ as well as a fixed *rate of substitution* $\mu(e) = \mu \in (0, \infty)$, for some constant μ. We assume that T_s is ultrametric, that is, from every node, the path lengths with respect to τ from that node to all its descendant leaves are equal.

Let $T_s = (V_s, E_s; r, \tau, \mu)$ be a fixed species phylogeny, we define the following

notations. By a *location* in T_s, we mean any position along T_s seen as a continuous object, that is, a point x along an edge $e \in E_s$ and denoted as $x \in e$ in that case. We denote the set of locations in T_s by $\mathcal{X}_s$. We say that $x \in \mathcal{X}_s$ is an *ancestor* of $y \in \mathcal{X}_s$ if x is on the path between y and r in T_s (in which case y is also a *descendant* of x). For any two locations x, y in $\mathcal{X}_s$, we let $\mathrm{MRCA}(x, y)$ be their most recent common ancestor (MRCA) in T_s.

Phylogeny is naturally equipped with notions of distance and time, that are useful in reconstructing phylogeny.

Definition 3 (Species metric)**.** A species phylogeny $T_s = (V_s, E_s; r, \tau, \mu)$ induces a metric d_s on the leaves defined as follows, for all $u, v \in L$:

$$d_s(u, v) = \sum_{e \in p(u,v)} \tau(e) \cdot \mu$$

where $p(u, v)$ is the unique path between leaves u, v in the phylogeny, viewed as a set of edges. We call d_s the species metric. We extend to all vertices naturally. Note that, given our assumption that τ is an ultrametric, d_s is also an ultrametric.

Definition 4 (Time)**.** Given a species phylogeny $T_s = (V_s, E_s; r, \tau, \mu)$, we assign a notion of time on T_s defined as follows, for all $x \in V_s$:

$$t(x) = \sum_{e \in p(r,x)} \tau(e)$$

where $p(r, x)$ is the unique path from the root r to x in the phylogeny, viewed as a set of edges. We extend to the locations along an edge linearly in a natural way, i.e. $t(x) < t(y)$ if x is an ancestor of y. Denoted as $t(r) = 0$. Note that, given our assumption that τ is an ultrametric, $t(u) = t(v)$ for all leaves $u, v \in L$ and we denote it as t_L.

We say that two locations x, y are *contemporaneous* if they correspond to the same time, that is,

$$t(x) = t(y).$$

We let

$$\mathcal{C}_x = \{y \in \mathcal{X}_s : t(y) = t(x) < +\infty\}$$

be the set of locations contemporaneous to x on T_s. Let $\lambda : [0, t_L] \to (0, \infty)$ be the LGT rate assigned to T_s along time. For each edge $e = (x, y) \in E_s$ with $t(x) < t(y)$, $\Lambda(e) = \int_{t(x)}^{t(y)} \lambda(z)dz$ gives the expected number of LGT events on edge e. Further, we let

$$\Lambda_{tot} = \sum_{e \in E_s} \Lambda(e)$$

be the *total LGT weight* of the phylogeny.

To infer the species phylogeny, we first reconstruct gene trees, that is, trees of ancestor-descendant relationships for orthologous genes(or, more generally, loci). Phylogenomic studies have revealed extensive discordance between gene trees(eg. [Bap+05], [DB07]).

Definition 5 (Gene tree)**.** A rooted gene tree $T_g = (V_g, E_g; r, \omega_g)$ for gene g is a directed tree rooted at r with vertex set V_g, edge set E_g and the same leaf-set $L = \{1, \ldots, n\}$ as the species phylogeny such that 1) the degree of every internal vertex is either 2 or 3, and 2) the edges are assigned branch lengths $\omega_g : E_g \to (0, +\infty)$. We let $\mathcal{T}_g = \mathcal{T}[T_g]$ be the topology of T_g where each internal vertex of degree 2 is suppressed except the root r which has degree 2. And we let $\mathcal{T}_g^{unrooted}$ be the topology of T_g where each internal vertex of degree 2 is suppressed including the root r.

As will become clear from the description of the LGT process below (see also Figure 3.1), each edge e of the gene tree T_g corresponds to a full or a partial edge

of the species phylogeny T_s. In particular, there exists a mapping $(\eta, \zeta_b, \zeta_f) : E_g \to E_s \times \mathbb{R}_+ \times \mathbb{R}_+$, mapping an edge $e \in E_g$ to an edge $\eta(e) \in E_s$ and a pair of times $0 \leq \zeta_b(e) \leq \zeta_f(e) \leq \tau(\eta(e))$. The quantities $\zeta_b(e)$ and $\zeta_f(e)$ represent times of LGT events on edge $\eta(e)$, as we will define below.

Definition 6 (Stochastic model of LGT)**.** Let $T_s = (V_s, E_s; r, \tau, \mu)$ be a fixed species phylogeny. A gene tree T_g is generated according to the following continuous-time stochastic process, which gradually modifies the species phylogeny starting at the root. There are two components to the process:

1. **LGT locations.** The recipient and donor locations of LGT events are selected as follows:

   - *Recipient locations.* Starting from the root, along each branch e of T_s, locations are selected as recipient of a genetic transfer according to a continuous-time Poisson process with rate λ. Equivalently, the total number of LGT events is Poisson with mean Λ_{tot} and each such event is located independently according to the following density. For a location x on branch e, the density at x is $\Lambda(e)/\Lambda_{tot}$.

   - *Donor locations.* If x is selected as a recipient location, the corresponding donor location y is chosen uniformly at random in $\mathcal{C}_x$. The LGT transfer $\sigma(x, y)$ is then obtained by performing an SPR move from x to y, that is, the subtree below x in T_s is moved to y in T_g.

   The probability that a recipient or donor location coincides with a node of T_s is 0. If that happens, we associate the recipient/donor to one of the the adjacent edges arbitrarily.

2. **Executing the LGT Process:** We perform genetic transfers chronologically

from the root:

- We initialize the gene tree as follows: $V_g = V_s$, $E_g = E_s$.
- We also initialize the mappings (η, ζ_b, ζ_f) as follows, for all $e \in E_g$: $\eta(e) = e$; $\zeta_b(e) = 0$; $\zeta_f(e) = \tau(e)$.
- We process the LGT events chronologically as follows:
  (a) Suppose the next event to process has $x \in e \in E_s$ as recipient location and $y \in e' \in E_s$ as donor location.
  (b) We find the unique edges $e_x, e_y \in E_g$ such that:
    - $\eta(e_x) = e$ and $\eta(e_y) = e'$; and
    - $\zeta_b(e_x) \leq \tau_x \leq \zeta_f(e_x)$ and $\zeta_b(e_y) \leq \tau_y \leq \zeta_f(e_y)$;

    where τ_x is the time between x and its most recent ancestor in T_s, and similarly for τ_y.
  (c) We introduce a new node v, splitting e_y into two consecutive edges, e_{y_1} and e_{y_2}, with the following features:
    - $\eta(e_{y_1}) = \eta(e_{y_2}) = e'$;
    - $\zeta_b(e_{y_1}) = \zeta_b(e_y)$; $\zeta_f(e_{y_1}) = \tau_y$;
    - $\zeta_b(e_{y_2}) = \tau_y$; $\zeta_f(e_{y_2}) = \zeta_f(e_y)$.
  (d) If $e_x = (u, w)$, we update it to $e_x = (v, w)$, and change $\zeta_b(e_x) = \tau_x$.

After all LGT events have been processed, the weights on the resulting gene tree T_g are defined as follows. For all $e \in E_g$, $\omega_g(e) = \big(\zeta_f(e) - \zeta_b(e)\big) \cdot \mu$.

We call $\sigma(e_r, e_d)$ the LGT transfer events with recipient location $x \in e_r$ and donor location $y \in e_d$. Observe that LGT process keep the same leaves on the phylogeny and LGT events may disconnect subtrees of the species phylogeny from their original roots, connecting them to other branches of the gene tree, thereby creating nodes of degree 2 in the gene tree. We allow internal vertices of degree 2 in a gene tree to potentially delineate between two consecutive species phylogeny edges, see Figture 3.1.

3.2.2 Species phylogeny inference in the presence of LGT

Let $T_s = (V_s, E_s; r, \tau, \mu)$ be an unknown species phylogeny and $\mathcal{T}_s = (V_s, E_s)$ be the topology of T_s. We say the species phylogeny is *identifiable* if we can identify T_s from the distribution of gene trees under a leaf labeled respecting isomorphism. We assume that N independent gene trees $T_{g_1}, \cdots, T_{g_N}$, corresponding to homologous genes $g_1, \cdots, g_N$, were generated according to the process of Definition 6. We aim to *reconstruct* the species phylogeny, given N gene trees (or their topologies).

However, we assume that we have *imperfect* knowledge of the gene trees as these trees are ultimately reconstructed from genetic sequences. Namely, we suppress all nodes of degree 2 (may also including the root), and we are missing all correspondence between the other internal nodes of the gene trees with the nodes or edges of the species phylogeny that derived them.

We organize this section in the following: we first introduce the identifiability result in section 3.2.2; then we provide a new reconstruction algorithm in section 3.2.2; we state the reconstruction result with add-on pairwise distance between leaves in section 3.2.2 .

Identifying the species phylogeny

Given the distribution of topology of gene tree T_g under the process of Definition 6, we focus on identifying the topology $\mathcal{T}_s$ of the phylogeny, namely identify the rooted tree (V_s, E_s) up to a leaf-label respecting isomorphism, for some constant LGT rate. Since the root is usual impossible to be identified from the sequential data, we will try to identify the unrooted topology of species trees from the distribution of the topology of unrooted gene trees.

Theorem 3.2.1 (Identifiability)**.** We can identify the species phylogeny without root from the distribution of the topology of unrooted gene tree generated under the process of Definition 6 with transfer rate $\lambda \in (0, \infty)$.

To proof Theorem 3.2.1, we view the LGT process in Definition 6 as a non-homogeneous continuous time Markov chain on graph space $\mathcal{G}$ containing all possible binary species trees with n labeled leaves, we reverse the Markov process to recover the species phylogeny backward in time from the isolated single leaf clusters. The speciation time corresponding to the internal vertices can be identified since it's the first time that two clusters merge, which leads to a shrink in distribution. We will provide more details in section 3.3.

Furthermore, since we recover the phylogeny from bottom up, with the extra root information on the gene trees, we can identify the species phylogeny with root by modifying the graph space $\mathcal{G}$ in the proof of Theorem 3.2.1.

Corollary 3.2.1 (Identifiability with root)**.** *We can identify the species phylogeny (with root) from the distribution of topology of rooted gene tree generated under the process of Definition 6 with transfer rate* $\lambda \in (0, \infty)$.

The transfer rate λ can be relaxed to more general case. In the proof of

Theorem 3.2.1, we reverse the Markov process backward in time, so it sufficient to require λ be an integrable function of time.

Corollary 3.2.2 (General transfer rate)**.** *Theorem 3.2.1 and Corollary 3.2.1 hold for* $\lambda : [0, t_L] \to (0, \infty)$ *which is a function of time such that* $\int_0^{t_L} \lambda(t)dt < \infty$*, where* $t_L := t(L)$ *is the time of the leaf-set* L *of the species phylogeny.*

Reconstructing $\mathcal{T}_s$ from finite gene trees

Observing the distribution of gene tree is impossible in practice, we introduce the reconstruction Algorithm 4 which is inspired by the proof of identifiability in Theorem 3.2.1. Given N independent gene trees $T_{g_1}, \cdots, T_{g_N}$ generated under the process of Definition 6 with no root, Algorithm 4 aims to reconstruct the topology $\mathcal{T}_s$ of the phylogeny without root from empirical distribution (3.1) with high probability up to a leaf-label respecting isomorphism, and more details can be found in section 3.3.

$$\left(\underline{p}^N\right)_i = \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}_{\{\mathcal{T}_{g_j} \cong S_i\}},\ i \in [c_n]. \tag{3.1}$$

where $S_i \in \mathcal{G}$ and $c_n := |\mathcal{G}| = (2n-5)!!$ is the number of possible unrooted binary species trees with n labeled leaves (see *e.g* [SS+03]).

Theorem 3.2.2 (Algorithm result)**.** Under the stochastic model of LGT, it's possible to reconstruct the unrooted topology of the species phylogeny w.h.p from N independent gene trees with no root, for any fixed transfer rate $\lambda \in (0, \infty)$.

To prove Theorem 3.2.2, we need to address the errors come from empirical distribution, the effect of time discretization, the guarantee of merging clusters and accumulation errors from recursive steps.

Next, we show the majority vote fails in the presence of LGT events with constant

Input: N gene trees $\{\mathcal{T}_{g_i}\}_{i=1}^N$ independently generated under the LGT process.

Procedure:

1. **Initialize** $\mathcal{N} = \mathcal{G}$, $k = 1$, $l_N = \frac{\log N}{\sqrt{N}}$, $\mathcal{L} = \{\{u\} : u \in L\}$ and empirical distribution $\underline{p}_k^N$ defined in (3.1).
2. For each state $S_i \in \mathcal{G}$, E_i denotes the leaf-edges on the tree S_i.
3. **while** $|\mathcal{L}| > 2$ **do**

   i. For each state S_i, $i \in \mathcal{N}$, define

$$Q_{ij} = \begin{cases} |\{(e_r, e_d) \in E_i^2 : S_j \cong S_i\,[\sigma(e_r, e_d)]\}| & \text{if } j \neq i \\ -\sum_{l \neq i} Q_{il} & \text{if } j = i \\ 0 & \text{else} \end{cases}$$

   where LGT events $\sigma(e_r, e_d)$ is defined below Definition 6.

   iii. Compute

$$s = \min\left\{m \in \mathbb{Z} : \exists\, i \in \mathcal{N}\ s.t. \left|\left(\exp(-mQ^\top)\underline{p}_k^N\right)_i\right| < l_N\right\}.$$

   iv. Let

$$\underline{p}_{k+1}^N = \exp\left(-sQ^\top\right)\underline{p}_k^N, \quad \mathcal{N} = \left\{i : \left(\underline{p}_{k+1}^N\right)_i > l_N\right\}.$$

   v. For every two distinct clusters $(A, B) \in \mathcal{L}^2$ s.t $S_i|A \cup B$ are leafsomorphic for all $i \in \mathcal{N}$,

   - On each state S_i, update E_i by adding

$$\left\{e^i = \left(\rho^i, y\right) : y \notin S_i\middle|A \cup B\right\},$$

     and removing edges

$$\left\{e = \left(\rho^i, x\right) : x \in S_i\middle|A \cup B\right\},$$

     where ρ^i be the root of the subtree $S_i\big|A \cup B$.
   - Update $\mathcal{L}$ by removing A, B and adding $A \cup B$.

   vi. Update $l_N = l_N\sqrt{\log N}$, $\left(\underline{p}_{k+1}^N\right)_i = 0$, $i \notin \mathcal{N}$ and $k = k + 1$.

Output: The phylogeny $\mathcal{T}_s = \mathcal{N}$.

Algorithm 4: Phylogeny reconstruction

transfer rate.

Theorem 3.2.3 (Stationary distribution with long leaf-edges)**.** For species tree T_s with leaves $L = [n]$, there exists a unique stationary distribution $\underline{\pi}$ of gene tree topology generated under the process of Definition 6 with fixed transfer rate $\lambda \in (0, \infty)$. Let $\mathring{V}_i$ denotes the set of interior vertices of S_i, and for each $v \in \mathring{V}_i$, let n_v denotes the number of elements of $\mathring{V}_i$ that are descendants of v. Then

$$\underline{\pi}(i) = \frac{2^{n-1}}{n!} \prod_{v \in \mathring{V}_i} n_v^{-1}, \qquad i \in [c_n].$$

Reconstructing $\mathcal{T}_s$ by providing pairwise distances between leaves on T_g

In this section, we state another Algorithm result to reconstruct the phylogeny if providing pairwise distances between leaves on the gene trees. To formalize this further, we introduce the following definitions.

Definition 7 (Leafsomorphic trees)**.** Given two leaf-labeled rooted (or unrooted), directed trees $T = (V, E)$ and $T' = (V', E')$ we call them *leafsomorphic*[1] if there exists a leaf labeled respecting isomorphism between the trees $\tilde{T}$ and $\tilde{T}'$ obtained from T and T' respectively, after replacing all maximal directed paths $\langle u, u_1, \ldots, u_k, v\rangle$ whose internal vertices have in- and out-degree 1 by a single directed edge $\langle u, v\rangle$.

Given N independent gene tree $T_{g_1}, \cdots, T_{g_N}$ generated under the process of Definition 6, we aim to reconstruct the topology $\mathcal{T}_s$ of the phylogeny up to a leaf-label respecting isomorphism, for all constant LGT rate.

To derive asymptotic results, we need the following model in [DR13b], which is

[1]This definition differs from the standard notion of isomorphism between leaf-labeled trees (see e.g. [SS+03]) in that we ignore degree-2 vertices.

related to a common assumption in the mathematical phylogenetics literature.

Definition 8 (Bounded-rates model)**.** Let $0 \leq \rho_\lambda \leq 1$, $0 < \rho_\tau \leq 1$ and $0 < \bar{\tau}, \bar{\lambda} < +\infty$ be constants. Under the bounded-rates model, we consider the set of phylogenies $T_s = (V_s, E_s; r, \tau, \mu)$ with n extant leaves such that the following conditions are satisfied, $\forall e \in E_s$: $\underline{\lambda} \equiv \rho_\lambda \bar{\lambda} \leq \lambda(e) \leq \bar{\lambda}$; $\underline{\tau} \equiv \rho_\tau \bar{\tau} \leq \tau(e) \leq \bar{\tau}$. Recall that μ is held constant on all edges.

Finally, our asymptotic result is the following:

Theorem 3.2.4 (Algorithm result)**.** Fix constants $0 \leq \rho_\lambda \leq 1$, $0 < \rho_\tau \leq 1$ and $0 < \bar{\tau}, \bar{\lambda} < \infty$. Under the bounded-rates model, for any $\epsilon > 0$, there exists large enough $N = \Omega(e^n)$, and it is possible to reconstruct the topology of the species phylogeny with probability at least $1 - \epsilon$ from the N independent gene trees generated under the process of Definition 6.

3.3 Identifying the species phylogeny from the distribution of the gene trees

In this section, we provide the proof of our main result Theorem 3.2.1 and its corollary. We view the LGT process in Definition 6 as a non-homogeneous continuous time Markov chain on graph space and reconstruct the species phylogeny backward in time from leaves L recursively.

Let $T_s = (V_s, E_s; r, \tau, \mu)$ be an unknown fixed species phylogeny with n labeled leaves $L = [n]$. Before introducing the Markov Chain, we need some notations.

First, we describe *levels* on species phylogeny T_s. We order the vertices V_s by the increasing values of time and split them into groups. There are $2n - 1$ vertices

on T_s and we denote the time of the i-th vertex v_i as t_i^0, $i = \{0, 1, \cdots, 2n-2\}$. Let $K = \left|\{t_i^0\}_{i=0}^{2n-2}\right| - 1$ and $K \leq n-1$ by the ultrametric property of τ. Let $\{t_i\}_{i=0}^{K}$ be the longest strictly increasing subsequence of $\{t_i^0\}_{i=0}^{2n-2}$ and we say $\{t_i\}_{i=0}^{K}$ are *speciation times* (see Figure 3.2 as an example). We define

$$\Gamma_i = \{x \in V_s : t(x) = t_i\}, \qquad i = 0, 1, \cdots, K.$$

Then $\{\Gamma_i\}_{i=0}^{K}$ form *levels* on the phylogeny T_s. In particular, $\Gamma_0 = \{r\}$ and $\Gamma_K = L$. Let

$$\gamma_i = |\Gamma_i|, \qquad i = 0, 1, \cdots, K. \tag{3.2}$$

$1 \leq \gamma_i \leq n$ by construction. On a binary tree T_s, for each location $x \in \mathcal{X}_s$ with

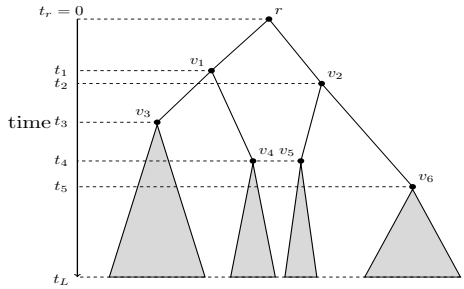

Figure 3.2: The levels and speciation times on the phylogeny.

$t(x) \in (t_{i-1}, t_i]$, the number of contemporaneous locations to x is

$$d_i := \sum_{j=0}^{i-1} \gamma_j \in \{0, 1, \cdots, n-1\}, \quad \text{for all } i \in [K]. \tag{3.3}$$

We denote $d_t = d_i$.

Next, we define the *truncation* of a phylogeny, following a similar definition in [DR16]. We truncate the species phylogeny at time ν, thus producing a forest. We will need the following notation. Given a rooted tree T and a subset of its leaves L', denote by $T|L'$ the restriction of T to leaf-set L', *i.e* the smallest connected subgraph of T that

contains $L' \cup \{\text{MRCA}(L')\}$.

Definition 9 (Truncation)**.** Given a phylogeny $T_s = (V_s, E_s; r, \tau, \mu)$ with labeled leaf-set $L = [n]$, the truncation of T_s at time ν is a leaf-labeled forest $\mathcal{T}_s^{\nu} = (V_s', E_s')$ with leaf-set L satisfying the properties:

- Disjoint forest. For some $m \leq n$, $\mathcal{T}_s^{\nu}$ comprises m rooted trees, with disjoint leaf-sets $L_1, L_2, \cdots, L_m$ which, further, correspond to clusters in the species phylogeny, that is, for all $1 \leq i \leq j \leq m$:

$$\text{MRCA}_{T_s}(L_i \cup L_j) \neq \text{MRCA}_{T_s}(L_i), \text{MRCA}_{T_s}(L_j).$$

- Truncation. Every pair of leaves $u, v \in L$, such that $\text{MRCA}_{T_s}(u, v) = x$ with $t(x) \geq \nu$ belongs to the same L_j, and every pair of leaves $u, v \in L$ such that $\text{MRCA}_{T_s}(u, v) = x$ with $t(x) < \nu$ belongs to different L_j's.

- Faithfulness. For all $j \in [m]$, the leaf-labeled tree $\mathcal{T}_s^{\nu}|L_j$ is isomorphic to the leaf-labeled tree $T_s|L_j$ under a leaf-labeled respecting isomorphism.

We formalize the Markov chain as below. Let $S_1, S_2, \cdots, S_{c_n}$ be an ordering of the c_n unrooted binary trees with n labeled leaves L in $\mathcal{G}$. We can view the LGT process in Definition 6 as M_t^n, $t \in [0, t_K]$, which is a non-homogeneous continuous-time Markov chain on states $\{S_i\}_{i=1}^{c_n}$, with initiate state $M_0^n = \{\mathcal{T}_s\}$ and transition rate matrix Q_t. Observe that Q_t depends on the topology of $\{S_i\}_{i=1}^{c_n}$ and the LGT transfer events at time t. More precisely, Q_t changes only when we cross a speciation time since we perform generic transfers chronologically from the root and each LGT transfer is an SPR move between edges. Because the receipt location defined in Definition 6 is sampled according to the Poisson distribution with same rate across all locations at time t, and the donor location is chosen uniformly form all contemporaneous, Q_t remains unchanged for all

$t \in (t_{k-1}, t_k]$, $k \in [K]$, *i.e* $Q_t = C \cdot Q_k$ where C is some positive constant and Q_k is defined explicitly in Claim 1.

We denote the distribution of M_t^n as $\underline{p}(t)$, $\underline{p}(t)$ is a c_n dimensional probability vector with exactly c_{d_t+1} nonzero elements, where $d_t + 1$ represents the number of branches at time t in (3.3). $\underline{p}(0) = \underline{p}(t_1) = \delta(\{\mathcal{T}_s\})$. Let $\mathcal{L}^k = \{L_j\}_{j=1}^m$ be the disjoint leaf-sets of $\mathcal{T}_s^k$, where $\mathcal{T}_s^k$ is the truncation of T_s at time t_k in Definition 9 (see Figure 3.3 as an example). Let $\mathcal{N}_k = \text{supp}\left(M_{t_k}^n\right)$ be the support of the Markov chain M_t^n at time $t = t_k$, *i.e* $\mathcal{N}_k = \left\{S_i : (\underline{p}(t_k))_i > 0\right\}$. Each state S_i, $i \in \mathcal{N}_k$ contains a *recovered subgraph* $\mathcal{T}_i^k$ that is isomorphic to the leaf-labeled forest $\mathcal{T}_s^k$ under a leaf-labeled respecting isomorphism. More precisely, $\mathcal{T}_i^k$ comprises d_k+1 rooted trees with disjoint leaf-sets $\mathcal{L}_i^k$, where $\mathcal{L}_i^k = \mathcal{L}^k$ for all $i \in \mathcal{N}_k$, and $\mathcal{T}_i^k = \left\{S_i | L_j : L_j \in \mathcal{L}^k\right\}$. We define *active edges* $E_{i,k} \in S_i$ be the collection of edges on S_i which disconnect $\mathcal{T}_i^k$ to the rest part of S_i when we remove these edges.

$$E_{i,k} = \left\{e = (x, y) : x \in \mathcal{T}_i^k,\, y \notin \mathcal{T}_i^k\right\}. \tag{3.4}$$

The proof of reconstructing $\mathcal{T}_s$ contains two main recursive steps. Fix $k \in [K]$, given the distribution of $M_{t_k}^n$ and the truncation $\mathcal{T}_s^k$, we reconstruct the distribution of $M_{t_{k-1}}^n$ and the truncation $\mathcal{T}_s^{k-1}$ as following:

1. **Recovering the distribution** $\underline{p}(t_{k-1})$. Given the distribution of $M_{t_k}^n$ and the truncation $\mathcal{T}_s^k$ at time t_k, we can identify recovered sub-graph $\mathcal{T}_i^k$ and active edges $E_{i,k}$ under a leaf-labeled respecting isomorphism on each state $S_i \in \mathcal{N}_k$. Next, for $t \in (t_{k-1}, t_k]$, we can construct the transition rate matrix Q_t by considering all possible SPR between the edges in $E_{i,k}$, and $\underline{p}(t_{k-1})$ follows from the Markov property since the transition matrix is invertible, and t_{k-1} is the first time the support of distribution shrink, *i.e* $\min_{s>0} |\text{supp}(p(t_k - s))| < |\mathcal{N}_k|$.

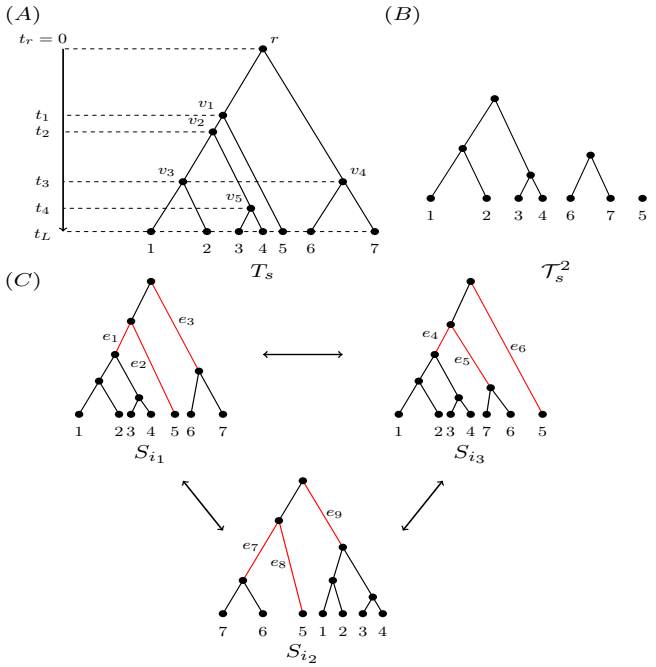

Figure 3.3: (A) A phylogeny T_s with labelled leaves $L = [7]$. (B) The truncation of T_s at time t_2. (C) The three elements in $\mathcal{N}_2$ with active edges in red.

2. **Merging the clusters**. On each state $S_i \in \text{supp}(p(t_{k-1}))$, S_i contains a subgraph $\mathcal{T}_i^{k-1}$ that is isomorphic to the leaf-labeled forest $\mathcal{T}_s^{k-1}$ under a leaf-labeled respecting isomorphism, so we can reconstruct $\mathcal{T}_s^{k-1}$ by finding the largest common subgraph of all states in the support of $M_{t_{k-1}}^n$ under a leaf-labeled respecting isomorphism.

Claim 1 shows the recovering the distribution step and Claim 2 shows the merging the clusters step.

Claim 1. Given $\underline{p}(t_k)$ and $\mathcal{L}^k$, we can uniquely solve for $\underline{p}(t_{k-1})$.

Proof of Claim: We first recall the support of distribution $\mathcal{N}_k = \{S_i : (\underline{p}(t_k))_i > 0\}$, on each state $S_i \in \mathcal{N}_k$, we can construct $\mathcal{T}_i^k = \{S_i|L : L \in \mathcal{L}^k\}$ and active edges $E_{i,k}$ in (3.4).

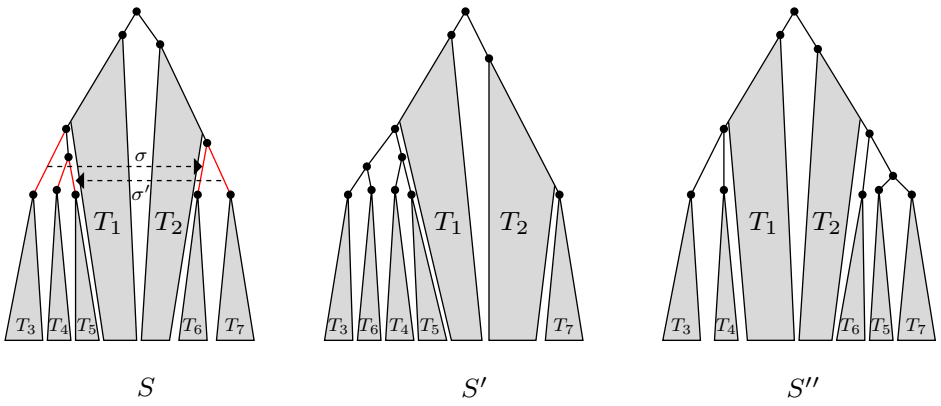

Figure 3.4: An example of constructing the generator matrix Q.

Then we define the generator matrix λQ_k of continuous-time Markov chain M_t^n by considering the LGT events between the pairs of edges in $E_{i,k}$, $t \in (t_{k-1}, t_k]$. Since Q_t is determined by LGT events between the edges at time t on T_s, for all $i \in \mathcal{N}_k$

$$(Q_k)_{ij} = \begin{cases} \left|\left\{(e_r, e_d) \in E_{i,k}^2 : S_j \cong S_i[\sigma(e_r, e_d)]\right\}\right| & \text{if } j \neq i, \\ -\sum_{l \neq i} (Q_k)_{il} & \text{if } j = i, \\ 0 & \text{others,} \end{cases}$$

where the transfer events $\sigma(e_r, e_d)$ are defined below Definition 6. we say tree $S \cong S'[\sigma(e_r, e_d)]$ if there exists a LGT transfer $\sigma(e_r, e_d)$ and S, S' are leafsomorphic after applying $\sigma(e_r, e_d)$ on S' ($e.g$ in Figure 3.4, $S' \cong S[\sigma]$ and $S'' \cong S[\sigma']$).

In the example in Figure 3.3, we consider the LGT events between the pairs of edges in $E_{i_1,2}, E_{i_2,2}$, and $E_{i_3,2}$ respectively, and get $Q_{i_l,i_j} = 2$ for all $l, j = 1, 2, 3$ and $j \neq l$, $Q_{i_l,i_l} = -4$ for $l = 1, 2, 3$ and $Q_{ij} = 0$ for others.

Then by standard Markov chain theory [Ste94b],

$$\underline{p}(t_k)^\top = \underline{p}(t_{k-1})^\top \exp\left[\lambda Q_k (t_k - t_{k-1})\right] \tag{3.5}$$

Since for all square matrix A, $\exp(A)$ is invertible and its inverse is $\exp(-A)$, (3.5) implies

$$\underline{p}(t_{k-1})^\top = \underline{p}(t_k)^\top \exp\left[-\lambda Q_k (t_k - t_{k-1})\right]. \tag{3.6}$$

Then, we define

$$s_k := \inf\left\{s > 0 : \left|\mathrm{supp}\left(\underline{p}(t_k)^\top \exp(-sQ_k)\right)\right| < \left|\mathrm{supp}\left(\underline{p}(t_k)\right)\right|\right\}, \tag{3.7}$$

where $\mathrm{supp}\left(\underline{p}\right) = \{i : p_i > 0\}$ is the support of the vector. We want to show $s_k = \lambda(t_k - t_{k-1})$.

Recall $\left|\mathrm{supp}\left(\underline{p}(t)\right)\right| = c_{d_k+1}$, where $d_k + 1$ is the number of lineages on T_s at time t. Firstly, $\left|\mathrm{supp}\left(\underline{p}(t_{k-1})\right)\right| < \left|\mathrm{supp}\left(\underline{p}(t_k)\right)\right|$ and by (3.6), we have

$$s_k \leq \lambda(t_k - t_{k-1}).$$

For the other direction, $s > 0$, $\left|\mathrm{supp}\left(\underline{p}(t_{k-1} + s)\right)\right| \geq \left|\mathrm{supp}\left(\underline{p}(t_k)\right)\right|$ as $d_{t_{k-1}+s} \geq d_k$, which implies

$$s_k \geq \lambda(t_k - t_{k-1}).$$

Combining with (3.6), we have

$$\underline{p}(t_{k-1})^\top = \underline{p}(t_k)^\top \exp\left[-s_k Q_k\right]. \tag{3.8}$$

■

Claim 2. Given $\underline{p}(t_k)$ and $\mathcal{L}^k$, we can reconstruct $\mathcal{L}^{k-1}$.

Proof of Claim: Given $\underline{p}(t_k)$ and $\mathcal{L}^k$, we can compute $\underline{p}(t_{k-1})$ by (3.8) in Claim 1 and define $\mathcal{N}_{k-1} = \text{supp}\left(M^n_{t_{k-1}}\right)$. At divergence time t_{k-1}, there exists $\gamma_{k-1} \geq 1$ distinct pairs $(A, B) \in (\mathcal{L}^k)^2$, such that $S_i|A \cup B$ are leafsomorphic for all $S_i \in \mathcal{N}_{k-1}$, where γ_{k-1} is defined in (3.2). We reconstruct $\mathcal{L}^{k-1}$ by removing all pairs (A, B) and adding $A \cup B$ on $\mathcal{L}^k$. ■

Proof of Theorem 3.2.1. Now, we can prove Theorem 3.2.1 by applying Claim 1 and Claim 2 recursively by levels backward in time. Let $\underline{p}(t_K) := \underline{p}(t_L)$ be the distribution of the topologies $\mathcal{T}_g$ of the gene tree generated under the process in Definition 6. We summarize our proof into the following steps:

1. Initialize $k = n$, take $\mathcal{N}_k = \mathcal{G}$ be the whole graph space and $\mathcal{L} = \{\{u\} : u \in [n]\}$ be single-leaf clusters.

2. On each $S_i \in \mathcal{G}$, define active edges E_i as the collection of leaf-edges on the tree S_i.

3. Until $|\mathcal{L}| = 2$,

   (a) For each $S_i \in \mathcal{N}_k$, we consider the transfer events between edges in E_i and

construct the rate matrix

$$(Q_k)_{ij} = \begin{cases} \left|\left\{(e_r, e_d) \in E^2_{i,k} : S_j \cong S_i\left[\sigma(e_r, e_d)\right]\right\}\right| & \text{if } j \neq i, \\ -\sum_{l \neq i} (Q_k)_{il} & \text{if } j = i, \\ 0 & \text{Else ,} \end{cases}$$

where $\sigma(e_r, e_d)$ represents the LGT events that is defined below the Definition 6.

(b) Compute the scaled inter-speciation time

$$s_k = \inf\left\{s > 0 : \left|\text{supp}\left(\underline{p}(t_k)^\top \exp(-sQ_k)\right)\right| < \left|\text{supp}(\underline{p}(t_k)\right|\right\}$$

and recover the distribution at previous speciation time

$$\underline{p}(t_{k-1})^\top = \underline{p}(t_k)^\top \exp\left[-s_k Q_k\right].$$

(c) Update $\mathcal{N}_{k-1} = \left\{S_i : \left(\underline{p}(t_{k-1}\right)_i > 0\right\}$.

(d) For each pair $(A, B) \in (\mathcal{L})^2$, $A \neq B$ s.t $S_i|A \cup B$ are leafsomorphic for all $i \in \mathcal{N}_{k-1}$.

- On each state $S_i \in \mathcal{N}_{k-1}$, update the actives edges E_i by adding edge

$$\left\{e = \left(\rho^i, y\right) : y \notin S_i|A \cup B\right\}$$

and removing the edges

$$\left\{e = \left(\rho^i, x\right) : x \in S_i | A \cup B\right\},$$

where ρ^i is the root of the new subtree $S_i|A \cup B$ with degree 3.

- Update $\mathcal{L}$ by removing A, B and adding $A \cup B$.

(e) Update $k = k - 1$.

By Claim 1 and Claim 2, as $|\mathcal{L}| = 2$, we can conclude that there is a unique element in the support and it is isomorphic to T_s under a leaf-labeled isomorphism. □

Proof of Corollary 3.2.2. The proof of Corollary 3.2.2 directly follows from the proof of Theorem 3.2.1 by modifying equation (3.6) in Claim 1 as

$$\underline{p}(t_{k-1})^\top = \underline{p}(t_k)^\top \exp\left[-\left(\int_{t_k-1}^{t_k} \lambda(s)ds\right) Q_k\right].$$

Since $\int_0^{t_L} \lambda(z)dz < \infty$ and $\lambda(\cdot) > 0$, all the rest arguments in Claim 1 hold. □

3.4 Statistical Consistency

In this section, we proof the statistical consistency of Algorithm 4. We begin with some claims. We first show the nonzero entries of the distribution is bounded away from zero sufficiently in claim 3.

Claim 3. [Lower bound for nonzero entries] There exists $p_0 \in (0, 1)$ s.t

$$\min_{k\in[K], S_i\in\mathcal{N}_k} \left\{\left(\underline{p}(t_k)\right)_i\right\} \geq p_0.$$

Moreover, there exists constant $C = C(n, \lambda, t_K) > 0$ independent of N, s.t

$$\big(\underline{p}(s)\big)_i \geq \begin{cases} Cp_0 & \text{if } s \in (t_{k-1}, t_k], S_i \in \mathcal{N}_{k-1}, \\ Cg_N p_0 & \text{if } s \in (t_{k-1} + g_N/\lambda, t_k], S_i \in \mathcal{N}_k \setminus \mathcal{N}_{k-1}, \end{cases}$$

where $g_N \downarrow 0$ as $N \to \infty$, provided N is larger than a sufficiently large constant.

Proof of Claim: We pick the minimum over a finite set,

$$\min_{k\in[K],\, S_i\in\mathcal{N}_k} \big(\underline{p}(t_k)\big)_i \geq p_0 > 0. \tag{3.9}$$

Fix $k \in \{2, \cdots, K\}$, at time $s \in (t_{k-1}, t_k]$, there are d_k+1 lineages on T_s, where $d_k \leq n-1$ is defined in (3.3). For $S_i \in \mathcal{N}_{k-1}$, then

$$\begin{aligned} \big(\underline{p}(s)\big)_i &\geq \big(\underline{p}(t_{k-1})\big)_i \cdot \mathbb{P}\,(\text{ No LGT transfer event during } (t_{k-1}, s]) \\ &\geq p_0 e^{-\lambda(d_k+1)(s-t_{k-1})} \\ &\geq p_0 e^{-n\lambda(t_k-t_{k-1})}. \end{aligned}$$

By (3.5) and taking the Taylor expansion of $\exp(\cdot)$ at $s = t_{k-1} + g_N/\lambda$,

$$\underline{p}(s) = \underline{p}(t_{k-1}) + \lambda(s - t_{k-1})Q_k^\top \underline{p}(t_{k-1}) + \frac{\lambda^2 c^2}{2}(Q_k^\top)^2 \underline{p}(t_{k-1}),$$

for some $c \in [0, g_N/\lambda]$ and N sufficiently large. For $S_i \in \mathcal{N}_k \setminus \mathcal{N}_{k-1}$, recall $(\underline{p}(t_{k-1}))_i = 0$. By performing an SPR between $d_k + 1$ lineages over S_i, there are $d_k + 1$ choices of cut edge and each has d_k possible edges to which the subtree will be regrafted. If the edge to which the subtree will be regrafted is adjacent to the cut edge, the SPR returns the

same ranked tree, and there are $(d_k - d_{k-1})$ pairs. We show

$$\sum_{j \neq i} (Q_k)_{ij} = d_k(d_k + 1) - 2(d_k - d_{k-1}).$$

Thus, we have

$$\begin{aligned}(\underline{p}(t_{k-1} + g_N/\lambda))_i &= g_N \sum_{j \neq i} (Q_k)_{ji} (\underline{p}(t_{k-1}))_j + \frac{c^2}{2} \sum_{j \neq i} (Q_k^2)_{ji} (\underline{p}(t_{k-1}))_j \\ &\geq g_N p_0 [d_k(d_k + 1) - 2(d_k - d_{k-1})]/2\end{aligned}$$

where we use the fact that the second term is dominated by the first term in the inequality, provided g_N is sufficiently small.

At time $s > t_{k-1} + g_N/\lambda$, we have

$$\begin{aligned}\big(\underline{p}(s)\big)_i &\geq \big(\underline{p}(t_{k-1} + g_N/\lambda)\big)_i \cdot \mathbb{P}\,(\text{ No LGT transfer during } (t_{k-1} + g_N/\lambda, s]) \\ &\geq g_N p_0 e^{-\lambda(d_k+1)(s - t_{k-1} - g_N/\lambda)} [d_k(d_k + 1) - 2(d_k - d_{k-1})]/2 \\ &\geq g_N p_* e^{-n\lambda(t_k - t_{k-1})}\end{aligned}$$

where we apply the fact that $d_k \geq 2$ in the inequality. We are done by taking $C = \min_{k \in [K]} \{e^{-n\lambda(t_k - t_{k-1})}\}$. ■

Next, we show the concentration of sample probability vectors for all levels in claim 4.

Claim 4. [Concentration] For N gene trees generated independently under the process in Definition 6 with the fixed unknown species tree T_s and divergence times $\{t_k\}_{k=1}^K$,

there exists a sequence $\{\underline{p}_k^N\}_{k=1}^K \subset [0,1]^{c_n}$ such that

$$\left\|\underline{p}_k^N - \underline{p}(t_k)\right\|_\infty \leq \sqrt{\frac{(\log N)^{K-k+1}}{N}}$$

simultaneously for all $k \in [K]$, with probability at least $1 - N^{-1}$, provided N is larger than a sufficiently large constant.

Proof of Claim: We proof by induction. For N gene trees generated independently under the process in Definition 6, we define the empirical probability

$$\left(\underline{p}_K^N\right)_i := \frac{1}{N}\sum_{j=1}^N \mathbf{1}_{\{\mathcal{T}_{g_j} \cong S_i\}},\ i \in [c_n].$$

We first show the basic case that

$$\left\|\underline{p}_K^N - \underline{p}(t_K)\right\|_\infty \leq \sqrt{\frac{\log N}{N}}$$

with probability at least $1 - N^{-1}$. We denote this event as $\mathcal{E}^0$.

Fix $i \in [c_n]$, $\left(\underline{p}_K^N\right)_i$ is an average of N independent indicators whose expectation is $\left(\underline{p}(t_K)\right)_i$, by Hoefdding's inequality

$$\mathbb{P}\left(\left|\left(\underline{p}_K^N\right)_i - \left(\underline{p}(t_K)\right)_i\right| > \sqrt{\frac{\log N}{N}}\right) \leq 2\exp\left(-2N\left(\sqrt{\frac{\log N}{N}}\right)^2\right) = \frac{2}{N^2}.$$

We are done after taking a union bound over all $i \in [c_n]$. And let $\mathcal{N}_K^N = \mathcal{G}$.

We restrict on the event $\mathcal{E}^0$ from now on. We define $l_N^k = \sqrt{\frac{(\log N)^{K-k+1}}{N}}$, $k \in [K]$ and pick $N_0 > 0$, such that $(\log N)^{1/2} > Cp_0$ for all $N \geq N_0$ and C is the constant in Claim 3. By adapting equations (3.7) and (3.6) in Claim 1, we construct $\{\underline{p}_k^N\}_{k\in[K]}$ backward in time from $\underline{p}_K^N$. First, we truncate the inter-splitting time s_k and the

distribution $\underline{p}_k$ at l_N^{k-1} as following:

$$\begin{aligned}
s_k^N &:= \min\left\{m \in \mathbb{Z} : \exists\, i \in \mathcal{N}_k^N \; s.t. \left|\left(\exp(-mQ_k^\top)\underline{p}_k^N\right)_i\right| < l_N^{k-1}\right\} \\
\underline{q}_k^N &:= \exp[-s_k^N Q_k^\top]\underline{p}_k^N \\
(\underline{p}_{k-1}^N)_i &:= (\underline{q}_k^N)_i \mathbf{1}_{\left\{i:\left|(\underline{q}_k^N)_i\right|>l_N^{k-1}\right\}} \\
\mathcal{N}_{k-1}^N &:= \left\{S_i : (\underline{p}_{k-1}^N)_i > 0\right\}
\end{aligned}$$

where Q_k is the generator matrix defined in Claim 1.

Observe that if $\left\|\underline{p}_k^N - \underline{p}(t_k)\right\|_\infty < l_N^k$, then $\mathcal{N}_k^N = \mathcal{N}_k$ by construction, thus we can recovery the clusters $\mathcal{L}_k$ and the generator matrix Q_k from the same procedure in Claim 1 and 2.

In order to show the induction step $\left\|\underline{p}_{k-1}^N - \underline{p}(t_{k-1})\right\|_\infty < l_N^{k-1}$, we need to show

$$\lfloor N(s_k - l_N^{k-1})\rfloor \le s_k^N \le \lceil N s_k\rceil \text{ if } \left\|\underline{p}^N(t_k) - \underline{p}(t_k)\right\|_\infty \le l_N^k,$$

where $s_k = \lambda(t_k - t_{k-1})$ is the scaled inter-speciation time.

Recall that for any two matrices $X, Y \in \mathbb{R}^{n\times n}$,

$$\left\|e^{X+Y} - e^X\right\| \le \|Y\| e^{\|X\|+\|Y\|},$$

where $\|\cdot\|$ denotes an arbitrary matrix norm and it follows from that the exponential map is continuous and Lipschitz continuous on compact subsets of the space of all $n \times n$ matrices. Applying this fact on $Q_k^\top$ with $s, t \ge 0$, we have

$$\left\|\exp\left((s+t)Q_k^\top\right) - \exp\left(sQ_k^\top\right)\right\| \le t\|Q_k\| e^{(s+t)\|Q_k\|} \tag{3.10}$$

For $S_i \in \mathcal{N}_k$ and $s \geq 0$, by Cauchy–Schwarz inequality

$$\begin{aligned}\left|(\exp(-sQ_k^\top)(\underline{p}^N(t_k) - \underline{p}(t_k))_i\right| &\leq \left\|\left(\exp\left(-sQ_k^\top\right)\right)_{i\cdot}\right\|_2 \left\|\underline{p}^N(t_k) - \underline{p}(t_k)\right\|_2 \\ &\leq \sqrt{c_n}\exp\left(s\|Q_k\|_2\right)\left\|\underline{p}^N(t_k) - \underline{p}(t_k)\right\|_\infty \end{aligned} \tag{3.11}$$

For $S_i \in \mathcal{N}_k \setminus \mathcal{N}_{k-1}$, we break $\left|\left(\exp\left(-(s_k + N^{-1})Q_k^\top\right)\underline{p}^N(t_k)\right)_i\right|$ into three pieces by as following:

$$\begin{aligned}&\left|\left(\exp\left(-(s_k + N^{-1})Q_k^\top\right)\left(\underline{p}^N(t_k) - \underline{p}(t_k)\right)_i\right| + \left|\left(\exp\left(-s_kQ_k^\top\right)\underline{p}(t_k)\right)_i\right| \\ &\quad + c_n\left\|\exp\left(-(s_k + N^{-1})Q_k^\top\right) - \exp\left(-s_kQ_k^\top\right)\right\|_\infty \left\|\underline{p}(t_k)\right\|_\infty\end{aligned}$$

where the first term is controlled by (3.11), the second term equals $(\underline{p}(t_{k-1}))_i$ which is 0 since $S_i \notin \mathcal{N}_{k-1}$ and the last term is controlled by 3.10.

$$\begin{aligned}&\leq \sqrt{c_n}e^{(s_k+N^{-1})\|Q_k\|_2}\left\|\underline{p}^N(t_k) - \underline{p}(t_k)\right\|_\infty + (\underline{p}(t_{k-1}))_i + \frac{c_n}{N}\|Q_k\|_\infty e^{(s_k+N^{-1})\|Q_k\|_\infty} \\ &\leq 2\sqrt{c_n}e^{(s_k+N^{-1})\|Q_k\|_2}\left\|\underline{p}^N(t_k) - \underline{p}(t_k)\right\|_\infty \\ &< 2\sqrt{c_n}e^{(s_k+N^{-1})\|Q_k\|_2}\cdot l_N^k \\ &< l_N^{k-1}\end{aligned}$$

which implies $s_k^N/N \leq s_k + 1/N$, provided N is sufficiently large.

By Claim 3 with $g_N = l_N^{k-1}$ and $e^{-sQ_k}e^{sQ_k} = I$, for $m < N(s_k - l_N^{k-1}) = N\lambda(t_k - t_{k-1} - l_N^{k-1}/\lambda)$, $S_i \in \mathcal{N}_k$

$$\begin{aligned}\left|\left(\exp\left(-\frac{m}{N}Q_k^\top\right)\underline{p}(t_k)\right)_i\right| &= \left|\left(\exp\left(\left(\lambda(t_k - t_{k-1}) - \frac{m}{N}\right)Q_k^\top\right)\underline{p}(t_{k-1})\right)_i\right| \\ &\geq \begin{cases} Cp_0 & \text{if } S_i \in \mathcal{N}_{k-1} \\ Cp_0l_N^{k-1} & \text{if } S_i \in \mathcal{N}_k \setminus \mathcal{N}_{k-1}\end{cases} \\ &\geq Cp_0g_N^{k-1}\end{aligned}$$

Then

$$\begin{aligned}
\left|\left(\exp(-\frac{m}{N}Q_k^\top)\underline{p}_k^N\right)_i\right| &\geq \left|\left(\exp(-mQ_k^\top)\underline{p}(t_k)\right)_i\right| \\
&\quad - \left|\left(\exp(-mQ_k^\top)(\underline{p}^N(t_k)-\underline{p}(t_k)\right)_i\right| \\
&\geq Cp_0 g_N^{k-1} - \sqrt{c_n}\exp\left(s_k\|Q_k\|_2\right)\left\|\underline{p}^N(t_k)-\underline{p}(t_k)\right\|_\infty \\
&\geq Cp_0 g_N^{k-1} - \sqrt{c_n}\exp\left(s_k\|Q_k\|_2\right) l_N^k \\
&\geq Cp_0 g_N^{k-1}/2
\end{aligned}$$

which implies $s_k^N \geq \lfloor N(s_k - g_N^{k-1})\rfloor$, for N sufficiently large. And we have the following recursive relation,

$$\begin{aligned}
\left\|\underline{q}_{k-1}^N - \underline{p}(t_{k-1})\right\|_\infty &= \left\|\exp\left[-s_k^N Q_k^\top\right]\underline{p}_k^N - \exp\left[-s_k Q_k^\top\right]\underline{p}(t_k)\right\|_\infty \\
&\leq \left\|\left(\exp\left[-s_k^N Q_k^\top\right] - \exp\left[-s_k Q_k^\top\right]\right)\underline{p}_k^N\right\|_\infty \\
&\quad + \left\|\exp\left[-s_k Q_k^\top\right]\left(\underline{p}_k^N - \underline{p}(t_k)\right)\right\|_\infty \\
&\leq |s_k^N/N - s_k|\|Q_k\|\exp(s_k\|Q_k\|) \\
&\quad + \sqrt{c_n}\exp\left(\lambda(t_k - t_{k-1})\|Q_k\|_2\right)\left\|\underline{p}^N(t_k)-\underline{p}(t_k)\right\|_\infty \\
&\leq 2g_N^{k-1}\|Q_k\|\exp(s_k\|Q_k\|) \\
&\leq l_N^{k-1}
\end{aligned}$$

■

With N sampled gene trees. Now, we can prove Theorem 3.2.1 by applying Claim 1 and Claim 2 recursively by levels backward in time. Following the Algorithm 4, it stops as $|\mathcal{L}| = 2$, then we can conclude that there is a unique element in $\mathcal{N}$ and it is isomorphic to T_s under a leaf-labelled isomorphism. □

3.4.1 Proof of Theorem 3.2.3

If we run the process backward in time and take the length of leaf edges arbitrary large, it can be viewed as a coalescence process. We recall the known result for the Yule model (see *e.g* [SM01]) in Claim 5. We will show it indeed is the stationary distribution with arbitrary long leaf edges for LGT process.

Claim 5 (Stationary distribution for Yule model)**.** Let T be a rooted phylogenetic tree with n labeled leaves and let $P_Y(T)$ denotes the probability of generating T under the Yule model.

$$P_Y(T) = \frac{2^{n-1}}{n!} \prod_{v \in \mathring{V}_i} n_v^{-1},$$

where $\mathring{V}$ is the set of interior vertices of T. For each $v \in \mathring{V}$, let n_v denotes the number of elements of $\mathring{V}$ that are descendants of v(including itself).

Proof of Theorem 3.2.3. We first consider the LGT process on ranked binary tree with n labeled leaves, *i.e* the binary trees with ordering interior vertices. This process can also be viewed as a Markov Chain $\tilde{M}_t^n$ on the extended graph space $\tilde{\mathcal{G}} = \{\tilde{S}_i\}$ with generator matrix $\lambda\tilde{Q}_k$, where $\tilde{\mathcal{G}}$ containing all ranked binary trees and $|\tilde{\mathcal{G}}| = n!(n-1)!/2^{(n-1)}$ (see *e.g* [SM01]). On the leaf level $t \in (t_{K-1}, t_K]$, $\mathrm{supp}(\tilde{M}_t^n) = \tilde{\mathcal{G}}$ and on each state $\tilde{S}_i$, all leaf edges are the active edges $\{E_i\}$ defined in (3.4). The generator matrix is defined as

$$(\tilde{Q}_K)_{ij} = \begin{cases} \left|\left\{(e_r, e_d) \in E_i^2 : \tilde{S}_j \cong \tilde{S}_i[\sigma(e_r, e_d)]\right\}\right| & \text{if } j \neq i, \\ -\sum_{l \neq i} (\tilde{Q}_K)_{il} & \text{if } j = i, \\ 0 & \text{others,} \end{cases}$$

where the transfer events $\sigma(e_r, e_d)$ are defined below Definition 6.

In Claim 6, we show the LGT process on ranked binary tree tends to a uniform distribution with long enough leaf edges. Then we are done by considering the projected process over unranked trees and recalling the number of ranked functions of S_i is $(n-1)!\prod_{v\in\mathring{V}_i} n_v^{-1}$ (see *e.g* [SM01]), where $\mathring{V}$ is the set of interior vertices of S_i and n_v denotes the number of elements of $\mathring{V}$ that are descendants of v(including itself), for each $v \in \mathring{V}$. □

Claim 6 (Stationary distribution on ranked binary trees)**.** The process $\tilde{M}_t$ is uniformly distributed on $\tilde{\mathcal{G}}$ with arbitrary long leaf edges, that is

$$\frac{2^{n-1}}{n!(n-1)!}\mathbf{1}^\top \tilde{Q}_K = 0$$

where $\mathbf{1}$ is the $n!(n-1)!/2^{n-1}$ dimensional one vector.

Proof of Claim: On the leaf level $t \in (t_{K-1}, t_K]$, $\mathrm{supp}(\tilde{M}_t^n) = \tilde{\mathcal{G}}$ and on each state $\tilde{S}_i$, all leaf edges are active $\{E_i\}$ defined in (3.4). It is equivalent to show that

$$-\tilde{Q}_K(i,i) = \sum_{j\neq i} \tilde{Q}_K(j,i), \text{ for all } i.$$

By performing an SPR between n leaf edges over S_i, there are n choices of cut edge and each has $n-1$ possible edges to which the subtree will be regrafted. If the edge to which the subtree will be regrafted is adjacent to the cut edge, the SPR returns the same ranked tree. We show

$$-\tilde{Q}_K(i,i) = \sum_{j\neq i} \tilde{Q}_K(i,j) = n(n-1) - 2.$$

We observe that for all pairs (i, j),

$$\tilde{Q}_K(j,i) := \left|\left\{(e_r, e_d) \in E_j^2 : \tilde{S}_i \cong \tilde{S}_j[\sigma(e_r, e_d)]\right\}\right| \leq 2.$$

Since S_j is leafsomorphic to the ranked binary tree S_i after applying a SPR between a pair of leaf edges, there are at most 2 possible choices of cut edges over S_j.

In the case that the regrafted edge is separated by exactly one edge from the cut edge, *i.e* S_j are different with S_i only on one triplet up to a leaves labeled isomorphism, $\tilde{Q}_K(j,i) = 2$.

There are 3 different topologies for a rooted triplet, thus

$$|\{j \neq i : \tilde{Q}_K(j,i) = 2\}| = 2.$$

For other cases that $\{j \neq i : \tilde{Q}_K(j,i) = 1\}$, there are 2 choices of cut edges and each can come from $\sum_{k=2}^{n-2} k$ possible segments between neighboring $\{x \in \mathcal{X}_s : t(x) \in \{t_i\}_{i=0}^K\}$, where $\sum_{k=2}^{n-2} k$ comes from the possible ordering of internal vertices and $\{t_i\}_{i=0}^K$ are speciation times. If the edge to which the subtree will be regrafted is adjacent to the cut edge, the SPR returns the same ranked tree. We can conclude that

$$\sum_{j \neq i} \tilde{Q}_K(j,i) = 2\left(\sum_{k=2}^{n-2} k - 2\right) + 2 \cdot 2 = n(n-1) - 2.$$

■

3.5 Reconstructing the species phylogeny with pairwise distances on gene trees

In this section, we provide the proof of Theorem 3.2.4. We assume that N independent gene trees $T_{g_1}, \cdots, T_{g_N}$, corresponding to homologous genes $g_1, \cdots, g_N$, were generated according to the process of Definition 6. The main idea is identifying the pairwise distances $\{d_s(u,v)\}_{u,v\in L}$ on the phylogeny T_s by the property of continuous-time Poisson process, and then reconstructing $\mathcal{T}_s$ by single-linkage clustering.

Let $D := \{d_s(u,v) : u, v \in L\}$ be the collection of all pairwise distances between leaves on the phylogeny. Recall the multi-nodes metric

$$d(a_1, a_2, \cdots, a_m) := \big(d(a_1,a_2), \cdots, d(a_1,a_m), d(a_2,a_3), \cdots, d(a_{n-1},a_m)\big),$$

We define

$$d(L) := d(1, 2, \cdots, n) \in \mathbb{R}^{n(n-1)/2}.$$

On each gene tree T_g generated under the process of Definition 6, for each pair $u, v \in L$, we show the pairwise distances of leaves $d_g(u,v) = d_s(u,v)$ with positive probabilities in Claim 7; we show the pairwise distances of leaves $d_g(u,v) = k \notin D$ with probability 0 in Claim 8; then we show $d_s(L)$ can be approximated by the mode of $\{d_{g_i}(L)\}_{i=1}^N$ in Claim 9.

Claim 7 (Positive probability distances)**.** For a gene tree T_g generated under the process in Definition 6,

$$p_{uv} := \mathbb{P}\big[d_g(u,v) = d_s(u,v)\big] \in (0,1)$$

for all pairs $u, v \in L$. Moreover,

$$p_L := \mathbb{P}\big[d_g(L) = d_s(L)\big] = \Theta(e^{-n}).$$

Proof of Claim: Fix $u, v \in L$, under the process in Definition 6, if there exists LGT transfer event $\sigma = (x, y)$ with recipient location $x \in \mathcal{X}_s$ and donor location $y \in \mathcal{X}_s$, such that $\{x, y\} \in p(u, v)$ on the path between u and v, then

$$d_g(u, v) \neq d_s(u, v).$$

Therefore, we get

$$\begin{aligned}
\mathbb{P}\left[d_g(u,v) = d_s(u,v)\right] &= \mathbb{P}\left[\text{ No LGT recipient locations on } p(u,v)\right] \\
&= \prod_{e \in p(u,v)} \mathbb{P}\left[\text{ No LGT recipient locations on edge } e\right] \\
&= \prod_{e=(x,y)\in p(u,v)} \exp\left[-\int_{\min\{t(x),t(y)\}}^{\max\{t(x),t(y)\}} \lambda(z)dz\right] \\
&= \exp\left[\sum_{e=(x,y)\in p(u,v)} -\int_{\min\{t(x),t(y)\}}^{\max\{t(x),t(y)\}} \lambda(z)dz\right] \\
&\geq \exp\left[-\frac{\bar{\lambda}}{\mu} d_s(u,v)\right] \in (0,1)
\end{aligned}$$

Considering all pairs of leaves on T_s,

$$\mathbb{P}\left[d_g(L) = d_s(L)\right] = \mathbb{P}\left[\text{ No LGT transfer events }\right] = e^{-\sum_{e\in E_s}\Lambda(e)} = \Theta(e^{-n})$$

where we use $|E_s| = \Theta(n)$. ■

Claim 8 (Zero probability distances)**.** For a gene tree T_g generated under the process

in Definition 6,

$$\mathbb{P}\big[d_g(u,v) = k\big] = 0 \text{ for all } k \notin D$$

for all pairs $u, v \in L$.

Proof of Claim: Under the process in Definition 6, we pick recipient locations according to a continuous-time Poisson process with $\Lambda(e)$ on each branch $e \in T_s$ starting from the root. Fix $k \notin D$ and $u, v \in L$, $d_g(u,v) = k$, there exists LGT transfer with recipient location $x \in \mathcal{X}_s$, such that $d_s(u,x) = k/2$ or $d_s(v,x) = k/2$. Thus

$$\mathbb{P}\big[d_g(u,v) = k\big] \leq \mathbb{P}\big[d_s(u,x) = k/2\big] + \mathbb{P}\big[d_s(v,x) = k/2\big] = 0.$$

■

Claim 9 (Majority vote)**.** For any $\varepsilon > 0$, there exists $N = \Omega(e^n)$ large enough, such that with probability at least $1 - \varepsilon$,

$$\text{Mode}_{i=1,\cdots,N}\, d_{g_i}(L) = d_s(L).$$

We denote this event as $\mathcal{E}$.

Proof of Claim: On each gene tree T_{g_i} that is independently generated under the process in Definition 6, the event $\{d_{g_i}(L) = d_s(L)\}$ happens independently with probability p_L defined in Claim 7. By Hoeffding's inequality,

$$\begin{aligned}\mathbb{P}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}_{\{d_{g_i}(L)=d_s(L)\}} - p_L < -\sqrt{\frac{\log N}{N}}\right] &\leq \exp\left(-2N\left[-\sqrt{\frac{\log N}{N}}\right]^2\right)\\ &= N^{-2}\end{aligned}$$

For the noise distances, by Claim 8,

$$\begin{aligned}
\mathbb{P}\left[\exists\, \underline{d} \neq d_s(L), \sum_{i=1}^{N} \mathbf{1}_{\left\{d_{g_i}(L)=\underline{d}\right\}} > 2\right] &\leq \mathbb{P}\left[\exists\, i \neq j,\, d_{g_i}(L) = d_{g_j}(L) \neq d_s(L)\right] \\
&\leq N^2 \mathbb{P}\left[d_{g_1}(L) = d_{g_2}(L) \neq d_s(L)\right] \\
&= N^2 \mathbb{P}\left[\exists\, u, v \in L,\, d_{g_1}(u,v) = d_{g_2}(u,v) \notin D\right] \\
&\leq N^2 n^2 \mathbb{P}\left[d_{g_1}(1,2) = d_{g_2}(1,2) \notin D\right] \\
&= N^2 n^2 \mathbb{P}\left[\mathbb{P}\left[d_{g_1}(1,2) = k \,|\, d_{g_2}(1,2) = k \notin D\right]\right] \\
&= 0
\end{aligned}$$

Thus we can conclude $\sum_{i=1}^{N} \mathbf{1}_{\left\{d_{g_i}(L)=\underline{d}\right\}} \leq 2$ simultaneously for all $\underline{d} \neq d_s(L)$.

For any $\varepsilon > 0$, taking $N = \Omega(e^n)$ large enough, with probability at least $1 - \varepsilon$,

$$\sum_{i=1}^{N} \mathbf{1}_{\left\{d_{g_i}(L)=d_s(L)\right\}} - \sum_{i=1}^{N} \mathbf{1}_{\{d_{g_i}(L)=\underline{d}\}} > Np_L - \sqrt{N \log N} - 2 > 0$$

simultaneously for all $\underline{d} \in \mathbb{R}^{n(n-1)/2}$. We finish the proof and clearly the Mode here is uniquely defined. ■

Proof of Theorem 3.2.4. Let

$$\hat{d}(u,v) = \mathrm{Mode}_{i=1,\cdots,N}\left\{d_{g_i}(L)\right\}|_{(u,v)}, \tag{3.12}$$

$$\hat{d}(A,B) = \min_{a \in A, b \in B} \hat{d}(a,b), \qquad A, B \subset L, A \cap B = \emptyset. \tag{3.13}$$

$\{\hat{d}\}$ is well-defined by Claim 9 and condition on event $\mathcal{E}$, we reconstruct the topology of species phylogeny $\mathcal{T}_s$ use the single-linkage clustering as following:

1. Let $\mathcal{L} = \{\{u\} : u = [n]\}$, set $\hat{d}$ as in (3.12) and let $T_{\{u\}}$ be the tree composed of

only u, for all $u \in [n]$.

2. Until $\mathcal{L} = \{[n]\}$:

   (a) Let A,B be two clusters in $\mathcal{L}$ achieving the minimum $\hat{d}-$distance in (3.13).

   (b) Base on T_A with root r_A and T_B with root r_B, we construct the tree $T_{A\cup B}$ by introducing a new root node $r_{A\cup B}$ and two edges $d(r_{A\cup B}, r_A) = d(r_{A\cup B}, r_B) = \frac{1}{2}\hat{d}(A, B)$ to $\{T_A \cup T_B\}$.

   (c) Update $\mathcal{L}$ by removing A, B and adding $A \cup B$.

□

Chapter 4

Inference the phylogeny from multi-loci using fmulti-sites

4.1 Introduction

Estimating the common evolutionary history of n species using sequential data from multiple genes or loci is a fundamental question in modern phylogenomics. The increasing availability of genomic-scale datasets of many aligned gene sequences has made clear that the phylogenetic trees inferred for individual genes often differ from one another [Pol+06; GD08; Cra+09; Nak13]. Although gene tree discordance might be due to errors in gene tree inference, there are also important biological processes that can cause it, which should be taken into account through appropriate modeling.

One well-recognized source of gene tree conflict is incomplete lineage sorting and it was studied in the multispecies coalescent (MSC) model [PN88][RY03], combined with standard models of sequence evolution by base substitutions. Under the multispecies coalescent model, the identifiability of phylogeny was studied in [DNR14][CK15][ALR19]. We provide a simpler method than [CK15] and [ALR19] by avoiding the detailed computation of site pattern probabilities. We improve the idea in [DNR14] by using the symmetric property of MSC between lineages, and we also allow the different rates across the loci which is not held in Metal[DNR14].

This chapter aims to show that the phylogenetic tree inference from multi-loci will be simplified with multiple independent sites per gene. Under the Jukes-Cantor substitution model, using two independent sites per locus, one can construct the estimator from sequential data to detect the correct quartets. Then the whole phylogeny can be constructed from any quartets-based algorithm like the Dyadic Closure Method (DCM) in[Erd+99]. In general time reversible model(GTR), we propose a more general estimator using the determinant of site-pattern frequency matrix. The site-pattern frequency matrix can be estimated from k independent sites per locus, where k represents the number of state under substitution model.

This chapter is organized as follows: In section 4.2.1, we start with an example of phylogeny inference under identical tree mixture model. Section 4.2.2 provides the idenfiability result under MSC, combined with standard sequence evolution model Jukes-Cantor. In section 4.2.3, we show the idenfiability result under general time-reversible (GTR) substitution model, using the determinants of transition matrices. In section 4.3, we extend the identifiability result to level-1 network by using the same estimator.

4.2 Phylogenetic inference under mixture of Gene trees

Definition 10 (Species tree (Phylogeny))**.** A species tree $T = (V, E; r, \tau)$ is a directed binary tree rooted at r and n labelled leaves $L = [n] = \{1, 2, \cdots, n\}$. Each edge are assigned inter-speciation times $\tau : E \to (0, +\infty)$. $\mathcal{T} = (V, E)$ be the topology of T after suppressing the root.

We say the species tree T is *identifiable* if we can infer its topology $\mathcal{T}$ from the distribution of gene trees (or its topologies) under a leaf-label respecting isomorphism.

In this chapter, we aim to show the identifiability of species tree T, moreover, we can infer the phylogeny from multiple loci with multiple independent sites, that is, to learn the structure $\mathcal{T} = (V, E)$ given the data $\{\vec{\mathcal{X}}_{ij}\}_{i\in[N],j\in[2m]}$ which is an $n \times N \times 2m$ array composed of k states in substitution model with state space S (*e.g.* $S = \{0, 1\}$ under two-state symmetric model or $S = \{A, C, G, T\}$ under Jukes-Cantor model), where $\{\mathcal{X}_{ij}\}_{j\in[m]}$ is the data generated from the random gene tree $\mathcal{G}^{(i)}$.

4.2.1 Phylogeny inference under identical tree mixture model

Masten & Steel [MS07] shows the non-identifiability under the two-state symmetric model from multi-genes with one site per locus, that is, we are unable to identify the structure of T if mixing the random gene tree $\{\mathcal{G}^{(i)}\}_{i\in[N]}$ generated from the two-state symmetric model with different scaled branch lengths. However, this problem can be solved by introducing two sites per locus, that is, T is identifiable if there are at least two independent sites per locus.

Identical tree mixtures of the two-state symmetric substitution model

We first introduce the *two-state symmetric model* or the *Cavender-Farris-Neyman (CFN) model* given the fixed species tree $T = (V, E, \tau)$. Each vertices $v \in V$ is assigned with a binary random variables 0 or 1 and each edge is $e \in E$ is assigned with a weight $p_e \in [0, \frac{1}{2}]$. The descendent vertex is assigned to different states with its parent vertex with probability p_e if these two vertices are connected by a directed edge $e \in E$.

Next, we introduce the *r-component identical tree mixtures* of the two-state symmetric substitution model on n-leaf binary species trees $\mathbf{T} = (T_1, \cdots, T_r)$ with mixing vector $\pi = (\pi_1, \cdots, \pi_r)$, where $\sum_{i=1}^{r} \pi_i = 1$ and T_i has the same topology with T for all i.

The sequence data with multiple independent sites per locus is generated as

follows: for each locus, we first sample a class i from the distribution π; then for each site, the bases for each $x \in L$ are sampled under the two-state symmetric substitution model independently on phylogenetic tree T_i.

In this chapter, we discuss the identifiability using two independent sites per locus. Suppose there are at least $2m$ independent sites per locus ($m \geq 1$) with an arbitrary ordering, we first split the sites evenly into two groups M_1 and M_2 such that $|M_1| = |M_2| = m(e.g$ by even or odd), then we consider the pairs $\mathbf{j} = (j_1, j_2) \in M_1 \times M_2$ where j_1, j_2 are in increasing order. For fixed gene $\mathcal{G}^{(i)}$, we define

$$\hat{\sigma}_{ij}^{xy} = \begin{cases} 1 & \text{if } \mathcal{X}_{ij}^{x} = \mathcal{X}_{ij}^{y}, \\ -1 & \text{if } \mathcal{X}_{ij}^{x} \neq \mathcal{X}_{ij}^{y}, \end{cases} \qquad j = \{1, 2, \cdots, m\},\ x, y \in L,$$

which can be viewed as an indicator of the agreement over taxa x and y at the site j of loci i. $\hat{\sigma}_{ij}^{xy} \equiv 1$, if $x = y \in L$.

We construct the estimator over two sites as follows: for every two pairs of leaves $(x, y), (z, w) \in L^2$,

$$\hat{l}(x, y; z, w) = -\frac{1}{mN} \sum_{i=1}^{N} \sum_{(j_1, j_2)} \hat{\sigma}_{ij_1}^{xy} \hat{\sigma}_{ij_2}^{zw}. \tag{4.1}$$

Observe that all gene trees generated under the two-state symmetric model are same distributed, thus same for $\{\hat{\sigma}_{i\cdot}\}$. For simplicity, we denote the common random variable as $\{\hat{\sigma}_{\cdot}\}$ after suppressing the subscript i. We define the following idealized version of $\hat{l}$,

$$l(x, y; z, w) = -\left[\mathbb{E}\left(\hat{\sigma}_1^{xy} \hat{\sigma}_2^{zw}\right)\right]. \tag{4.2}$$

We aim to show the structure of phylogeny $\mathcal{T}$ is identifiable from l. Firstly, we define the following four points test, which is a modified version of the classic four points condition (see [Bun74]) for two independent sites per locus.

Definition 11 (Split-equivalent Function)**.** We say function f on L^4 is a **Split-equivalent Function** if it satisfies the following conditions: for all $x, y, z, w \in L$(not necessary distinct),

1. (interchangeability)

$$f(x, y; z, w) = f(y, x; z, w) = f(x, y; w, z) = f(z, w; x, y);$$

2. (Nonnegative) $f(x, y; z, w) \geq \max\{f(x, x; z, w), f(x, y; z, z)\} \geq 0$ and the equality holds if and only if $x = y$ and $z = w$;

Definition 12 (Four-point Condition(4FC))**.** We say a Split-equivalent function f satisfies the **four points condition** with respect to T if for any four taxon $x, y, z, w \in L$, two of the three terms in the following list are equal and greater than the third:

$$f(w, x; y, z), \quad f(w, y; x, z), \quad f(w, z; x, y).$$

That is, suppose the leaves $w, x, y, z \in L$ are such that either $((w, x), (y, z))$ or $(((w, x), y), z)$ holds on T, then

$$f(w, x; y, z) < f(w, y; x, z) = f(w, z; x, y).$$

Next, Theorem 4.2.1 shows that $\mathcal{T}$ is identifiable from the collection of all restricted subtrees S with $|S| \leq 4$.

Theorem 4.2.1 (Quartet theorem [SS+03])**.** Let $\mathcal{T}, \mathcal{T}'$ be two phylogenies with same leaves L, then $\mathcal{T}$, $\mathcal{T}'$ are isomorphic if and only if $\mathcal{T}|S$, $\mathcal{T}'|S$ are isomorphic for all $S \subset L$ with $|S| \leq 4$.

Since all unrooted trees with a size less than 4 are isomorphic (see [SS+03]), it suffices to check quartets only. We show in Theorem 4.2.2 that we can identify all restricted subtrees with size 4 from l, which implies the phylogeny $\mathcal{T}$ is identifiable from Split-equivalent function l.

Theorem 4.2.2 (4PC)**.** l is a Split-equivalent function and it satisfies the Four points condition with respect to T.

Theorem 4.2.2 also tells us that one can ignore the fact that there are different rates across loci, then the gene tree estimated from this "concatenated molecular sequence by groups" has the same topology as T.

In light of Theorem 4.2.2, we use any Quartets-based algorithm to reconstruct $\mathcal{T}$ (like Dyadic Closure Method(DCM) in[Erd+99]).

Proof of Theorem 4.2.2. Suppose leaves $A, B, C, D \in L$ are distinct and the topology of T restricted to these leaves is $((A, B), (C, D))$ or $(((A, B), C), D)$, we want to check

$$f(A, B; C, D) < f(A, C; B, D) = f(A, D; B, C).$$

It is equivalent to show

$$\mathbb{E}\left(\hat{\sigma}_1^{AB}\hat{\sigma}_2^{CD}\right) > \mathbb{E}\left(\hat{\sigma}_1^{AC}\hat{\sigma}_2^{BD}\right) = \mathbb{E}\left(\hat{\sigma}_1^{AD}\hat{\sigma}_2^{BC}\right). \tag{4.3}$$

Recall the following properties of two-state symmetric model (see [Ney71]),

$G|\{A,B,C,D\}$ and $T|\{A,B,C,D\}$ are isomorphic. Condition on a fixed gene tree G, we have

$$\mathbb{P}(\hat{\sigma}_j^{xy} = \pm 1|G) = \frac{1 \pm e^{-2d_G(x,y)}}{2}, \qquad \mathbb{E}(\hat{\sigma}_j^{xy}|G) = e^{-2d_G(x,y)},\ j = 1, 2,$$

where $d_G(x,y) = \sum_{e \in \text{path}(x,y;G)} \mu_e \tau_e$ represents the scaled length between taxa x and y on gene tree G.

Applying total expectations, we have

$$\begin{aligned}
\mathbb{E}(\hat{\sigma}_1^{xy}\hat{\sigma}_2^{zw}) &= \mathbb{E}\left(\mathbb{E}\left(\hat{\sigma}_1^{xy}\hat{\sigma}_2^{zw}|G\right)\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(\hat{\sigma}_1^{xy}|G\right)\mathbb{E}\left(\hat{\sigma}_2^{zw}|G\right)\right) \\
&= \mathbb{E}\left(e^{-2d_G(x,y)-2d_G(z,w)}\right) \qquad (4.4)
\end{aligned}$$

for any $(x,y),(z,w) \in L^2$, where the second equality holds by the independence of two sites.

First, we show the inequality part in equation (4.3). Applying (4.4) on pairs $(A,B),(C,D)$ and pairs $(A,C),(B,D)$, we have

$$\begin{aligned}
\mathbb{E}(\hat{\sigma}_1^{AB}\hat{\sigma}_2^{CD}) &= \mathbb{E}\left(e^{-2d_G(A,B)-2d_G(C,D)}\right) \\
&> \mathbb{E}\left(e^{-2d_G(A,C)-2d_G(B,D)}\right) \\
&= \mathbb{E}(\hat{\sigma}_1^{AC}\hat{\sigma}_2^{BD})
\end{aligned}$$

where the inequality follows from the four points condition, that is, $d_G(A,B) + d_G(C,D) < d_G(A,C) + d_G(B,D)$, since the restricted subtree on G has the structure $((A,B),(C,D))$ or $(((A,B),C),D)$.

Next, We show the equality part in equation (4.3). Applying (4.4) on pairs

$(A, D), (B, C)$ and pairs $(A, C), (B, D)$, we have

$$\begin{aligned}\mathbb{E}(\hat{\sigma}_1^{AC}\hat{\sigma}_2^{BD}) &= \mathbb{E}\left(e^{-2d_G(A,C)-2d_G(B,D)}\right) \\ &= \mathbb{E}\left(e^{-2d_G(A,D)-2d_G(B,C)}\right) \\ &= \mathbb{E}(\hat{\sigma}_1^{AD}\hat{\sigma}_2^{BC}).\end{aligned}$$

where we apply the four points condition on G, that is, $d_G(A, D) + d_G(B, C) = d_G(A, C) + d_G(B, D)$.

□

4.2.2 Phylogenetic inference under multispecies coalescent

In this section, we consider the identifiability from more complicate gene tree structures caused by multispecies coalescent which models the gene tree incongruence from population level.

Multispecies coalescent model

Multispecies Coalescent(MSC) model provides a framework for inferring species phylogeny while accounting for ancestral polymorphism and gene tree-species tree conflict, which is also referred as Incomplete Lineage Sorting(ILS).

Consider a species tree $T = (V, E; r, \tau)$ with n leaves. The branch lengths $\{\tau_e\}_{e\in E}$ are expressed in "coalescent time units". The species tree topology, $\mathcal{T}$, is an unknown parameter in a distribution $\mathcal{D}(\mathcal{T})$ [1]. We have N i.i.d. realizations of $\mathcal{D}(\mathcal{T})$. These realizations are referred to as **gene trees**, and are created by the following process:

The coalescence of any two branches is distributed as Exp(1), independently from

[1] $\mathcal{D}$ is the probability distribution governing the MSC

all other pairs of branches. When two branches merge in the species tree, we assume the lineages of the corresponding populations also merge. The genes are assumed to be unlinked.

The realization of this model for N independent genes is

$$\prod_{j=1}^{N}\prod_{e\in E}\exp\left(-\binom{O_j^e}{2}p_{\max}\gamma_j^{e,O_j^e+1}-\gamma_j^{e,O_j^e}\right)\times\prod_{l=1}^{I_j^e-O_j^e}\exp\left(-\binom{l}{2}p_{\max}\gamma_j^{e,l}-\gamma_j^{e,l-1}\right)$$

where for gene j and branch e, I_j^e is the number of lineages entering e, O_j^e is the number of lineages exiting e, and $\gamma_j^{e,l}$ is the l^{th} coalescence time in e.

Identifiability results under multispecies coalescent

Recall the following classic metric on Jukes-Cantor(JC) substitution model (see *e.g*, [SS+03],[DNR14]).

Definition 13 (Tree metric for Jukes-Cantor model)**.** Given a species tree T, we define the dissimilarity

$$d_T(x,y)=-\frac{3}{4}\log\left(1-\frac{4}{3}\mathbb{E}[\hat{p}_{xy}]\right),\ \text{for } x,y\in L.$$

If there are at least $2m$ ($m\geq 1$) independent sites per locus, the classic empirical measure on leaves L is given by

$$\hat{p}_{xy}=\frac{1}{2mN}\sum_{i\in[N],j\in[2m]}\mathbf{1}_{\{\mathcal{X}_{ij}^x\neq\mathcal{X}_{ij}^y\}},\quad x,y\in L,$$

which can be thought of as the normalized hamming distance between the concatenated molecular sequences corresponding to taxon x and y.

Following the same construction in section 4.2.1, we first split the $2m$ sites evenly into two groups with $m \geq 1$. We introduce the estimator $\hat{H}^{JC}$ which is inspired by tree metric in Definition 13. For every two pairs of leaves $(x, y), (z, w) \in L^2$,

$$\hat{H}^{JC}(x, y; z, w) := -\frac{1}{mN} \sum_{i=1}^{N} \sum_{(j_1, j_2)} \left[1 - \frac{4}{3} \mathbf{1}_{\{\mathcal{X}_{ij_1}^x \neq \mathcal{X}_{ij_1}^y\}}\right] \left[1 - \frac{4}{3} \mathbf{1}_{\{\mathcal{X}_{ij_2}^z \neq \mathcal{X}_{ij_2}^w\}}\right].$$

By Theorem 4.2.1, it suffices to show its idealized version H^{JC} satisfies the Four points condition.

$$H^{JC}(x, y; z, w) = -\mathbb{E}\left[e^{-\frac{4}{3}[d_G(x,y)+d_G(z,w)]}\right]$$

where $d_G(x, y)$ is the distance between x and y on G and $\mathbb{E}$ is the expectation respect to the gene trees.

Lemma 4.2.1. For all $(x, y), (z, w) \in L^2$, we have

$$H^{JC}(x, y; z, w) = \mathbb{E}\hat{H}^{JC}(x, y; z, w)$$

Proof of Lemma 4.2.1. Applying total expectation and by the independence between sites, we have

$$\begin{aligned}
H^{JC}(x, y; z, w) &= -\mathbb{E}\left[\hat{H}^{JC}(x, y; z, w)\right] \\
&= -\mathbb{E}\left[\mathbb{E}\left[\left[1 - \frac{4}{3}\mathbf{1}_{\{\mathcal{X}_{ij_1}^x \neq \mathcal{X}_{ij_1}^y\}}\right]\left[1 - \frac{4}{3}\mathbf{1}_{\{\mathcal{X}_{ij_2}^z \neq \mathcal{X}_{ij_2}^w\}}\right] | G\right]\right] \\
&= -\mathbb{E}\left[\mathbb{E}\left[1 - \frac{4}{3}\mathbf{1}_{\{\mathcal{X}_{ij_1}^x \neq \mathcal{X}_{ij_1}^y\}} | G\right]\left[1 - \frac{4}{3}\mathbf{1}_{\{\mathcal{X}_{ij_2}^z \neq \mathcal{X}_{ij_2}^w\}} | G\right]\right] \\
&= -\mathbb{E}\left[e^{-\frac{4}{3}(d_G(x,y)+d_G(z,w))}\right]
\end{aligned}$$

where $d_G(x, y) = \sum_{e \in \text{path}(x,y;G)} \mu_e \tau_e$ represents the scaled length between taxa x and y

on gene tree G. □

Theorem 4.2.3 (4PT under MSC)**.** H^{JC} is a Four-taxon function and it satisfies the Four points condition respect to the phylogeny T under the MSC.

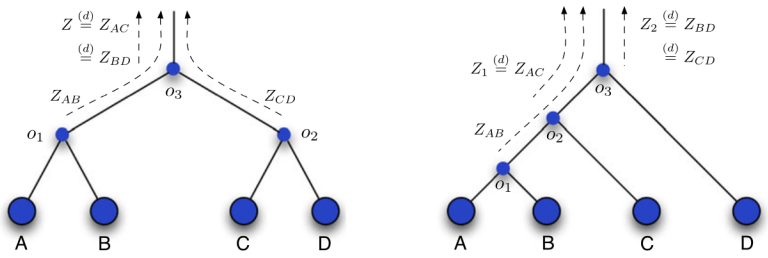

Figure 4.1: two possible tree topologies of T_4

Lemma 4.2.2 (Symmetric case)**.** Suppose distinct four leaves $x, y, z, w \in L$ that form a tree T_4 such that $((x, y)(z, w))$ with respect to $\mathcal{T}$ (Fig 4.1 (a)), then

$$\mathbb{E}\left[e^{\alpha\cdot[d_G(x,y)+d_G(z,w)]}\right] > \mathbb{E}\left[e^{\alpha\cdot[d_G(x,z)+d_G(y,w)]}\right].$$

where constant $\alpha < 0$.

Proof of Lemma 4.2.2. Let o_1, o_2, and o_3 be the common ancestors of (x, y), (z, w) and (x, z) respectively. Let $\mathcal{E}_1$ be the event that at least two lineages corresponding to x, y, z and w coalesce in the segment (o_1, o_3) or (o_2, o_3) of the tree and let $\overline{\mathcal{E}_1}$ be the event that this does not occur. From now on, for vertices $u, v \in T_4$, let μ_{uv} denotes $\sum_{e\in \mathrm{path}(u,v;T_4)} \mu_e \tau_e$.

Observe that on event $\mathcal{E}_1$, the gene tree G has the correct unrooted species tree topology, thus four points condition $d_G(x, y) + d_G(z, w) < d_G(x, z) + d_G(y, w)$ implies

$$\mathbb{E}\left[e^{\alpha d_G(x,y)+\alpha d_G(z,w)} - e^{\alpha d_G(x,z)+\alpha d_G(y,w)}|\mathcal{E}_1\right] > 0 \tag{4.5}$$

with $alpha < 0$.

Applying total expectation, we have

$$\begin{aligned}
&\mathbb{E}\left[e^{\alpha[d_G(x,y)+d_G(z,w)]}\right] \\
=&\mathbb{E}\left[e^{\alpha d_G(x,y)+\alpha d_G(z,w)}|\mathcal{E}_1\right]\mathbb{P}(\mathcal{E}_1)+\mathbb{E}\left[e^{\alpha d_G(x,y)+\alpha d_G(z,w)}|\overline{\mathcal{E}_1}\right]\mathbb{P}(\overline{\mathcal{E}_1}) \\
=&\mathbb{E}\left[e^{\alpha d_G(x,y)+\alpha d_G(z,w)}-e^{\alpha d_G(x,z)+\alpha d_G(y,w)}|\mathcal{E}_1\right]\mathbb{P}(\mathcal{E}_1) \\
&\quad+\mathbb{E}\left[e^{\alpha d_G(x,z)+\alpha d_G(y,w)}|\mathcal{E}_1\right]\mathbb{P}(\mathcal{E}_1)+\mathbb{E}\left[e^{\alpha d_G(x,y)+\alpha d_G(z,w)}|\overline{\mathcal{E}_1}\right]\mathbb{P}(\overline{\mathcal{E}_1}) \\
>&\mathbb{E}\left[e^{\alpha d_G(x,z)+\alpha d_G(y,w)}|\mathcal{E}_1\right]\mathbb{P}(\mathcal{E}_1)+\mathbb{E}\left[e^{\alpha d_G(x,z)+\alpha d_G(y,w)}|\overline{\mathcal{E}_1}\right]\mathbb{P}(\overline{\mathcal{E}_1}) \\
=&\mathbb{E}\left[e^{\alpha[d_G(x,z)+d_G(y,w)]}\right] \qquad (4.6)
\end{aligned}$$

where the second equality holds by linearity, and the inequality follows from (4.5) and the fact that on $\overline{\mathcal{E}_1}$ one always get

$$d_G(x,y)+d_G(z,w) \overset{dist}{=} d_G(x,z)+d_G(y,w)$$

since they both equal to the sum $\mu_{xo_1}+\mu_{yo_1}+\mu_{zo_2}+\mu_{wo_2}$ plus a term that depends on the gene tree topology above the root o_3, but that last one is symmetric in the leaf labels. □

Lemma 4.2.3 (Asymmetric case)**.** Suppose distinct four leaves $x, y, z, w \in L$ that form a tree T_4 such that $(((x,y),z),w)$ with respect to $\mathcal{T}$ (Fig 4.1 (b)), then

$$\mathbb{E}\left[e^{\alpha\cdot[d_G(x,y)+d_G(z,w)]}\right] > \mathbb{E}\left[e^{\alpha\cdot[d_G(x,z)+d_G(y,w)]}\right].$$

where constant $\alpha < 0$.

Proof of Lemma 4.2.3. We write o_1, o_2, o_3 to denote the most recent common ancestors

of (x, y), (x, z) and (x, w) respectively. In this case, we let $\mathcal{E}_2$ denote the event that the lineages corresponding to x and y coalesce in the branch (o_1, o_2) in Fig 4.1.

Similar to balance case, observe that on event $\mathcal{E}_2$, the gene tree G has the correct unrooted species tree topology, thus four points condition $d_G(x, y) + d_G(z, w) < d_G(x, z) + d_G(y, w)$ further implies

$$\mathbb{E}\left[e^{\alpha d_G(x,y)+\alpha d_G(z,w)} - e^{\alpha d_G(x,z)+\alpha d_G(y,w)}|\mathcal{E}_2\right] > 0 \tag{4.7}$$

with $\alpha < 0$.

Again, applying total expectation, we have

$$\begin{aligned}
&\mathbb{E}\left[e^{\alpha[d_G(x,y)+d_G(z,w)]}\right] \\
=&\mathbb{E}\left[e^{\alpha d_G(x,y)+\alpha d_G(z,w)}|\mathcal{E}_2\right]\mathbb{P}(\mathcal{E}_2) + \mathbb{E}\left[e^{\alpha d_G(x,y)+\alpha d_G(z,w)}|\overline{\mathcal{E}_2}\right]\mathbb{P}(\overline{\mathcal{E}_2}) \\
=&\mathbb{E}\left[e^{\alpha d_G(x,y)+\alpha d_G(z,w)} - e^{\alpha d_G(x,z)+\alpha d_G(y,w)}|\mathcal{E}_2\right]\mathbb{P}(\mathcal{E}_2) \\
&\quad + \mathbb{E}\left[e^{\alpha d_G(x,z)+\alpha d_G(y,w)}|\mathcal{E}_2\right]\mathbb{P}(\mathcal{E}_2) + \mathbb{E}\left[e^{\alpha d_G(x,y)+\alpha d_G(z,w)}|\overline{\mathcal{E}_2}\right]\mathbb{P}(\overline{\mathcal{E}_2}) \\
>&\mathbb{E}\left[e^{\alpha d_G(x,z)+\alpha d_G(y,w)}|\mathcal{E}_2\right]\mathbb{P}(\mathcal{E}_2) + \mathbb{E}\left[e^{\alpha d_G(x,z)+\alpha d_G(y,w)}|\overline{\mathcal{E}_2}\right]\mathbb{P}(\overline{\mathcal{E}_2}) \\
=&\mathbb{E}\left[e^{\alpha[d_G(x,z)+d_G(y,w)]}\right]
\end{aligned} \tag{4.8}$$

where the inequality follows from (4.7) and the fact that on event $\overline{\mathcal{E}_2}$ we always get

$$d_G(x, y) + d_G(z, w) \overset{dist}{=} d_G(x, z) + d_G(y, w)$$

since they both equal to the sum $\mu_{xo_1} + \mu_{yo_1} + \mu_{zo_2} + \mu_{wo_3} + 2\mu_{o_1o_2} + \mu_{o_2o_3}$ plus two terms that depend on the gene tree topology above the vertex o_2, but these last two are both symmetric in the leaf labels.

□

Proof of Theorem 4.2.3. We first show that for any distinct four leaves $x, y, z, w \in L$ that form a tree T_4 such that either $((x, y)(z, w))$ or $(((x, y), z), w)$ with respect to $\mathcal{T}$, then

$$H^{JC}\left((x, z), (y, w)\right) > H^{JC}\left((x, y), (z, w)\right).$$

It follows from Lemma 4.2.2 and Lemma 4.2.3 with constant $\alpha = -\frac{4}{3}$.

Using similar techniques, we will next establish that

$$H^{JC}\left((x, w), (y, z)\right) = H^{JC}\left((x, z), (y, w)\right).$$

It follows from the fact that in both cases $((x, y)(z, w))$ and $(((x, y), z), w)$ in Fig. 4.1, exchanging the closest taxon y and x has no affect on the distribution of gene tree under MSC, which is symmetric in the leaf labels, and by four points condition

$$d_G(x, z) + d_G(y, w) \stackrel{dist}{=} d_G(x, w) + d_G(y, z).$$

This concludes the proof. □

Next, we show the 4PC also holds after considering different rate across loci to the model. Let $\{\lambda_G\} \in \mathbb{R}_+$ be a random scaling parameters assigned to each sampled gene tree $\mathcal{G}$ with density function f, that is,

$$\tilde{H}^{JC}(x, y; z, w) = -\mathbb{E}\left[\int e^{-\frac{4}{3}\lambda_G[d_G(x,y)+d_G(z,w)]} df(\lambda_G)\right] \tag{4.9}$$

Corollary 4.2.1. *$\tilde{H}^{JC}$ satisfies the four points condition in Definition 12 with respect to T.*

4.2.3 Phylogenetic Inference with Determinants

For modeling the evolution of sequences composed of k bases, we use a continuous-time Markov process, with a $k \times k$ instantaneous rate matrix Q such that the off-diagonal entries of Q are nonnegative, the rows of Q sum to 0, and Q has stationary distribution π, with positive entries and $\pi Q = 0$.

In this section, we consider the identifiablity on this general Markov model with estimator with determinant (see [Ste94a]).

Definition 14 (logDet distance)**.** Let $\hat{F}_{G,xy}$ be a $k \times k$ matrix of empirical site-pattern frequencies, obtained by normalizing the site pattern count matrix for $x, y \in L$, so that its entries sum to 1.

$$d_G(x,y) = -\ln\left[\left|\det(\hat{F}_{G,xy})\right|\right].$$

Suppose there are at least $2km$ independent sites per loci ($m \geq 1$) with an arbitrary ordering, we first split the sites evenly into two groups M_1, M_2, such that $|M_1| = |M_2| = km$, then we consider the pairs $(\mathbf{l}_1, \mathbf{l}_2) \subset M_1 \times M_2$, where $\left\{\mathbf{l}_i = (l_i^{(1)}, \cdots, l_i^{(k)})\right\}_{p=1}^{m}$ forms a equal size partition of M_i, $i = 1, 2$.

Next, we introduce the estimator $\hat{H}^{det}$ over 2 pairs of leaves $(x,y),(w,z) \in L^2$ as

$$\hat{H}^{det}\left((x,y),(z,w)\right) := -\frac{1}{mN}\sum_{i=1}^{N}\sum_{(\mathbf{l}_1,\mathbf{l}_2)}\left[\det\left(\hat{F}^{xy}_{G_i\mathbf{l}_1}\hat{F}^{zw}_{G_i\mathbf{l}_2}\right)\right],$$

where $\hat{F}^{xy}_{G_i\mathbf{l}}$ is $k \times k$ matrix of empirical site-pattern frequencies between x and y from $\mathbf{l}$ sites on loci G_i, that is,

$$\hat{F}^{xy}_{G_i\mathbf{l}_1} := \left[\hat{f}^{xy}_{G_i\mathbf{l}_1}(\alpha,\beta)\right]_{\alpha\beta} = \left[\mathbf{1}\{\mathcal{X}^{x}_{il_1^{o_\alpha}} = \alpha, \mathcal{X}^{y}_{il_1^{o_\alpha}} = \beta\}\right]_{\alpha\beta},$$

$$\hat{F}^{zw}_{G_{i\mathbf{l}_2}} := \left[\hat{f}^{zw}_{G_{i\mathbf{l}_2}}(\alpha,\beta)\right]_{\alpha\beta} = \left[\mathbf{1}\{\mathcal{X}^{x}_{il_2^{o_\alpha}} = \alpha, \mathcal{X}^{y}_{il_2^{o_\alpha}} = \beta\}\right]_{\alpha\beta},$$

where $o : S \to [k]$ is a natural ordering of S, for example, $o_A = 1, o_C = 2, o_G = 3, o_T = 4$ in JC model.

Note that for all $k \times k$ matrix $\mathbf{P}$, we can estimate its determinant as following

$$\det(\mathbf{P}) = \sum_{\sigma \in S_k} \operatorname{sgn}(\sigma) \prod_{u=1}^{k} (\mathbf{P})_{u\sigma_u}, \tag{4.10}$$

where S_k is the collection of all permutation on $[k]$ and $\sigma_u := \sigma(u)$, $u \in [k]$.

Remark. By the independence between sites, the expression of determinants in (4.10) under expectation can be rewritten as the product of expectation by construction of $\hat{F}_{G_i}$ since we pick the same site $o(\alpha)$ on entry (α, β). In this construction, the estimator depends only on $2k$ independent sites per locus instead of $2k^2$ sites if naively picking different sites on all entries.

Since all loci are generated independently, we consider the idealized version of H^{det}, that is

$$H^{det}\left((x,y),(z,w)\right) := -\mathbb{E}\left[\det(\hat{F}^{xy}_{G}\hat{F}^{zw}_{G})\right] \tag{4.11}$$

where $\mathbb{E}$ is the expectation respect to the gene trees and $\hat{F}_{G,\cdot}$ is defined with a random gene tree G and its random $2k$ sites.

The general time reversible(GTR) model includes the additional assumption that $\operatorname{diag}(\pi)Q$ is symmetric. At the root of a gene tree, sites in the ancestral sequence have bases chosen independently with distribution π.

Theorem 4.2.4 (4PC under GTR)**.** H^{det} is a Split-equivalent function and it satisfies

the Four points condition respect to the phylogeny T.

Before the proof, we introduce some properties of GTR model (see [ALR19]).

Lemma 4.2.4 (lemma 2.2 [ALR19])**.** Let Q be a GTR rate matrix with stationary distribution π. Then $Q = S\Lambda S^{-1}$, where $S = \text{diag}(\pi)^{-1/2}U$ for some orthogonal matrix U, and $\Lambda = \text{diag}(\lambda)$ with $\lambda_1 = 0$, $\lambda_i \leq 0$. If $Q \neq 0$, then $\lambda_i < 0$ for some i.

For a vector v, we use $\exp(v)$ to denote the entrywise application of the exponential function.

Lemma 4.2.5 (lemma 2.3 [ALR19])**.** Let $Q = S\,\text{diag}(\lambda)S^{-1}$ be the diagonalization of a GTR rate matrix, and let $\mu(t)$ be a scalar-valued rate function. Then the Markov transition matrix $M(x) = M(\mu, Q, x)$ describing cumulative base substitutions with rate $\mu(t)Q$ for $t \in [0, x]$ is

$$M(x) = S\,\text{diag}\left(\exp(s(x)\lambda)\right)S^{-1},$$

where $s(x) = \int_0^x \mu(t)dt$. Thus the pairwise pattern frequency array $F = \text{diag}(\pi)M$ is symmetric positive definite.

Proof of Theorem 4.2.4. Following the same idea in previous proof, we want to show that for any distinct four leaves $x, y, z, w \in L$ that form a tree T_4 such that either $((x, y)(z, w))$ or $(((x, y), w), z)$ with respect to $\mathcal{T}$, then

$$H^{det}\left((x, w), (y, z)\right) = H^{det}\left((x, z), (y, w)\right) > H^{det}\left((x, y), (z, w)\right).$$

By lemma 4.2.5, we have

$$\mathbb{E}(\hat{F}_{G,xy}|G) = S\text{diag}(\exp(\lambda d_G(x,y)))S^{-1}$$

where $Q = S\text{diag}(\lambda)S^{-1}$ is the eigen decomposition of Q, $\lambda = (\lambda_1, \cdots, \lambda_k)$ are the eigenvalues of rate matrix Q and $d_G(x,y) = \sum_{e \in \text{path}(G;x,y)} \mu_e \tau_e$ is the rate scaled distance between taxa x and y. Applying product property of determinant $\det(AB) = \det(A)\det(B)$, $\det(S) = \det(S^{-1}) = 1$ and trace identity $\det(e^A) = e^{\text{tr}(A)}$, we have

$$\det\left(\mathbb{E}(\hat{F}_{G,xy}|G)\right) = \exp\left(\sum_{i=1}^{k} \lambda_i d_G(x,y)\right) \tag{4.12}$$

Combining 4.12 with equation (4.10), we have

$$\begin{aligned}
&\det(\hat{F}_{G,xy}\hat{F}_{G,zw}) \\
&= \det(\hat{F}_{G,xy})\det(\hat{F}_{G,zw}) \\
&= \left(\sum_{\sigma^1 \in S_k} \text{sgn}(\sigma^1) \prod_{u=1}^{k} (\hat{F}_{G,xy})_{u\sigma_u^1}\right)\left(\sum_{\sigma^2 \in S_k} \text{sgn}(\sigma^2) \prod_{v=1}^{k} (\hat{F}_{G,zw})_{v\sigma_v^2}\right) \\
&= \sum_{\sigma^1 \in S_k} \sum_{\sigma^2 \in S_k} \text{sgn}(\sigma^1)\text{sgn}(\sigma^2) \prod_{u=1}^{k}\prod_{v=1}^{k} \hat{f}_{G,xy}(u,\sigma_u^1)\hat{f}_{G,zw}(v,\sigma_v^2).
\end{aligned}$$

Applying total expectation, we have

$$\begin{aligned}
&\mathbb{E}\left[\det(\hat{F}_{G,xy}\hat{F}_{G,zw})\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\det(\hat{F}_{G,xy}\hat{F}_{G,zw})|G\right]\right] \\
&= \mathbb{E}\left[\sum_{\sigma^1 \in S_k} \sum_{\sigma^2 \in S_k} \text{sgn}(\sigma^1)\text{sgn}(\sigma^2)\mathbb{E}\left[\prod_{u=1}^{k}\prod_{v=1}^{k} \hat{f}_{G,xy}(u,\sigma_u^1)\hat{f}_{G,zw}(v,\sigma_v^2)|G\right]\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[\sum_{\sigma^1\in S_k}\sum_{\sigma^2\in S_k}\operatorname{sgn}(\sigma^1)\operatorname{sgn}(\sigma^2)\prod_{u=1}^{k}\prod_{v=1}^{k}\mathbb{E}\left[\hat{f}_{G,xy}(u,\sigma^1_u)\hat{f}_{G,zw}(v,\sigma^2_v)|G\right]\right]\\
&= \mathbb{E}\left[\sum_{\sigma^1\in S_k}\sum_{\sigma^2\in S_k}\operatorname{sgn}(\sigma^1)\operatorname{sgn}(\sigma^2)\prod_{u=1}^{k}\prod_{v=1}^{k}\mathbb{E}\left[\hat{f}_{G,xy}(u,\sigma^1_u)|G\right]\mathbb{E}\left[\hat{f}_{G,zw}(v,\sigma^2_v)|G\right]\right]\\
&= \mathbb{E}\left[\det\left(\mathbb{E}\left[\hat{F}_{G,xy}|G\right]\right)\det\left(\mathbb{E}\left[\hat{F}_{G,zw}|G\right]\right)\right]\\
&= \mathbb{E}\left[\det\left(\mathbb{E}\left[\hat{F}_{G,xy}|G\right]\mathbb{E}\left[\hat{F}_{G,zw}|G\right]\right)\right]\\
&= \mathbb{E}\left(\exp\left(\sum_{i=1}^{k}\lambda_i[d_G(x,y)+d_G(z,w)]\right)\right) \qquad (4.13)
\end{aligned}
$$

where the second equality holds by linearity, the third and fourth equality hold by independence between sites, and the last equality follows from equation (4.12) and property $\det(AB) = \det(A)\det(B)$.

Applying equation (4.13) on pairs $(xw; yz)$, $(xz; yw)$ and $(xy; zw)$, by lemma 4.2.2 and lemma 4.2.3 with the fact $\sum_{c=1}^{k}\lambda_c < 0$ by lemma 4.2.4, we have

$$\mathbb{E}\left[\det(\hat{F}^{xw}_G\hat{F}^{yz}_G)\right] = \mathbb{E}\left[\det(\hat{F}^{xz}_G\hat{F}^{yw}_G)\right] < \mathbb{E}\left[\det(\hat{F}^{xy}_G\hat{F}^{zw}_G)\right]$$

We complete the proof after substituting back to $H^{\det}$. □

Corollary 4.2.2 (4PC under MSC)**.** *H^{det} is a Split-equivalent function and it satisfies the Four points condition respect to the phylogeny T under MSC.*

Next, we assign different rates across genes. Let $\{\lambda_G\} \in \mathbb{R}_+$ be a random scaling parameters assigned to each sampled gene tree $G \in \mathcal{G}$ with density function f, that is,

$$\tilde{H}^{det}(x,y;z,w) = -\mathbb{E}\left[\int \lambda_G \det(\hat{F}^{xy}_G\hat{F}^{zw}_G)df(\lambda_G)\right]$$

Corollary 4.2.3. *$\tilde{H}^{\det}$ satisfies the Four points condition with respect to T.*

4.3 Phylogenetic level-1 network inference

4.3.1 Definitions with level-1 networks

Definition 15 (Phylogenetic Network)**.** A topological binary **rooted phylogenetic network** $N = (V, E, \gamma)$(we will usually use "network" or N for convenience) on labelled leaves set V_L(we also use X denote general leaves set) is a connected directed acyclic graph, where V is set of all nodes, E is set of all edges, γ is hybrid rate map. We denote r as the root. $V = \{r\} \cup V_L \cup V_H \cup V_{tree}$ is disjoint union,

1. The root r has indegree 0 and outdegree 2.

2. A leaf $v \in V_L$ has indegree 1 and outdegree 0.

3. A tree node $v \in V_{tree}$ has indegree 1 and outdegree 2.

4. A hybrid node $v \in V_H$ has indegree 2 and outdegree 1.

5. A hybrid edge $e \in E_H$ is an edge whose child is a hybrid node.

For each pair of hybrid edges (e_1, e_2) which share same hybrid-node child, are assigned hybridization parameters $\gamma : E_H \to (0, 1)$ satisfying $\gamma(e_1) + \gamma(e_2) = 1$. We also denote $n_L = |V_L|$ and $n_H = |V_H|$. $\bar{\gamma} = \max_{e \in E_H} \gamma(e)$, $\underline{\gamma} = \min_{e \in E_H} \gamma(e)$.

Definition 16 (Level-1 network)**.** The phylogenetic network N is level-1 if no two cycles share a vertex.

Definition 17 (LSA)**.** For two nodes a, b in a rooted network N, we write $a \leq b$ and say that a is *above* b if there is a directed path from a to b. We write $a < b$ if $a \leq b$

and $a \neq b$. For a set of nodes W in a rooted network N, let D be the set of nodes that lie on all paths from the root to the elements of W. The greatest element of D (i.e. the node $s \in D$ such that $s \geq t$ for all $t \in D$) is called the *lowest stable ancestor* of W, or LSA(W) [Ste16, p.263].

Definition 18 (semidirected network)**.** A *semidirected graph* $G^- = (V, E)$ is a tuple where V is the set of nodes, and $E = E_D \sqcup E_U$ with a set E_D of *directed edges* (also referred to as *hybrid edges)* and a set E_U of *undirected edges* (also referred to as *tree edges)*. E_D consists of ordered pairs (a, b) where $a, b \in V$. In contrast, E_U consists of unordered pairs $\{a, b\}$, such that if $\{a, b\} \in E_U$, then $(a, b) \notin E_D$, i.e. an edge cannot be both directed and undirected.

Let (N^+, f) be a rooted network on X. The *topological semidirected phylogenetic network induced from* (N^+, f) is a tuple (N^-, f), where N^- is the semidirected graph obtained by:

1. removing all the edges and nodes above LSA(X);

2. undirecting all tree edges $e \in E_T$, but keeping the direction of hybrid edges;

3. suppressing $s = \text{LSA}(X)$ if it has degree 2: if s is incident to two tree edges, then remove s and replace the two edges with a single undirected edge; if s is incident to one tree edge and one hybrid edge, then remove s, and replace the two edges by a directed edge with the same direction as the original hybrid edge.

For a semidirected graph M^- with vertex set V and labelling function $g : X \to V$, (M^-, g) is a *topological semidirected phylogenetic network* if it is the semidirected network induced from some rooted network.

Definition 19 (Mixed network)**.** A *mixed network* is a semidirected graph where

undirected edges are partitioned into two sets: tree edges E_T and split edges E_S; and where E_S is itself partitioned into a set of classes. When the graph is embedded in a Euclidean space, split edges within the same class are represented as parallel segments. A *metric* (ℓ, γ) on a mixed network M is such that $\ell : E \to \mathbb{R}_{\geq 0}$ assigns the same length to all edges in the same class of split edges; and $\gamma : E \to [0, 1]$ assigns $\gamma(e) = 1$ if e is undirected and $\gamma(e) \in (0, 1)$ if e is directed.

4.3.2 Identifiability on level-1 Network

In this section, we want to show the identifiability result on level-1 Network by viewing the distance on network as a linear combination of it on displayed trees.

Theorem 4.3.1 (Identifiability)**.** Under NMSC, the mixed network $\mathcal{N}^+$ is identifiable from d_1^N, d_2^N and H^{JC}.

Before the proof, we first introduces split system we used in the proof.

Definition 20 (Generalized Isolation Index and f-splits)**.** For f be a split-symmetric function on $L^2 \times L^2$. The **generalized isolation index** $\alpha_f(S)$ of a split $S = S1|S2$ over L is given by

$$\alpha_f(S) = \min\{\tilde{\alpha_f}(x_1, y_1; x_2, y_2) : x_1, y_1 \in S_1, x_2, y_2 \in S_2\}$$

where

$$\tilde{\alpha_f}(x_1, y_1; x_2, y_2) = \frac{1}{2}\left[\max\{f(x_1, y_1; x_2, y_2), f(x_1, x_2; y_1, y_2), f(x_1, y_2; y_1, x_2)\} - f(x_1, y_1; x_2, y_2)\right]$$

We say that S is a f-split if $\alpha_f(S) > 0$.

Let's claim one of the most important results of this chapter, about the sign of

f-splits in tree and network. Denote $\bar{\mathcal{S}}$ is the collection of all possible splits on L.

Lemma 4.3.1 (f-split on Tree)**.** Let T to be a species tree with leaf set L and exists split-symmetric function f^T satisfies the 4PC w.r.t. T, then all f^T-splits coincides with $\mathcal{S}^T$, that is

$$\begin{cases} \alpha_{f^T}(S) > 0, \ \forall S \in \mathcal{S}^T \\ \alpha_{f^T}(S) = 0, \ \forall S \in \bar{\mathcal{S}}/\mathcal{S}^T \end{cases} \tag{4.14}$$

Proof. We first show $\alpha_f(S)$ is positive, $\forall S = S_1|S_2 \in \mathcal{S}(T)$. Picking arbitrary four points $x_1, x_2 \in S_1$ and $y_1, y_2 \in S_2$ (not necessary distinct), both x_1, x_2 and y_1, y_2 are sisters on restrict tree $T|_{\{x_1,x_2,y_1,y_2\}}$, then $f(x_1, x_2; y_1, y_2) < f(x_1, y_1; x_2, y_2)$ by 4PC, thus $\tilde{\alpha}_f(x_1, x_2|y_1, y_2) > 0$ and $\alpha_f(S) > 0$ after taking minimum over all choices over S_1 and S_2.

Next, we show $\alpha_f(S)$ is zero, if $S = S_1|S_2 \notin \mathcal{S}(T)$. There exists $x_1, x_2 \in S_1$ and $y_1, y_2 \in S_2$ such that x_1, y_1 is sisters on restrict tree $T|_{\{x_1,x_2,y_1,y_2\}}$, then $f(x_1, x_2; y_1, y_2) > f(x_1, y_1; x_2, y_2)$ by 4PC, then $\alpha_f(S) \leq \tilde{\alpha}_f(x_1, x_2|y_1, y_2) = 0$. □

Theorem 4.3.2 (f-split on Network)**.** Let N to be a general phylogenetic network(not necessarily level-1), and call T to be a displayed tree of N, if T is a tree generated by deleting edges and suppressing degree 2 nodes on N. If there exists a split-symmetric function f^T holds 4PC w.r.t. T for all displayed tree T on N, and we denote

$$f^N = \mathbb{E}^{dis}[f^T] \tag{4.15}$$

with $\mathbb{E}^{dis}$ is the probability measure of displayed tree, and $\mathbb{P}^{dis}[T] > 0$ for all $T \in \mathcal{G}(N)$, then all f^N-splits coincides with $\mathcal{S}(N)$.

Proof. If $S = S_1|S_2 \in \mathcal{S}(N)$, observe that $\forall x_1, x_2 \in S_1$ and $y_1, y_2 \in S_2$, we have

$$f^T(x_1, x_2; y_1, y_2) \leq \max\{f^T(x_1, y_1; x_2, y_2), f^T(x_1, y_2; y_1, x_2)\}$$

on all $T \in \mathcal{G}(N)$ by 4PC and the strict inequality holds on at least one T. By linearity, we have

$$\mathbb{E}^{dist}\left[f^T(x_1, x_2; y_1, y_2)\right] < \max\{\mathbb{E}^{dist}\left[f^T(x_1, y_1; x_2, y_2)\right], \mathbb{E}^{dist}\left[f^T(x_1, y_2; y_1, x_2)\right]\}$$

where we apply $\sum w_i a_i \geq \max\{\sum w_i b_i, \sum w_i c_i\}$ with $w_i \geq 0$ and $a_i, b_i, c_i > 0$, if $a_i \geq \max\{b_i, c_i\}$, which implies $\alpha_{f^N}(x_1, x_2|y_1, y_2) > 0$. Thus $\alpha_{f^N}(S) > 0$ after taking minimum over all possible choices.

If $S = S_1|S_2 \notin \mathcal{S}(N)$, we have $S \notin \mathcal{S}(T)$ for all $T \in \mathcal{G}(N)$. Then on a displayed tree T_0, there exists $x_1, x_2 \in S_1$ and $y_1, y_2 \in S_2$ such that

$$f^{T_0}(x_1, x_2; y_1, y_2) \geq \max\{f^{T_0}(x_1, y_1; x_2, y_2), f^{T_0}(x_1, y_2; y_1, x_2)\}$$

that is, x_1, x_2 are not sisters on $T_0|_{\{x_1,x_2,y_1,y_2\}}$ by 4PC. Observe that x_1, x_2 are not sisters on $T|_{\{x_1,x_2,y_1,y_2\}}$ for all $T \in \mathcal{G}(N)$, since deleting edges or suppressing vertex with degree 2 will not change the circular order. Thus

$$f^T(x_1, x_2; y_1, y_2) \geq \max\{f^T(x_1, y_1; x_2, y_2), f^T(x_1, y_2; y_1, x_2)\}$$

for all $T \in \mathcal{G}(N)$ by 4PC. Thus

$$\begin{aligned} f^N(x_1, x_2; y_1, y_2) &= \mathbb{E}^{dist}\left[f^T(x_1, y_1; x_2, y_2)\right] \\ &\geq \max\{\mathbb{E}^{dist}\left[f^T(x_1, y_1; x_2, y_2)\right], \mathbb{E}^{dist}\left[f^T(x_1, y_2; y_1, x_2)\right]\} \end{aligned}$$

by observing $\sum w_i a_i \geq \max\{\sum w_i b_i, \sum w_i c_i\}$ with $w_i \geq 0$ and $a_i, b_i, c_i > 0$, if $a_i \geq \max\{b_i, c_i\}$. Thus $\alpha_{f^N}(S) \leq \alpha_{f^N}(x_1, x_2|y_1, y_2) = 0$.

□

Chapter 5

Discussion

In chapter 2, we proposed a new way to estimate the number of latent dimensions in a graph k using the concept of cross-validated eigenvalues. Through edge splitting and thanks to a simple central limit theorem, the estimation of cross-validated eigenvalues is efficient for very large graphs. The paper also provides theoretical justification showing that the estimator is consistent. Our simulations and empirical data application validate the theory and further demonstrate the efficacy of the proposed method. In addition to being quickly computable, a key advantage of cross-validated eigenvalues is our rigorous understanding of their behavior outside of the asymptotic setting where all k dimensions can be estimated. Theorem 2.3.1 encodes this rigorous understanding into a p-value. This theorem requires very little of the population matrix P; it does not presume that it is from a Degree-Corrected Stochastic Blockmodel, nor does it presume the actual rank of P or the order of the eigenvalue being tested. Of course, this level of ease and generality comes with a price. In particular, we only get to compute the eigenvectors with $1 - \varepsilon$ fraction of the edges. The natural possibility is to estimate k with a fraction of the edges and then recompute the eigenvectors with the full graph. Going forward, we hope others will join us in crafting new estimators for $\lambda_P(\hat{x}_j)$ that do not require leaving out edges.

In chapter 3, we show the most topological features of an ultrametric phylogeny can be identified from the distribution of gene trees under LGT with arbitrary constant transfer rate. More generally, the argument also works for the unrooted gene trees, which

works better for the sequence data. Beyond the theoretical identifiability question, we also provide a consistent reconstruction algorithm. As we state in the introduction, we view the LGT process as a Markov process on graph space $\mathcal{G}$ and reconstruct the phylogeny backward in time from the leaves. The computation cost is expansive since we need to compute the inverse matrix at each level. Moreover, there are a number of questions can be studied in the future. We are interested in extending to more complex model, for example in the present of extinction, incorporating gene duplication and loss. Since the leaf set on the gene trees will be affected in generalized model, there is not guarantee to recover the generator matrix recursively from the leaf-edges.

In chapter 4, we proposed a new estimator to show the identifiability of level-1 phylogenetic networks under some common DNA substitution models including Jukes-Cantor and General time reversible model. However, the idea in the proof of lemma 4.2.2 and 4.2.3 that applying the four points condition on some gene tree with corrected structure is not helpful in finding a suitable gap to compute the sample complexity. So more work need to do next step when considering the construction algorithm and its sample complexity. We also proposed a new estimator defined on $2k$ many independent sites per gene under the GTR model, where k represents the number of states under the substitution model. However, k may not be the smallest number of nonzero entries to compute a nontrivial determinant of a matrix, one could reduce the number by more advance algebraic tools. In this dissertation, when extending to the network, we use the natural generation NMSC model which indicates the lineages picks their hybrid parents total dependent, however in biology, it's more common to pick their parents independently across the hybrid edges, so more simulation work to compare the difference between these two Multispecies coalescence on networks need to be explored.

Bibliography

[Abb+20] Emmanuel Abbe et al. “Entrywise eigenvector analysis of random matrices with low expected rank”. In: *Annals of Statistics* 48.3 (2020), pp. 1452–1474.

[Abb18] Emmanuel Abbe. “Community Detection and Stochastic Block Models”. English. In: *Foundations and Trends® in Communications and Information Theory* 14.1-2 (June 2018). Publisher: Now Publishers, Inc., pp. 1–162.

[ABH16] E. Abbe, A. S. Bandeira, and G. Hall. “Exact Recovery in the Stochastic Block Model”. In: *IEEE Transactions on Information Theory* 62.1 (Jan. 2016). Conference Name: IEEE Transactions on Information Theory, pp. 471–487.

[AC10] Sylvain Arlot and Alain Celisse. “A survey of cross-validation procedures for model selection”. In: *Statistics Surveys* 4.none (Jan. 2010). Publisher: Amer. Statist. Assoc., the Bernoulli Soc., the Inst. Math. Statist., and the Statist. Soc. Canada, pp. 40–79.

[ADZ14] Karim M. Abadir, Walter Distaso, and Filip Zikes. “Design-free estimation of variance matrices”. en. In: *Journal of Econometrics* 181.2 (Aug. 2014), pp. 165–180.

[AEK17] Oskari H. Ajanki, László Erdős, and Torben Krüger. “Universality for general Wigner-type matrices”. en. In: *Probability Theory and Related Fields* 169.3 (Dec. 2017), pp. 667–727.

[Ahl21] Thomas D. Ahle. “Sharp and Simple Bounds for the raw Moments of the Binomial and Poisson Distributions”. In: *arXiv:2103.17027 [math, stat]* (Apr. 2021). arXiv: 2103.17027.

[Air+08] Edoardo Maria Airoldi et al. “Mixed membership stochastic blockmodels”. In: *Journal of machine learning research* (2008).

[Ald85] David J Aldous. “Exchangeability and related topics”. In: *École d’Été de Probabilités de Saint-Flour XIII—1983*. Springer, 1985, pp. 1–198.

[Ale+14] Bloemendal Alex et al. “Isotropic local laws for sample covariance and generalized Wigner matrices”. EN. In: *Electronic Journal of Probability* 19 (2014). Publisher: The Institute of Mathematical Statistics and the Bernoulli Society.

[ALR19] Elizabeth S Allman, Colby Long, and John A Rhodes. “Species tree inference from genomic sequences using the log-det distance”. In: *SIAM journal on applied algebra and geometry* 3.1 (2019), pp. 107–127.

[AM07] Dimitris Achlioptas and Frank McSherry. “Fast computation of low-rank matrix approximations”. In: *Journal of the ACM (JACM)* 54.2 (2007), 9–es.

[Amm+18] Waleed Ammar et al. “Construction of the literature graph in semantic scholar”. In: *arXiv preprint arXiv:1805.02262* (2018).

[Arn97] Michael L Arnold. *Natural hybridization and evolution*. Oxford University Press on Demand, 1997.

[AS15] Emmanuel Abbe and Colin Sandon. “Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms”. In: *arXiv:1503.00609 [cs, math]* (Apr. 2015). arXiv: 1503.00609.

[Ath+13] Avanti Athreya et al. “A central limit theorem for scaled eigenvectors of random dot product graphs”. In: *arXiv preprint arXiv:1305.7388* (2013).

[Bap+05] Eric Bapteste et al. “Do orthologous gene phylogenies really support tree-thinking?” In: *BMC Evolutionary Biology* 5.1 (2005), p. 33.

[BB15] Sharmodeep Bhattacharyya and Peter J. Bickel. "Subsampling bootstrap of count features of networks". EN. In: *Annals of Statistics* 43.6 (Dec. 2015). Publisher: Institute of Mathematical Statistics, pp. 2384–2411.

[BBK19] Florent Benaych-Georges, Charles Bordenave, and Antti Knowles. "Largest eigenvalues of sparse inhomogeneous Erdős–Rényi graphs". EN. In: *Annals of Probability* 47.3 (May 2019). Publisher: Institute of Mathematical Statistics, pp. 1653–1676.

[BBK20] Florent Benaych-Georges, Charles Bordenave, and Antti Knowles. "Spectral radii of sparse random matrices". In: *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*. Vol. 56. 3. Institut Henri Poincaré. 2020, pp. 2141–2161.

[BH95] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

[BLM15] C. Bordenave, M. Lelarge, and L. Massoulié. "Non-backtracking Spectrum of Random Graphs: Community Detection and Non-regular Ramanujan Graphs". In: *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*. ISSN: 0272-5428. Oct. 2015, pp. 1347–1357.

[BM21] Douglas Bates and Martin Maechler. *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.3-3. 2021.

[BN03] Mikhail Belkin and Partha Niyogi. "Laplacian eigenmaps for dimensionality reduction and data representation". In: *Neural computation* 15.6 (2003), pp. 1373–1396.

[Bou+13] Bastien Boussau et al. "Genome-scale coestimation of species and gene trees". In: *Genome research* 23.2 (2013), pp. 323–330.

[BS16] Peter J. Bickel and Purnamrita Sarkar. “Hypothesis testing for automated community detection in networks”. In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 78.1 (2016). Publisher: [Royal Statistical Society, Wiley], pp. 253–273.

[Bun74] Peter Buneman. “A note on the metric properties of trees”. In: *Journal of Combinatorial Theory, Series B* 17.1 (1974), pp. 48–50.

[BW09] Mohamed-Ali Belabbas and Patrick J Wolfe. “Spectral methods in machine learning and new strategies for very large datasets”. In: *Proceedings of the National Academy of Sciences* 106.2 (2009), pp. 369–374.

[CCB16] Diana Cai, Trevor Campbell, and Tamara Broderick. “Edge-exchangeable graphs and sparsity”. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 2016, pp. 4249–4257.

[CCH20] Arijit Chakrabarty, Sukrit Chakraborty, and Rajat Subhra Hazra. “Eigenvalues Outside the Bulk of Inhomogeneous Erdős–Rényi Random Graphs”. en. In: *Journal of Statistical Physics* 181.5 (Dec. 2020), pp. 1746–1780.

[CD18] Harry Crane and Walter Dempsey. “Edge exchangeable models for interaction networks”. In: *Journal of the American Statistical Association* 113.523 (2018), pp. 1311–1326.

[Cha15] Sourav Chatterjee. “Matrix estimation by Universal Singular Value Thresholding”. In: *The Annals of Statistics* 43.1 (Feb. 2015). Publisher: Institute of Mathematical Statistics, pp. 177–214.

[Che+21a] Fan Chen et al. “Estimating Graph Dimension with Cross-validated Eigenvalues”. In: *arXiv preprint arXiv:2108.03336* (2021).

[Che+21b] Yuxin Chen et al. “Spectral methods for data science: A statistical perspective”. In: *Foundations and Trends® in Machine Learning* 14.5 (2021), pp. 566–806.

[CK15] Julia Chifman and Laura Kubatko. "Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites". In: *Journal of theoretical biology* 374 (2015), pp. 35–47.

[CL18] Kehui Chen and Jing Lei. "Network Cross-Validation for Determining the Number of Communities in Network Data". In: *Journal of the American Statistical Association* 113.521 (Jan. 2018). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2016.1246365, pp. 241–251.

[CR09] Emmanuel J Candès and Benjamin Recht. "Exact matrix completion via convex optimization". In: *Foundations of Computational mathematics* 9.6 (2009), p. 717.

[CR20] Fan Chen and Karl Rohe. "A new basis for sparse PCA". In: *arXiv preprint arXiv:2007.00596* (2020).

[Cra+09] Karen A Cranston et al. "Species trees from highly incongruent gene trees in rice". In: *Systematic Biology* 58.5 (2009), pp. 489–500.

[CZR19] Fan Chen, Yini Zhang, and Karl Rohe. "Targeted sampling from massive block model graphs with personalized PageRank". In: *arXiv preprint arXiv:1910.12937* (2019).

[DB07] W Ford Doolittle and Eric Bapteste. "Pattern pluralism and the Tree of Life hypothesis". In: *Proceedings of the National Academy of Sciences* 104.7 (2007), pp. 2043–2049.

[DNR14] Gautam Dasarathy, Robert Nowak, and Sebastien Roch. "Data requirement for phylogenetic inference from multiple loci: a new distance method". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 12.2 (2014), pp. 422–432.

[DR13a] Constantinos Daskalakis and Sebastien Roch. "Alignment-free phylogenetic reconstruction: Sample complexity via a branching process analysis". In: *The Annals of Applied Probability* 23.2 (2013), pp. 693–721.

[DR13b] Constantinos Daskalakis and Sebastien Roch. “Alignment-free phylogenetic reconstruction: Sample complexity via a branching process analysis”. EN. In: *Annals of Applied Probability* 23.2 (Apr. 2013). Publisher: Institute of Mathematical Statistics, pp. 693–721.

[DR16] Constantinos Daskalakis and Sebastien Roch. “Species trees from gene trees despite a high rate of lateral genetic transfer: A tight bound”. In: *Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms*. SIAM. 2016, pp. 1621–1630.

[Dur19] Rick Durrett. *Probability: Theory and Examples*. 5th ed. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[DZ19] Ioana Dumitriu and Yizhe Zhu. “Sparse general Wigner-type matrices: Local law and eigenvector delocalization”. In: *Journal of Mathematical Physics* 60.2 (Feb. 2019). Publisher: American Institute of Physics, p. 023301.

[Erd+12] László Erdős et al. “Spectral Statistics of Erdős-Rényi Graphs II: Eigenvalue Spacing and the Extreme Eigenvalues”. en. In: *Communications in Mathematical Physics* 314.3 (Sept. 2012), pp. 587–640.

[Erd+13] László Erdős et al. “The local semicircle law for a general class of random matrices”. EN. In: *Electronic Journal of Probability* 18 (2013). Publisher: The Institute of Mathematical Statistics and the Bernoulli Society.

[Erd+99] Péter L Erdős et al. “A few logs suffice to build (almost) all trees (I)”. In: *Random Structures & Algorithms* 14.2 (1999), pp. 153–184.

[Fis36] Ronald A Fisher. “The use of multiple measurements in taxonomic problems”. In: *Annals of eugenics* 7.2 (1936), pp. 179–188.

[FP+20] Cheryl Flynn, Patrick Perry, et al. “Profile likelihood biclustering”. In: *Electronic Journal of Statistics* 14.1 (2020), pp. 731–768.

[Gal07] Nicolas Galtier. "A Model of Horizontal Gene Transfer and the Bacterial Phylogeny Problem". In: *Systematic Biology* 56.4 (2007), pp. 633–642.

[GD08] Nicolas Galtier and Vincent Daubin. "Dealing with incongruence in phylogenomic analyses". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 363.1512 (2008), pp. 4023–4029.

[GS17] Alden Green and Cosma Rohilla Shalizi. "Bootstrapping Exchangeable Random Graphs". In: *arXiv:1711.00813 [stat]* (Nov. 2017). arXiv: 1711.00813.

[HC17] Douglas D Heckathorn and Christopher J Cameron. "Network sampling: From snowball and multiplicity to respondent-driven sampling". In: *Annual review of sociology* 43 (2017), pp. 101–119.

[Hib+21] Samuel GS Hibdige et al. "Widespread lateral gene transfer among grasses". In: *New Phytologist* 230.6 (2021), pp. 2474–2486.

[HLY20] Jong Yun Hwang, Ji Oon Lee, and Wooseok Yang. "Local law and Tracy–Widom limit for sparse stochastic block models". EN. In: *Bernoulli* 26.3 (Aug. 2020). Publisher: Bernoulli Society for Mathematical Statistics and Probability, pp. 2400–2435.

[Hoo89] DN Hoover. "Tail fields of partially exchangeable arrays". In: *Journal of Multivariate Analysis* 31.1 (1989), pp. 160–163.

[Hot33] Harold Hotelling. "Analysis of a complex of statistical variables into principal components." In: *Journal of educational psychology* 24.6 (1933), p. 417.

[HRH02] Peter D Hoff, Adrian E Raftery, and Mark S Handcock. "Latent space approaches to social network analysis". In: *Journal of the american Statistical association* 97.460 (2002), pp. 1090–1098.

[HRS10] Daniel H Huson, Regula Rupp, and Celine Scornavacca. *Phylogenetic networks: concepts, algorithms and applications*. Cambridge University Press, 2010.

[JC+69] Thomas H Jukes, Charles R Cantor, et al. “Evolution of protein molecules”. In: *Mammalian protein metabolism* 3 (1969), pp. 21–132.

[JC14] Abigail Z Jacobs and Aaron Clauset. “A unified view of generative models for networks: models, methods, opportunities, and challenges”. In: *arXiv preprint arXiv:1411.4070* (2014).

[Jin+20] Jiashun Jin et al. “Estimating the number of communities by Stepwise Goodness-of-fit”. In: *arXiv:2009.09177 [math, stat]* (Sept. 2020). arXiv: 2009.09177.

[JY16] Antony Joseph and Bin Yu. “Impact of regularization on spectral clustering”. EN. In: *Annals of Statistics* 44.4 (Aug. 2016). Publisher: Institute of Mathematical Statistics, pp. 1765–1791.

[KMO09] Raghunandan Keshavan, Andrea Montanari, and Sewoong Oh. “Matrix completion from noisy entries”. In: *Advances in neural information processing systems* 22 (2009).

[KN11] Brian Karrer and M. E. J. Newman. “Stochastic blockmodels and community structure in networks”. In: *Physical Review E* 83.1 (Jan. 2011). Publisher: American Physical Society, p. 016107.

[Kru64] Joseph B Kruskal. “Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis”. In: *Psychometrika* 29.1 (1964), pp. 1–27.

[Krz+13] Florent Krzakala et al. “Spectral redemption in clustering sparse networks”. en. In: *Proceedings of the National Academy of Sciences* 110.52 (Dec. 2013). Publisher: National Academy of Sciences Section: Physical Sciences, pp. 20935–20940.

[Lam+18] Sangeet Lamichhaney et al. “Rapid hybrid speciation in Darwin’s finches”. In: *Science* 359.6372 (2018), pp. 224–228.

[Lam16] Clifford Lam. “Nonparametric eigenvalue-regularized precision or covariance matrix estimator”. In: *The Annals of Statistics* 44.3 (June 2016). Publisher: Institute of Mathematical Statistics, pp. 928–953.

[Lar+10] Bret R Larget et al. “BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis”. In: *Bioinformatics* 26.22 (2010), pp. 2910–2911.

[Lei16] Jing Lei. “A goodness-of-fit test for stochastic block models”. EN. In: *Annals of Statistics* 44.1 (Feb. 2016). Publisher: Institute of Mathematical Statistics, pp. 401–424.

[Liu+19] Yan Liu et al. “Community Detection Based on the L_∞ convergence of eigenvectors in DCBM”. In: *arXiv:1906.06713 [math, stat]* (June 2019). arXiv: 1906.06713.

[LKF07] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. “Graph evolution: Densification and shrinking diameters”. In: *ACM transactions on Knowledge Discovery from Data (TKDD)* 1.1 (2007), 2–es.

[LL19a] Can M. Le and Elizaveta Levina. “Estimating the number of communities in networks by spectral methods”. In: *arXiv:1507.00827 [cs, math, stat]* (Nov. 2019). arXiv: 1507.00827.

[LL19b] Keith Levin and Elizaveta Levina. “Bootstrapping Networks with Latent Space Structure”. In: *arXiv:1907.10821 [math, stat]* (July 2019). arXiv: 1907.10821.

[LLS20a] Qiaohui Lin, Robert Lunde, and Purnamrita Sarkar. “Higher-Order Correct Multiplier Bootstraps for Count Functionals of Networks”. In: *arXiv:2009.06170 [math, stat]* (Sept. 2020). arXiv: 2009.06170.

[LLS20b] Qiaohui Lin, Robert Lunde, and Purnamrita Sarkar. “On the Theoretical Properties of the Network Jackknife”. en. In: *International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, Nov. 2020, pp. 6105–6115.

[LLV17] Can M. Le, Elizaveta Levina, and Roman Vershynin. “Concentration and regularization of random graphs”. en. In: *Random Structures & Algorithms* 51.3 (2017). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/rsa.20713, pp. 538–561.

[LLZ20] Tianxi Li, Elizaveta Levina, and Ji Zhu. "Network cross-validation by edge sampling". In: *Biometrika* 107.2 (June 2020), pp. 257–276.

[Lov12] László Lovász. *Large networks and graph limits*. Vol. 60. American Mathematical Soc., 2012.

[LP07] Liang Liu and Dennis K Pearl. "Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions". In: *Systematic biology* 56.3 (2007), pp. 504–514.

[LRH07] Simone Linz, Achim Radtke, and Arndt von Haeseler. "A likelihood framework to measure horizontal gene transfer". In: *Molecular biology and evolution* 24.6 (2007), pp. 1312–1319.

[LS19] Robert Lunde and Purnamrita Sarkar. "Subsampling Sparse Graphons Under Minimal Assumptions". In: *arXiv:1907.12528 [math, stat]* (Aug. 2019). arXiv: 1907.12528.

[Mal07] James Mallet. "Hybrid speciation". In: *Nature* 446.7133 (2007), pp. 279–283.

[MC17] Bradon R McDonald and Cameron R Currie. "Lateral gene transfer dynamics in the ancient bacterial genus Streptomyces". In: *MBio* 8.3 (2017).

[McD+10] Lauren D McDaniel et al. "High frequency of horizontal gene transfer in the oceans". In: *Science* 330.6000 (2010), pp. 50–50.

[McS01] Frank McSherry. "Spectral partitioning of random graphs". In: *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*. IEEE. 2001, pp. 529–537.

[Mir+14] Siavash Mirarab et al. "ASTRAL: genome-scale coalescent-based species tree estimation". In: *Bioinformatics* 30.17 (2014), pp. i541–i548.

[MK06] Wayne P Maddison and L Lacey Knowles. "Inferring phylogeny despite incomplete lineage sorting". In: *Systematic biology* 55.1 (2006), pp. 21–30.

[MK09] Chen Meng and Laura Salter Kubatko. “Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model”. In: *Theoretical population biology* 75.1 (2009), pp. 35–45.

[MNS15] Elchanan Mossel, Joe Neeman, and Allan Sly. “Reconstruction and estimation in the planted partition model”. In: *Probability Theory and Related Fields* 162.3 (2015), pp. 431–461.

[MR08] Elchanan Mossel and Sebastien Roch. “Incomplete lineage sorting: consistent phylogeny estimation from multiple loci”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 7.1 (2008), pp. 166–171.

[MS07] Frederick A Matsen and Mike Steel. “Phylogenetic mixtures on a single tree can mimic a tree of another topology”. In: *Systematic Biology* 56.5 (2007), pp. 767–775.

[MSZ19] Shujie Ma, Liangjun Su, and Yichong Zhang. “Determining the Number of Communities in Degree-corrected Stochastic Block Models”. In: *arXiv:1809.01028 [stat]* (July 2019). arXiv: 1809.01028.

[Nak13] Luay Nakhleh. “Computational approaches to species phylogeny inference and gene tree reconciliation”. In: *Trends in ecology & evolution* 28.12 (2013), pp. 719–728.

[Nau+21] Zacharie Naulet et al. “Bootstrap estimators for the tail-index and for the count statistics of graphex processes”. EN. In: *Electronic Journal of Statistics* 15.1 (2021). Publisher: The Institute of Mathematical Statistics and the Bernoulli Society, pp. 282–325.

[Ney71] Jerzy Neyman. “Molecular studies of evolution: a source of novel statistical problems”. In: *Statistical decision theory and related topics*. Elsevier, 1971, pp. 1–27.

[PC84] Richard R. Picard and R. Dennis Cook. “Cross-Validation of Regression Models”. In: *Journal of the American Statistical Association* 79 (1984).

Publisher: [American Statistical Association, Taylor & Francis, Ltd.], pp. 575–583.

[Pea01] Karl Pearson. "LIII. On lines and planes of closest fit to systems of points in space". In: *The London, Edinburgh, and Dublin philosophical magazine and journal of science* 2.11 (1901), pp. 559–572.

[PN88] Pekka Pamilo and Masatoshi Nei. "Relationships between gene trees and species trees." In: *Molecular biology and evolution* 5.5 (1988), pp. 568–583.

[Pol+06] Daniel A Pollard et al. "Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting". In: *PLoS Genet* 2.10 (2006), e173.

[QM19] Yixuan Qiu and Jiali Mei. *RSpectra: Solvers for Large-Scale Eigenvalue and SVD Problems*. R package version 0.16-0. 2019.

[QR13] Tai Qin and Karl Rohe. "Regularized spectral clustering under the Degree-Corrected Stochastic Blockmodel". In: *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*. 2013, pp. 3120–3128.

[RCY11a] Karl Rohe, Sourav Chatterjee, and Bin Yu. "Spectral clustering and the high-dimensional stochastic blockmodel". In: *The Annals of Statistics* 39.4 (2011), pp. 1878–1915.

[RCY11b] Karl Rohe, Sourav Chatterjee, and Bin Yu. "Spectral clustering and the high-dimensional stochastic blockmodel". EN. In: *Annals of Statistics* 39.4 (Aug. 2011). Publisher: Institute of Mathematical Statistics, pp. 1878–1915.

[Rie97] Loren H Rieseberg. "Hybrid origins of plant species". In: *Annual review of Ecology and Systematics* 28.1 (1997), pp. 359–389.

[Roh+18] Karl Rohe et al. "A Note on Quickly Sampling a Sparse Matrix with Low Rank Expectation". In: *Journal of Machine Learning Research* 19.77 (2018), pp. 1–13.

[Roh19] Karl Rohe. “A critical threshold for design effects in network sampling”. In: *The Annals of Statistics* 47.1 (2019), pp. 556–582.

[RS13] Sebastien Roch and Sagi Snir. “Recovering the treelike trend of evolution despite extensive lateral genetic transfer: a probabilistic analysis”. In: *Journal of Computational Biology* 20.2 (2013), pp. 93–112.

[RS15] Sebastien Roch and Mike Steel. “Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent”. In: *Theoretical population biology* 100 (2015), pp. 56–62.

[RSM19] Maryam Rabiee, Erfan Sayyari, and Siavash Mirarab. “Multi-allele species reconstruction using ASTRAL”. In: *Molecular phylogenetics and evolution* 130 (2019), pp. 286–296.

[RY03] Bruce Rannala and Ziheng Yang. “Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci”. In: *Genetics* 164.4 (2003), pp. 1645–1656.

[RZ20] Karl Rohe and Muzhe Zeng. “Vintage Factor Analysis with Varimax Performs Statistical Inference”. In: *arXiv preprint arXiv:2004.05387* (2020).

[SB99] Tom AB Snijders and Stephen P Borgatti. “Non-parametric standard errors and tests for network statistics”. In: *Connections* 22.2 (1999), pp. 161–170.

[SM01] Mike Steel and Andy McKenzie. “Properties of phylogenetic trees generated by Yule-type speciation models”. In: *Mathematical biosciences* 170.1 (2001), pp. 91–112.

[Sos99] Alexander Soshnikov. “Universality at the Edge of the Spectrum in Wigner Random Matrices”. en. In: *Communications in Mathematical Physics* 207.3 (Nov. 1999), pp. 697–733.

[SR08] Mike Steel and Allen Rodrigo. “Maximum likelihood supertrees”. In: *Systematic biology* 57.2 (2008), pp. 243–250.

[SS+03] Charles Semple, Mike Steel, et al. *Phylogenetics*. Vol. 24. Oxford University Press on Demand, 2003.

[SS13] Andreas Sand and Mike Steel. "The standard lateral gene transfer model is statistically consistent for pectinate four-taxon trees". In: *Journal of theoretical biology* 335 (2013), pp. 295–298.

[SSJ03] Leo M Schouls, Corrie S Schot, and Jan A Jacobs. "Horizontal transfer of segments of the 16S rRNA genes between species of the Streptococcus anginosus group". In: *Journal of bacteriology* 185.24 (2003), pp. 7241–7246.

[Ste+13] Mike Steel et al. "Identifying a species tree subject to random lateral gene transfer". In: *Journal of theoretical biology* 322 (2013), pp. 81–93.

[Ste16] Mike Steel. *Phylogeny: Discrete and Random Processes in Evolution*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2016, p. 302.

[Ste94a] M. Steel. "Recovering a tree from the leaf colourations it generates under a Markov model". In: *Applied Mathematics Letters* 7.2 (1994), pp. 19–23.

[Ste94b] William J Stewart. *Introduction to the numerical solution of Markov chains*. Princeton University Press, 1994.

[SWZ19] Liangjun Su, Wuyi Wang, and Yichong Zhang. "Strong Consistency of Spectral Clustering for Stochastic Block Models". In: *arXiv:1710.06191 [stat]* (May 2019). arXiv: 1710.06191.

[Tav+86] Simon Tavaré et al. "Some probabilistic and statistical problems in the analysis of DNA sequences". In: *Lectures on mathematics in the life sciences* 17.2 (1986), pp. 57–86.

[Tho+16] Mary E. Thompson et al. "Using the bootstrap for statistical inference on random graphs". en. In: *Canadian Journal of Statistics* 44.1 (2016). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cjs.11271, pp. 3–24.

[TKF91] Jeffrey L Thorne, Hirohisa Kishino, and Joseph Felsenstein. "An evolutionary model for maximum likelihood alignment of DNA sequences". In: *Journal of Molecular Evolution* 33.2 (1991), pp. 114–124.

[TKF92] Jeffrey L Thorne, Hirohisa Kishino, and Joseph Felsenstein. "Inching toward reality: an improved likelihood model of sequence evolution". In: *Journal of molecular evolution* 34.1 (1992), pp. 3–16.

[Tro12] Joel A Tropp. "User-friendly tail bounds for sums of random matrices". In: *Foundations of computational mathematics* 12.4 (2012), pp. 389–434.

[TW94] Craig A. Tracy and Harold Widom. "Level-spacing distributions and the Airy kernel". en. In: *Communications in Mathematical Physics* 159.1 (Jan. 1994), pp. 151–174.

[Von07] Ulrike Von Luxburg. "A tutorial on spectral clustering". In: *Statistics and computing* 17.4 (2007), pp. 395–416.

[WB17] Y. X. Rachel Wang and Peter J. Bickel. "Likelihood-based model selection for stochastic block models". EN. In: *Annals of Statistics* 45.2 (Apr. 2017). Publisher: Institute of Mathematical Statistics, pp. 500–528.

[WC11] Jane Wiedenbeck and Frederick M Cohan. "Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches". In: *FEMS microbiology reviews* 35.5 (2011), pp. 957–976.

[YDN12] Yun Yu, James H Degnan, and Luay Nakhleh. "The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection". In: *PLoS Genet* 8.4 (2012), e1002660.

[YWS15] Yi Yu, Tengyao Wang, and Richard J Samworth. "A useful variant of the Davis–Kahan theorem for statisticians". In: *Biometrika* 102.2 (2015), pp. 315–323.

[ZA20] Zhixin Zhou and Arash A. Amini. "Optimal Bipartite Network Clustering". In: *Journal of Machine Learning Research* 21.40 (2020), pp. 1–68.

[ZG06] Mu Zhu and Ali Ghodsi. "Automatic dimensionality selection from the scree plot via the use of profile likelihood". In: *Computational Statistics & Data Analysis* 51.2 (2006), pp. 918–930.

[ZR18] Yilin Zhang and Karl Rohe. "Understanding Regularized Spectral Clustering via Graph Conductance". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18. Montréal, Canada: Curran Associates Inc., 2018, pp. 10654–10663.