The graph alignment problem: fundamental limits and efficient algorithms

Luca Ganassali

PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'École Normale Supérieure

The graph alignment problem: fundamental limits and efficient algorithms

Alignement de graphes : limites fondamentales et algorithmes efficaces

Soutenue par

Luca GANASSALI

Le 23 septembre 2022

Ecole doctorale n°386

Sciences Mathématiques de Paris Centre

Spécialité

Mathématiques

Composition du jury :

Christophe GIRAUD Professeur, Université Paris-Saclay	*Président*
Jiaming XU Assistant Professor, Duke University	*Rapporteur*
Negar KIYAVASH Associate Professor, EPFL	*Rapporteur*
Jean-François DELMAS Professeur, École des Ponts ParisTech	*Examinateur*
Laurent MASSOULIÉ Directeur de recherche, Inria	*Directeur de thèse*
Marc LELARGE Directeur de recherche, Inria	*Co-directeur de thèse*

ENS | PSL

Inria

The graph alignment problem: fundamental limits and efficient algorithms

Luca Ganassali

Inria, Département d'Informatique de l'ENS
PSL Research University, Paris, France

PhD thesis under the supervision of Laurent Massoulié and Marc Lelarge

September 20, 2022

Abstract This thesis focuses on statistical inference in graphs (or matrices) in high dimension and studies the graph alignment problem which aims to recover a hidden underlying matching between the nodes of two correlated random graphs.

Similarly to many other inference problems in planted models, we are interested in understanding the fundamental information-theoretical limits as well as the computational hardness of graph alignment.

First, we study the Gaussian setting, when the graphs are complete and the signal lies on correlated Gaussian edges weights. We prove that the exact recovery task exhibits a sharp information-theoretic threshold, characterize it, and study a simple and natural spectral method for recovery, `EIG1`, which consists in aligning the leading eigenvectors of the adjacency matrices of the two graphs.

While most of the recent work on the subject was dedicated to recovering the hidden signal in dense graphs, we next explore graph alignment in the sparse regime, where the mean degrees are constant, not scaling with the graph size. In this particularly challenging setting, for sparse Erdős-Rényi graphs, only a fraction of the nodes can be correctly matched by any algorithm. Our second contribution is an information-theoretical result which characterizes a regime where even this partial alignment is impossible, and gives upper bounds on the reachable overlap between any estimator and the true planted matching.

We next propose an algorithm that performs partial alignment, `NTMA`, which is based on a measure of similarity – called the tree matching weight – between tree-like neighborhoods of the nodes in the graphs.

Under this local approach in the sparse regime, we are brought to study a related problem: correlation detection in random unlabeled trees. This hypothesis testing problem consists in testing whether two trees are correlated or independent. The tree matching weight yields a first method for this question as well; another contribution is to study an optimal test based on the likelihood ratio. In a correlated Galton-Watson model, which is well-known to be the local approximation of sparse Erdős-Rényi graphs, we characterize the regimes of performance of this test.

Finally, we come back to graph alignment and propose a message-passing algorithm, `MPAlign`, naturally inspired by the study of the related problem on trees. This message-passing algorithm is analyzed and provably recovers a fraction of the planted signal in some regimes of parameters.

Keywords: statistical inference, random graphs, graph alignment, correlation detection in trees, message-passing algorithms, machine learning, probability.

Résumé Cette thèse a pour contexte l'inférence statistique sur des graphes (ou matrices) en grande dimension et étudie le problème d'alignement de graphes, qui consiste à retrouver un appariement sous-jacent entre les sommets de deux graphes aléatoires correlés.

Comme pour de nombreux problèmes d'inférence dans des modèles dit plantés, nous nous intéressons aux limites fondamentales ainsi qu'à la difficulté computationnelle du problème.

Nous étudions tout d'abord un modèle Gaussien dans lequel les graphes sont complets et où les poids des arêtes matchées sous la correspondence sous-jacente sont correlées. Nous établissons un seuil informationnel pour la tâche d'alignement exact dans ce modèle, et étudions un algorithme spectral simple, `EIG1`, consistant à aligner les vecteurs propres dominants des matrices d'adjacence des deux graphes.

Tandis que la grande majorité des travaux récents sur le sujet est dédiée à l'alignement exact dans les graphes denses, nous explorons dans la suite de la thèse un régime dit creux dans lequel le degré moyen des sommets est constant, indépendant de la taille du graphe. Pour les graphes d'Erdős-Rényi, dans ce régime ou l'alignement est plus difficile, n'importe quel algorithme ne pourra aligner seulement qu'une fraction des noeuds : on parle d'alignement partiel. Nous démontrons un résultat informationnel caractérisant un régime dans lequel l'alignement partiel est impossible, et donnant une borne supérieure sur la fraction de noeuds du graphe qu'un estimateur peut espérer correctement aligner.

Nous proposons par la suite un algorithme, `NTMA`, pour l'alignement partiel dans le régime où les graphes sont creux, définissant une mesure de similarité – le *tree matching weight* – entre les voisinages arborescents des sommets des deux graphes.

En étudiant l'alignement de graphes creux d'un point de vue local, un autre problème associé apparaît : la détection de corrélation dans les arbres. Ce problème de test d'hypothèses, non étudié auparavant, consiste à décider si deux arbres aléatoires sont correlés ou indépendants. Le *tree matching weight* donne une première méthode pour résoudre ce problème; dans une autre contribution, nous étudions un test optimal pour cette tâche de détection, basé sur le rapport de vraisemblance. Dans un modèle d'arbres de Galton-Watson corrélés, qui sont bien connus pour être les approximations locales des graphes d'Erdős-Rényi creux, nous caractérisons le régime de performance de ce test.

L'étude de ce problème sur les arbres donne ainsi naturellement un algorithme de passage de messages en temps polynomial pour notre tâche initiale d'alignement de graphes : `MPAlign`. Nous prouvons que cette méthode retrouve une fraction du signal planté dans certains régimes de paramètres.

Mots-clés: inférence statistique, graphes aléatoires, alignement de graphes, détection de corrélation dans des arbres, algorithmes de message-passing, machine learning, probabilités.

Remerciements

À l'heure où s'achève l'écriture de ce manuscrit, je réalise que ces quelques mots seront sans doute les plus lus, les mieux compris. Il est heureux qu'il en soit ainsi ; ils disent plus que tout le reste.

Mes premiers remerciements vont à mes directeurs de thèse. Je mesure la chance d'avoir été encadré par Laurent : ces travaux n'auraient jamais pu voir le jour sans ses grandes qualités scientifiques et humaines, qui m'ont bercé tout au long de ces trois années. Merci à Laurent, donc, pour ces heures passées sans compter autour de discussions, de calculs, pour sa disponibilité à toute épreuve, pour l'humilité qui le pousse à partager généreusement ses vastes connaissances, à faire montre à l'abord d'un problème scientifique d'une savante sérénité, d'un flegme sans froideur ; autant de sources d'inspiration desquelles j'espère savoir un jour me rendre digne.

Un grand merci à Marc de m'avoir accompagné, d'avoir laissé toujours ouverte à nos discussions la porte de son bureau ; merci pour ses conseils, ses questions qui m'ont bien souvent aidé à y voir plus clair, ses orientations scientifiques. Merci aussi d'avoir su transmettre avec tant de sérieux son aspiration pour l'art si délicat de ne pas trop se prendre au sérieux.

Je remercie Negar Kiyavash et Jiaming Xu d'avoir accepté d'être rapporteurs de cette thèse, ainsi que Jean-François Delmas et Christophe Giraud d'en constituer le jury. Merci également à Guilhem Semerjian avec qui j'ai eu l'honneur de travailler et dont j'apprends énormément à chaque discussion.

J'ai une pensée pour tous ceux qui ont pu nourrir en moi l'envie de faire de la recherche et d'enseigner les mathématiques, professeurs, chercheurs dont j'ai eu la chance de croiser la route tout au long de ma vie d'élève et d'étudiant : Christian Dufournet au collège, Sébastien Ravassard au lycée, Serge Francinou en classes préparatoires, puis plus tard Nicolas Curien et Jean-François Le Gall. Un merci particulier à Christophe Giraud pour sa bienveillance, de Polytechnique à Orsay, son écoute, sa pédagogie, et ses bons conseils qui m'ont tourné vers Laurent.

Je tiens à présent à remercier tous ceux qui ont contribué à faire de ces trois ans une heureuse aventure.

Tout d'abord, merci à mes amis du bureau. Merci pour toutes ces discussions toujours enthousiastes, les semaines de workshop ou de summer school ensemble, les moments d'amitié dans les rires et dans les doutes. Difficile de ne pas évoquer les interminables parties de rigolade et la construction d'un langage souterrain dont il serait bien long de détailler les sinuosités. Bastien, Éric, Mathieu et Matthieu, en ouvrant chaque matin la porte d'un bureau inondé de vos larges sourires – souvent ponctués d'un allègre *Eccoloqua!* – me venait à l'esprit cette même pensée : je n'aurais pas pu rêver meilleure compagnie. Vous savez que vous serez toujours les bienvenus. Matthieu et Mathieu, notre séminaire d'équipe est désormais entre de bonnes mains. En tout cas, on se le souhaite.

Un très grand merci à Hélène, pour n'avoir cessé de verser du rose sur mes années de thèse. Merci pour ton dévouement et ton implication dans ton travail. Merci pour ta profonde douceur, ta gentillesse et ta bonne humeur. Merci d'avoir adopté Ginny.

Merci à Ginny, notre ourse de bureau, pour sa bonne humeur quelque peu indolente, ses étreintes et discours réconfortants – bien que plutôt laconiques.

Merci à Bertille d'être souvent descendue d'un étage pour contribuer à faire battre le cœur de notre bureau, pour les gâteaux bien sûr, et pour tout le reste. Merci à Antonin.

Une reconnaissance joyeuse revient aussi au groupe des doctorants, Amaury, Cédric, Claire, David, Ilia, Jakob, Lucas, Romain, et Thomas ainsi qu'aux autres membres de l'équipe, qui ont su eux aussi insuffler le doux climat si propice au bon déroulé d'une thèse.

Merci à Alexandre et à Marion, dont l'humeur primesautière illumine le troisième étage. Merci pour votre humour moqueur mais jamais caustique, votre présence et votre amitié, vos rires qui bourdonnent en essaim, les soirées passées ensemble et celles à venir.

Mes remerciements vont aussi à Alix, Christine, Julien, Antoine et Lucile pour tous les bons moments, au service RH, à l'Agos et à la comm' qui nous ont tant choyé, ainsi qu'à Ortenca, Sandra et Eric pour leur accessibilité, et pour nos interactions toujours naturelles et bienveillantes.

Merci à tous les compagnons de route, doctorants et docteurs, Victor, Maria, Meriem, Perrine, Alice, Etienne, Antoine et Lucile, rencontrés en chemin lors d'escapades sugitones ou oléronaises. Au carrefour de ces dernières, merci à Quentin avec qui, j'espère, les échanges continueront. Merci à mes aînés de thèse – antiques ou grands cousins – avec qui j'ai pu parler de science, et qui ont su parfois me partager leurs conseils : Léo, Simon, Thomas, Hadrien, Raphaël et Ludovic.

Quelques mots également pour celles et ceux que l'on voit presque tous les jours pendant trois ans, qui nous ouvrent les portes, les referment, nous donnent les clés, contribuent à rendre les locaux agréables pour travailler sereinement, avec qui on commence d'abord par échanger quelques mots hasardeux qui tissent parfois jusqu'à de longues discussions de fin de journée. Merci à Mariam, Anthony, Mamadou, Muriel et Emeline.

Mes pensées s'envolent aussi par-delà la thèse, loin de l'Inria, pour mes amis de longue date. Il serait bien long de faire ici la liste de tout ce qui me rend heureux de vous avoir près de moi, aussi resterai-je succint. Merci en particulier à Baptiste, Marie-Lou, Virginie, Marion, Thierry, PA, Feng, Laura, Madeline. Merci à Paul de naviguer à mes côtés au gré des flots, et pour tout l'umami de nos conversations. Merci à Thomas et à Pierre pour tout le temps passé ensemble et pour tout celui qu'il nous reste. Merci à tous les autres membres de la Pouloc, Marthe, Elisa, Héloïse, Georg, pour ouvrir toujours pour un soir ou pour un mois votre maison du bonheur, à moi et à tous les autres. Merci à Xavier.

Merci à celles et ceux qui m'ont accompagné sur mon terrain favori, celui de la musique ; merci à Claudine Boymond, Fanny Vicens, Sophia Vaillant, et Carlos de Castellarnau. Merci à l'aréopage de *Volveria*, à Éric et à *Tout Pour Le Tout*, bien sûr.

Merci à Stéphane, Christophe et Isabelle Ravel, pour leur accueil généreux, leur humanité, et pour le sage écho de leurs paroles qui résonne loin de nos montagnes lacustres.

Jorge, Valérie, Margaux, merci pour l'affection immense dont vous m'avez honoré. Votre maison et nos cœurs ont toujours été ouverts. Votre sollicitude et votre générosité n'ont jamais ployé. Merci infiniment.

Merci à ma famille. Merci à mes grands-parents. Merci à ma mère pour son indéfectible présence contre tous les rouleaux de la vie. Merci à mes parents pour m'avoir donné le goût de la curiosité. Merci à ma sœur Léna et à mes frères Léo, Nino et Sacha, pour cet amour

que nous savons ; celui contre lequel on ne peut rien.

Enfin, merci à Julie.

Acknowledgments

This work was partially supported by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR19-P3IA-0001 (PRAIRIE 3IA Institute).

Contents

Contributions and outline

Chapter 1. This opening chapter is a general introduction to the dissertation. We start with a general framework for inference on graphs, and go over basic concepts of random graph theory stating general results that will be useful throughout the thesis. We give a general overview of inference problems in random graphs, with several iconic examples, and introduce the phase transition phenomena arising in the high-dimensional regime. We next motivate and describe the graph alignment problem, discuss elementary results and give a survey of prior techniques, methods and theoretical work on the subject. We also introduce the problem of detecting correlation in trees.

Chapter 2. This chapter, based on the paper [Gan22] published at *MSML 2021*, investigates information-theoretic limits for exact alignment in the Gaussian setting, when the graphs are complete and the signal lies on correlated Gaussian edges weights. This model is often viewed as an interesting playground for graph alignment. We prove that the exact recovery task exhibits a sharp fundamental threshold, and characterize it.

Chapter 3. We then continue the exploration of the Gaussian setting, studying a simple and natural spectral method for recovery which consists in aligning the leading eigenvectors of the adjacency matrices of the two graphs. We give theoretical guarantees for this algorithm, showing a zero-one law property in terms of the signal-to-noise ratio for this method to work. This chapter is based on the paper [GLM22], published in *Advances in Probability*.

Chapter 4. We focus in this chapter on the study of Erdős-Rényi graph alignment in the sparse regime, where the mean degrees of the graphs are constant, not scaling with the number of nodes. Based on the paper [GML21b] published at *COLT 2021*, we prove an information-theoretical result characterizing a regime where even partial alignment is impossible, and giving upper bounds on the reachable overlap between any estimator and the planted matching. The proof builds upon building automorphisms of the intersection graph exchanging copies of small tree components.

Chapter 5. This chapter investigates an algorithm for sparse graph alignment, which relies on a measure of similarity – called the tree matching weight – between tree-like neighborhoods of the nodes in the graphs. We give theoretical guarantees for this method to work in the Erdős-Rényi model, and propose along the way a test to decide whether two trees are correlated or independent. This chapter is based on the paper [GM20], published at *COLT 2020*.

Chapter 6. Following the previous local approach in the sparse regime, we are interested in a related problem: correlation detection in random unlabeled trees. For this hypothesis testing problem, we study an optimal test based on the likelihood ratio. In a correlated Galton-Watson model, which is well-known to be the local approximation of sparse Erdős-Rényi graphs, we characterize regimes of performance of this test. Then, we come back to graph alignment and propose a message-passing algorithm naturally inspired by the study of the related problem on trees. This message-passing algorithm is analyzed and provably recovers a fraction of the planted signal in some regimes of parameters. The chapter is based on the paper [GML21a], which short version is published at *ITCS 2021*.

Chapter 7 (Addendum). A last chapter is appended to the dissertation, and presents recent results for correlation detection in trees. These results are significantly improving on previous work and give a general understanding of the fundamental limits of the problem, as well as some perspectives discussed afterwards in the conclusion.

Notations

Basics

$i, j, k, \ell, m...$	non negative integers, most of the time
$[m]$	set $\{1, \dots, m\}$ of integers from 1 to m
$\lvert\mathcal{X}\rvert$	cardinal of a finite set $\mathcal{X}$
$\mathcal{S}_m$	set of permutations on $[m]$ (we often identify $\mathcal{S}_k$ to $\mathcal{S}_\mathcal{X}$ whenever $\lvert\mathcal{X}\rvert = k$)
$\mathcal{S}(A, B)$	set of injective mappings between finite sets A and B
$\mathcal{S}(k, \ell)$	set of injective mappings from $[k]$ to $[\ell]$
π, σ	permutations, most of the time
Π, Σ	permutation matrices, most of the time
$O, o, \Omega, \omega, \Theta, \sim$	standard Landau notations
$\mathbb{1}_C$	indicator function at event C; $\mathbb{1}_C = 1$ if C is satisfied, 0 otherwise

Graphs

$G = (V, E)$	a graph G with vertex set V and edge set E
$\longleftrightarrow$ or $\underset{G}{\longleftrightarrow}$	connectivity in undirected graph G (eluded if no ambiguity)
$A(G)$	adjacency matrix of graph G, sometimes A if no amibiguity
n	number of nodes of a graph, most of the time
u, v	nodes of a graph, most of the time
T, t, τ	a tree, most of the time
d	depth of a tree, most of the time
$\mathcal{X}_d$	set of unlabeled finite trees of depth at most d
$\mathbb{T}_k$	set of unlabeled trees of size k

Probability

General convention: sometimes lowercase characters are used to distinguish deterministic objects from random variables (uppercase).

$\sim$	is distributed according to (sometimes also denotes asymptotic equivalence)
$\overset{(\mathsf{d})}{=}$	equality in distribution
$\mathsf{Ber}(p)$	Bernoulli distribution with parameter $p \in [0, 1]$
$\mathsf{Bin}(n, p)$	Binomial distribution with parameters $n \geq 0$, $p \in [0, 1]$
$\mathsf{Poi}(\lambda)$	Poisson distribution with parameter $\lambda > 0$
$\mathsf{Exp}(\mu)$	exponential distribution with parameter $\mu > 0$
$\mathcal{N}(\mu, v)$	Gaussian distribution with mean μ and variance v
GOE	Gaussian Orthogonal Ensemble
$\mathsf{Wig}(n, \xi)$, $\mathsf{Wig}'(n, \rho)$	Correlated Gaussian Wigner model (or a variant thereof) with size $n \times n$, noise parameter $\xi > 0$ or correlation $\rho \in [0, 1]$.
$\mathsf{G}(n, p)$	Erdős-Rényi model with n nodes and edge probability p
$\mathsf{G}(n, q, s)$	correlated Erdős-Rényi model with n nodes, edge probability q and correlation s
$\mathsf{SBM}(n, \alpha, P)$	stochastic block model with n nodes, community distribution α and edge probabilities P
$\mathsf{GW}_d^{(\mu)}$	Galton-Watson model with offspring distribution $\mathsf{Poi}(\mu)$ up to depth d
$\mathbb{P}_d^{(\lambda, s)}$	joint distribution of correlated Galton-Watson trees
$\mathbb{P}_d^{(\lambda)}$	joint distribution of independent Galton-Watson trees

Chapter 1

Introduction

1.1. Context

A myriad of datasets found in real life can be represented as graphs, which are nowadays becoming more and more useful to model complex systems. Facebook users can be viewed in a graph where each edge encodes a friendship relationship; Netflix as a graph between users and movies, each edge carrying rating or browsing data. Examples of the overwhelming presence of graphs can be found across applications in a variety of different fields: visualizing interaction between proteins in an organism, representing cities or destinations for route optimization, extracting a mesh from a 3D object, analyzing the spread of epidemics or fake news on Twitter, finding similar patterns in data, etc.

This thesis focuses on statistical inference in graphs, which consists in extracting relevant information from the observation of graph-shaped data. We are interested in understanding the fundamental aspects of these problems, as well as designing and analyzing algorithms for the considered tasks, seeking to characterize the regimes in which they may suceed.

Statistical inference In its broadest sense, statistical inference aims to draw meaningful conclusions based upon the observation of data. Suppose that we are given samples (in the form of measurements in $\mathbb{R}^d$, graphs, matrices, etc.) assumed to be drawn according to some probability distribution. The statistician designs methods to recover some information about the latter, e.g. testing hypotheses or deriving estimates for some parameter $\theta \in \Theta$ of the distribution. The usual framework is as follows:

$$\text{parameter } \theta \in \Theta \longrightarrow \text{data } Y \longrightarrow \text{estimator } \hat{\theta}$$

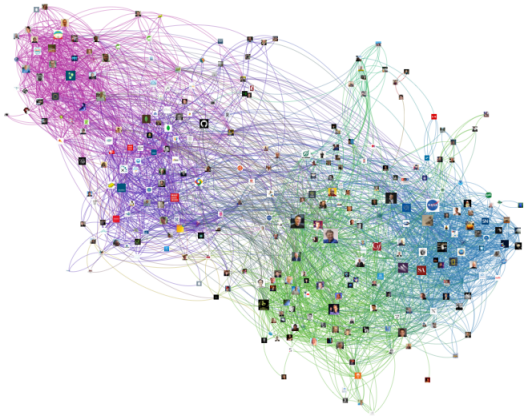
The main (informal) questions that arise in this setting are: 'how well can we discriminate between different models/hypotheses?', 'can we efficiently estimate the parameter θ?'

This thesis focuses on statistical inference in random graphs – or random matrices – in *high dimension*, when both the intrinsic dimension of the data and that of the parameter are large. This common assumption is particularly relevant for two main reasons: first, it fits well with real data, datasets being nowadays larger and larger; second, results in this asymptotic regime exhibit interesting and unexpected phenomena, see Section 1.2.3.

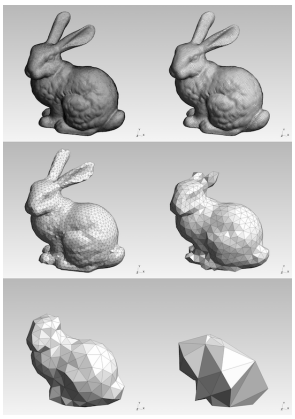
Planted models Some inference problems fall into the intuitive, conventional, and slightly different *planted framework*, where the observation is the result of a perturbation of some underlying *signal* of interest.

In this planted framework, the model – sometimes also referred to as *teacher-student model* in the machine learning and statistical physics communities – is as follows: some signal X is drawn according to some prior (we hence work in a *bayesian setting*), and 'planted' in

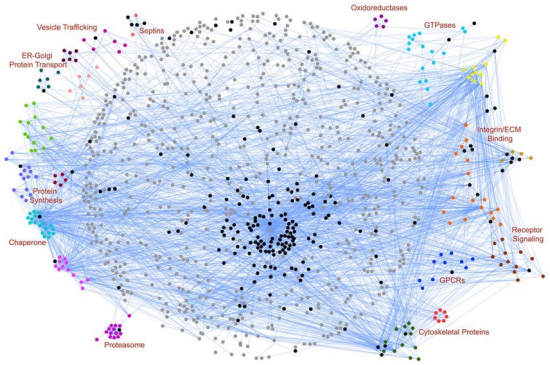
[1] sources: http://allthingsgraphed.com/2014/11/02/twitter-friends-network/ for (a), https://gmsh.info/ for (b), [QCM+09] for (c).



(a) – *a Twitter network with four communities*



(b) – *some 3D finite element meshes*



(c) – *a graph showing the platelet Protein-Protein Interaction network*

Figure 1.1 – *Some graphs*[1] *in real life.*

the data. Given the signal X, the observation Y is drawn according to some conditional distribution $p(\cdot \,|\, X)$. The framework is as follows:

$$\text{signal } X \sim p_X \longrightarrow \text{observation } Y \sim p(\cdot|X) \longrightarrow \text{estimator } \hat{X}$$

We refer to Section 1.2.2 for a closer look at planted models in the context of inference in random graphs.

Detection and reconstruction tasks The two informal questions stated earlier – testing hypotheses and estimating some parameter – can now be reformulated, and their counterparts are proper to the planted framework:

(i) Can we detect the presence of a planted signal in the data?

(ii) If yes, are we also able to recover the signal?

Let us more formally give a succinct mathematical formulation of questions (i) and (ii). Let us define n to be a generic dimension parameter; working in the high-dimensional regime here corresponds to making n tend to infinity.

Question (i) defines the *detection task*. Detecting the presence of signal exactly consists in discriminating a model with no planted signal – the *null model* – from the planted model, based on the observation of Y. In other words, detection task corresponds to the following hypothesis testing

$\mathcal{H}_0 :=$ "Y is drawn from the null model" versus $\mathcal{H}_1 :=$ "Y is drawn from the planted model"

Given a detection task, we are thus interested in designing a test $\mathcal{T}$ (i.e., a measurable function of Y taking values in $\{0,1\}$) for which we are able to give guarantees with probability tending to 1 asymptotically in n.

Question (*ii*) defines the *reconstruction* – or *recovery* – *task*. Recovering the signal now corresponds to designing an estimator $\hat{X}$ (i.e., a measurable function of Y) for which we are able to give guarantees (e.g. prove that $\hat{X}$ is somehow close to X) with probability tending to 1 asymptotically in n.

Let us pause for a moment after these broad definitions. Intuitively, the detection task is in general easier than the reconstruction task – even though some counter-examples can be found in [BMV+17] – and impossibility of detection almost always implies impossibility of reconstruction. Conversely, non-equivalence between detection and reconstruction (at least partial) may also seem rather counter-intuitive, when considering that the usual strategy for detecting some signal consists precisely in exhibiting the latter. However, these two problems are indeed definitely different; let us give hereafter an easy example of this fact.

A toy example (1/2): finding hay in a haystack Consider a bit sequence $(\xi_1, \dots, \xi_n)$ of length n made of i.i.d. entries taking the value 0 or 1 with equal probability. Let $k > 0$ that may depend on n.

Under the planted model, the observation $Y = (Y_1, \dots, Y_n)$ is generated as follows: we choose k positions $1 \leq i_1 < \dots < i_k \leq n$ uniformly at random, and set for all $i \in \{1, \dots, n\}$, $Y_i = 1$ if $i \in \{i_1, \dots, i_k\}$, and $Y_i = \xi_i$ otherwise. We denote $Y \sim \mathbb{P}_{1,k}$.

Under the null model, we simply set $Y_i = \xi_i$ for all $i \in \{1, \dots, n\}$, and denote $Y \sim \mathbb{P}_0$.

ξ	1	0	0	0	1	1	0	1	0	1	1	1	1	0	1	0	0	0
k positions	×		×	×						×			×	×			×	
Y	1	0	1	1	1	1	0	1	0	1	1	1	1	1	1	0	1	0

Figure 1.2 – *A realization with $n = 18$, $k = 7$. After transformation, in the sequence Y, the presence of a planted signal is highly probable; but where are the k extra ones?*

Detection. In this simple example the planted signal consists in extra ones somewhere in the data. It then easy to see that an optimal test for detection is simply based on counting occurrences of ones. Define $N_1(Y) := |\{i : Y_i = 1\}|$. Standard concentration inequalities (e.g. Hoeffding's inequality) straightaway give that with high probability – that is, with probability tending to 1 when $n \to \infty$:

$$N_1(Y) = \begin{cases} n/2 + \Theta(\sqrt{n}) & \text{under the null model } \mathbb{P}_0, \\ (n+k)/2 + \Theta(\sqrt{n-k}) & \text{under the planted model } \mathbb{P}_{1,k}. \end{cases}$$

Comparing these typical values shows that as soon as $k = \omega(\sqrt{n})$, extra ones can be detected, e.g. with a test $\mathcal{T}_n$ outputting 1 if and only if $N_1(Y)$ is greater than $n/2 + k/4$. Indeed, if $k = \omega(\sqrt{n})$, we will have $\mathbb{P}_0(\mathcal{T}_n = 0) \to 1$, $\mathbb{P}_{1,k}(\mathcal{T}_n = 1) \to 1$. Such a test is said to achieve *strong detection* (see Section 1.4.2). On the contrary, unreachability of strong detection when $k = O(\sqrt{n})$ is established e.g. by applying the central limit theorem – details are left to the reader.

Reconstruction. We are now in a position to understand why the two tasks are of different kind. Though it is rather simple to detect extra ones when $k = \omega(\sqrt{n})$, the reconstruction task would consist in recovering the exact positions of the extra ones. If one had no idea about the data, a naive – and somehow the worst – method would consist in choosing these k positions uniformly at random among the $N_1(Y)$ possibilities. It is easy to check that the number of positions that are correctly recovered – or, the *overlap* – with this method is of order k^2/n which is almost always very small compared to k even when $k = \omega(\sqrt{n})$.

A moment of thought shows that this naive method can never be outperformed. Indeed, the posterior distribution of the positions of extra ones is given by

$$\mathbb{P}_{1,k}(i_1,\dots,i_k \,|\, Y) = \frac{1}{\mathbb{P}_{1,k}(Y)} \mathbb{1}_{i_1<\dots<i_k} \mathbb{1}_{Y_{i_1}=\dots=Y_{i_k}=1} \binom{n}{k}^{-1} \left(\frac{1}{2}\right)^{n-k} .$$

The dependence on $i_1,\dots,i_k$ lying only in the terms $\mathbb{1}_{i_1<\dots<i_k}$ and $\mathbb{1}_{Y_{i_1}=\dots=Y_{i_k}=1}$, it is therefore the uniform distribution on the set of ordered lists of length k among the $N_1(Y)$ positions of ones.

In particular, if $k = \omega(\sqrt{n})$ and $k = o(n)$, then detection is easy but reconstruction is impossible, even partially, in the sense that no method can recover more than $o(k)$ of the planted extra ones. See Section 1.2.3 for the definition of a formalized context for these observations.

Organization of rest of the introduction

We start in Section 1.2 with some basics of random graph theory as well as general results and famous techniques that will be useful throughout this work. We then give a general overview on inference problems in random graphs through several widely studied examples, as well as the definition of the phase transition phenomena that crop up in these problems.

We introduce in Section 1.3 the graph alignment problem, lying at the very heart of this thesis and which various aspects will be discussed in the next chapters. We give insights on the motivations, discuss general related topics and give an overview of prior techniques and methods for this problem as well as theoretical guarantees, aside from our work.

We finally describe in Section 1.4 a related problem which will be the focus of Chapters 5, and mostly 6 and 7: correlation detection in random trees. This problem is interesting for its own sake, but has also a strong connection with graph alignment.

1.2. Inference on random graphs: a short tour

In this section, we will present the general framework of inference on randoms graphs, introducing some basic concepts and notations, and describing several – fundamental – examples for problems of this sort.

1.2.1. Basics of random graph theory

Graphs A (simple) *graph* $G = (V, E)$ is a discrete structure consisting in a vertex set V and an edge set E. Elements of V are called *vertices*, sometimes *nodes*.

In an *undirected* graph, E is a subset of $\binom{V}{2}$, the set of unordered pairs (or 2−sets) of distinct elements of V, and an edge e between nodes u and v is denoted by $\{u, v\}$. If the graph is *oriented*, then E contains ordered pairs (or 2−tuples) of elements of V, and they are denoted by (u, v).

All graphs considered throughout along this manuscript are finite (namely V and E are finite sets), and undirected, unless stated otherwise. If $u, v \in V$ are such that $\{u, v\} \in E$, we denote $u \underset{G}{\longleftrightarrow} v$ and $u \longleftrightarrow v$ when there is no ambiguity on the graph, and the vertices u and v are said to be *connected*, or *neighbors* in G.

Adjacency matrix, weighted graphs A graph $G = (V, E)$ with node set $V = [n]$ is often represented through its *adjacency matrix* $A = A(G) \in \mathbb{R}^{n\times n}$ defined as follows:

$$\forall\, u, v \in [n],\ A_{u,v} = \mathbb{1}_{\{u,v\}\in E}\,.$$

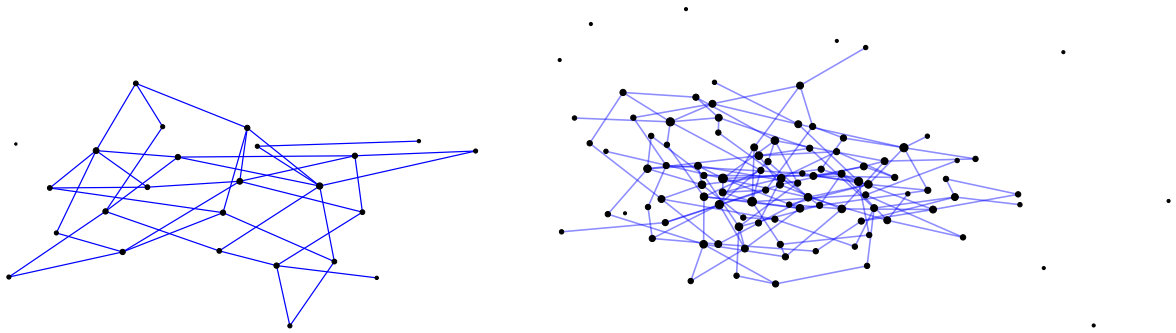
An undirected *weighted graph* $G = (V, E)$ is a graph with additional information on edges,

namely

$$\forall\, u, v \in [n],\ A_{u,v} = \mathbb{1}_{\{u,v\}\in E} W_{u,v}\,,$$

where the variables $W_{u,v} \in \mathbb{R}$ are *edge weights*.

The Erdős-Rényi model A simple, greatly celebrated, and widely used model of random graphs is the Erdős-Rényi model, introduced by Paul Erdős and Alfréd Rényi in 1959 [ER59]. In this model, denoted by $\mathsf{G}(n,p)$, the graph G has node set $V = [n]$ and each pair $\{u,v\}$ for $u \neq v \in [n]$ is present in E independently with probability p.



(a) – $n = 25, p = 0.14$ (b) – $n = 100, p = 0.03$

Figure 1.3 – *Some realizations of* $\mathsf{G}(n,p)$.

Note that the Erdős-Rényi model is in some sense the simplest model of random graphs one can ever think of: edges are drawn independently with the same probability, and there is no particular geometry in the graph. An immediate corollary of this absence of geometry is the following

Lemma 1.2.1. *Fix $n \geq 1$, $p \in (0,1)$ and $0 \leq m \leq \binom{n}{2}$. Let $G \sim \mathsf{G}(n,p)$, conditioned to have m edges. Then G is uniform among all graphs with node set $[n]$ and m edges.*

Many interesting results with high probability are known for Erdős-Rényi graphs, and literature investigating this model is abundant. For a general and thorough view on this very rich topic, the reader can refer to Bollobás [Bol01], Janson, Luczak and Rucinski [JLR00] and Hofstadt [Hof16].

High probability properties, first and second moment methods Some event A depending on a size (or dimension) parameter n is said to be verified *with high probability (w.h.p.)* if the probability of A tends to 1 when $n \to \infty$.

We will start with merely giving one of the most elementary – and famous – results for the Erdős-Rényi model, which proof will be the occasion to introduce the *first* and *second-moment* methods (see e.g. [AS16]) that are instrumental for solving many probabilistic questions in random graphs. Let us introduce them hereafter.

Lemma 1.2.2 (First moment method)**.** *Let X be a non-negative, integer-valued random variable. Then*

$$\mathbb{P}(X > 0) \leq \mathbb{E}[X].$$

Proof. This is a consequence of Markov's inequality : for all $b > 0$, $\mathbb{P}(X \geq b) \leq \frac{\mathbb{E}[X]}{b}$. Taking $b = 1$ gives the desired result. □

In particular, in the case where X depends on n, and $\mathbb{E}[X] \to 0$ when $n \to \infty$, then Lemma 1.2.2 implies that $X = 0$ with high probability.

Lemma 1.2.3 (Second moment method[2]). *Let X be a real random variable with positive mean and finite variance. Then for all $0 \leq c \leq 1$,*

$$\mathbb{P}\left(X \geq c\,\mathbb{E}\left[X\right]\right) \geq (1-c)^2 \frac{\mathbb{E}\left[X\right]^2}{\mathbb{E}\left[X^2\right]}.$$

Proof. Using Cauchy-Schwarz inequality,

$$\mathbb{E}[X] = \mathbb{E}\left[X\mathbb{1}_{X<c\,\mathbb{E}[X]}\right] + \mathbb{E}\left[X\mathbb{1}_{X\geq c\,\mathbb{E}[X]}\right] \leq c\,\mathbb{E}\left[X\right] + \mathbb{E}[X^2]^{1/2}\,\mathbb{P}(X \geq c\,\mathbb{E}\left[X\right])^{1/2},$$

which gives $\mathbb{E}[X]^2\,(1-c)^2 \leq \mathbb{E}[X^2]\,\mathbb{P}(X \geq c\,\mathbb{E}\left[X\right])$. □

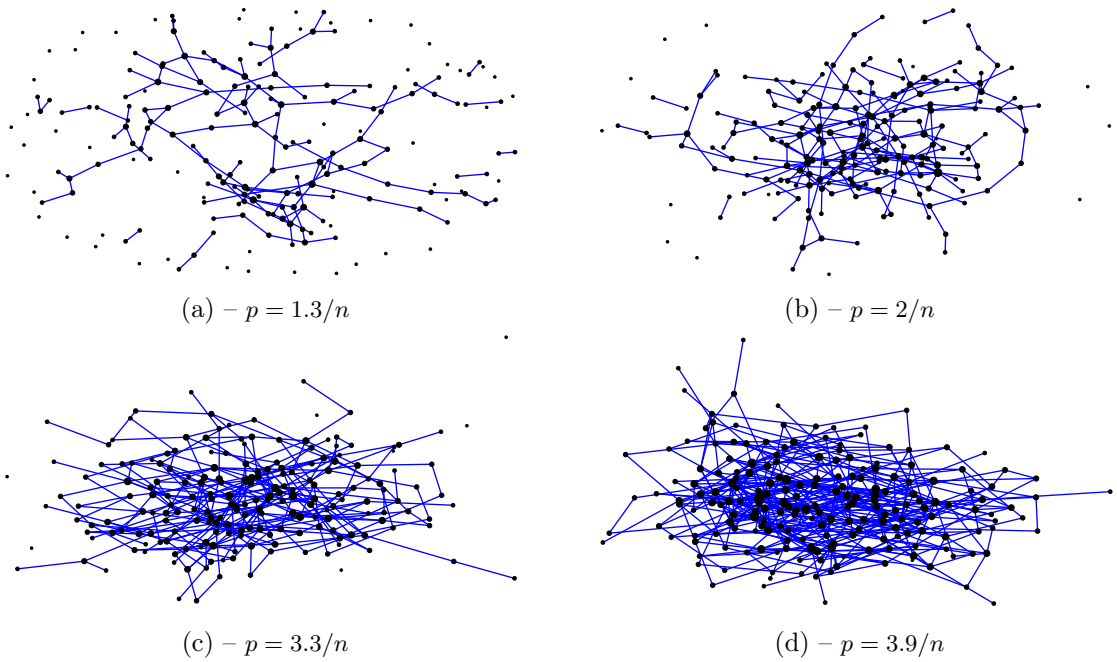
In particular, in the case where X depends on n, $\mathbb{E}\left[X\right] \to \infty$ and $\mathbb{E}\left[X^2\right] \sim \mathbb{E}\left[X\right]^2$ when $n \to \infty$, taking $c \to 0$ in Lemma 1.2.3 implies that $X \geq o(\mathbb{E}\left[X\right])$ with high probability and hence that $X \to \infty$ w.h.p.

Let us now state an elementary result that will be proved by appealing to these standard methods. A graph G is *connected* if for any $u \neq v \in G$, there exists a path from u to v made of edges of G. A node $u \in V$ is *isolated* if it has no neighbors in G.

Theorem 1.1 (Connectivity of Erdős-Rényi graphs). *Let $G \sim \mathsf{G}(n,p)$ with p depending on n. Then, with high probability,*

(*i*) *if $np \leq (1-\varepsilon)\log n$ for some $\varepsilon > 0$, then G contains isolated vertices and hence is not connected.*

(*ii*) *if $np \geq (1+\varepsilon)\log n$ for some $\varepsilon > 0$, then G is connected.*



(a) – $p = 1.3/n$ (b) – $p = 2/n$

(c) – $p = 3.3/n$ (d) – $p = 3.9/n$

Figure 1.4 – *Some realizations of $\mathsf{G}(n,p)$ with $n = 200$, connected and disconnected.*

[2]In this form, this result is known as the *Payley-Sigmund inequality.*

Proof. ***Proof of*** (i)**.** We will use the second moment method for the proof of point (i). Let us denote

$$X := |\{u \in V, u \text{ is isolated in } G\}| = \sum_{u \in V} \mathbb{1}_{u \text{ is isolated in } G}.$$

For any $u \in V$, $\mathbb{P}(u \text{ is isolated in } G) = (1-p)^{n-1}$, hence $\mathbb{E}[X] = n(1-p)^{n-1}$ which equals $\exp((1+o(1))[\log n - np]) \geq \exp((1+o(1))\varepsilon \log n) \to \infty$ under the assumption $np \leq (1-\varepsilon)\log n$.

Let us now check that we indeed have $\mathbb{E}\left[X^2\right] \sim \mathbb{E}\left[X\right]^2$ when $n \to \infty$.

$$\begin{aligned}
\mathbb{E}[X^2] &= \sum_{u,v \in V} \mathbb{E}\left[\mathbb{1}_{u,v \text{ are isolated in } G}\right] \\
&= \mathbb{E}[X] + \sum_{u,v \in V, u \neq v} \mathbb{P}\left(u, v \text{ are isolated in } G\right) \\
&= o(\mathbb{E}[X]^2) + n(n-1)(1-p)^{n-1+n-2} = (1+o(1))\mathbb{E}\left[X\right]^2.
\end{aligned}$$

Proof of (ii)**.** Point (ii) is proved with the first moment method. Let us assume that $np \geq (1+\varepsilon)\log n$. Define a *zero-cut* of G to be a partition of V into two sets which are crossed by no edges. It is clear that G is disconnected if and only if G admits a non trivial zero-cut, the trivial one being the partition $\{V, \varnothing\}$. Let Y be defined as the number of non trivial zero-cuts of G. For a given partition $\{S, V \setminus S\}$ of V into two sets of size k and $n-k$, $\{S, V \setminus S\}$ is a zero-cut with probability $(1-p)^{k(n-k)}$, hence

$$\mathbb{E}[Y] = \sum_{k=1}^{\lfloor n/2 \rfloor} \binom{n}{k} (1-p)^{k(n-k)}.$$

The right hand term being decreasing with p, we can assume without loss of generality that $p = (1+\varepsilon)\frac{\log n}{n}$. Using $\binom{n}{k} \leq (en/k)^k$, and splitting the last sum at $k = \alpha n$, where α is to be specified later, we get the following upper bound:

$$\begin{aligned}
\mathbb{E}[Y] &\leq \sum_{k=1}^{\alpha n} \exp\left(k\left[\log(en/k) - (1-\alpha)(1+\varepsilon)\log n\right]\right) + \sum_{k=\alpha n+1}^{\lfloor n/2 \rfloor} \binom{n}{k} (1-p)^{\alpha n^2/2} \\
&\overset{(a)}{\leq} \sum_{k=1}^{\alpha n} \exp\left(-(c+o(1))k \log n\right) + 2^n e^{-\alpha(1+\varepsilon)n\log n} = o(1),
\end{aligned}$$

where (a) holds as soon as $\alpha \in (0,1)$ and c are such that $(1-\alpha)(1+\varepsilon) - 1 > c > 0$, which is true whenever $\alpha \in (0, \varepsilon/(1+\varepsilon))$. □

Remark 1.2.1. *Note that result of Theorem 1.1 shows that the asymptotic probability of connectivity in an Erdős-Rényi graph abruptly jumps from* 0 *to* 1 *when* $np/(\log n)$ *begins to exceed* 1*: in this case we say that the connectivity property exhibits a* (sharp) threshold. *This remarkable fact occurs for a large range of properties in random graphs (see e.g. [Bol01, JLR00]). In inference problems, such underlying threshold phenomena involving parameters of the random models are often the cause of the emergence of so-called* phases *(impossible, hard or easy), see Section 1.2.3.*

1.2.2. The zoo of inference problems on graphs

Why planting signal? At first sight, the planted framework described earlier may leave the reader somewhat bemused; the question of finding some interesting information in data – in the broadest meaning – is very different from assuming that the data is *literally* constructed out of some underlying signal, and aiming to recover it. We would like to start by emphasizing

and elaborating on this important point.

Here are few words to clarify the above statement and release this apparent tension: for the overwhelming majority of inference problems in random graphs, the planted formulation is in fact a probabilistic rephrasing of an original deterministic combinatorial optimization problem, which we often refer to as the *worst-case* version. The planted approach differs from the initial problem, but has the advantage of carrying it its very essence a notion of *ground truth*, offering a comfortable framework for the evaluation of the performance of algorithms as well as a direct control on the signal-to-noise ratio. Also, theoretical guarantees can be obtained with high probability in planted models under less stringent constraints, taking into account the typical properties of data sampled from the generative model. Cris Moore echoes these statements in [Moo17], justifying this approach in the context of community detection in the following words:

> *"For the most part we are used to thinking about worst-case instances rather than random ones, since we want algorithms that are guaranteed to work on any instance. But why should we expect a community detection algorithm to work, or care about its results, unless there really are communities in the first place? And when Nature adds noise to a data set, isn't it fair to assume that this noise is random, rather than diabolically designed by an adversary?"*

This duality is believed to be fundamental and should be kept in mind when facing inference problems (on graphs). We will endeavour to shed light on the two flavours of the problems given as examples hereafter: a worst-case – deterministic – formulation, as well as a planted – probabilistic – counterpart, leading to different objectives and results.

A non-exhaustive bestiary Without further ado, we will now give three representative examples of inference problems on graphs.

(*a*) *Max-clique, planted clique.* A *clique* of a graph is a subset of vertices all adjacent to each other, i.e. a complete subgraph. The *max-clique* problem consists in finding the maximum clique in a graph $G = (V, E)$, that is solving the following

$$\underset{\substack{S \subset V \\ S \text{ is a clique}}}{\arg\max} \quad |S| \,. \tag{1.1}$$

The max-clique problem, as well as the problem of deciding whether the graph contains a clique of given size is NP-hard [Kar72], as well as some of its approximations [Hå99], unless $\mathsf{P} = \mathsf{NP}$.

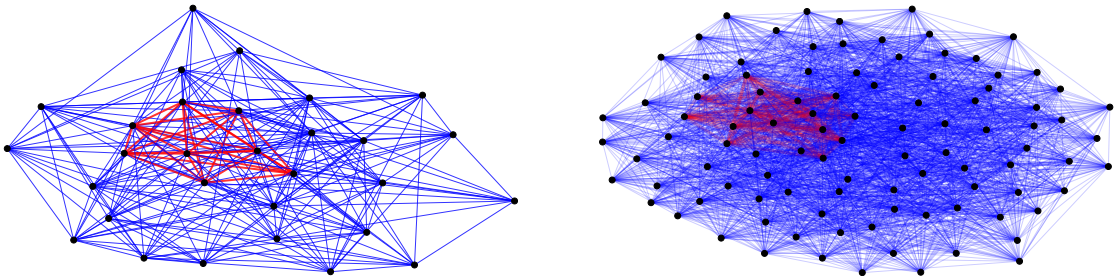
The planted version of max-clique, namely the *planted clique* problem, is defined as follows. Consider two integers n and $k \leq n$ possibly scaling with n. Let us generate a graph $G = (V, E)$ with vertex set $V = [n]$ as follows. First, a subset $K^\star \subset V$ of size k is chosen uniformly among the $k-$subsets of V. $K^\star$ will form a clique: all possible edges between vertices of $K^\star$ are added to E. Then, all possible remaining edges are drawn independently with probability $1/2$. We denote this model by $\mathsf{G}_k(n, 1/2)$.

Note that when $k = 0$, that is in the absence of a planted clique, G is simply distributed according to $\mathsf{G}(n, 1/2)$. Given k, the statistical test for the detection task is thus

$$\mathcal{H}_0 := \text{“}G \sim \mathsf{G}(n, 1/2)\text{”} \text{ versus } \text{“}\mathcal{H}_1 := G \sim \mathsf{G}_k(n, 1/2)\text{”} \,.$$

For reconstruction, the goal is to find an estimator $\hat{K} = \hat{K}(G)$ that recovers the whole true clique $K^\star$ with high probability, that is such that $\mathbb{P}(\hat{K}(G) = K^\star) \to 1$ when $n \to \infty$ and $G \sim \mathsf{G}_k(n, 1/2)$.

Note that under the planted model, the posterior distribution of $K^\star$ (conditionally on G)



(a) – $n = 30, k = 8$ (b) – $n = 100, k = 15$

Figure 1.5 – *Some realizations of* $\mathsf{G}_k(n, 1/2)$*, where the planted clique is highlighted in red*

is given by

$$\mathbb{P}\left(K^\star = K \,|\, G\right) = \frac{1}{\mathbb{P}(G)} \left(\frac{1}{2}\right)^{n-k} \mathbb{1}_{K \text{ is a clique in } G}\,,$$

which shows that, unsurprisingly, the posterior distribution of $K^\star$ is the uniform distribution among all cliques of size k in G. In the case where $k \geq (2+\varepsilon)\log_2(n)$ it can be shown [Mat72] that the only clique of size k in G is the planted one, and that it is the maximal clique, with high probability. Hence w.h.p. in this regime, the maximum a posteriori estimator of $K^\star$ is *precisely* the solution of the max-clique problem (1.1) in G. This last statement makes the worst-case/planted duality even more explicit. For more insights on the performance of simple algorithms for this problem, we refer to the lecture notes by Wu and Xu [WX19].

(*b*) *Min-bisection, community detection.* A *bisection* of a graph G is a partition of the vertex set V into two sets of equal size (we assume that $|V|$ is even). In a weighted graph $G = (V, E)$ with adjacency matrix A, the *min-bisection* problem corresponds to finding a bisection with minimal crossing edge weights, namely

$$\underset{(V_1,V_2) \text{ bisection of } G}{\arg\min} \sum_{u \in V_1,\, v \in V_2} A_{u,v}\,. \tag{1.2}$$

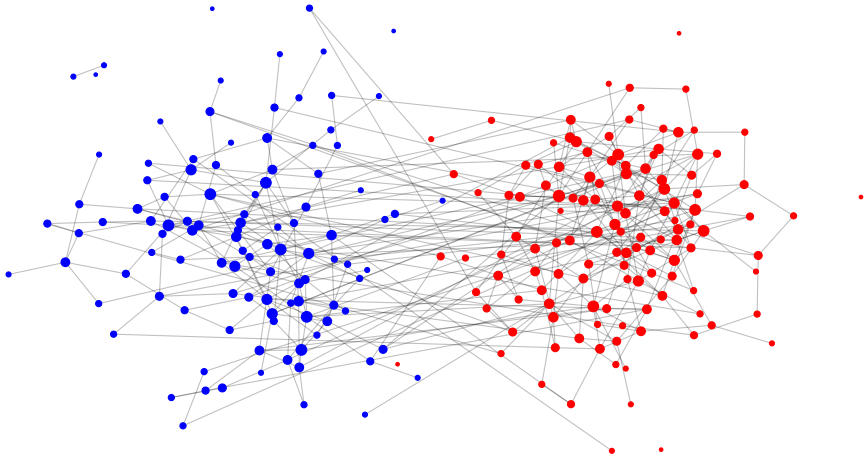
Finding the min-bisection of a graph is known to be NP-hard [GJS74]. In the planted version of min-bisection, the random graph $G = (V, E)$ has to satisfy the following property: there is an underlying optimal partition of V consisting in two subsets that are referred to as *communities*. Therefore, the problem amounts to recovering these communities, which can very well be more than two in a general setting. In order for the graph to satisfy this property, it is sampled according to the celebrated *stochastic block model*, originally introduced in [HLL83], widely studied in recent threads of research [DKMZ11, Mas14, MNS15, MNS18, Abb18].

For a number of nodes $n \geq 0$, a number of blocks $r \geq 1$, a distribution $\alpha = (\alpha_i)_{i\in[r]}$ on $[r]$ and a $r \times r$ symmetric matrix P with non-negative entries, the stochastic block model $\mathsf{SBM}(n, \alpha, P)$ is defined as follows. First, draw independently for every node $u \in V = [n]$ a community (or type) $\chi^\star(u) \sim \alpha$. Then, every edge $\{u, v\}$ for $u \neq v \in V$ is present independently with probability $P_{\chi^\star(u),\chi^\star(v)}$.

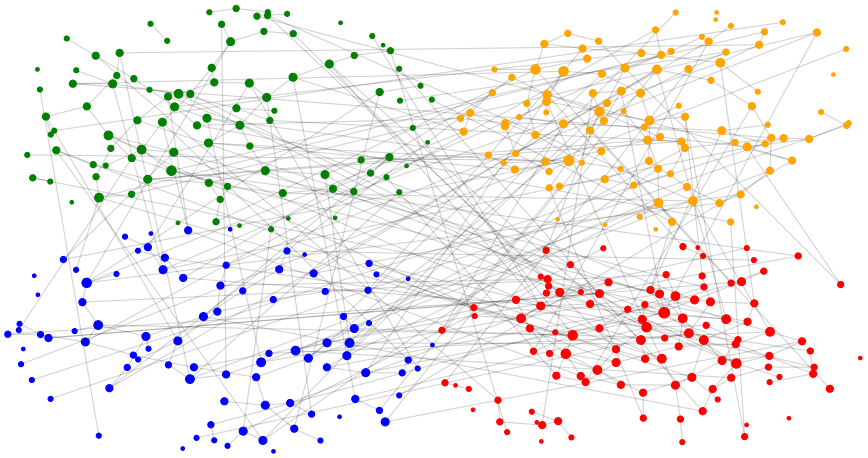
Note that when $r = 1$ (single community), $G \sim \mathsf{G}(n, p)$ with $p := P_{11}$. Given n and P, detection of planted communities consists in testing

$$\mathcal{H}_0 := \text{“}G \sim \mathsf{G}(n, p) \text{ for some } p \in (0, 1)\text{”} \text{ versus } \mathcal{H}_1 := \text{“}G \sim \mathsf{SBM}(n, \alpha, P)\text{”}\,.$$

For reconstruction, we assert the performance of an estimator $\hat{\chi} = \hat{\chi}(G) : V \to [r]$ of the communities through its rescaled *overlap* with the ground truth $\chi^\star$, defined by



(a) – $n = 230, r = 2, P_{c,c'} = 3.2 \cdot \mathbb{1}_{c=c'} + 0.7 \cdot \mathbb{1}_{c \neq c'}$



(b) – $n = 400, r = 4, P_{c,c'} = 3.3 \cdot \mathbb{1}_{c=c'} + 0.5 \cdot \mathbb{1}_{c \neq c'}$

Figure 1.6 – *Some realizations of* $\mathsf{SBM}(n, \alpha, P)$. *For both cases,* α *is the uniform distribution among the* r *communities, and nodes are colored and placed accordingly.*

$$\mathrm{ov}(\hat{\chi}, \chi^\star) := \frac{1}{n} \max_{\sigma \in \mathcal{S}_r} \sum_{u \in V} \mathbb{1}_{\sigma \circ \hat{\chi}(u) = \chi^\star(u)} - \sum_{i \in [r]} \alpha_i^2 \,.$$

The second term in the right-hand side in the above ensures that $\mathrm{ov}(\hat{\chi}, \chi^\star) > 0$ implies that the estimator $\hat{\chi}$ strictly outperforms random guess. Indeed, $\sum_{i \in [r]} \alpha_i^2$ is the expected fraction of good predictions achieved by the random guess estimator outputting communities drawn under the prior distribution α.

The connection between the maximum a posteriori (MAP) estimator and the min-bisection problem can be illustrated in the standard case of two symmetric communities, in the sparse regime with $\alpha = (1/2, 1/2)$ and $P = \begin{pmatrix} a/n & b/n \\ b/n & a/n \end{pmatrix}$ in the *assortative* setting where $0 < b < a$. In this case, denoting by $S^\star := \{u \in V, \chi^\star(u) = 1\}$, the posterior distribution of $S^\star$ under the stochastic block model writes

$$\mathbb{P}\left(S^\star = S \,|\, G\right) \propto \exp\left(\log\left(\frac{b/n}{a/n}\right) \sum_{u \in S, v \in V \setminus S} A_{u,v} + \log\left(\frac{1 - b/n}{1 - a/n}\right) \sum_{u \in S, v \in V \setminus S} (1 - A_{u,v}) \right).$$

where $\propto$ stands for proportionality up to terms that do not depend on S. Neglecting the effect of non-edges, which is fair in the sparse regime (see [Moo17]), since $0 < b/a < 1$ by assumption, and since $(S^\star, V \setminus S^\star)$ is w.h.p. close to a bisection of G, heuristically the MAP estimator of the two communities $(S^\star, V \setminus S^\star)$ is well approximated by the solution to the min-bisection problem (1.2) in graph G. For an excellent survey on the subject with a statistical physics approach, we refer to [Moo17].

(*c*) *Min-weight perfect matching, planted matching.* Let $G = (V, E)$ be a graph with $|V| = 2n$ and adjacency matrix A. Assume that there is a partition $\{V_0, V_1\}$ of V with $|V_0| = |V_1| = n$ such that every edge $\{u, v\} \in E$ satisfies $u \in V_0$ and $v \in V_1$ (we say that G is *bipartite*). A *perfect matching* (p.m. hereafter) of G is a set $M := \{e_1, \dots, e_n\}$ of n edges of E such that each node $u \in V$ appears exactly once in M. The *weight* of a matching M is defined by

$$\mathrm{weight}(M) := \sum_{e=\{u,v\} \in M} A_{u,v}$$

The *min-weight perfect matching* problem writes

$$\operatorname*{arg\,min}_{M \text{ p.m. of } G} \mathrm{weight}(M)\,. \tag{1.3}$$

Unlike the first two examples (*a*) and (*b*), the min-weight perfect matching problem is an instance of the Linear Assignment Problem (LAP) and can be solved in polynomial-time, e.g. by the Hungarian algorithm [Kuh55] which runs in $O(n^3)$ time. In the *planted matching* problem [SSZ20, DWXY21, MMX21], the graph G is taken to be a subgraph of a complete bipartite graph $K_{n,n}$, namely $V = [2n]$ and $E \subseteq \{\{u, v\},\ 1 \le u \le n,\ n+1 \le v \le 2n\}$. A planted matching $M^\star$ is first picked uniformly at random from the set of perfect matchings of $K_{n,n}$. The remaining possible $n^2 - n = n(n-1)$ edges are then sampled independently with probability p. Then, edge weights are drawn independently for all $e \in E$ from some distribution $\mathcal{P}$ if $e \in M^\star$ and from another distribution $\mathcal{Q}$ otherwise.

Reconstructing the planted matching refers to finding an estimator $\widehat{M} = \widehat{M}(G)$ such that the overlap $\frac{1}{n}\left|\widehat{M} \cap M^\star\right|$ is as large as possible. Let us here again derive the posterior distribution of the signal in the planted matching model. Denoting by $\propto$ proportionality up to terms that do not depend on M, we have

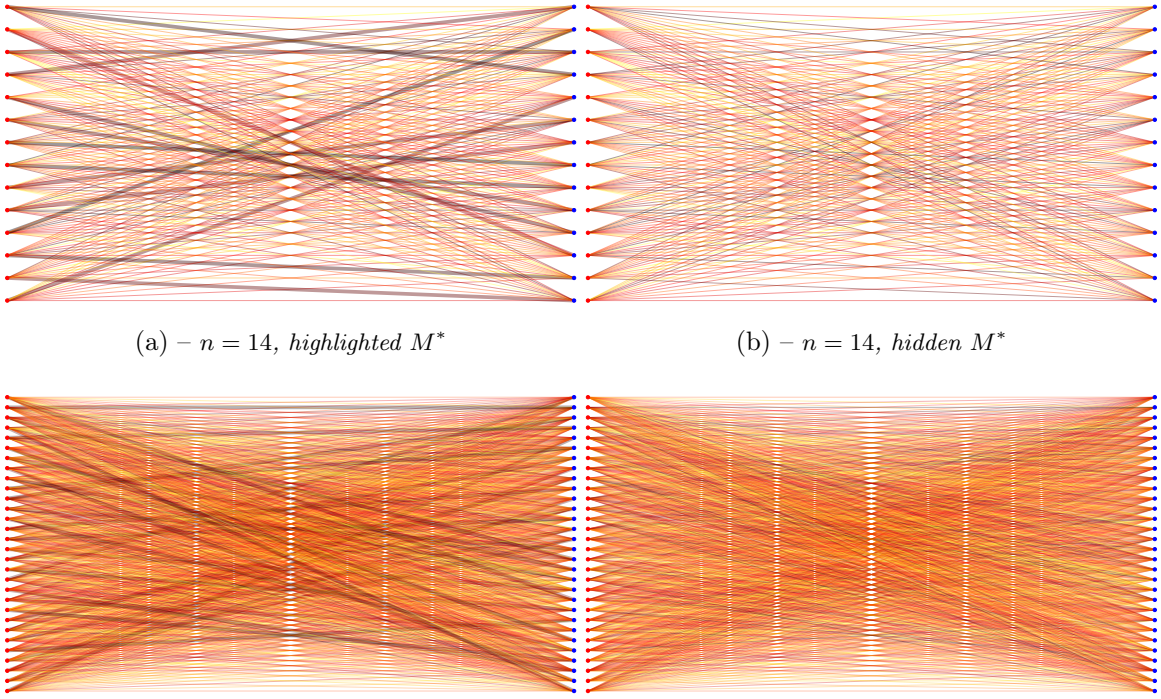
$$\begin{aligned}
\mathbb{P}\left(M^\star = M \,|\, G\right) &\propto \exp\left(\sum_{e=\{u,v\} \in M} \log \mathcal{P}(A_{u,v}) + \sum_{e=\{u,v\} \in E \setminus M} \log \mathcal{Q}(A_{u,v})\right) \mathbb{1}_{M \text{ is a p.m. of } G} \\
&\propto \exp\left(-\sum_{e=\{u,v\} \in M} \log \frac{\mathcal{Q}}{\mathcal{P}}(A_{u,v})\right) \mathbb{1}_{M \text{ is a p.m. of } G}.
\end{aligned}$$

This gives immediately that once again, the MAP estimator of $M^\star$ is precisely the solution to the worst-case instance (1.3) on the reweighted graph $\widetilde{G}$ such that for each edge $e = \{u, v\}$,

$$A(\widetilde{G})_{u,v} := \log \frac{\mathcal{Q}}{\mathcal{P}}(A(G)_{u,v}).$$

Note that in particular, if $\mathcal{P} = \mathsf{Exp}(\mu_P)$ and $\mathcal{Q} = \mathsf{Exp}(\mu_Q)$ with $\mu_Q < \mu_P$ – which is the model considered in [SSZ20, DWXY21, MMX21], see Figure 1.7 – then we exactly have $\log \frac{\mathcal{Q}}{\mathcal{P}}(A_{u,v}) = c + (\mu_P - \mu_Q) A_{u,v}$ with some constant c, and hence the MAP estimator is exactly the solution to the min-weight perfect matching problem (1.3) directly on G.

We close this short glimpse on the bestiary by mentioning other planted structures in



(a) – $n = 14$, *highlighted* M^* (b) – $n = 14$, *hidden* M^*

(c) – $n = 30$, *highlighted* M^* (d) – $n = 30$, *hidden* M^*

Figure 1.7 – *Some realizations of planted matchings on the complete bipartite graph* $K_{n,n}$ *(*$p = 1$*). For both cases,* $\mathcal{P} = \mathsf{Exp}(\mu_P)$*,* $\mathcal{Q} = \mathsf{Exp}(\mu_Q)$ *with* $\mu_P = 4.2$ *and* $\mu_Q = 1/n$*, and edges are colored according to their weights.*

graphs which have been recently studied, such as trees [MST19], colorings [DF16], or hamiltonian paths [BDT$^+$20].

After having given these classical examples, we are now ready to elaborate about some asymptotic (high-dimensional) phenomena that arise in these inference problems, namely the emergence of some regimes of model parameters scaling with the dimension n, where the task – reconstruction or detection – turns out to be impossible, hard or easy.

1.2.3. Impossible, hard and easy phases

Definitions Let us consider an inference task (e.g. detection, reconstruction) in a planted model where the data – not necessarily graphs – is sampled from a parametric distribution with parameters $\theta \in \Theta$.

- The *impossible phase* (or *impossible regime*) is defined as a subset $\Theta_{\text{impossible}}$ of the set of parameters Θ such that for all $\theta \in \Theta_{\text{impossible}}$, provably no algorithm can perform the task with high probability.
- The *easy phase* (or *easy regime*) is the regime of parameters Θ_{easy} where the task can provably be solved *by a polynomial-time algorithm*, with high probability.
- The *hard phase* (or *hard regime*) is the regime Θ_{hard} where some exhaustive, non-polynomial search provably works but where no polynomial-time is known to succeed with high probability.

The above definitions imply that $\{\Theta_{\text{impossible}}, \Theta_{\text{easy}}, \Theta_{\text{hard}}\}$ is a partition of Θ. The understanding – and the pinning down – of these three phases, gathered in a so-called *phase diagram*, is of course of paramount importance for the understanding of inference problems,

their related algorithms, and has thus been the subject of many recent threads of research (see e.g. [BBH18] for a unified view).

A toy example (2/2): finding hay in a haystack The remarks made earlier in our simple toy example (see 'A toy example (1/2)') can be specified to fit this context. For the (strong) detection task, the impossible phase covers the regime where $k = O(\sqrt{n})$. When $k = \omega(\sqrt{n})$, counting the occurrences of ones takes $O(n)$ time and enables to detect the presence of signal with high probability: the task is easy.

The (partial) reconstruction task has however a much larger impossible phase ($k = o(n)$). To complete the phase diagram, let us mention that in the – not so interesting – case where $k = \Theta(n)$, the optimal method for partial recovery still consists in choosing k positions at random among the positions of ones, which runs in polynomial time.

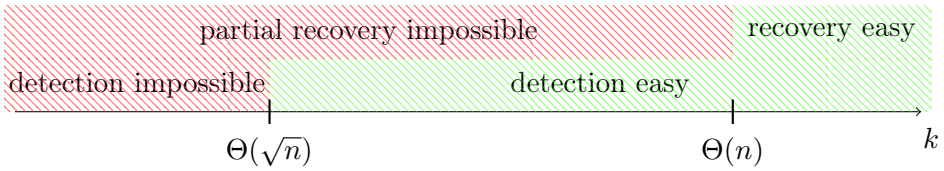


Figure 1.8 – *Phase diagram for detection and reconstruction in the 'find hay in a haystack' problem*

Note that in this (very simple) case, there is no hard phase neither for detection nor reconstruction. However, a considerable variety of inference problems are conjectured to exhibit a hard phase in their phase diagram. Planted clique (see Section 1.2.2, (a)) may be the most appealing example. A significant amount of recent contributions [GZ19, DM13, FR10, Jer92, BHK$^+$16] has agreed on the fact that no polynomial-time algorithm is known to recover a planted clique smaller than $\Theta(\sqrt{n})$, even though as discussed in previous section, an exhaustive – non polynomial – search recovers a planted clique of size k as soon as $k \geq (2+\varepsilon)\log_2(n)$. The phase diagram for reconstruction in the planted clique problem is hence as follows:

Figure 1.9 – *Phase diagram for reconstruction in planted clique – see Section 1.2.2,* (a)

For another example where this phase transition of this type also appears, we can mention community detection for three or more communities (see e.g. Abbé's survey [Abb18] on the subject).

Understanding the phase diagram is also a fundamental and central question for the graph alignment problem, which we are now ready to introduce.

1.3. Graph alignment

After this short introduction to the general topic of inference in random graphs, let us now dive into the core of this thesis, namely the graph alignment problem, which will be the subject of interest in the upcoming chapters.

1.3.1. Motivations

Graph alignment[3] (or network alignment) aims to answer the following (informal) question: *'what is the best way to match the nodes of two graphs?'* Providing an answer to this

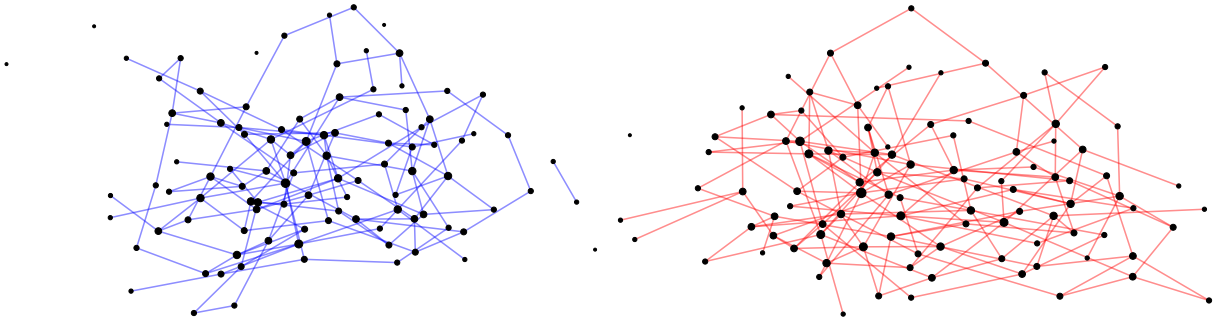


Figure 1.10 – *Graph alignment consists in the following informal question: 'what is the best way to match the nodes of two graphs?'*

very general query corresponds to exhibiting a vertex correspondence, or *alignment*, between two – labeled or unlabeled – graphs so that the aligned – labeled – versions of the graphs are resembling or close enough, for some well-defined distance.

Motivations for the investigation of this problem are numerous, since many questions from various fields can be phrased as graph alignment problems. Let us now give a short overview of several applications.

- *(De-)anonymization.* De-anonymization problems in networks aroused great interest when Narayanan and Shmatikov [NS08] were able to de-anonymize an unlabeled dataset of film ratings (subsampled from the Netflix dataset [NET]) with the help of auxiliary information given by the observation of a publicly available database (namely IMDb, the Internet Movie Database [IMB]). The authors proposed a simple method relying on the correspondence (or correlation) between movies and ratings across the databases, that were able to match some pairs of records and thus to recover the entire movie viewing history of a given subscriber, which may be in turn used as input to uncover political preferences or other sensitive information.

  Since then, de-anonymization problems have been studied in recent literature in several versions and reformulations: related topics such as quantifying privacy issues related to databases [Dwo08] or social networks [NS09] have been investigated.

- *Image processing and pattern recognition.* Some recognition tasks in image processing such as shape matching and object recognition [BBM05] can be achieved by finding correspondences between feature points across two (or several) images. The similarities are based on correspondence between vertex features as well as the cost of geometric transformation between pairs of nodes.

  Some popular algorithms for graph matching have been widely proposed for pattern recognition (see [CFVS04] for a broad survey) in many areas since the late seventies: 2D/3D image analysis [Mad16], document processing, video analysis, biometric identification as well as biomedical/biological applications. All these fields have in common that some structured information is represented by graphs, and the goal is to find a correspondence that somehow ensures that substructures in the first graph are mapped to similar substructures in the other.

[3]The same problem is sometimes found under the name *graph matching*. However, for the sake of clarity, we will only refer to graph alignment throughout the manuscript, in order not to confuse the reader with the different problem of planted matching evoked earlier (see Section 1.2.2, example (*b*)).

- *Protein interaction networks in computational biology.* Authors of [SXB07, SXB08] study protein-protein interaction (PPIs) networks represented as labeled graphs, where the nodes are proteins and edges represent interaction. These networks are observed across different species. They provide an algorithm, IsoRank, encoded as an eigenvalue problem, which performs a global alignment of two or more PPIs networks, using both the network structure of the data and sequence similarity.

  Aligning these networks proves to be a very valuable tool: first, it provides a phylogenetic function-oriented comparison of proteins across different species, identifying those that may play the same role, thus transferring knowledge and insights across species; second, it can be used to perform *ortholog prediction*, that is being able to spot genes that derive from the same ancestor.

  Some following works elaborated on approximations [KG16] or refined versions [LLB$^+$09] of IsoRank, and also developed further competitive methods for this problem [KSMG13, EKHK15].

- *Natural language processing and semantic entailment.* A fundamental task in natural language processing (NLP) is the recognition of *semantic entailment*, that is, given a piece of text, whether an hypothesis can be concluded by logical implication, or simply by general world-knowledge.

  In [HNM05], the authors use a representation of sentences as directed, labeled graphs between words and phrases, originally introduced in [LP01] (for a recent general survey on graph representations in NLP, see [OB20]). In these small networks, edges encode underlying dependency relationships in the sentence. Given a sentence and an hypothesis, the proposed strategy in order to identify entailment is to represent both the sentence and the hypothesis as graphs, and then to measure similarity between them, that is to find a mapping in the two graphs minimizing a score built from both the semantic resemblance of the matched vertices and how well the edges (namely, the relationships) are preserved by the mapping.

Graph alignment recently grew some new interest in other applied fields, including computational neurosciences [FCC$^+$21], analysis of data from diffusion magnetic resonance imaging [OSA16], and cross-lingual knowledge alignment [CTYZ17].

We refer to Section 1.3.4 for a brief history of theoretical aspects and results in graph alignment.

1.3.2. The quadratic assignment problem

Given two graphs $G = (V, E)$, $G' = (V', E')$ with same number of vertices $n = |V| = |V'|$, the problem of graph alignment consists in identifying a bijective mapping, or alignment $\pi : V \to V'$ that minimizes

$$\sum_{i,j \in V} \left(\mathbb{1}_{\{i,j\} \in E} - \mathbb{1}_{\{\pi(i),\pi(j)\} \in E'} \right)^2 , \tag{1.4}$$

that is the number of disagreements between adjacencies in the two graphs under the alignment π. In the case where the two graphs are isomorphic, the two node sets V and V' can be matched perfectly: an isomorphism between G and G' achieves zero cost in (1.4).

However, we are interested in graph alignment for general, non necessarily isomorphic graphs: the problem can hence be viewed as a noisy version of the isomorphism problem.

Given the adjacency matrices A and B of the two graphs G and G', the graph matching problem can be phrased as an instance of the quadratic assignment problem (QAP) [PRW94]

which is the following

$$\arg\min_{\Pi\in\mathcal{S}_n} \|A - \Pi B\Pi^\top\|^2 = \arg\max_{\Pi\in\mathcal{S}_n} \langle A, \Pi B\Pi^\top\rangle\,, \tag{1.5}$$

where Π ranges over all $n\times n$ permutation matrices, $\langle\cdot,\cdot\rangle$ denotes the matrix Frobenius inner product, i.e. $\langle C, D\rangle := \mathrm{Tr}(C^\top D)$, and $\|\cdot\|$ is the associated norm.

In a more general setting, including that of applications discussed in Section 1.3.1, the loss function can also take into account a matching cost for pairs of vertices, and the problem becomes

$$\arg\max_{\Pi\in\mathcal{S}_n} \langle A, \Pi B\Pi^\top\rangle + \langle C, \Pi\rangle\,, \tag{1.6}$$

where C is a $n\times n$ matrix such that the cost for matching vertex $u\in V$ and $u'\in V'$ is given by $-C_{u,u'}$.

Under its general formulation, QAP is known to be a NP-hard problem, as well as some of its approximations [PRW94, MMS14]. These hardness results are applicable in the worst case, where the observed graphs are designed by an adversary. In line with the worst-case/planted duality detailed earlier in Section 1.2.2, a natural idea is then to study the planted formulation, when A and B are random instances.

1.3.3. Planted graph alignment

We now study the planted version of graph alignment, namely the *planted graph alignment problem*, where the pair of graphs (G, H) is sampled according to the following general procedure. We generate a pair (G, G') of graphs, (or adjacency matrices (A, A')) with same node set such that G and G' are edge correlated, and relabel the nodes of G' with some uniform random permutation $\pi^\star\in\mathcal{S}_n$, independent from everything else, to form H.

Henceforward, we will always refer to graph alignment for planted graph alignment.

Let us describe several models of random correlated graphs that will be studied in the sequel: the Gaussian model, where the graph is complete and the signal lies on the edge weights, and the correlated Erdős-Rényi model, where the correlated graphs both have Erdős-Rényi marginal distributions.

Correlated Gaussian Wigner model The *correlated Gaussian Wigner model* was first introduced by Ding et al. [DMWX21] as a simple playground for graph alignment, and has been further investigated for its own sake in some recent works (see Section 1.3.4).

Under this model, the graphs are complete and the signal lies in the weights of edges between all pairs of nodes. The correlated weighted adjacency matrices A and A' are simply sampled as follows: first, A is drawn from the Gaussian Orthogonal Ensemble (GOE), namely, independently for all $1\le u\le v\le n$,

$$A_{u,v} = A_{v,u} \sim \begin{cases} \mathcal{N}(0, 1/n) & \text{if } u\neq v\,,\\ \mathcal{N}(0, 2/n) & \text{if } u = v\,. \end{cases} \tag{1.7}$$

Given H an independent copy of A, we define

$$A' = A + \xi H\,, \tag{1.8}$$

where $\xi > 0$ is the noise parameter. We denote $(A, A')\sim\mathsf{Wig}(n,\xi)$. Under this model, coefficients of A and A' are pairwise correlated with correlation parameter $\frac{1}{\sqrt{1+\xi^2}}$. This model is the subject of Chapter 3.

A natural variant of the model is to ensure that the two marginals are the same, and to remove self-loops in the graphs (i.e. diagonal coefficients). All pairs of edge weights

$(A_{u,v}, A'_{u,v})_{1\le u<v\le n}$ can be taken to be i.i.d. couples of normal variables with zero mean, unit variance and correlation parameter $\rho \in [0,1]$. An equivalent sampling procedure is to generate matrix A' from A as follows:

$$A' = \rho \cdot A + \sqrt{1-\rho^2} \cdot H, \tag{1.9}$$

where H is an independent copy of A. We denote $(A, A') \sim \mathsf{Wig}'(n, \rho)$. This model, very close to $\mathsf{Wig}(n, \xi)$, is the subject of Chapter 2.

Correlated Erdős-Rényi model As the simplest, most natural model of correlated random graphs, the *correlated Erdős-Rényi model* has naturally been the focus of recent threads of research (see Section 1.3.4) for the study of graph alignment. We refer to Section 1.2.1 for the definition of the non-correlated Erdős-Rényi model. This model will be studied in detail in Chapters 4, 5 and 6.

For a number of nodes n, edge probability $q \in [0,1]$ and correlation parameter $s \in [0,1]$ such that $s \ge q$, the correlated Erdős-Rényi model, denoted $\mathsf{G}(n,q,s)$, consists of two random graphs G, G' with symmetric adjacency matrices A, A', with same node set $V = [n]$, where $\{(A_{u,v}, A'_{u,v})\}_{u<v\in[n]}$ are i.i.d. pairs of correlated Bernoulli random variables such that

$$(A_{u,v}, A'_{u,v}) = \begin{cases} (1,1) & \text{with probability } qs \\ (1,0) & \text{with probability } q(1-s) \\ (0,1) & \text{with probability } q(1-s) \\ (0,0) & \text{with probability } 1-q(2-s). \end{cases} \tag{1.10}$$

Note that in this setting, G and G' both have $\mathsf{G}(n,q)$ marginal distributions.

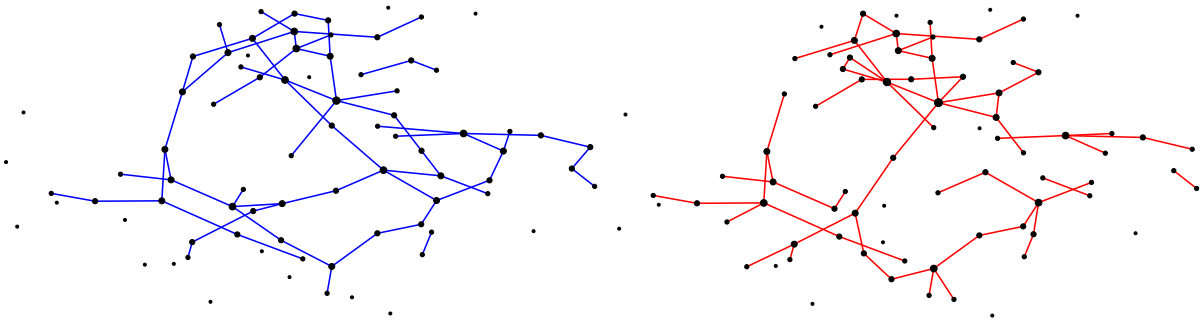


Figure 1.11 – *Two samples from* $\mathsf{G}(n,q,s)$ *with* $n = 80$, $p = 1.9/n$ *and* $s = 0.8$.

Remark 1.3.1. *Note that if* $s = 1$, *the graphs* G *and* G' *are identical, and in the case* $s = q$, *the two graphs are independent, that is* $\mathsf{G}(n,q,q) \overset{(\mathsf{d})}{=} \mathsf{G}(n,q) \otimes \mathsf{G}(n,q)$.

Remark 1.3.2. *Another equivalent sampling procedure for the correlated Erdős-Rényi model is as follows. Starting from a parent graph* $F \sim \mathsf{G}(n, q/s)$, G *and* G' *are obtained by two independent* $s-$*subsamplings of* F *(a* $s-$*subsampling of* F *consists in keeping each edge of* F *independently with probability* s*).*

Planting the alignment As mentioned earlier, in the planted model for graph alignment, after having generated two labeled correlated graphs G and G', the last step is to draw the planted permutation $\pi^\star$ uniformly at random in $\mathcal{S}_n$.

We then relabel the second graph G' according to this permutation $\pi^\star$, forming the graph H with adjacency matrix B such that for all $1 \le u, v \le n$,

$$B_{\pi^\star(u),\pi^\star(v)} = A'_{u,v}\,, \tag{1.11}$$

or, equivalently, $B = (\Pi^\star)^\top A' \Pi^\star$ where $\Pi^\star$ is the $n \times n$ matrix representation of permutation $\pi^\star$, that is $\Pi^\star_{u,v} = \mathbb{1}_{v=\pi^\star(u)}$.

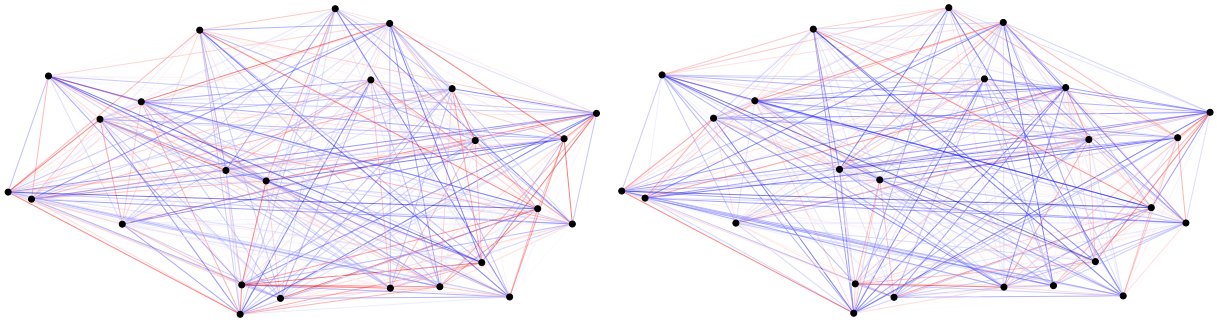


Figure 1.12 – *A sample from model* $\mathsf{Wig}'(n,\rho)$ (1.9) *with* $n = 25$ *and* $\rho = 0.8$. *Edges are colored according to their weights.*

Reconstruction Given the two graphs (G, H) generated from the planted model described above, the reconstruction task consists in finding an estimator $\hat{\pi}$ of the planted solution $\pi^\star$ upon observing G and H (or equivalently A and B).

The performance of any estimator $\hat{\pi} = \hat{\pi}(G, H) : [n] \to [n]$ will be assessed through its *overlap* with the unknown planted permutation $\pi^\star$, defined as

$$\mathrm{ov}(\hat{\pi}, \pi^\star) := \frac{1}{n} \sum_{u \in [n]} \mathbb{1}_{\hat{\pi}(u) = \pi^\star(u)} \,. \tag{1.12}$$

The overlap (1.12) is now the measure of performance which we seek to optimize when performing graph alignment in this planted setting, and differs from that of the non-planted case (1.5).

We can however straightaway note that in the Erdős-Rényi setting, the posterior distribution of the planted alignment $\Pi^\star$ is given by

$$\begin{aligned}
\mathbb{P}(\Pi^* = \Pi | A, B) &\propto \mathbb{P}\left(\Pi^* = \Pi, A, B\right) \\
&= \frac{1}{n!} (qs)^{\frac{1}{2}\langle A, \Pi B \Pi^\top \rangle} (q(1-s))^{\frac{1}{2}\langle A, \mathbb{1} \rangle + \frac{1}{2}\langle \mathbb{1}, B \rangle - \langle A, \Pi B \Pi^\top \rangle} (1 - q(2-s))^{\frac{1}{2}\langle \mathbb{1} - A, \mathbb{1} - \Pi B \Pi^\top \rangle} \\
&\propto \left(\frac{s(1-q(2-s))}{q(1-s)^2} \right)^{\frac{1}{2}\langle A, \Pi B \Pi^\top \rangle} ,
\end{aligned} \tag{1.13}$$

where $\mathbb{1}$ denotes the all-ones matrix. Since $s \geq q$, $\frac{s(1-q(2-s))}{q(1-s)^2} \geq 1$, and the maximum-a-posteriori estimator of $\pi^\star$ given G, H is thus exactly the solution of the QAP (1.5). This is again unsurprisingly in accordance with the worst-case/planted duality (see Section 1.2.2). The same computations show that this result also holds in the Gaussian Wigner models $\mathsf{Wig}(n, \xi)$ (1.7) and $\mathsf{Wig}'(n, \rho)$ (1.8).

We now specify different types of reconstruction tasks that will be referred to in the rest of the thesis. A sequence of estimators $\{\hat{\pi}_n\}_n$ (i.e. measurable functions of G, H) – omitting the dependence in n – is said to achieve

- *Exact recovery* if $\mathbb{P}(\hat{\pi} = \pi^\star) \underset{n\to\infty}{\longrightarrow} 1$,
- *Almost exact recovery* if $\mathbb{P}(\mathrm{ov}(\hat{\pi}, \pi^\star) = 1 - o(1)) \underset{n\to\infty}{\longrightarrow} 1$,
- *Partial recovery* if there exists some $\varepsilon > 0$ such that $\mathbb{P}(\mathrm{ov}(\hat{\pi}, \pi^\star) > \varepsilon) \underset{n\to\infty}{\longrightarrow} 1$.

Remark 1.3.3. *Partial recovery consists in ensuring that the estimator $\hat{\pi}$ matches a non-vanishing fraction of nodes – we hope this fraction to be as large as possible. Though simple to formulate, from an application standpoint it may however be of little use to know that one has a permutation with 30% of correctly matched nodes if one does not have a clue about which pairs are correctly matched. Motivated in part by the following remark, we will introduce in Section 1.4.4, Chapters 4 and 6 another slightly different recovery task, namely one-sided partial recovery, which is believed to be more relevant both for theory and practice.*

1.3.4. A summary of related work

We give in the following a overview of related algorithmic and theoretical contributions, aside from our work.

Seeded graph alignment An interesting and widely studied setting is graph alignment with presence of side information, namely *seeds*, that are correct pre-mapped vertex pairs. The idea of seeded alignment methods is that vertices u of G and u' of H will have more witnesses – that is, correct pairs (w, w') such that $u \underset{G}{\longleftrightarrow} w$ and $u' \underset{H}{\longleftrightarrow} w'$ – if they are matched than if they are not.

It is proved in [LFP14a] that when the graphs are dense enough, a logarithmic number of correct seeds is sufficient to recover the whole alignment. To perform this task, several methods are proposed [PG11, FAP$^+$19, SGE17, YG13, MX19, ABT22], some of them relying on percolation techniques [JLTV12, YG13], large neighborhoods statistics [MX19], or projected power method [ABT22].

An interesting line of work extends the problem in the case where some of them are likely to be incorrect, e.g. since they may be provided by seedless methods. The NoisySeeds algorithm, also built on a percolation procedure, is proposed in [KHG15], and [LS18] uses 1-hop witnesses to recover the full alignment. The recent paper [YXL21] establishes information-theoretic results for seeded alignment in the noisy case, and proposes a method which considers both 1-hop and 2-hop witnesses, improving on previous theoretical guarantees.

Information-theoretic results First fundamental results for Erdős-Rényi graph alignment are due to Pedarsani and Grossglauser [PG11], followed by Cullina and Kiyavash [CK17] who prove that under some mild sparsity constraints, feasibility of exact alignment exhibits a sharp threshold at $nqs \simeq \log n$. Their approach is based on the analysis of the maximum a posteriori estimator for the positive side, and the impossibility result is the consequence of the large number[4] of automorphisms of an Erdős-Rényi graph with mean degree less than $(1-\varepsilon)\log n$ [Bol01].

Results for almost-exact recovery proved in [CKMP18] establish that almost-exact recovery is feasible if and only if $nqs \to +\infty$, under some mild sparsity assumptions.

These reconstruction thresholds are sharpened by the recent work [WXY21]. This paper shows, among other results, a sharp all-or-nothing phenomenon in the Gaussian setting at $n\rho^2 = 4\log n$. If $n\rho^2 > (4+\varepsilon)\log n$, exact alignment is feasible, whereas below the threshold even partial recovery is infeasible.

For dense Erdős-Rényi graphs with $q/s = n^{-o(1)}$, another phase transition arises between infeasibility of partial alignment and possibility of almost-exact alignment, at

$$\frac{nqs(\log(s/q) - 1 + q/s)}{\log n} = 2\,. \tag{1.14}$$

Partial recovery was first studied by Hall and Massoulié [HM20], who showed that $nqs \to 0$ is an impossibility condition, whereas $nqs > C$ (with a large, non-explicit constant C), together with some additional sparsity constraints, ensures feasibility. These results are improved in [WXY21], where the authors show that if $q = \lambda/n$, and s is an constant, then

[4]Note that the illustrative result in Theorem 1.1 proved earlier on gives a short proof of this fact.

partial recovery is shown to be feasible when $nqs > 4 + \varepsilon$. The impossibility result requires $nq/s = \omega(\log^2 n)$ and $nqs < 1 - \varepsilon$.

At the time this manuscript is being completed, a very recent contribution [DD22] sharpens this last result in the case where $q/s = n^{-\alpha+o(1)}$, showing a sharp threshold for partial recovery at $nqs = \lambda^\star(\alpha)$, where $\lambda^\star(\alpha)$ is given as the asymptotic maximal edge-vertex ratio over all nonempty subgraphs of an Erdős-Rényi graph $\mathsf{G}(n, \frac{1}{\alpha n})$.

Algorithms for exact recovery in the dense case Hitherto, the vast majority of prior work focused on exact recovery with no side information, seeking to provide polynomial-time (or quasi-polynomial time) algorithms that (sometimes provably) recover the entire permutation $\pi^\star$ under some conditions on the parameters n, q, s (or n, ρ, ξ is the Gaussian setting).

Spectral methods. A first spectral method for recovery is due to [Ume88], and uses spectral decompositions and relaxation of the QAP on the orthogonal group. Another spectral, rank-reduction method is proposed by Feizi et al. [FQM+16], and an another algorithm, `GRAMPA`, is proposed and analyzed in [FMWX19a, FMWX19b], both for the Wigner and the Erdős-Rényi model. `GRAMPA` builds a similarity matrix based on outer products between pairs of eigenvectors of the two graphs, and outputs a matching via a rounding procedure.

QAP relaxations. A class of algorithms designed for recovery follows a Frank-Wolfe approach (see [ABGL02, VCL+11, ZBV09]) , which consists in relaxing the integer programming formulation of the QAP (1.5) to a continuous optimization problem on which iterative linearized procedures are used, and then projecting the final iterate on the space of solutions. Every linear optimization step at each iteration is a *linear assignment problem (LAP)* of the form

$$\arg\max_{\Pi \in \mathcal{S}_n} \langle \Pi u, v \rangle \,, \tag{1.15}$$

with $u, v \in \mathbb{R}^n$, which can be solved using the Hungarian algorithm [Kuh55] in $O(n^3)$ time complexity. In the same vein, authors of [ZBV09] study a path following algorithm on a concave relaxation of the QAP.

The question of giving theoretical guarantees on the performance of such relaxations of the QAP is interestingly discussed in [LFP14b]. In this paper, the common convex relaxation of the QAP which consists in minimizing $\|AD - DB\|^2$ over all doubly-stochastic matrices D is proved to almost always fail, whereas the indefinite relaxed graph alignment problem[5] which minimizes $-\langle AD, DB \rangle$ over all doubly-stochastic matrices D almost always discovers the true permutation, if solved exactly. Though non-convex quadratic programming is NP-hard in general, this indefinite relaxation can still be efficiently approximately solved with the Frank-Wolfe methodology evoked here above.

Relevant to QAP relaxation methods is the recent contribution [DML17] which proposes a convex quadratic programming relaxation, proved to be more accurate than the classical double-stochastic and spectral relaxations, although with same time complexity as the former.

Methods using network topology. Another class of methods are based on the exploration of the network topology, in order to design vertex signatures that can efficiently recover the matched pairs. Ding et al. [DMWX21] introduced a matching procedure based on degree profiles, that is the empirical distribution of the degrees of neighbors. A method proposed in [BCL+19] relies on counting copies of subgraphs adjacent to a given node, for a well-chosen family of graphs. In recent contributions, Mao, Rudelson and Thikhomirov design an method involving a two-generation partitioning procedure [MRT21b] and another algorithm [MRT21a] based on comparison of partition trees associated with the graph vertices. These

[5] Note that since $\|DA\|^2 \neq \|A\|^2$ in general, these two relaxations have now different solutions. Moreover, the objective in this second relaxation is no more convex in D, the Hessian being indefinite.

methods are shown to improve the previously state-of-the art performances in terms of noise robustness (see below).

Theoretical guarantees. From the methods for exact recovery cited here above, those giving rigorous theoretical guarantees all require a mean degree at least $nq \geq \text{polylog}\, n$, and $1 - s \leq 1/(\text{polylog}\, n)$ in the Erdős-Rényi setting or $1 - \rho^2 \leq 1/(\text{polylog}\, n)$ in the Gaussian setting, that is a correlation close enough to 1. The only exceptions for exact recovery are the recent works of Mao, Rudelson and Thikhomirov: in the Erdős-Rényi model, [MRT21b] tolerates a noise $1 - s$ up to $(\log \log n)^{-c}$, and [MRT21a] can tolerate up to constant noise – the constant being unspecified. The recent algorithm proposed in [ABT22] can also tolerate constant noise for exact recovery in the seeded setting.

Note that these methods are not proved to work in the sparse setting where both the correlation and the mean degree are constant, setting on which we will focus in Chapters 4, 5 and 6.

Detection problem Aside from the reconstruction tasks, the detection has less been studied until very recently. Given n, q, s, the associated hypothesis problem is as follows:

$$\mathcal{H}_0 := \text{“}(G, H) \text{ are two independent } \mathsf{G}(n, q) \text{ graphs”}$$

versus

$$\mathcal{H}_1 := \text{“}(G, H) \text{ are drawn under the Erdős-Rényi planted model”}\,.$$

Wu, Xu and Yu [WXY20] give fundamental results for the detection problem. They establish a sharp threshold for detection in the Gaussian model at $n\rho^2/\log n = 4$, and show for the Erdős-Rényi model that in the dense case $q/s = n^{-o(1)}$, the sharp threshold for partial/almost-exact alignment given in (1.14) also holds for detection. The picture in the sparser case $q/s = n^{-\Omega(1)}$ is however less clear, but in the case where $q = \lambda/n$ and s is a constant, their result implies that strong detection is feasible if $\lambda s > 2$ and infeasible if $\lambda s < 1$ and $s < 0.01$.

Improving on a previous work [BCL$^+$19], the state-of-the art algorithm for this detection task is proposed in [MWXY21] which consider a test based on counting trees in the two graphs. This algorithm runs in $O(n^{2+o(1)})$ time and is proved to succeed with high probability if $n \min(q, 1-q) \geq n^{-o(1)}$ (this assumption is very mild) and the correlation coefficient (which is asymptotically s if $q \to 0$) is greater that $\sqrt{\alpha} \sim 0.58$, where α is Otter's constant [Ott48], defined as the inverse of the exponential growing rate of the number of unlabeled trees with K edges. This algorithm also improves on previous informational results and will be the object of further discussion in Chapter 7.

1.4. Correlation detection in random trees

In this Section, we introduce a problem which will be at the heart of Chapters 5 and 6: correlation detection in random trees.

1.4.1. Problem statement

The problem of detecting correlation in random trees is a fundamental statistical task, consisting in deciding whether two rooted trees are correlated up to a relabeling of the nodes, that is if they contain a common planted subtree, or if they are independent.

This problem could very well be defined per se and studied as such; we nevertheless explain briefly how this problem arises from the study of sparse graph alignment. Let us imagine that we are given correlated graphs G, H from the Erdős-Rényi planted model, and that one would like to know whether node $u \in V(G)$ is matched to $u' \in V(H)$, namely if $u' = \pi^\star(u)$. An answer to this question can be to build an estimator $\hat{\pi}$ such that $\hat{\pi}(u) = u'$ if

and only if the local structure of graph G in the neighborhood of node u is somehow 'close' to the local structure of graph H in the neighborhood of node u'.

In the sparse regime, it is well known that the neighborhoods up to distance d of node u (resp. u') in G (resp. G'), are both asymptotically distributed as Galton-Watson branching trees[6]. More specifically, if $u' = \pi^\star(u)$, then the pair of neighborhoods are asymptotically jointly distributed as correlated Galton-Watson branching trees (distribution denoted $\mathbb{P}_d^{(\lambda,s)}$). On the other hand, for pairs of nodes (u, u') taken at random in $[n]$, the neighborhoods are asymptotically independent Galton-Watson branching trees (distribution denoted $\mathbb{P}_d^{(\lambda)}$). Hence, we are now left with the problem of detecting correlation in random trees.

Rooted labeled trees A *labeled rooted tree* $t = (V, E)$ is an undirected graph with node set V and edge set E which is connected and contains no cycle. The *root* of t is a given distinguished node $\rho \in V$, and the *depth* of a node u is defined as its graph distance to the root ρ. The depth of tree t is given as the maximum depth of all nodes in t.

In a rooted tree t, each node u at depth $d \geq 1$ has a unique *parent* in t, which can be defined as the unique node at depth $d-1$ on the path from u to the root ρ. Similarly, the *children* of a node u of depth d are all the neighbors of u at depth $d+1$. For any node u of the tree t, we denote by t_u the subtree of t rooted at node u, that is the tree obtained by deleting the edge between u and its parent and keeping the connected component of u.

Models of random trees, hypothesis testing We describe hereafter models of random trees that will be useful in the sequel. For more detailed definitions we refer to Chapter 5, Section 5.2.4.

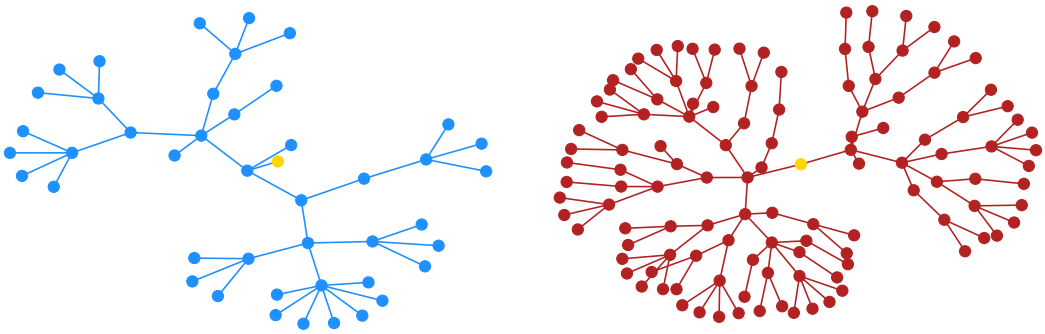
Galton-Watson trees with Poisson offspring. The *Galton-Watson tree with offspring* $\mathsf{Poi}(\mu)$ *up to depth* d, denoted by $\mathsf{GW}_d^{(\mu)}$, is defined recursively as follows. First, the distribution $\mathsf{GW}_0^{(\mu)}$ is a Dirac at the trivial tree only consisting in the root. Then, for $d \geq 1$, sample a number $Z \sim \mathsf{Poi}(\mu)$ of independent $\mathsf{GW}_{d-1}^{(\lambda)}$ trees, and attach each of them as children of the root, to form a tree of depth at most d.

Tree augmentation. For $\lambda > 0$ and $s \in [0, 1]$, a (random) $(\lambda, s)-$*augmentation* of a given tree $\tau = (V, E)$, denoted by $\mathsf{Aug}_d^{(\lambda,s)}(\tau)$, is defined as follows. First, attach to each node u in V at depth $< d$ a number Z_u^+ of additional children, where the Z_u^+ are i.i.d. of distribution $\mathsf{Poi}(\lambda(1-s))$. Let V^+ be the set of these additional children. To each $v \in V^+$ at depth $d_v \in [d]$, we attach another random tree of distribution $\mathsf{GW}_{d-d_v}^{(\lambda)}$, independently of everything else.

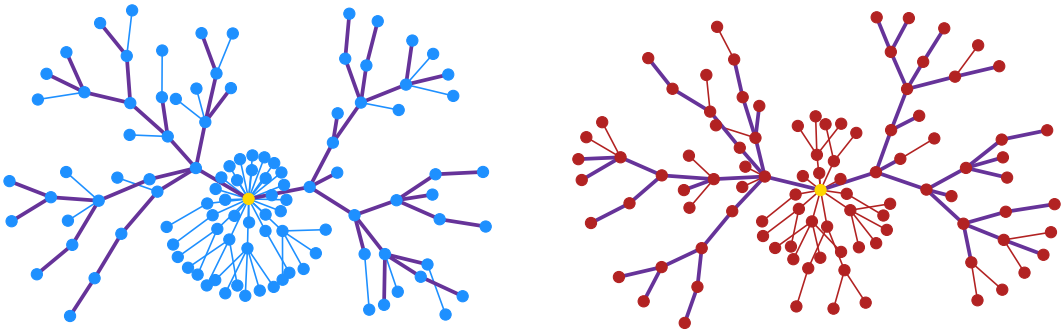
We are now ready to describe the two models $\mathbb{P}_d^{(\lambda)}$ and $\mathbb{P}_d^{(\lambda,s)}$ in simple words. Under the independent model $\mathbb{P}_d^{(\lambda)}$, T and T' are two independent $\mathsf{GW}_d^{(\lambda)}$ trees. The correlated model $\mathbb{P}_d^{(\lambda,s)}$ is built as follows. Starting from an *intersection tree* $\tau^\star \sim \mathsf{GW}_d^{(\lambda s)}$, and T and T' are obtained as two independent $(\lambda, s)-$augmentations of $\tau^\star$.

In both models, the labels of the trees are always forgotten, or randomly uniformly re-sampled. We however still distinguish the roots of the two trees. It can easily be verified that the marginals of T and T' are the same under $\mathbb{P}_d^{(\lambda)}$ and $\mathbb{P}_d^{(\lambda,s)}$, namely $\mathsf{GW}_d^{(\lambda)}$. The parameters are λ, the mean number of children, and the correlation s.

[6]This convergence in fact happens in the sense of *Benjamini-Schramm*, see [BS11].



(a) – *Samples T, T' from $\mathbb{P}_d^{(\lambda)}$.*



(b) – *Samples T, T' from $\mathbb{P}_d^{(\lambda,s)}$. The common subtree $\tau^\star$ is drawn thick and purple.*

Figure 1.13 – *Samples from models $\mathbb{P}_d^{(\lambda)}$ and $\mathbb{P}_d^{(\lambda,s)}$, with $\lambda = 1.8$, $s = 0.8$, and $d = 5$. The root node is highlighted in yellow. Labels are forgotten.*

1.4.2. Hypothesis testing, one-sided test

The corresponding hypothesis test can be formalized as follows: given the observation of a pair of trees (T, T') of depth at most d, we want to test

$$\mathcal{H}_0 = \text{"}T, T' \text{ are drawn under } \mathbb{P}_d^{(\lambda)}\text{"} \quad \text{versus} \quad \mathcal{H}_1 = \text{"}T, T' \text{ are drawn under } \mathbb{P}_d^{(\lambda,s)}\text{"}. \tag{1.16}$$

In statistical detection problems (see Section 1.2), the commonly considered tasks are that of

- *strong detection*, i.e. designing tests $\mathcal{T}_d$ that verify

$$\lim_{d\to\infty} \left[\mathbb{P}_d^{(\lambda)}\left(\mathcal{T}_d(T,T') = 1\right) + \mathbb{P}_d^{(\lambda,s)}\left(\mathcal{T}_d(T,T') = 0\right)\right] = 0,$$

- *weak detection*, i.e. tests $\mathcal{T}_n$ that verify

$$\limsup_{d\to\infty} \left[\mathbb{P}_d^{(\lambda)}\left(\mathcal{T}_d(T,T') = 1\right) + \mathbb{P}_d^{(\lambda,s)}\left(\mathcal{T}_d(T,T') = 0\right)\right] < 1,$$

In other words, strong detection corresponds to exactly discriminating w.h.p. between $\mathbb{P}_d^{(\lambda)}$ and $\mathbb{P}_d^{(\lambda,s)}$, whereas weak detection corresponds to strictly outperforming random guessing. We here argue that neither strong detection nor weak detection are relevant for our problem.

First, because of the event that the intersection tree does not survive, which is of positive probability under $\mathbb{P}_d^{(\lambda,s)}$: we always have $\mathbb{P}_d^{(\lambda,s)}(t,t') \geq C \cdot \mathbb{P}_d^{(\lambda)}(t,t')$ for some $C = C(\lambda,s) > 0$. This implies that $\mathbb{P}_d^{(\lambda)}$ is always absolutely continuous w.r.t. $\mathbb{P}_d^{(\lambda,s)}$, hence strong detection

can never be achieved.

Second, weak detection is always achievable as soon as $s > 0$: with the same notations as here above, the difference of the degree of the root in T and that of the root in T' is always centered but has different variance under $\mathbb{P}_d^{(\lambda)}$ and under $\mathbb{P}_d^{(\lambda,s)}$, hence these two distributions can be weakly distinguished, without any further assumption than $s > 0$.

Moreover, if we want our test to be relevant for partial alignment – for which we know that only a fraction of the nodes can be recovered – it is natural to require a positive power (i.e., being able to detect matched nodes with some positive probability), but also to ensure that the output of the algorithm contain almost no wrong pair (i.e. imposing a vanishing type I error).

We are thus interested in being able to ensure the existence of an asymptotic *one-sided test*, that is a test $\mathcal{T}_d : \mathcal{X}_d \times \mathcal{X}_d \to \{0, 1\}$ such that $\mathcal{T}_d$ chooses hypothesis $\mathcal{H}_0$ under $\mathbb{P}_d^{(\lambda)}$ with probability $1-o(1)$, and chooses $\mathcal{H}_1$ under $\mathbb{P}_d^{(\lambda,s)}$ with some positive uniformly lower-bounded probability.

1.4.3. Two methods

We now give the outline of two methods for detection of correlation in random trees, that will be the object of Chapters 5 and 6.

Tree matching weight In Chapter 5, we build a test based on a measure of similarity between two trees: the *tree matching weight*.

Matching weight of two rooted trees. For any $d \geq 0$, let $\mathcal{A}_d$ denote the collection of rooted trees whose leaves are all of depth d. Given two rooted trees t and t' of depth at most d, let $M(t, t')$ denote the collection of trees $\tau \in \mathcal{A}_d$ such that there exist injective embeddings of τ in t and t' that preserve the rooted tree structure, that is the depth of the nodes and the child-parent relationship. The *matching weight of trees t and t' at depth d* is then defined as:

$$\mathcal{W}_d(t, t') := \sup_{\tau \in M(t,t')} |\mathcal{L}_d(\tau)|, \tag{1.17}$$

where $\mathcal{L}_d(\tau)$ is the number of leaves at depth d of tree τ. In other words, the tree matching weight of a pair of trees is defined as the maximal size of a common subtree, measured in terms of number of leaves.

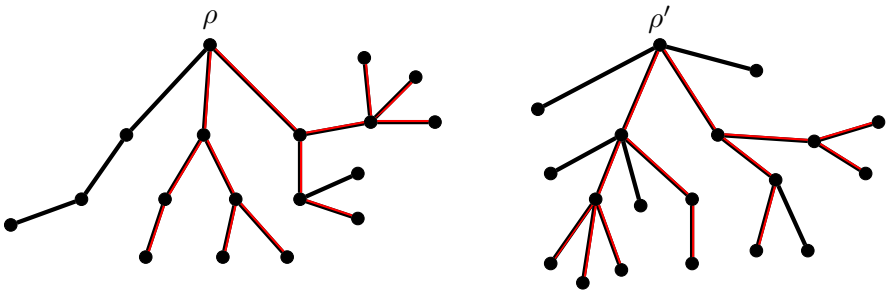


Figure 1.14 – *Example of two trees t, t' with $\mathcal{W}_3(t, t') = 7$, where an optimal $t \in \mathcal{A}_3$ is drawn in red.*

Recursive computation of $\mathcal{W}_d$. From the previous definition, a first step conditioning

yields a recursion formula for the matching weight $\mathcal{W}_d$, of the following form:

$$\mathcal{W}_d(t,t') = \sup_{\mathfrak{m}\,:\,\mathcal{M}(C_t,C_{t'})} \sum_{(u,u')\in\mathfrak{m}} \mathcal{W}_{d-1}(t_u, t'_{u'})\,, \tag{1.18}$$

where the supremum is taken over all matchings, that is one-to-one mappings $\mathfrak{m}$ from a subset of the root's children set C_t in t to the root's children set $C_{t'}$ in t'. This recursion formula (1.18) is at the heart of the analysis of this statistic as well as the design of related algorithms. The general idea is that if the trees are correlated, with positive probability they will tend to have a significantly higher matching weight than if they are independent. We refer to Section 5.2.2 for a proof of (1.18) and more generally to Chapter 5 for the study of this method.

The likelihood ratio In Chapter 6, we are interested in studying the existence of one-sided tests for detection, which we recall are tests guarantying an asymptotic vanishing type I error and non vanishing power. According to the Neyman-Pearson Lemma, optimal one-sided tests are based on the *likelihood ratio* L_d of the distributions under the distinct hypotheses $\mathbb{P}_d^{(\lambda,s)}$ and $\mathbb{P}_d^{(\lambda)}$. For a pair of trees (t,t'), this likelihood ratio is given by

$$L_d(t,t') := \frac{\mathbb{P}_d^{(\lambda,s)}(t,t')}{\mathbb{P}_d^{(\lambda)}(t,t')}\,. \tag{1.19}$$

Recursive computation of L_d. This likelihood ratio also satisfies a nice recursive property:

$$L_d(t,t') = \sum_{k=0}^{c\wedge c'} \psi(k,c,c') \sum_{\substack{\sigma:[k]\to[c]\\ \sigma':[k]\to[c']}} \prod_{i=1}^{k} L_{n-1}(t_{\sigma(i)}, t'_{\sigma'(i)})\,, \tag{1.20}$$

where c (resp. c') is the degree of the root in t (resp. in t'), and the second sum of the RHS is taken over injective mappings σ and σ'. The coefficients $\psi(k,c,c')$ are given by

$$\psi(k,c,c') = e^{\lambda s} \times \frac{s^k \bar{s}^{c+c'-2k}}{\lambda^k k!}\,.$$

The idea here again is that with positive probability the likelihood ratio is going to be significantly larger for correlated trees than for independent pairs, hence a test $\mathcal{T}_d$ of the form $\mathcal{T}_d(t,t') = \mathbb{1}_{L_d(t,t')>\beta_d}$ for a well chosen threshold β_d should solve one-sided detection whenever possible. We refer to Section 6.3.1 for the details and proof of (1.20), and more generally to Chapter 6 for a thorough study of this method.

1.4.4. Heuristics for partial graph alignment

We briefly state the results that establish a link between tree correlation detection and graph alignment. Given that the tests considered above are one-sided tests, we are going to perform *one-sided partial recovery.*

One-sided partial recovery In order to define this notion, we are left to consider estimators of $\pi^\star$ that are no longer necessarily permutations, but only one-to-one functions from a subset $\mathcal{C}\subset[n]$ of the node set of G to the node set of H – which we recall is also $[n]$.

For any subset $\mathcal{C}\subset[n]$, the performance of any one-to-one estimator $\hat{\pi}:\mathcal{C}\to[n]$ is still assessed through its overlap $\mathrm{ov}(\hat{\pi},\pi^\star)$, defined as in (1.12) by:

$$\mathrm{ov}(\hat{\pi},\pi^\star) = \frac{1}{n}\sum_{u\in\mathcal{C}} \mathbb{1}_{\hat{\pi}(u)=\pi^\star(u)}\,.$$

Note that the estimator may not be in $\mathcal{S}_n$, and only consist in a partial matching. We also

need to define the *error fraction* of $\hat{\pi}$ with the unknown permutation $\pi^\star$:

$$\mathrm{err}(\hat{\pi}, \pi^\star) := \frac{1}{n} \sum_{u \in \mathcal{C}} \mathbb{1}_{\hat{\pi}(u) \neq \pi^\star(u)} = \frac{|\mathcal{C}|}{n} - \mathrm{ov}(\hat{\pi}, \pi^\star). \tag{1.21}$$

A sequence of injective estimators $\{\hat{\pi}_n\}_n$ – omitting the dependence in n – is said to achieve one-sided partial recovery if there exists some $\varepsilon > 0$ such that w.h.p. $\mathrm{ov}(\hat{\pi}, \pi^\star) > \varepsilon$ and also $\mathrm{err}(\hat{\pi}, \pi^\star) = o(1)$.

We end this Section and the introduction with an informal statement relating the existence of a one-sided test for tree correlation detection and one-sided partial graph alignment, which will be discussed further in Chapter 6.

(Informal Statement). *For given (λ, s), if there exists a one-sided test for tree correlation detection, then one-sided partial alignment in the correlated Erdős-Rényi model $\mathcal{G}(n, \lambda/n, s)$ is achieved in polynomial time by the* `MPAlign` *algorithm defined in Chapter 6.*

Chapter 2

Alignment of graph databases with Gaussian weights: fundamental limits

In this chapter, we study the fundamental limits for reconstruction in weighted graph (or matrix) database alignment. We consider the Wigner model $\mathsf{Wig}'(n,\rho)$ (1.9), and we prove that there is a sharp threshold for exact recovery of $\pi^\star$: if $n\rho^2 \geq (4+\varepsilon)\log n + \omega(1)$ for some $\varepsilon > 0$, there is an estimator $\hat{\pi}$ – namely the MAP estimator – based on the observation of databases A, B that achieves exact reconstruction with high probability. Conversely, if $n\rho^2 \leq 4\log n - \log\log n - \omega(1)$, then any estimator $\hat{\pi}$ verifies $\hat{\pi} = \pi$ with probability $o(1)$.

This result shows that the information-theoretic threshold for exact recovery is the same as the one obtained for detection in [WXY20]: in other words, for Gaussian weighted graph alignment, the problem of reconstruction is, fundamentally, not more difficult than that of detection.

The proofs build upon the analysis of the MAP estimator and the second moment method – introduced earlier in Section 1.2.1 – together with the study of the correlation structure of energies of permutations.

This chapter is based on the paper *Sharp threshold for alignment of graph databases with gaussian weights* [Gan22], published at *MSML 2021.*

2.1. Introduction

2.1.1. Aligning databases

We address the following problem: suppose that we have two databases consisting in weighted graphs represented by their adjacency matrices A and B. For simplicity, assume that the two graphs have same size and that each individual appears in both graphs. For a given individual, its attached signal consists in weighted edges with all other users. Across databases, edges that correspond to pairs of matched individuals are correlated. We consider the following question: *if the graphs are shown unlabeled (that is, if users are anonymized), is it possible to recover the corresponding matching between databases by aligning them at the sight of their correlation structure?*

Intuitively, when the matrices are correlated enough, one can learn the true matching between individuals present in the databases. We investigate the precise conditions on correlation under which exact reconstruction (or perfect de-anonymization) is feasible with high probability.

As mentioned in Section 1.3.1, *de-anonymization problems* aroused great interest when [NS08] were able to de-anonymize an unlabeled dataset of film ratings with the observation of a publicly available database, using correlations between the ratings. Since then, some authors have sought to quantify privacy issues related to databases [Dwo08] or social networks [NS09], one of the starting points of the widespread attention given on the more general *graph*

alignment problem. We refer to Section 1.3.1 for further applications and to Section 1.3.4 for a survey of theoretical results.

Vector-shaped and graph-shaped databases From the theoretical point of view, fundamental limits for the deanonymisation problem are now well understood when data only consists in vectors u, v of size n (or more generally, rectangular databases of size $n \times k$) [CMK18, ECK19], that is when each user has its own signal, regardless of its connections with others. In this setting, the problem can be phrased in terms of a *Linear Assigment Problem (LAP)*:

$$\arg\max_{\Pi} \langle \Pi u, v \rangle, \tag{2.1}$$

where the maximum runs over all permutation matrices of size n. As mentioned earlier in the introduction, LAP can be solved efficiently in $O(n^3)$ steps using the classical Hungarian algorithm ([Kuh55]).

Another related problem is that of linear regression with an unknown permutation, studied in [PWC16]: this time, one observes $y = \Pi^\star A x^\star + w$, where $x^\star \in \mathbb{R}^d$ is an unknown vector, $\Pi^\star$ is an unknown $n \times n$ permutation matrix, and $w \in \mathbb{R}^n$ is additive Gaussian noise. Here again, the permutation $\Pi^\star$ applies only on the left side of A, which corresponds to row permutation.

On the other hand, we recall that when the databases are graphs, the problem is different and can be phrased this time in terms of a *Quadratic Assigment Problem (QAP)*:

$$\arg\max_{\Pi} \langle A, \Pi B \Pi^T \rangle. \tag{2.2}$$

We recall that a significant difference with the previous vector-shaped setting is that this problem is known to be NP-hard in the worst case, as well as some of its approximations [MMS14, PRW94]. In the case where the signal lies in the graph structure itself – that is, when the pairs $(A_{u,v}, B_{\pi^\star(i),\pi^\star(j)})$ are correlated pairs of Bernoulli variables – [CK17] shows that there exists a sharp threshold for exact recovery, where the signal-to-noise ratio can be expressed in the correlated Erdős-Rényi model in terms of the size n of both graphs, the marginal edge probability q and the correlation parameter s between edges of the two graphs. Namely, this sharp threshold is at $nqs \sim \log n$.

This chapter focuses on the case where signal lies in weights on edges between all pairs of nodes. We recall hereafter the correlated Gaussian Wigner model $\mathsf{Wig}'(n, \rho)$ defined in (1.9).

Model of Gaussian Wigner matrices In the correlated Gaussian Wigner model $\mathsf{Wig}'(n, \rho)$ (1.9), the weighted adjacency matrices A and B of the two graphs G and H are symmetric, and sampled as follows: first draw the planted permutation $\pi^\star$ uniformly at random in $\mathcal{S}_n$. Then all pairs of edge weights $(A_{u,v}, B_{\pi^\star(u),\pi^\star(v)})_{1 \le i < j \le n}$ are i.i.d. couples of normal variables with zero mean, unit variance and correlation parameter $\rho \in [0, 1]$. Since all Gaussian variables are independent from $\pi^\star$, matrix B can also be drawn from A as follows:

$$B = \rho \cdot \Pi^{\star\top} A \Pi^\star + \sqrt{1 - \rho^2} \cdot H, \tag{2.3}$$

where H is an independent copy of A, and $\Pi^\star$ is the $n \times n$ matrix representation of permutation $\pi^\star$, that is $\Pi^\star_{u,v} = \mathbb{1}_{v = \pi^\star(u)}$.

Detection problem A most recent paper ([WXY20]) studies fundamental limits for detection, both in correlated Gaussian weighted and correlated Erdős-Rényi graphs. This time, the problem is as follows: *given A, B, are we able to distinguish between model* (2.3) *and a null model, where the two graphs are just independent Gaussian weighted graphs?* Intuitively, this problem is less demanding than that of exact alignment, since the task is to detect – wherever in the graph – the presence of a hidden planted alignment. Under the same model (2.3), Y. Wu, J. Xu and S. Yu showed that strong detection is feasible with high probability if $n\rho^2 \ge 4 \log n$, whereas it is impossible if $n\rho^2 \le (4 - \varepsilon) \log n$ for some $\varepsilon > 0$. Their

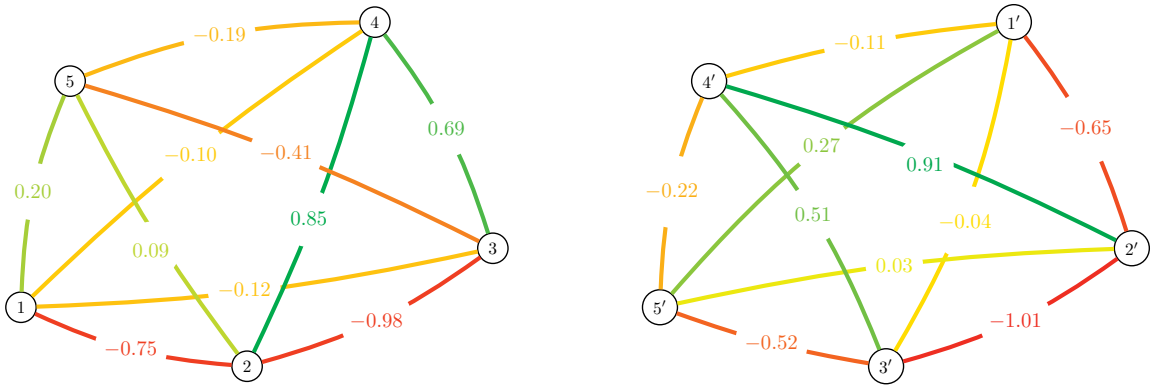


Figure 2.1 – *A sample from model* (2.3) *with* $n = 5$. *For representation, edges are colored according to their weights, and the underlying alignment is* $u \mapsto u'$ *for* $u \in \{1, 2, 3, 4, 5\}$.

study builds on an analysis of the likelihood ratio, as often done in detection problems. The contribution of this chapter is to show that this sharp detection threshold is also that of exact reconstruction. Interestingly, for Gaussian weighted graph alignment, the problem of reconstruction is in fact fundamentally not more difficult than that of detection.

After this study was completed, the author was made aware of recent and independent work conducted by [WXY21], which also obtains – among other things – the results of this study, albeit with different proof techniques.

2.1.2. Main results

In the sequel, we work with the correlated Gaussian Wigner model described in (2.3), and establish the precise (sharp) threshold for exact recovery of $\pi^\star$ in this model.

Theorem 2.1 (Achievability part)**.** *If for* n *large enough*

$$\rho^2 \geq \frac{(4+\varepsilon)\log n}{n} \tag{2.4}$$

for some $\varepsilon > 0$*, then there is an estimator (namely, the MAP estimator)* $\hat{\pi}$ *of* $\pi^\star$ *given* A, B *such that* $\hat{\pi} = \pi^\star$ *with probability* $1 - o(1)$.

Theorem 2.2 (Converse part)**.** *Conversely, if*

$$\rho^2 \leq \frac{4\log n - \log\log n - \omega(1)}{n} \tag{2.5}$$

then any estimator $\hat{\pi}$ *of* π *given* A, B *verifies* $\hat{\pi} = \pi^\star$ *with probability* $o(1)$.

Computational limits of exact recovery For the correlated Gaussian Wigner model (2.3), several algorithms have been studied, usually as a first step in order to analyze further graph alignment algorithms. The state-of-the-art polynomial-time algorithms are either based on *degree profiles* [DMWX21], or on a spectral method [FMWX19a]. In both cases, these methods require the noise parameter $\sqrt{1-\rho^2}$ to be $O\left(\log^{-1} n\right)$. In Chapter 3, we will study a simpler algorithm with lower computational complexity, requiring $\sqrt{1-\rho^2}$ to be $O(n^{-7/6})$ [GLM22]. In any case, ρ needs to tend to 1, and the regimes in which these methods work well are far from the fundamental limits established in Theorems 2.1 and 2.2. The main result of this chapter thus corroborates the idea that matrix alignment may be computationally hard even in the feasibility regime. In other words, the hard phase can be conjectured to be wide

for this reconstruction problem. Proving a result of that form however remains a very thorny question.

Organization of the chapter We first some notations at the beginning of Section 2.2, and then establish a control on correlations between energies of permutations, using Hanson-Wright inequality. The achievability result is proved in Section 2.3: after showing that the classical first moment method fails, we take advantage of the correlation structure established before to handle the sharp bound. Then, second moment method is applied in Section 2.4 to show that lots of small perturbations of the true underlying permutation have lower energies, establishing the converse bound. Finally, some additional proofs are deferred to Appendix 2.A. The proof techniques are not far from those used by [ECK19], the main novelty being the use of correlation of energies, which is essential to both achievability and impossibility result.

2.2. Preliminaries

2.2.1. Definitions and notations

Recall that for any positive integer n, $[n] := \{1, 2, \dots, n\}$. For two positive sequences $\{u_n\}$ and $\{v_n\}$, denote $u_n = O(v_n)$ if there exists $C > 0$ such that $u_n \leq Cv_n$ for all n. We will also write $u_n = o(v_n)$ (resp. $u_n = \omega(v_n)$) if $u_n/v_n \to 0$ (resp. $v_n/u_n \to 0$). All limits considered are taken when $n \to \infty$.

Linear algebra. We work with the canonical euclidean norm $\|\cdot\|$ on $\mathbb{R}^n$, and $\langle \cdot, \cdot \rangle$ the canonical inner product on $\mathbb{R}^n$ or $\mathbb{R}^{n\times n}$. For any $n \times n$ matrix M with real entries, its *Frobenius norm* $\|M\|_F$ and its *operator norm* $\|M\|_{\mathrm{op}}$ are defined as follows:

$$\|M\|_F := \left(\sum_{1\leq u,v\leq n} A_{u,v}^2 \right)^{1/2} \quad \text{and} \quad \|M\|_{\mathrm{op}} := \sup_{X\in\mathbb{R}^n\setminus\{0\}} \frac{\|MX\|}{\|X\|}.$$

Note that for any normal matrix (that is, if $M^TM = MM^T$), then $\|M\|_{\mathrm{op}}$ equals $\rho(M)$, the spectral radius of M.

Probability. When working with model (2.3), we will denote by $\mathbb{P}_A$ (resp. $\mathbb{E}_A$) the conditional probability (resp. the conditional expectation) with respect to the random matrix A. We recall that $\mathcal{N}(\mu, v)$ denotes a Gaussian variable (resp. vector) with mean μ and variance (resp. covariance matrix) v. Such a Gaussian variable (resp. vector) is called *standard* if $\mu = 0$ and $v = 1$ (resp. v is the identity matrix). We say that an event $\mathcal{A}_n$ happens *with high probability (w.h.p)* if $\mathbb{P}(\mathcal{A}_n) \to 1$ when $n \to \infty$.

Permutations. We denote by $\mathcal{S}_m$ the set of permutations of $[m]$. To any permutation $\sigma \in \mathcal{S}_m$, we can associate its $m \times m$ matrix representation Σ defined by $\Sigma_{u,v} = \mathbb{1}_{v=\sigma(u)}$. Define $\mathcal{F}_\sigma$ the set of *fixed points* of σ:

$$\mathcal{F}_\sigma := \{u \in [m], \sigma(u) = u\}, \tag{2.6}$$

and denote $f_\sigma := |\mathcal{F}_\sigma|$. Similarly, we define the set of *unfixed points* of σ:

$$\mathcal{D}_\sigma := [m] \setminus \mathcal{F}_\sigma = \{i \in [m], \sigma(i) \neq i\}, \tag{2.7}$$

and we denote $d_\sigma := |\mathcal{D}_\sigma|$. For any $d \in \{0, \dots, m\}$ we define $\mathcal{S}_{m,d}$ the set of permutations of

$\mathcal{S}_m$ with exactly d unfixed points. Note that $|\mathcal{S}_{m,1}| = 0$ and that we have the inequality

$$|\mathcal{S}_{m,d}| = \binom{m}{m-d} |\{\sigma \in \mathcal{S}_d, F_\sigma = 0\}| \leq \binom{m}{m-d} d! \leq m^d. \tag{2.8}$$

We recall that similarity between two permutations $\sigma, \sigma' \in \mathcal{S}_n$ is measured by their *overlap*:

$$\mathrm{ov}(\sigma, \sigma') := \frac{1}{n} \sum_{u=1}^{n} \mathbf{1}_{\sigma(u)=\sigma'(u)} = \frac{1}{n} f_{\sigma^{-1}\circ\sigma'}.$$

Observe that on a graph of size n, each permutation σ of the vertices $[n]$ has a natural extension to a canonical permutation on edges $\sigma^{\mathrm{E}} : \binom{[n]}{2} \to \binom{[n]}{2}$ defined as follows:

$$\sigma^{\mathrm{E}} : e = \{u, v\} \mapsto \sigma^{\mathrm{E}}(e) = \{\sigma(u), \sigma(jv)\}.$$

Note that the mapping $\sigma \mapsto \sigma^{\mathrm{E}}$ is one-to-one as soon as $n \geq 3$, since for all $u \in [n]$ and $v \neq v' \in [n] \setminus \{u\}$, edges $\sigma^{\mathrm{E}}(\{u, v\})$ and $\sigma^{\mathrm{E}}(\{u, v'\})$ have only one node in common, which is $\sigma(u)$. We will use the notation $\mathcal{F}_\sigma^{\mathrm{E}} := \mathcal{F}_{\sigma^{\mathrm{E}}}$ (resp. $\mathcal{D}_\sigma^{\mathrm{E}} := \mathcal{D}_{\sigma^{\mathrm{E}}}$) the set of *fixed edges* (resp. *unfixed edges*) of σ. Similarly we denote $f_\sigma^{\mathrm{E}} := f_{\sigma^{\mathrm{E}}}$ and $f_\sigma^{\mathrm{E}} := d_{\sigma^{\mathrm{E}}}$, for brievity.

Note that d_σ^{E} and are d_σ are closely tied, since for all $\sigma \in \mathcal{S}_n$, we have the inequality

$$d_\sigma \left(n - \frac{d_\sigma}{2} \right) \leq d_\sigma^{\mathrm{E}} \leq d_\sigma \left(n - \frac{d_\sigma - 1}{2} \right). \tag{2.9}$$

Indeed, observe that

(*i*) the number of fixed edges is at least the number of pairs of fixed points, and

(*ii*) the number of fixed edges is exactly the number of pairs of fixed points plus the number of pairs $(u, v), u < v$ that are exchanged by σ (that is, the number of transpositions), this number being at most $d_\sigma/2$.

These remarks give that

$$\binom{n - d_\sigma}{2} \leq \binom{n}{2} - d_\sigma^{\mathrm{E}} \leq \binom{n - d_\sigma}{2} + \frac{d_\sigma}{2},$$

which directly implies (2.9).

Remark 2.2.1. *Note that inequality* (2.9) *gives the almost sure equivalents* $d_\sigma^{\mathrm{E}} \sim d_\sigma n$ *when* $d_\sigma = o(n)$*, and* $d_\sigma^{\mathrm{E}} \sim \frac{1}{2}\alpha(2-\alpha)n^2$ *when* $d_\sigma = \alpha n$*. In any case,* $d_\sigma^{\mathrm{E}} \in \left[\frac{1}{2} d_\sigma n, d_\sigma n\right]$.

2.2.2. MAP estimation, relative energy of permutations

Since $\pi^\star$ is uniformly chosen, we work in a Bayesian setting: let us evaluate the posterior probability density of $\pi^\star$ given A, B:

$$\begin{aligned} p_{\pi^\star|A,B}(\pi|a,b) &\propto p_{\pi^\star,A,B}(\pi, a, b) \\ &\propto \exp\left(-\frac{1}{2(1-\rho^2)} \sum_{1 \leq i < j \leq n} \left(B_{\pi(u),\pi(v)} - \rho A_{u,v} \right)^2 \right), \end{aligned}$$

where $\propto$ indicates equality up to some factors that do not depend on σ. Define the *loss function*

$$\mathcal{E}(\pi, A, B) := \sum_{1 \leq i < j \leq n} \left(B_{\pi(u),\pi(v)} - \rho A_{u,v} \right)^2. \tag{2.10}$$

This loss function can also be viewed as the *energy* associated with permutation π. Note that the posterior distribution is a Gibbs measure corresponding to this energy $\mathcal{E}$, with inverse temperature $\beta = \frac{1}{2(1-\rho^2)}$. The MAP (maximum a posteriori) estimator is thus

$$\hat{\pi}_{\mathrm{MAP}} := \arg\max_{\pi} p_{\pi^\star|A,B}\left(\pi|A,B\right) = \arg\min_{\pi} \mathcal{E}(\pi, A, B), \tag{2.11}$$

where the minimum is taken over all permutations $\pi \in \mathcal{S}_n$. As previously stated in Section 1.3.3, the above formulation (2.11) is standard in the literature of graph alignment and meets the classical QAP formulation (2.2), since

$$\arg\min_{\pi} \mathcal{E}(\pi, A, B) = \arg\max_{\Pi} \langle A, \Pi B \Pi^T \rangle.$$

Theory from Bayesian optimal estimation guarantees that the best possible estimator for our exact reconstruction problem, in the Bayes risk sense, is $\hat{\pi}_{\mathrm{MAP}}$. Thus, if MAP estimator fails with high probability, then no estimator can succeed. This is why this estimator is often studied in exact reconstruction problems, as already done in previous works ([CK17, CMK18, ECK19]).

From now on we work conditionally on $\pi^\star$ which can always be assumed to be id without loss of generality. More precisely, we will make the variable change $\sigma = \pi^\star \circ \pi^{-1}$; writing B as a function of σ, A and H, (2.10) becomes

$$\begin{aligned}\mathcal{E}(\sigma, A, H) = \rho^2 \sum_{1\le i<j\le n} \left(A_{u,v} - A_{\sigma(u),\sigma(v)}\right)^2 - 2\rho\sqrt{1-\rho^2} \sum_{1\le i<j\le n} H_{u,v}\left(A_{u,v} - A_{\sigma(u),\sigma(v)}\right) \\ + (1-\rho^2) \sum_{1\le i<j\le n} H_{u,v}^2.\end{aligned}$$

The loss function $\mathcal{E}$ applied to the ground truth $\pi = \pi^\star$ – that is $\sigma = \mathrm{id}$ – gives the energy reference $(1-\rho^2)\sum_{1\le i<j\le n} H_{u,v}^2$. In order to compare any π with $\pi^\star$ – or any σ with id – we further define the *relative energy* of a permutation $\sigma \in \mathcal{S}_n$:

$$\begin{aligned}\delta(\sigma) &:= \mathcal{E}(\sigma, A, H) - \mathcal{E}(\mathrm{id}, A, H) \\ &= \rho^2 \sum_{1\le i<j\le n} \left(A_{u,v} - A_{\sigma(u),\sigma(v)}\right)^2 - 2\rho\sqrt{1-\rho^2} \sum_{1\le i<j\le n} H_{u,v}\left(A_{u,v} - A_{\sigma(u),\sigma(v)}\right). \end{aligned} \tag{2.12}$$

We next omit in our notations the dependency on A and H of $\delta(\sigma)$.

Remark 2.2.2. *This relative energy δ, also introduced by [CK17] for Erdős-Rényi graph alignment, is a measurement of the quality of a proposed alignment: $\delta(\sigma) \le 0$ means that $\sigma^{-1} \circ \pi^\star$ is a better alignment than $\pi^\star$ for A and B in the posterior sense. A crucial set is then*

$$\mathcal{Q} := \{\sigma \in \mathcal{S}_n,\, \delta(\sigma) \le 0\}.$$

Points of $\mathcal{Q}$ are alignments on which the posterior distribution puts important weights – at least greater weights than that of the ground truth – or equivalently points of low energy. Note that $\mathrm{id} \in \mathcal{Q}$.

In view of (2.12), conditionally on A, $\delta(\sigma)$ is as follows:

$$\delta(\sigma) = \rho^2 v_\sigma - 2\rho\sqrt{1-\rho^2} X_\sigma, \tag{2.13}$$

where

$$v_\sigma := \sum_{1\le i<j\le n} \left(A_{u,v} - A_{\sigma(u),\sigma(v)}\right)^2,$$

and $X = (X_\sigma)_{\sigma \in \mathcal{S}_n}$ is a Gaussian vector, centered, with covariance given by

$$\mathrm{Cov}(X_\sigma, X_{\sigma'}) = \sum_{1 \le i < j \le n} \left(A_{u,v} - A_{\sigma(u),\sigma(v)}\right)\left(A_{u,v} - A_{\sigma'(i),\sigma'(j)}\right) := c_{\sigma,\sigma'}.$$

Note that for all $\sigma \in \mathcal{S}_n$, $c_{\sigma,\sigma} = v_\sigma$. Elaborating on the correlation structure of these relative energies is the object of the end of this section.

2.2.3. Control of covariance structure of relative energies

For all $\sigma, \sigma' \in \mathcal{S}_n$, $c_{\sigma,\sigma'}$ can be written as follows

$$c_{\sigma,\sigma'} = \sum_{e \in \binom{[n]}{2}} \left(A_e - A_{\sigma^E(e)}\right)\left(A_e - A_{\sigma'^E(e)}\right)$$

and satisfies

$$\mathbb{E}\left[c_{\sigma,\sigma'}\right] = \left|\mathcal{D}_\sigma^{\mathrm{E}} \cap \mathcal{D}_{\sigma'}^{\mathrm{E}}\right| + \left|\mathcal{D}_\sigma^{\mathrm{E}} \cap \mathcal{D}_{\sigma'}^{\mathrm{E}} \cap \mathcal{F}_{\sigma^{-1}\circ\sigma'}^{\mathrm{E}}\right|.$$

In particular,

$$\mathbb{E}\left[v_\sigma\right] = d_\sigma^{\mathrm{E}} + d_\sigma^{\mathrm{E}} = 2d_\sigma^{\mathrm{E}}.$$

Random variables $c_{\sigma,\sigma'}$ only depend on the entries of A, which are Gaussian. Moreover, $c_{\sigma,\sigma'}$ being a quadratic form evaluated on a Gaussian vector, it can be controlled using Hanson-Wright inequality:

Lemma 2.2.1 (Hanson-Wright inequality ([HW71]))**.** *Let X be a standard Gaussian vector, and M a deterministic matrix. Then there exists a universal constant $c > 0$ such that with probability at least $1 - 2\delta$:*

$$\left|X^T M X - \mathrm{Tr} M\right| \le c\left(\|M\|_F \sqrt{\log(1/\delta)} + \|M\|_{\mathrm{op}} \log(1/\delta)\right). \tag{2.14}$$

We refer to [HW71] for a proof. Inequality (2.14) used in our context leads to the following

Corollary 2.2.1. *There exists a universal constant $C > 0$ such that with high probability, for every $d \in \{2, \dots, n\}$, for all $\sigma, \sigma' \in \mathcal{S}_{n,d}$,*

$$\left|c_{\sigma,\sigma'} - \left|\mathcal{D}_\sigma^{\mathrm{E}} \cap \mathcal{D}_{\sigma'}^{\mathrm{E}}\right| - \left|\mathcal{D}_\sigma^{\mathrm{E}} \cap \mathcal{D}_{\sigma'}^{\mathrm{E}} \cap \mathcal{F}_{\sigma^{-1}\circ\sigma'}^{\mathrm{E}}\right|\right| \le Cd\sqrt{n \log n}.$$

Proof. We first make the following observation: for any $\sigma, \sigma' \in \mathcal{S}_n$,

$$\begin{aligned} c_{\sigma,\sigma'} &= \sum_e \left(A_e - A_{\sigma(e)}\right)\left(A_e - A_{\sigma'(e)}\right) \\ &= A^T (I_N - \Sigma)^T (I_N - \Sigma') A, \end{aligned}$$

where $A = (A_e)_e$ is viewed as a standard Gaussian vector of size $N = \binom{n}{2}$, and Σ (resp. Σ') is the $N \times N$ permutation matrix associated with σ^E (resp. σ'^E). Note that

$$\begin{aligned} \mathrm{Tr}((I_N - \Sigma)^T (I_N - \Sigma')) &= N - f_\sigma^{\mathrm{E}} - f_{\sigma'}^{\mathrm{E}} + f_{\sigma^{-1}\circ\sigma'}^{\mathrm{E}} \\ &\overset{(a)}{=} \left|\mathcal{D}_\sigma^{\mathrm{E}} \cap \mathcal{D}_{\sigma'}^{\mathrm{E}}\right| + \left|\mathcal{D}_\sigma^{\mathrm{E}} \cap \mathcal{D}_{\sigma'}^{\mathrm{E}} \cap \mathcal{F}_{\sigma^{-1}\circ\sigma'}^{\mathrm{E}}\right|, \end{aligned}$$

where (a) is obtained by noticing that

$$\left|\mathcal{D}_\sigma^{\mathrm{E}} \cap \mathcal{D}_{\sigma'}^{\mathrm{E}}\right| + \left|\mathcal{D}_\sigma^{\mathrm{E}} \cap \mathcal{D}_{\sigma'}^{\mathrm{E}} \cap \mathcal{F}_{\sigma^{-1}\circ\sigma'}^{\mathrm{E}}\right| = d_\sigma^{\mathrm{E}} + d_{\sigma'}^{\mathrm{E}} - \left|\mathcal{D}_\sigma^{\mathrm{E}} \cup \mathcal{D}_{\sigma'}^{\mathrm{E}}\right| + f_{\sigma^{-1}\circ\sigma'}^{\mathrm{E}} - \left|\mathcal{F}_\sigma^{\mathrm{E}} \cup \mathcal{F}_{\sigma'}^{\mathrm{E}}\right|$$

and that $\left|\mathcal{D}_\sigma^{\mathrm{E}} \cup \mathcal{D}_{\sigma'}^{\mathrm{E}}\right| + \left|\mathcal{F}_\sigma^{\mathrm{E}} \cup \mathcal{F}_{\sigma'}^{\mathrm{E}}\right| = N$. For a fixed d and $\sigma, \sigma' \in \mathcal{S}_{n,d}$, one has

$$\begin{aligned}\|(I_N - \Sigma)^T (I_N - \Sigma')\|_F \leq \|(I_N - \Sigma')\|_F + \|\Sigma^T (I_N - \Sigma')\|_F &= 2\|(I_N - \Sigma')\|_F \\ &\leq 2\sqrt{2d_{\sigma'}^{\mathrm{E}}} \\ &\leq 2\sqrt{2dn},\end{aligned}$$

where we used (2.9) in the last step. One also has

$$\begin{aligned}\|(I_N - \Sigma)^T (I_N - \Sigma')\|_{\mathrm{op}} &\leq \rho(I_N - \Sigma) \times \rho(I_N - \Sigma') \\ &\leq 2 \times 2 = 4.\end{aligned}$$

Taking $\delta = n^{-(2d+2)}$, Lemma 2.2.1 gives that with probability at least $1 - 2\delta$,

$$\begin{aligned}\left|c_{\sigma,\sigma'} - \left|\mathcal{D}_\sigma^{\mathrm{E}} \cap \mathcal{D}_{\sigma'}^{\mathrm{E}}\right| - \left|\mathcal{D}_\sigma^{\mathrm{E}} \cap \mathcal{D}_{\sigma'}^{\mathrm{E}} \cap \mathcal{F}_{\sigma^{-1}\circ\sigma'}^{\mathrm{E}}\right|\right| &\leq c\left(2\sqrt{2}\sqrt{d(2d+2)}\sqrt{n\log n} + 4(2d+2)\log n\right) \\ &\leq Cd\sqrt{n\log n}, \qquad (2.15)\end{aligned}$$

for some universal constant $C > 0$. The proof is concluded by checking that this inequality holds w.h.p. for all d and $\sigma, \sigma' \in \mathcal{S}_{n,d}$: the probability that at least one pair (σ, σ') contradicts (2.15) is upper bounded by

$$n \times |\mathcal{S}_{n,d}|^2 \times 2\delta \leq 2n^{1+2d-2d-2} = o(1).$$

□

In the rest of the chapter we define the event

$$\mathcal{A} := \left\{\forall d \in [n], \forall \sigma, \sigma' \in \mathcal{S}_{n,d}, \left|c_{\sigma,\sigma'} - \left|\mathcal{D}_\sigma^{\mathrm{E}} \cap \mathcal{D}_{\sigma'}^{\mathrm{E}}\right| - \left|\mathcal{D}_\sigma^{\mathrm{E}} \cap \mathcal{D}_{\sigma'}^{\mathrm{E}} \cap \mathcal{F}_{\sigma^{-1}\circ\sigma'}^{\mathrm{E}}\right|\right| \leq Cd\sqrt{n\log n}\right\}, \qquad (2.16)$$

which happens with probability $1 - o(1)$ by Corollary 2.2.1.

2.3. Achievability result

In this section, we establish the result of Theorem 2.1.

2.3.1. Failure of first moment method

For the achievability result, the first strategy is to use the union bound (or first moment method) to show that under condition (2.4) of Theorem 2.1,

$$\mathbb{P}\left(\text{MAP fails}\right) = \mathbb{P}\left(\hat{\pi}_{\mathrm{MAP}} \neq \pi\right) = o(1).$$

As described hereafter, this naive method does not give the correct bound. Indeed, let us evaluate $\mathbb{P}\left(\delta(\sigma) \leq 0\right)$ for a given $\sigma \neq \mathrm{id}$. In view of the conditional distribution (2.13) of $\delta(\sigma)$ we have

$$\begin{aligned}\mathbb{P}\left(\delta(\sigma) \leq 0\right) &= \mathbb{E}\left[\mathbb{E}_A\left[\mathbb{1}_{\delta(\sigma)\leq 0}\right]\right] = \mathbb{E}\left[\mathbb{P}_A\left(\rho^2 v_\sigma - 2\rho\sqrt{1-\rho^2}X_\sigma \leq 0\right)\right] \\ &= \mathbb{E}\left[\mathbb{P}_A\left(\rho^2 v_\sigma - 2\rho\sqrt{1-\rho^2}\sqrt{v_\sigma}\cdot\mathcal{N}(0,1) \leq 0\right)\right] \\ &= \mathbb{E}\left[\mathbb{P}_A\left(\mathcal{N}(0,1) \geq \frac{\rho\sqrt{v_\sigma}}{2\sqrt{1-\rho^2}}\right)\right] \leq \mathbb{E}\left[\exp\left(-\frac{\rho^2}{8(1-\rho^2)}v_\sigma\right)\right],\end{aligned}$$

where we used standard Gaussian concentration in the last inequality: $\mathbb{P}\left(\mathcal{N}(0,1) \geq t\right) \leq \exp(-t^2/2)$. Note that on event $\mathcal{A}$ defined in (2.16) and inequality (2.9),

$$\forall d \in [n], \forall \sigma \in \mathcal{S}_{n,d},\, v_\sigma \geq 2d_\sigma^{\mathrm{E}} - Cd_\sigma\sqrt{n\log n} \geq d_\sigma^{\mathrm{E}}\left(2 - 2\varepsilon_n\right),$$

setting $\varepsilon_n = 2C\sqrt{\log n/n}$. Union bound then gives

$$\begin{aligned}
\mathbb{P}\left(\text{MAP fails}\right) &\leq \mathbb{P}\left(\exists \sigma \in \mathcal{S}_n \setminus \{\mathrm{id}\},\, \delta(\sigma) \leq 0\right) \\
&\leq o(1) + \sum_{\sigma \in \mathcal{S}_n \setminus \{\mathrm{id}\}} \mathbb{E}\left[\exp\left(-\frac{\rho^2}{8(1-\rho^2)} v_\sigma\right) \mathbb{1}_{\mathcal{A}}\right] \\
&\leq o(1) + \sum_{\sigma \in \mathcal{S}_n \setminus \{\mathrm{id}\}} \exp\left(-\frac{\rho^2}{8(1-\rho^2)}(2-2\varepsilon_n) d_\sigma^{\mathrm{E}}\right) \\
&\leq o(1) + \sum_{\sigma \in \mathcal{S}_n \setminus \{\mathrm{id}\}} \exp\left(-\frac{\rho^2}{4}(1-\varepsilon_n) d_\sigma^{\mathrm{E}}\right),
\end{aligned}$$

where we used $1/(1-\rho^2) > 1$ in the last step. Let us now study the last sum, distinguishing the terms according to $d := d_\sigma$:

- As long as $d = o(n)$, by Remark 2.2.1, the terms behave like $\exp\left(-\frac{\rho^2}{4}(1-\varepsilon_n)dn\right)$. By (2.8), $\log|\mathcal{S}_{n,d}| \leq d\log n$ so the partial sum is small if $\frac{\rho^2}{4}(1-\varepsilon_n)n - \log n > 0$, which gives the necessary condition $\rho^2 \geq 4\frac{\log n}{n}$.

- However, the situation is different when it comes to large values of d. For instance, let us study the contribution of *derangements* to the sum (that is, σ such that $d_\sigma = n$). Note that these derangements are very numerous (their number is $\sim e^{-1}n!$). Again by Remark 2.2.1, their contribution is thus of order

  $$e^{-1}n!\exp\left(\rho^2(1-\varepsilon_n)n^2/8(1-o(1))\right) = \exp\left(\left(n\log n - \rho^2 n^2/8\right)(1-o(1))\right),$$

  which gives a more restrictive condition: $\rho^2 \geq 8\frac{\log n}{n}$.

As seen here-above, this naive first moment method enables to ensure feasibility of exact reconstruction only in the regime where $\rho^2 \geq 8\frac{\log n}{n}$, which is not the optimal one. This bound is actually quite rough here, because the variables are substantially correlated when d gets large and their contributions make the first moment explode. We take advantage of these correlations in the next section in order to get access to the sharp bound.

2.3.2. Improving the first moment method with correlations.

For all $d \in \{2, \ldots, n\}$, define $\mathcal{E}_d$ the event:

$$\mathcal{E}_d := \left\{\exists \sigma \in \mathcal{S}_{n,d},\, \delta(\sigma) \leq 0\right\}.$$

In this Section we will assume that

$$\rho \geq (2+\varepsilon)\sqrt{\frac{\log n}{n}},$$

for some $\varepsilon > 0$. Recall that we work on the event $\mathcal{A}$ defined in (2.16), and that conditionally on entries of matrix A, we can write

$$\delta(\sigma) = \rho^2 v_\sigma - 2\rho\sqrt{1-\rho^2}X_\sigma, \tag{2.17}$$

where $X = (X_\sigma)_{\sigma\in\mathcal{S}_{n,d}}$ is a Gaussian vector, centered, with covariance given by $\mathrm{Cov}(X_\sigma, X_{\sigma'}) = c_{\sigma,\sigma'}$. Also note that on event $\mathcal{A}$, for all $d \le \alpha n$ and $\sigma \in \mathcal{S}_{n,d}$, inequality (2.9) gives

$$v_\sigma = (1 - o(1))2dn(1 - \alpha/2). \tag{2.18}$$

In view of (2.18), as previously done in Section 2.3.1, naive first moment method may suffice for $d \le \alpha n$:

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{2\le d\le \alpha n} \mathcal{E}_d\right) &\le o(1) + \sum_{d=2}^{\alpha n} |\mathcal{S}_{n,d}| \times \mathbb{P}\left(\mathcal{N}(0,1) \ge \frac{\rho\sqrt{v_\sigma}}{2\sqrt{1-\rho^2}} \cap \mathcal{A}\right) \\
&\le o(1) + \sum_{d=2}^{\alpha n} |\mathcal{S}_{n,d}| \times \mathbb{P}\left(\mathcal{N}(0,1) \ge (1+\varepsilon/2)\sqrt{2d\log n(1-\alpha/2)}(1-o(1))\right) \\
&\le o(1) + \sum_{d=2}^{\alpha n} \exp\left(d\log n - d\log n(1+\varepsilon)(1-\alpha/2) + o(d\log n)\right),
\end{aligned}$$

which is $o(1)$ as soon as $\alpha < \alpha_0 := \frac{2\varepsilon}{1-\varepsilon/2}$. It then remains to control the probabilities $\mathbb{P}(\mathcal{E}_d)$ for $d \ge \alpha_0 n$. As mentioned earlier, we take advantage of the correlation structure in (2.17). More precisely, we show that all variables X_σ at a given level $d = \alpha n$ have substantial positive covariance when compared to their variance – of order $\alpha(2-\alpha)n^2$ on $\mathcal{A}$ by (2.18) – as shown in Figure 2.2. To do so, we derive an appropriate lower bound for $c_{\sigma,\sigma'}$ for $\sigma, \sigma' \in \mathcal{S}_{n,\alpha n}$. This is the scope of the following Lemma:

Lemma 2.3.1. *With high probability, there exists a universal constant $C_1 > 0$ such that for any $d = \alpha n$ with fixed $\alpha > 0$ and $\sigma, \sigma' \in \mathcal{S}_{n,\alpha n}$:*

$$\mathrm{Cov}(X_\sigma, X_{\sigma'}) = c_{\sigma,\sigma'} \ge f(\alpha)n^2 - C_1 n^{3/2}\log^{1/2} n,$$

with

$$f(\alpha) := \begin{cases} \alpha^2 & \text{if } \alpha < 1/2 \\ \alpha^2 - \frac{1}{2}(2\alpha-1)^2 & \text{if } \alpha \ge 1/2 \end{cases} \tag{2.19}$$

Thus for any $\varepsilon' > 0$, with high probability, for any $d = \alpha n$ with fixed $\alpha > 0$,

$$\max_{\sigma\in\mathcal{S}_{n,\alpha n}} X_\sigma \le \sqrt{2\alpha\left(\alpha(2-\alpha) - f(\alpha)\right)}n^{3/2}\log^{1/2} n + (2+\varepsilon')n\log^{1/2} n.$$

The proof of this Lemma is obtained by working on event $\mathcal{A}$ defined in (2.16), and establishing a lower bound on $\left|\mathcal{D}^{\mathrm{E}}_\sigma \cap \mathcal{D}^{\mathrm{E}}_{\sigma'}\right|$, which is simply the number of edges that are deranged both by σ^E and σ'^E. It can be found in Appendix 2.A.1.

Then, since $f(\alpha) \le \alpha(2-\alpha)$ with elementary computations, according to Lemma 2.3.1, there is an event $\mathcal{B}$ of probability $1 - o(1)$ such that

$$\max_{\sigma\in\mathcal{S}_{n,d}} X_\sigma \le (1+o(1))\sqrt{2\alpha\left(\alpha(2-\alpha) - f(\alpha)\right)}n^{3/2}\log^{1/2} n$$

holds for all $d = \alpha n$ with $\alpha > \alpha_0$. Note that on event $\mathcal{A}\cap\mathcal{B}$, for all $d = \alpha n$ and $\sigma \in \mathcal{S}_{n,d}$,

$$\begin{aligned}
\rho^{-1}\delta(\sigma) &\ge \rho v_\sigma - 2\sqrt{1-\rho^2}\max_{\sigma\in\mathcal{S}_{n,d}} X_\sigma \\
&\ge (1+o(1))n^{3/2}\log^{1/2} n\left[(2+\varepsilon)\alpha(2-\alpha) - 2\sqrt{2\alpha\left(\alpha(2-\alpha) - f(\alpha)\right)}\right] \\
&\ge (1+o(1)) \times 2 \times \left[\alpha(2-\alpha) - \sqrt{2\alpha\left(\alpha(2-\alpha) - f(\alpha)\right)}\right] n^{3/2}\log^{1/2} n \ge 0,
\end{aligned}$$

for n large enough, since it can be easily checked (see Appendix 2.A.3) that

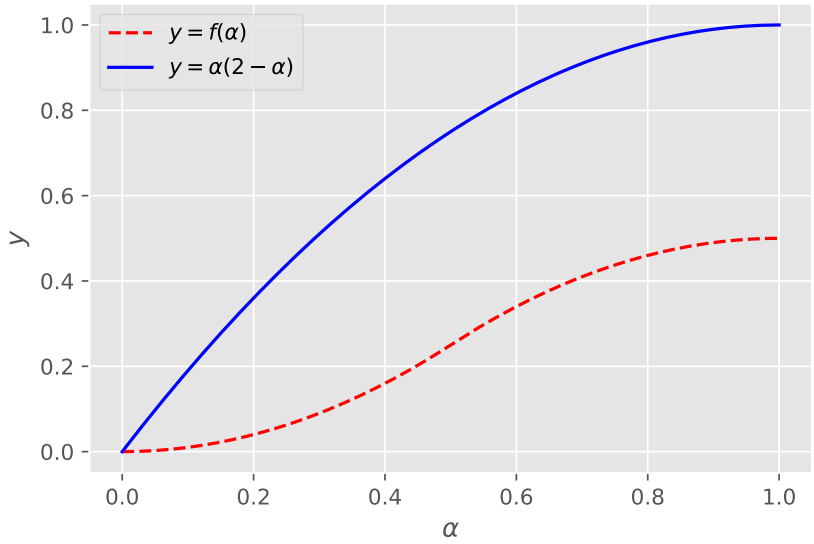


Figure 2.2 – *Plot on* $[0,1]$ *of normalized variance* $\alpha(2-\alpha)$*, together with the lower bound on the normalized covariance (function* f*) defined by* (2.19).

Lemma 2.3.2. *For every* $\alpha \in [0,1]$*,*

$$\alpha(2-\alpha) - \sqrt{2\alpha\left(\alpha(2-\alpha) - f(\alpha)\right)} \geq 0. \tag{2.20}$$

Previous computations hence give that $\mathbb{P}\left(\bigcup_{d\geq\alpha n} \mathcal{E}_d\right) \leq 1 - \mathbb{P}(\mathcal{A}\cap\mathcal{B}) = o(1)$, and ends the proof of Theorem 2.1.

2.4. Converse bound: second moment method for transpositions

In this section, we prove Theorem 2.2. As claimed in the introduction, theory from Bayesian optimal estimation guarantees that the best possible estimator for our exact reconstruction problem, in the Bayes risk sense, is $\hat{\pi}_{\mathrm{MAP}}$. We will show that under assumption (2.5) of Theorem 2.2, this MAP estimator fails with high probability, which implies that no estimator can succeed.

This converse bound is obtained by a second moment argument, showing that with high probability, there are lots of permutation $\tau \neq \mathrm{id}$ – in fact, transpositions – such that $\delta(\tau)$ is negative, that is, $\tau^{-1}\circ\pi^\star$ is a substantially better alignment than $\pi^\star$, with lowest energy. Let us denote $\mathcal{T}_n \subset \mathcal{S}_n$ the set of all permutations of $[n]$ that are transpositions. For all $\tau \in \mathcal{T}_n$, we have $d_\tau^{\mathrm{E}} = 2(n-2)$. Corollary 2.2.1 gives that the event

$$\mathcal{C} := \left\{\forall \tau,\tau' \in \mathcal{T}_n, \left|c_{\tau,\tau'} - \left|\mathcal{D}_\tau^{\mathrm{E}} \cap \mathcal{D}_{\tau'}^{\mathrm{E}}\right| - \left|\mathcal{D}_\tau^{\mathrm{E}} \cap \mathcal{D}_{\tau'}^{\mathrm{E}} \cap \mathcal{F}_{\tau\circ\tau'}^{\mathrm{E}}\right|\right| \leq C\sqrt{n\log n}\right\}$$

happens with probability $1-o(1)$ for $C>0$ large enough. In particular, on $\mathcal{C}$, for $C>0$ large enough,

$$\forall \tau \in \mathcal{T}_n,\ |v_\tau - 4n| \leq C\sqrt{n\log n}.$$

In this section we are working under the assumption (2.5) that we recall here:

$$\rho^2 \leq \frac{4\log n - \log\log n - \omega(1)}{n}$$

We are about to show the following: under condition (2.5), with high probability,

$$|\{\tau \in \mathcal{T}_n,\, \delta(\tau) < 0\}| = \omega(1). \tag{2.21}$$

To do so, we use the classical Paley-Zygmund inequality (Lemma 1.2.2 of Section 1.2.1) that implies (taking $c \to 0$ in Lemma 1.2.2) that if Y is a positve random variable such that $\mathbb{E}\left[Y^2\right] \sim \mathbb{E}\left[Y\right]^2$, then $Y \geq o(\mathbb{E}\left[Y\right])$ with high probability. Define

$$X := \sum_{\tau \in \mathcal{T}_n} \mathbf{1}_{\delta(\tau)<0}. \tag{2.22}$$

Using a standard coupling argument in (2.22), one can see that X is decreasing with ρ, thus we can assume without loss of generality that

$$\rho^2 = \frac{4\log n - \log\log n - a_n}{n}, \tag{2.23}$$

with a sequence $(a_n)_n$ such that $a_n = \omega(1)$ and $a_n = o(\log\log n)$, e.g. $a_n = \log\log\log n$. We compute the first moment of X, in view of the conditional distribution of $\delta(\tau)$ given in (2.13):

$$\begin{aligned}
\mathbb{E}\left[X\right] \geq \mathbb{E}\left[X\mathbf{1}_{\mathcal{C}}\right] &= \frac{n(n-1)}{2}\mathbb{E}\left[\mathbb{P}_A\left(\mathcal{N}(0,1) \geq \frac{\rho\sqrt{v_\tau}}{2\sqrt{1-\rho^2}} \cap \mathcal{C}\right)\right] \\
&\geq \frac{n(n-1)}{2}\mathbb{E}\left[(1-o(1))\mathbb{P}_A\left(\mathcal{N}(0,1) \geq \frac{1}{2}\sqrt{4\log n - \log\log n - a_n}\sqrt{4 - Cn^{-1/2}\log^{1/2} n}\right)\right] \\
&= \frac{n(n-1)}{2}\mathbb{E}\left[(1-o(1))\mathbb{P}_A\left(\mathcal{N}(0,1) \geq \sqrt{4\log n - \log\log n - a_n} - o(1)\right)\right] \\
&\sim \frac{n^2}{4\sqrt{2\pi}\sqrt{\log n}} \exp\left(-2\log n + \frac{\log\log n}{2} + \frac{a_n}{2}\right) \\
&= \frac{1}{4\sqrt{2\pi}} \exp\left(\frac{a_n}{2}\right) \to \infty.
\end{aligned}$$

Note that (2.23) is thus precisely the condition ensuring that $\mathbb{E}\left[X\mathbf{1}_{\mathcal{C}}\right] \to \infty$. The second moment argument computation being a little more technical, we encapsulate it into the following Lemma:

Lemma 2.4.1 (Second moment computation of $X\mathbf{1}_{\mathcal{C}}$)**.** *Let $Y := X\mathbf{1}_{\mathcal{C}}$. Under assumption* (2.23)*,*

$$\mathbb{E}\left[Y^2\right] \leq (1+o(1))\mathbb{E}\left[Y\right]^2.$$

Proof of Lemma 2.4.1. We represent a transposition τ by its only 2−cycle $(i\ j)$ with $i < j$. We then distinguish two cases in couples $\tau = (i\ j) \neq \tau' = (k\ \ell) \in \mathcal{T}_n$:

- We write $\tau \cap \tau' = \varnothing$ when τ and τ' have no common point in their 2−cycle: $i \neq k$ and $j \neq l$. When $\tau \in \mathcal{T}_n$ is fixed, note that

$$\left|\left\{\tau' \in \mathcal{T}_n,\, \tau \cap \tau' = \varnothing\right\}\right| = \frac{(n-2)(n-3)}{2}.$$

- We write $\tau \cap \tau' \neq \varnothing$ when τ and τ' are different but share one common point: for instance $\tau = (3\ 5)$ and $\tau = (5\ 11)$ verify $\tau \cap \tau' \neq \varnothing$. When $\tau \in \mathcal{T}_n$ is fixed, note that

$$\left|\left\{\tau' \in \mathcal{T}_n,\, \tau \cap \tau' \neq \varnothing\right\}\right| = 2(n-2).$$

Note that

$$\mathbb{E}\left[Y^2\right] = \mathbb{E}\left[Y\right] + \sum_{\tau\in\mathcal{T}_n}\sum_{\tau',\tau\cap\tau'=\varnothing} \mathbb{P}(\delta(\tau)<0,\delta(\tau')<0,\mathcal{C}) + \sum_{\tau\in\mathcal{T}_n}\sum_{\tau',\tau\cap\tau'\neq\varnothing} \mathbb{P}(\delta(\tau)<0,\delta(\tau')<0,\mathcal{C}).$$

We now evaluate these two sums. For this, we will need the following Lemma, which proof is deferred to Appendix 2.A.4.

Lemma 2.4.2 (Control of deviation probabilities for correlated Gaussians)**.** *Let Z_1, Z_2 be two Gaussian variables with mean 0, variance 1 and correlation $\alpha_n \in [0,1]$. For any t_n such that $t_n \to \infty$,*

(i) If $\alpha_n t_n \to 0$, then for n large enough

$$\mathbb{P}\left(Z_1 > t_n, Z_2 > t_n\right) \leq e^{-2t_n^2} + (1+o(1))\mathbb{P}\left(Z_1 > t_n\right)\mathbb{P}\left(Z_2 > t_n\right). \tag{2.24}$$

(ii) More generally,

$$\mathbb{P}\left(Z_1 > t_n, Z_2 > t_n\right) \leq (1+o(1))\frac{1+\alpha_n}{\sqrt{2\pi}\,t_n}\exp\left(-\frac{t_n^2}{1+\alpha_n}\right). \tag{2.25}$$

First case: $\tau\cap\tau'=\varnothing$**.** Without loss of generality we can assume that $\tau = (1\ 2)$ and $\tau' = (3\ 4)$. The following diagram shows the simple action of τ and τ' on an interesting (overlapping) subset of edges.

$$\begin{array}{ccc} \{1,3\} & \overset{\tau}{\longleftrightarrow} & \{2,3\} \\ {\scriptstyle\tau'}\updownarrow & & \updownarrow{\scriptstyle\tau'} \\ \{1,4\} & \overset{\tau}{\longleftrightarrow} & \{2,4\} \end{array}$$

We then see that $\left|\mathcal{D}_\tau^{\mathrm{E}}\cap\mathcal{D}_{\tau'}^{\mathrm{E}}\right| + \left|\mathcal{D}_\tau^{\mathrm{E}}\cap\mathcal{D}_{\tau'}^{\mathrm{E}}\cap\mathcal{F}_{\tau\circ\tau'}^{\mathrm{E}}\right| = 4+0 = 4$. So, denoting $\alpha_{\tau,\tau'} := \frac{c_{\tau,\tau'}}{\sqrt{v_\tau v_{\tau'}}}$, on $\mathcal{C}$,

$$\left|\alpha_{\tau,\tau'}\right| \leq \frac{C\sqrt{n\log n}+4}{4n - C\sqrt{n\log n}} = O\left(\sqrt{\frac{\log n}{n}}\right).$$

In view of the conditional distribution of $\delta(\tau)$ given in (2.13):

$$\sum_{\tau\in\mathcal{T}_n}\sum_{\tau',\tau\cap\tau'=\varnothing} \mathbb{P}(\delta(\tau)<0,\delta(\tau')<0,\mathcal{C}) = (1-o(1))\sum_{\tau\in\mathcal{T}_n}\sum_{\tau',\tau\cap\tau'=\varnothing} \mathbb{P}\left(Z_\tau > t_n,\, Z_{\tau'} > t_n\right), \tag{2.26}$$

with $t_n = \sqrt{4\log n - \log\log n - a_n}$, where $Z_\tau, Z_{\tau'}$ are two Gaussian variables of mean 0, with correlation coefficient α_n of order $O(\log^{1/2} n^{-1/2})$. Since $\alpha_n t_n \to 1$, by lemma 2.4.2 case (i), the sum in (2.26) is upper bounded by

$$\begin{aligned}&(1-o(1))\frac{n(n-1)}{2}\times\frac{(n-2)(n-3)}{2}\times\left[Ce^{-2t_n^2} + (1-o(1))\mathbb{P}\left(Z_1>t_n\right)\mathbb{P}\left(Z_2>t_n\right)\right]\\ &\leq (1+o(1))\mathbb{E}\left[Y\right]^2.\end{aligned}$$

Second case: $\tau\cap\tau'\neq\varnothing$**.** Without loss of generality we can assume that $\tau = (1\ 2)$ and $\tau' = (2\ 3)$. We can immediately deduce that $\left|\mathcal{D}_\tau^{\mathrm{E}}\cap\mathcal{D}_{\tau'}^{\mathrm{E}}\right| + \left|\mathcal{D}_\tau^{\mathrm{E}}\cap\mathcal{D}_{\tau'}^{\mathrm{E}}\cap\mathcal{F}_{\tau\circ\tau'}^{\mathrm{E}}\right| = (n-2)+0 = n-2$. So, denoting $\alpha_{\tau,\tau'} := \frac{c_{\tau,\tau'}}{\sqrt{v_\tau v_{\tau'}}}$, on $\mathcal{C}$,

$$\left|\alpha_{\tau,\tau'}\right| \leq \frac{C\sqrt{n\log n}+n-2}{4n - C\sqrt{n\log n}} \sim \frac{1}{4}.$$

Again, in view of the conditional distribution of $\delta(\tau)$ given in (2.13):

$$\sum_{\tau\in\mathcal{T}_n}\sum_{\tau',\tau\cap\tau'\neq\varnothing}\mathbb{P}(\delta(\tau)<0,\delta(\tau')<0,\mathcal{C})=(1-o(1))\sum_{\tau\in\mathcal{T}_n}\sum_{\tau',\tau\cap\tau'\neq\varnothing}\mathbb{P}\left(Z_\tau>t_n,\,Z_{\tau'}>t_n\right), \tag{2.27}$$

with $t_n=\sqrt{4\log n-\log\log n-a_n}$, where $Z_\tau, Z_{\tau'}$ are two Gaussian variables of mean 0, with correlation coefficient $\alpha_n\sim 1/4$. By Lemma 2.4.2 case (ii), the sum in (2.27) is upper bounded by

$$\begin{aligned}&(1-o(1))\frac{n(n-1)}{2}\times 2(n-2)\times\left[(1+o(1))\frac{1+\alpha_n}{\sqrt{2\pi}\,t_n}\exp\left(-\frac{t_n^2}{1+\alpha_n}\right)\right]\\&\quad\leq C''n^3\log^{-1/2}(n)\exp\left(-\frac{16}{5}\log n+o(\log n)\right)=o(1)=o(\mathbb{E}\,[Y]^2).\end{aligned}$$

□

Lemma 2.4.1 together with Payley-Zigmund inequality (Lemma 1.2.2 of Section 1.2.1) implies that $Y\geq o\left(\mathbb{E}[Y]\right)$ with high probability and thus proves (2.21) and the converse result of Theorem 2.2.

Remark 2.4.1. *We have shown here that under condition* (2.5)*, there is with high probability a great number of negative relative energy points near the ground truth, none of them being of significant interest to recover* exactly *our permutation. We may also study this relative energy far from the planted permutation, which would be interesting to address the problem of almost exact (resp. partial) alignment, which consists in finding an estimator $\hat{\pi}$ that coincides with π on at least $n-o(n)$ (resp. some positive fraction of n) points. In the light of our result which shows that exact recovery is not more difficult than detection, we can also conjecture that the same threshold $n\rho^2/\log n=4$ is sharp for the tasks of almost exact and partial recovery.*

Appendix of Chapter 2

2.A. Additional proofs

2.A.1. Proof of Lemma 2.3.1: lower bound on correlations of relative energies

Proof. Recall that we work under event $\mathcal{A}$. Fix $\alpha \in (0,1]$ and take $d = \alpha n$ and $\sigma, \sigma' \in \mathcal{S}_{n,d}$. The proof is obtained by establishing a fine lower bound on $\left|\mathcal{D}_\sigma^{\mathrm{E}} \cap \mathcal{D}_{\sigma'}^{\mathrm{E}}\right|$, which is simply the number of edges that are deranged both by σ^E and σ'^E. In order to establish this lower bound, let us assume that σ and σ' have $|\mathcal{D}_\sigma \cap \mathcal{D}_{\sigma'}| = \beta n$ common unfixed points, with $\beta \in [0, \alpha]$. We then form edges in $\mathcal{D}_\sigma^{\mathrm{E}} \cap \mathcal{D}_{\sigma'}^{\mathrm{E}}$ in the following way:

- First, by taking all pairs but the pairs made of points in the complement of $\mathcal{D}_\sigma \cap \mathcal{D}_{\sigma'}$ and those made of pairs (i,j) that are transpositions of σ or σ', we obtain at least $\frac{1}{2}\beta(2-\beta)n^2 - \alpha n$ edges.
- Then, add new edges made of one extremity in $\mathcal{D}_\sigma \setminus \mathcal{D}_{\sigma'}$ and one in $\mathcal{D}_{\sigma'} \setminus \mathcal{D}_\sigma$. Since $\mathcal{D}_\sigma$ (resp $\mathcal{D}_\sigma$) is stable by σ (resp. by σ'), all these $(\alpha-\beta)^2 n^2$ edges are in $\mathcal{D}_\sigma^{\mathrm{E}} \cap \mathcal{D}_{\sigma'}^{\mathrm{E}}$.

Finally we formed $g(\alpha,\beta)n^2 - \alpha n$ edges, with

$$g(\alpha,\beta) := \frac{1}{2}\beta^2 + (1-2\alpha)\beta + \alpha^2, \tag{2.28}$$

which is minimal on $[0,\alpha]$ at $\beta = 2\alpha - 1$ if $\alpha \geq 1/2$, or at $\beta = 0$ if $\alpha < 1/2$. In any case, this minimum is $f(\alpha)$. The first inequality is established by applying inequality (2.16) of event $\mathcal{A}$.

For the second part, consider a centered vector $Z = (Z_\sigma)_{\sigma \in \mathcal{S}_{n,\alpha n}}$ such that all Z_σ have same variance v_α and $\mathrm{Cov}(Z_\sigma, Z_{\sigma'}) = c_\alpha$ for $\sigma \neq \sigma'$, with v_α, c_α defined as follows:

$$\begin{aligned} v_\alpha &:= \alpha(2-\alpha)n^2 - C_1 n^{3/2}\log^{1/2} n, \\ c_\alpha &:= f(\alpha)n^2 - C_1 n^{3/2}\log^{1/2} n. \end{aligned}$$

for some $C_1 > 0$ large enough. Note that on event $\mathcal{A}$, for all $\alpha \in (0,1]$, all $\sigma, \sigma' \in \mathcal{S}_{n,\alpha n}$,

$$\mathrm{Cov}(Z_\sigma, Z_{\sigma'}) \leq \mathrm{Cov}(X_\sigma, X_{\sigma'}),$$

so one has that for all $t > 0$,

$$\mathbb{P}\left(\max_{\sigma \in \mathcal{S}_{n,\alpha n}} X_\sigma > t \ \cap \mathcal{A}\right) \leq \mathbb{P}\left(\max_{\sigma \in \mathcal{S}_{n,\alpha n}} Z_\sigma > t\right). \tag{2.29}$$

We now control the right-hand side of (2.29) with this classical Lemma, which proof is find hereafter in Appendix 2.A.2:

Lemma 2.A.1 (Maximum of totally correlated Gaussian variables)**.** *Let Z be a centered Gaussian vector of size N, such that all Z_i have same variance v and $\mathrm{Cov}(Z_i, Z_j) = c$ for*

$i \neq j$. *Then*

$$\mathbb{P}\left(\max_{1\leq i\leq N} Z_i > \sqrt{2(v-c)\log N} + 2\sqrt{v\log\log N}\right) \leq \frac{2}{\log N}. \tag{2.30}$$

Note that for v_α, c_α previously defined, one has

$$\sqrt{2(v_\alpha - c_\alpha)\log|\mathcal{S}_{n,\alpha n}|} \leq \sqrt{2\alpha(\alpha(2-\alpha)-f(\alpha))}n^{3/2}\log^{1/2} n, \tag{2.31}$$

and for n large enough,

$$2\sqrt{v_\alpha \log\log|\mathcal{S}_{n,\alpha n}|} \leq 2\sqrt{\alpha(2-\alpha)}n\sqrt{\log n + \log\log n} \leq (2+\varepsilon')n\log^{1/2} n. \tag{2.32}$$

Finally, we use equations (2.29)–(2.32) to conclude that for n large enough:

$$\begin{aligned}
&\mathbb{P}\left(\exists d = \alpha n, \alpha > \alpha_0, \max_{\sigma\in\mathcal{S}_{n,d}} X_\sigma > \sqrt{2\alpha\left(\alpha(2-\alpha)-f(\alpha)\right)}n^{3/2}\log^{1/2} n + (2+\varepsilon')n\log^{1/2} n\right)\\
&\leq 1 - \mathbb{P}(\mathcal{A}) + \sum_{d=\alpha n,\,\alpha>\alpha_0} \mathbb{P}\left(\max_{\sigma\in\mathcal{S}_{n,\alpha n}} Z_i > \sqrt{2(v_\alpha - c_\alpha)\log|\mathcal{S}_{n,\alpha n}|} + 2\sqrt{v_\alpha\log\log|\mathcal{S}_{n,\alpha n}|}\right)\\
&\leq o(1) + \sum_{d=\alpha n,\,\alpha>\alpha_0} \frac{2}{\log|\mathcal{S}_{n,\alpha n}|} \leq o(1) + \frac{2n}{\log|\mathcal{S}_{n,\alpha_0 n}|} = o(1) + \frac{2}{\alpha_0\log n} = o(1),
\end{aligned}$$

and Lemma 2.3.1 is proved. □

2.A.2. Proof of Lemma 2.A.1: maximum of totally correlated Gaussian variables

Proof. Let us make a change of variables which preserves the joint distribution:

$$(Z_1, Z_2, \ldots, Z_N) = \left(\sqrt{c}\,\xi_0 + \sqrt{v-c}\,\xi_1, \ldots, \sqrt{c}\,\xi_0 + \sqrt{v-c}\,\xi_N\right),$$

where $\xi_0, \ldots, \xi_N$ are independent standard Gaussian random variables. The maximum thus writes

$$\max_{1\leq i\leq N} Z_i = \sqrt{c}\,\xi_0 + \sqrt{v-c}\max_{1\leq i\leq N}\xi_i$$

Then, with the classical inequality $\mathbb{P}\left(\mathcal{N}(0,1)\geq t\right) \leq e^{-t^2/2}$, then with probability at least $1 - 1/(\log N)$, one has:

$$\sqrt{c}\,\xi_0 \leq \sqrt{2c\log\log N}, \quad \text{and} \quad \sqrt{v-c}\max_{1\leq i\leq N}\xi_i \leq \sqrt{2(v-c)\log N\left(1 + \frac{\log\log N}{\log N}\right)},$$

so with probability at least $1 - 2/(\log N)$:

$$\begin{aligned}
\max_{1\leq i\leq N} Z_i &\leq \sqrt{2(v-c)\log N} + \sqrt{2\log\log N}\left(\sqrt{c} + \sqrt{v-c}\right)\\
&\leq \sqrt{2(v-c)\log N} + 2\sqrt{v\log\log N},
\end{aligned}$$

where we used $\sqrt{c} + \sqrt{v-c} \leq \sqrt{2v}$ in the last step. □

2.A.3. Proof of Lemma 2.3.2

Proof. For $\alpha \in (0,1]$,

$$\begin{aligned}
(2.20) &\iff \alpha^2(2-\alpha)^2 \geq 2\alpha\left(\alpha(2-\alpha) - f(\alpha)\right)\\
&\iff f(\alpha) \geq \alpha^2 - \alpha^3/2.
\end{aligned}$$

The inequality is verified for $\alpha < 1/2$. To conclude the proof of (2.20), it remains to check that for $1 \geq \alpha \geq 1/2$, $f(\alpha) \geq \alpha^2 - \alpha^3/2$, which is equivalent to

$$\begin{aligned}\alpha^2 - \frac{1}{2}(2\alpha-1)^2 \geq \alpha^2 - \alpha^3/2 &\iff \alpha^3 - 4\alpha^2 + 4\alpha - 1 \geq 0 \\ &\iff (\alpha-1)(\alpha^2 - 3\alpha + 1) \geq 0 \\ &\iff \alpha^2 - 3\alpha + 1 \leq 0 \iff \alpha \geq \frac{3-\sqrt{5}}{2} \sim 0.382...\end{aligned}$$

□

2.A.4. Proof of Lemma 2.4.2: control of deviation probabilities for correlated Gaussians

Proof. Let us first make a change of variable which preserves the joint distribution:

$$(Z_1, Z_2) = (Z, \alpha_n Z + \sqrt{1-\alpha_n^2} Z'),$$

with Z, Z' two independent standard Gaussian variables.

Proof of (i)**.** Note that standard Gaussian concentration gives $\mathbb{P}\left(Z > 2t_n \middle| Z > t_n\right) \sim \frac{1}{2} e^{-3t_n^2/2}$. Thus, for n large enough

$$\begin{aligned}\mathbb{P}\left(Z_1 > t_n, Z_2 > t_n\right) &\leq \mathbb{P}\left(Z > t_n\right) e^{-3t_n^2/2} + \mathbb{P}\left(Z > t_n\right) \mathbb{P}\left(\alpha_n Z + \sqrt{1-\alpha_n^2} Z' > t_n, Z \leq 2t_n \middle| Z > t_n\right) \\ &\leq e^{-2t_n^2} + \mathbb{P}\left(Z > t_n\right) \mathbb{P}\left(Z' > t_n - 2\alpha_n t_n + O(t_n \alpha_n^2)\right) \\ &\leq e^{-2t_n^2} + \mathbb{P}\left(Z > t_n\right) \mathbb{P}\left(Z' > t_n - o(1)\right) \\ &\leq e^{-2t_n^2} + (1+o(1)) \mathbb{P}\left(Z > t_n\right) \mathbb{P}\left(Z' > t_n\right) \\ &= e^{-2t_n^2} + (1+o(1)) \mathbb{P}\left(Z_1 > t_n\right) \mathbb{P}\left(Z_2 > t_n\right).\end{aligned}$$

Proof of (ii)**.** For any (s_n) such that $s_n \leq t_n$ for all n, one has

$$\begin{aligned}\mathbb{E}\left[e^{s_n Z} \,\middle|\, Z > t_n\right] &= \frac{1}{\sqrt{2\pi}} \int_{t_n}^{+\infty} e^{s_n z - z^2/2} dz \left(\frac{1}{\sqrt{2\pi}} \int_{t_n}^{+\infty} e^{-z^2/2} dz\right)^{-1} \\ &= e^{s_n^2/2} \int_{t_n - s_n}^{+\infty} e^{-z^2/2} dz \left(\int_{t_n}^{+\infty} e^{-z^2/2} dz\right) \\ &\sim \frac{t_n}{t_n - s_n} \exp\left(s_n^2/2 - (t_n - s_n)^2/2 + t_n^2/2\right) = \frac{t_n}{t_n - s_n} e^{s_n t_n}.\end{aligned}$$

Using independence of Z, Z' and Chernoff bound, we get, taking s_n such that $\alpha s_n = u t_n$ with $u < 1$, for n large enough,

$$\begin{aligned}\mathbb{P}\left(\alpha Z + \sqrt{1-\alpha^2} Z' > t_n \middle| Z > t_n\right) &\leq (1+o(1)) \frac{t_n}{t_n - \alpha s_n} \exp\left(\alpha s_n t_n + \frac{1-\alpha^2}{2} s_n^2 - s_n t_n\right) \\ &\leq (1+o(1)) \frac{1}{1-u} \exp\left(\left(u + \frac{u^2(1-\alpha^2)}{2\alpha^2} - \frac{u}{\alpha}\right) t_n^2\right) \\ &\overset{(a)}{\leq} (1+o(1))(1+\alpha) \exp\left(-\frac{1-\alpha}{1+\alpha} \cdot \frac{t_n^2}{2}\right)\end{aligned}$$

where we took $u = \frac{\alpha}{1+\alpha} < 1$ in (a). The proof follows from this last inequality, together with the bound $\mathbb{P}\left(Z > t_n\right) \leq \frac{1}{\sqrt{2\pi} t_n} \exp\left(-\frac{t_n^2}{2}\right)$. □

Chapter 3

Alignment of graph databases with Gaussian weights: analysis of a spectral method

In this chapter, we analyze a simple spectral method (`EIG1`) for the problem of matrix alignment, consisting in aligning their leading eigenvectors: given two adjacency matrices A and B, `EIG1` aligns v_1 and v_1', their two corresponding leading eigenvectors (up to the sign of v_1') and outputs the corresponding permutation.

We will consider the Gaussian model $\mathsf{Wig}(n, \xi)$ defined earlier in (1.7): A belongs to the Gaussian Orthogonal Ensemble (GOE) of size $n \times n$, and B is a noisy version of A where all nodes have been relabeled according to some planted permutation $\pi^\star$. We show the following zero-one law: with high probability, under the condition $\xi n^{7/6+\epsilon} \to 0$ for some $\varepsilon > 0$, `EIG1` recovers all but a vanishing part of the underlying permutation π, whereas if $\xi n^{7/6-\epsilon} \to \infty$, this method cannot recover more than $o(n)$ correct matches.

This result gives an understanding of the simplest and fastest spectral method for matrix alignment (or complete weighted graph alignment), and involves proof methods and techniques which could be of independent interest.

This chapter is based on the paper *Spectral alignment of correlated gaussian matrices* [GLM22], published in *Advances in Applied Probability*, a joint work with M. Lelarge and L. Massoulié.

3.1. Introduction

3.1.1. The `EIG1` algorithm

As in Chapter 2, we are interested in alignment of Gaussian databases, which is one of the instances of the graph alignment problem. For a general overview, we refer here again to the introduction of this manuscript, to Section 1.3.1 for applications and to Section 1.3.4 for theoretical results.

Related work: spectral methods for graph alignment Some general spectral methods for random graph alignment are introduced in [FQM$^+$16], based on representation matrices and low-rank approximations. These methods are tested over synthetic graphs and real data; however no precise theoretical guarantee – e.g. an error control of the inferred mapping depending on the signal-to-noise ratio – can be found for such techniques.

Most recently, a spectral method for matrix and graph alignment (`GRAMPA`) was proposed in [FMWX19a, FMWX19b] and computes a similarity matrix which takes into account all pairs of eigenvalues (λ_i, μ_j) and eigenvectors (u_i, v_j) of matrices A and B. The authors study the regime in which the method exactly recovers the underlying vertex correspondence: this method can tolerate a noise ξ up to $O\left(1/\mathrm{polylog}\, n\right)$ to recover the entire underlying vertex correspondence. Since the computations of all eigenvectors is required, the time complexity of `GRAMPA` is at least $O(n^3)$.

It is important to note that the signs of eigenvectors are ambiguous: in practice, it is necessary to test over all possible signs of eigenvectors. This additional complexity has no consequence when reducing A and B to rank-one matrices, but becomes costly when the reduction made is of rank $k \gg 1$. This combinatorial observation makes implementation and analysis of general rank-reduction methods (as the ones proposed in [FQM+16]) more difficult. We therefore focus on the analysis of the rank-one reduction (`EIG1` hereafter) which is the simplest and most natural spectral alignment method, where only the leading eigenvectors of A and B are computed, with time complexity $O(n^2)$, which is significantly less than `GRAMPA`.

Model and method Let us recall the model $\mathsf{Wig}(n, \xi)$ defined in (1.8). In this model, A is a matrix from the normalized Gaussian Orthogonal Ensemble (GOE), i.e. for all $1 \leq u \leq v \leq n$,

$$A_{u,v} = A_{v,u} \sim \begin{cases} \mathcal{N}(0, 1/n) & \text{if } u \neq v, \\ \mathcal{N}(0, 2/n) & \text{if } u = v, \end{cases} \tag{3.1}$$

and H is an independent copy of A. We define

$$B = \Pi^{\star\top} (A + \xi H) \Pi^{\star} \tag{3.2}$$

where $\Pi^\star$ is a random uniform matrix of a permutation $\pi^\star$ – e.g. random uniform – of $[n]$ and $\xi = \xi(n)$ is the *noise parameter*.

Given two vectors $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ having all distinct coordinates, the permutation ρ which *aligns* x and y is the permutation such that for all $1 \leq i \leq n$, the rank (for the usual order) of $y_{\rho(i)}$ in y is the rank of x_i in x.

Remark 3.1.1. *Note that in our model, all the probability distributions are absolutely continuous with respect to Lebesgue measure, thus the eigenvectors of A and B all have almost surely pairwise distinct coordinates.*

We recall that the aim is to infer the underlying permutation $\Pi^\star$ given the observation of A and B. We now introduce our simple spectral algorithm derived from [FQM+16], which we call `EIG1`, that consists in computing and aligning the leading eigenvectors v_1 and v_1' of A and B. This very natural method can be thought of as the relaxation of the QAP formulation (1.5) when reducing A and B to rank-one matrices $\lambda_1 v_1 v_1^\top$ and $\lambda_1' v_1' v_1'^T$. Indeed, as soon as v_1 and v_1' have pairwise distinct coordinates, is it easy to see that

$$\arg\max_{\Pi \in \mathcal{S}_n} \langle \lambda_1 v_1 v_1^\top, \Pi \lambda_1' v_1' v_1'^T \Pi^\top \rangle = \arg\max_{\Pi \in \mathcal{S}_n} \pm v_1^\top \Pi v_1' = \rho,$$

where ρ is the aligning permutation of v_1 and $\pm v_1'$. Computing the two normalized leading eigenvectors (i.e. corresponding to the highest eigenvalues) v_1 and v_1' of A and B, the `EIG1` algorithm returns the aligning permutation of v_1 and $\pm v_1'$. The method then decides which permutation to output according to the scores.

The aim of this chapter is to find the regime in which `EIG1` achieves almost exact recovery, i.e. recovers all but a vanishing fraction of nodes of the planted ground truth $\Pi^\star$.

3.1.2. Main results and proof scheme

We start by introducing specific notations and recall some useful basic definitions. Throughout the chapter, all limits are taken when $n \to \infty$, and the dependency in n will most of the time be eluded, as an abuse of notation.

Eigenvalues, eigenvectors. In the following, $(v_1, v_2, \dots, v_n)$ (resp. $(v_1', v_2', \dots, v_n')$) denote two orthonormal bases of eigenvectors of A (resp. of B) with respect to the (real) eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ of A (resp. $\lambda_1' \geq \lambda_2' \geq \dots \geq \lambda_n'$ of B). Through all the study,

Algorithm 3.1: `EIG1` Algorithm for matrix alignment

```
Compute v_1 a normalized leading eigenvector of A;
Compute v'_1 a normalized leading eigenvector of B;
Compute Π_+ the permutation aligning v_1 and v'_1;
Compute Π_− the permutation aligning v_1 and −v'_1;
if ⟨A, Π_+ B Π_+^⊤⟩ ≥ ⟨A, Π_− B Π_−^⊤⟩ then
    return Π_+
else
    return Π_−
end
```

the sign of v_1' is fixed such that $\langle \Pi^\star v_1, v_1' \rangle > 0$.

Overlap. For any (matrix) estimator $\hat{\Pi}$ of $\Pi^\star$ its overlap is defined as follows

$$\mathrm{ov}(\hat{\Pi}, \Pi^\star) := \mathrm{ov}(\hat{\pi}, \pi^\star) = \frac{1}{n} \sum_{u=1}^{n} \mathbb{1}_{\hat{\pi}(u)=\pi^\star(u)}\,, \tag{3.3}$$

where $\hat{\pi}$ (resp. $\pi^\star$) is the permutation corresponding to matrix $\hat{\Pi}$ (resp. $\Pi^\star$).

Probability. The equality $\overset{(d)}{=}$ will refer to equality in distribution. Some event A_n is said to hold *with high probability* (we will use the abbreviation "w.h.p."), if $\mathbb{P}(A_n)$ converges to 1 when $n \to \infty$.

For two random variables $u = u_n$ and $v = v_n$, we will use the notation $u = o_{\mathbb{P}}(v)$ if $u_n/v_n \xrightarrow{\mathbb{P}} 0$ when $n \to \infty$. We also use this notation when $X = X_n$ and $Y = Y_n$ are $n-$dimensional random vectors: $X = o_{\mathbb{P}}(Y)$ if $\|X_n\|/\|Y_n\| \xrightarrow{\mathbb{P}} 0$ when $n \to \infty$.

Define

$$\mathcal{F} := \left\{ f : \mathbb{N} \to \mathbb{R} \mid \forall t > 0, n^t f(n) \to \infty, n^{-t} f(n) \to 0 \right\}. \tag{3.4}$$

For two random variables $u = u(n)$ and $v = v(n)$, $u \asymp v$ refers to equivalence with high probability up to some sub-polynomial factor, meaning that there exists a function $f \in \mathcal{F}$ such that

$$\mathbb{P}\left(\frac{v(n)}{f(n)} \le u(n) \le f(n) v(n) \right) \to 1. \tag{3.5}$$

Main results, proof scheme The main result of this chapter can be stated as follows: there exists a condition – a threshold – on ξ and n under which the `EIG1` method enables us to recover $\Pi^\star$ almost exactly, in terms of the overlap defined in (3.3). Above this threshold, we show that `EIG1` Algorithm cannot recover more than a vanishing part of Π.

Theorem 3.1 (Zero-one law for `EIG1` method)**.** *For all n, Π_n denotes an arbitrary permutation of size n, $\hat{\Pi}_n$ is the estimator obtained with Algorithm `EIG1`, for A and B of model (3.2), with permutation $\Pi_n^\star$ and noise parameter ξ. We have the following zero-one law:*

(i) If there exists $\epsilon > 0$ such that $\xi = o(n^{-7/6-\epsilon})$ then

$$\mathrm{ov}(\hat{\Pi}_n, \Pi_n^\star) \xrightarrow{L^1} 1.$$

(ii) If there exists $\epsilon > 0$ such that $\xi = \omega(n^{-7/6+\epsilon})$ then

$$\mathrm{ov}(\hat{\Pi}_n, \Pi_n^\star) \xrightarrow{L^1} 0.$$

Results of Theorem 3.1 are illustrated on Figure 3.1 showing the zero-one law at $\xi \asymp n^{-7/6}$. Note that the convergence to the step function appears to be slow.

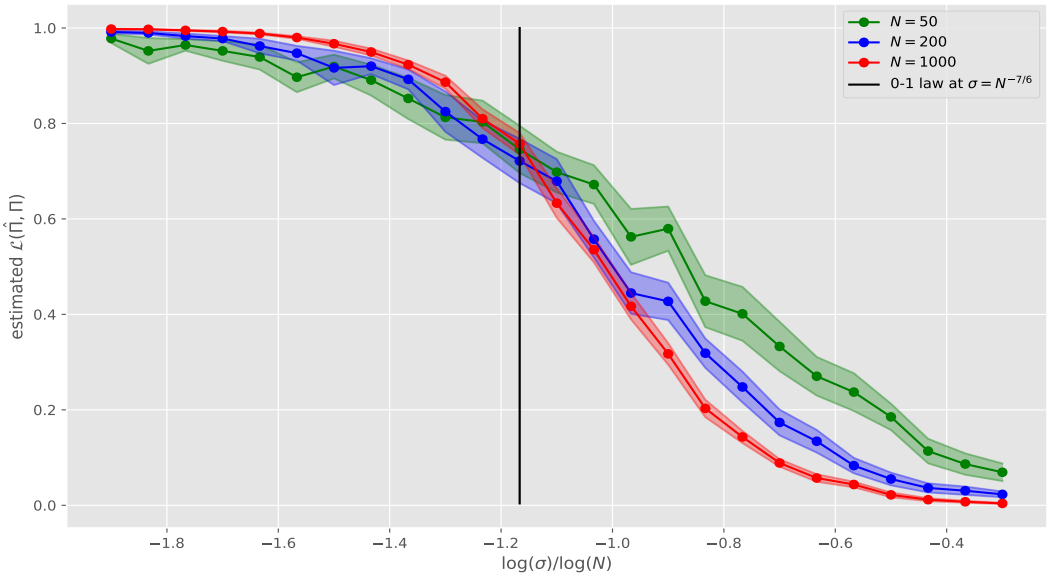


Figure 3.1 – *Estimated overlap* $\mathrm{ov}(\hat{\Pi}, \Pi^\star)$ *reached by* `EIG1` *in model* (3.2)*, for varying* n *and* ξ*. With* 95% *confidence intervals.*

Remark 3.1.2. *We can now underline that without loss of generality, we can assume that* $\Pi^\star = I_n$*, the identity matrix. Indeed, one can return to the general case applying transformations* $A \to \Pi^\star A \Pi^{\star,\top}$ *and* $H \to \Pi^\star H \Pi^{\star,\top}$*. From now on we will assume in the rest of the chapter that* $\Pi^\star = I_n$*.*

In order to prove this theorem, it is necessary to establish two intermediate results along the way, which could also be of independent interest. First, we study the behavior of v_1' with respect to v_1, showing that under some conditions on ξ and n, the difference $v_1 - v_1'$ can be approximated by a renormalized Gaussian standard vector, multiplied by a variance term $\mathbf{S}$, where $\mathbf{S}$ is a random variable which behavior is well understood in terms of n and ξ when $n \to \infty$. For this we work under the following assumption:

$$\exists\, \alpha > 0,\ \xi = o\left(n^{-1/2-\alpha}\right), \tag{3.6}$$

Proposition 3.1.1. *Under assumption* (3.6)*, there exists a standard Gaussian vector* $Z \sim \mathcal{N}(0, I_n)$ *independent from* v_1 *and a random variable* $\mathbf{S} \asymp \xi n^{1/6}$*, such that*

$$v_1' = (1 + o_{\mathbb{P}}(1)) \left(v_1 + \mathbf{S} \frac{Z}{\|Z\|} \right).$$

Remark 3.1.3. *This assumption (3.6) (or a tighter formulation) arises when studying the diffusion trajectories of eigenvalues and eigenvectors in random matrices, and corresponds to the microscopic regime in [ABB14]. This assumption ensures that all eigenvalues of* B *are close enough to the eigenvalues of* A*. This comparison term is justified from the random matrix theory (*$n^{-1/2}$ *is the typical amplitude of the spectral gaps* $\sqrt{n}(\lambda_i - \lambda_{i+1})$ *in the bulk, which are the smaller ones).*

Eigenvectors diffusions in similar models (diffusion processes dawn with the scaling $\xi = \sqrt{t}$*) are studied in [ABB14], where the main tool is the Dyson Brownian motion (see e.g. [AGZ09]) and its formulation for eigenvectors trajectories, giving stochastic differential equations for the evolutions of* $v_j'(t)$ *with respect to vectors* $v_i = v_i'(0)$*. These equations lead to a*

system of stochastic differential equations for the overlaps $\langle v_i, v_j'(t)\rangle$*, which is quite difficult to analyze rigorously. In this work a more elementary method to get a expansion of* v_1' *around* v_1*, for which this very condition (3.6) also appears.*

Note that here, spectral gaps at the edge are of order $n^{-1/6}$ *so assumption* (3.6) *may not optimal for our study, and we expect Proposition 3.1.1 to hold up to* $\xi = o\left(n^{-1/6-\alpha}\right)$*. However, since the positive result of Theorem 3.1 holds in a way more restrictive regime – see condition* (i)*, condition* (3.6) *is enough for our purpose and allows a short and simple proof.*

Proposition 3.1.1 suggests the study of v_1' as a Gaussian perturbation of v_1. The main question is now formulated as follows: *what is the probability that the perturbation on* v_1 *has an impact on the overlap of the estimator* $\hat{\Pi}$ *from the* `EIG1` *method?* To answer this question, we introduce a correlated Gaussian vectors model (or *toy model* hereafter) of parameters n and $s > 0$. In this model, we draw a standard Gaussian vector X of size n and $Y = X + sZ$ where Z is an independent copy of X. We will use the notation $(X, Y) \sim \mathsf{J}(n, s)$.

Define r_1 the function that associates to any vector $T = (t_1, \dots, t_p)$ the rank of t_1 in T (for the usual decreasing order). For $(X, Y) \sim \mathsf{J}(n, s)$ we evaluate

$$p(n, s) := \mathbb{P}\left(r_1(X) = r_1(Y)\right).$$

Our second result shows that there is a zero-one law for the property of rank preservation in the toy model $\mathsf{J}(n, s)$.

Proposition 3.1.2 (Zero-one law for $p(n, s)$)**.** *In the correlated Gaussian vectors model we have the following:*

(i) *If* $s = o(1/n)$ *then*

$$p(n, s) \underset{n\to\infty}{\longrightarrow} 1.$$

(ii) *If* $s = \omega(1/n)$ *then*

$$p(n, s) \underset{n\to\infty}{\longrightarrow} 0.$$

These results are illustrated on Figure 3.2, showing the zero-one law at $s \asymp n^{-1}$.

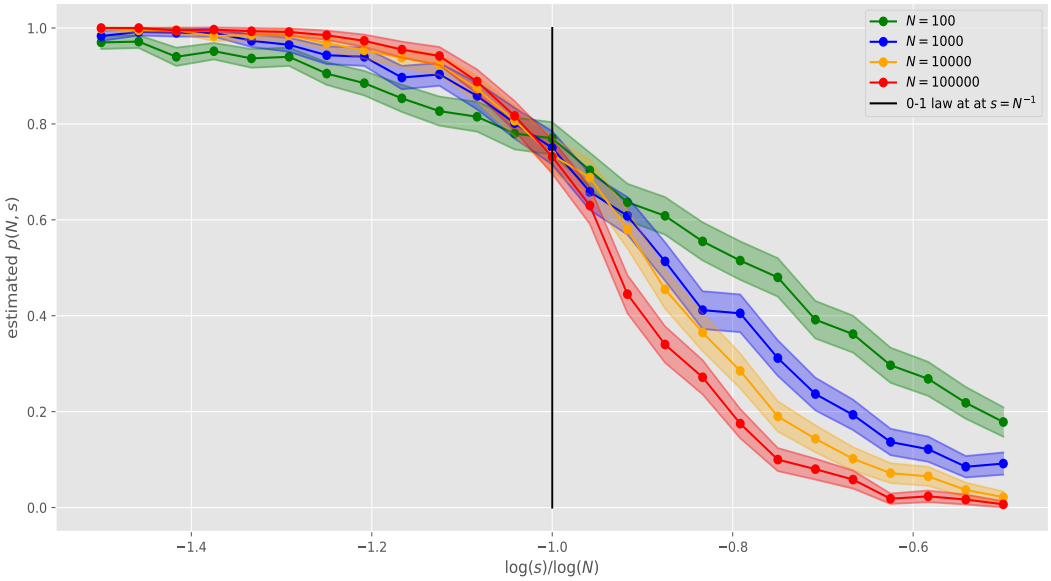


Figure 3.2 – *Estimated* $p(n, s)$ *in the toy model* $\mathsf{J}(n, s)$*. With* 95% *confidence intervals.*

Organization of the chapter The gaussian approximation of $v_1 - v_1'$ is established in Section 3.2 with the proof of Proposition 3.1.1. The toy model defined here above is studied in Section 3.3 where Proposition 3.1.2 is established. Finally, we gather results of Propositions 3.1.1 and 3.1.2 in Section 3.4 to show Theorem 3.1. Some additional proofs are deferred to Appendices 3.A and 3.B.

3.2. Behavior of the leading eigenvectors of correlated matrices

The main idea of this section is to find a first order expansion of v_1' around v_1. Recall that we use the notations $(v_1, v_2, \dots, v_n)$ for normalized eigenvectors of A, corresponding to the eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$. Similarly, $(v_1', v_2', \dots, v_n')$ and $\lambda_1' \geq \lambda_2' \geq \dots \geq \lambda_n'$ will refer to eigenvectors and eigenvalues of $B = A + \xi H$. Since A and B are symmetric, all these eigenvalues are real and the vectors $\{v_i\}_i$ (resp. $\{v_i'\}_i$) are pairwise orthogonal. We also recall that v_1' is taken such that $\langle v_1, v_1' \rangle > 0$.

3.2.1. Computation of a leading eigenvector of B

Recall now that we are working under assumption (3.6):

$$\exists\, \alpha > 0,\ \xi = o\left(n^{-1/2-\alpha}\right).$$

Let w' be an (non normalized) eigenvector of B for the eigenvalue λ_1' of the form

$$w' := \sum_{i=1}^{n} \theta_i v_i,$$

where we assume that $\theta_1 = 1$. Such an assumption can be made a.s. since any hyperplane of $\mathbb{R}^n$ has a null Lebesgue measure in $\mathbb{R}^n$ (see Remark 3.1.1).

The defining eigenvector equations projected on vectors v_i give

$$\begin{cases} \theta_1 &= 1, \\ \forall i > 1,\ \theta_i &= \dfrac{\xi}{\lambda_1' - \lambda_i} \sum_{j=1}^{n} \theta_j \langle H v_j, v_i \rangle, \\ \lambda_1' - \lambda_1 &= \xi \sum_{j=1}^{n} \theta_j \langle H v_j, v_1 \rangle. \end{cases} \tag{3.7}$$

The strategy is then to approximately solve (3.7) with an iterative scheme, leading to the following expansion:

Proposition 3.2.1. *Under the assumption* (3.6) *one has the following:*

$$w' = v_1 + \xi \sum_{i=2}^{n} \frac{\langle H v_i, v_1 \rangle}{\lambda_1 - \lambda_i} v_i + o_{\mathbb{P}}\left(\xi \sum_{i=2}^{n} \frac{\langle H v_i, v_1 \rangle}{\lambda_1 - \lambda_i} v_i \right). \tag{3.8}$$

We refer to Appendix 3.A.1 for the details regarding the definition of the mentioned iterative scheme, as well as a proof of Proposition 3.2.1. The proof uses assumption (3.6) an builds upon some standard results on the distribution of eigenvalues in the GOE.

Remark 3.2.1. *The above proposition could easily be extended for all eigenvectors of B, under assumption* (3.6)*. Based on the studies of the trajectories of the eigenvalues and eigenvectors in the* GUE *[ABB14] and the* GOE *[AB14], since we are only interested here in the leading eigenvectors, we expect the result of Proposition 3.2.1 to hold under the weaker assumption $\xi n^{1/6+\alpha} \to 0$, for $n^{-1/6}$ is the typical spectral gap $\sqrt{n}(\lambda_1 - \lambda_2)$ on the edge. However, as explained before (see Remark 3.1.3), our analysis doesn't require this more optimal assumption. We also know that the expansion* (3.8) *doesn't hold as soon as $\xi = \omega(n^{-1/6})$. A result*

proved by Chatterjee ([Cha14], Theorem 3.8) shows that the eigenvectors corresponding to the highest eigenvalues v_1 of A and v_1' of $B = A + \xi H$, when A and H are two independent matrices from the GUE*, are delocalized (in the sense that $\langle v_1, v_1' \rangle$ converges in probability to 0 as $n \to \infty$), when $\xi = \omega(n^{-1/6})$.*

3.2.2. Gaussian representation of $v_1' - v_1$

We still work under assumption (3.6). After renormalization, we have $v_1' = \frac{w'}{\|w'\|}$. We are now able to study the behavior of the overlap $\langle v_1', v_1 \rangle$:

$$\langle v_1', v_1 \rangle = \left(1 + \xi^2 (1 + o_{\mathbb{P}}(1)) \sum_{i=2}^{n} \frac{\langle H v_i, v_1 \rangle^2}{(\lambda_1 - \lambda_i)^2} \right)^{-1/2}$$

Hence

$$\langle v_1', v_1 \rangle = 1 - \frac{\xi^2}{2} \sum_{i=2}^{n} \frac{\langle H v_i, v_1 \rangle^2}{(\lambda_1 - \lambda_i)^2} + o_{\mathbb{P}} \left(\xi^2 \sum_{i=2}^{n} \frac{\langle H v_i, v_1 \rangle^2}{(\lambda_1 - \lambda_i)^2} \right). \tag{3.9}$$

Let us give the heuristic to evaluate the first sum in the right-hand side of (3.9): since the GOE distribution is invariant by rotation (see e.g. [AGZ09]), the random variables $\langle H v_i, v_1 \rangle$ are zero-mean Gaussian, with variance $1/n$. Moreover, it is well known [AGZ09] that the eigenvalue gaps $\lambda_1 - \lambda_i$ are of order $n^{-1/6}$ when i is small, and $n^{-1/2}$ in the bulk (when i is typically of order n). These considerations lead to the following:

Lemma 3.2.1. *We have the following concentration*

$$\sum_{i=2}^{n} \frac{\langle H v_i, v_1 \rangle^2}{(\lambda_1 - \lambda_i)^2} \asymp n^{1/3}. \tag{3.10}$$

We refer to Appendix 3.A.6 for a rigorous proof of this result. With this Lemma, we are now able to give the first order expansion of $\langle v_1', v_1 \rangle$ with respect to ξ:

$$\langle v_1', v_1 \rangle = 1 - \frac{\xi^2}{2} n^{1/3} + o_{\mathbb{P}} \left(\xi^2 n^{1/3} \right). \tag{3.11}$$

Remark 3.2.2. *The comparison between ξ and $n^{1/6}$ made in [Cha14] naturally reappears here, as $\xi^2 n^{1/3}$ is the typical shift of v_1' with respect to v_1.*

The intuition is that the scalar product $\langle v_1', v_1 \rangle$ is sufficient to derive a Gaussian representation of v_1' w.r.t. v_1. We formalize this in the following

Lemma 3.2.2. *Given v_1, when writing the decomposition $w' = v_1 + w$, with*

$$w := \sum_{i=2}^{n} \theta_i v_i,$$

the distribution of w is invariant by rotation in the orthogonal complement of v_1. This implies in particular that given v_1, $\|w\|$ and $\frac{w}{\|w\|}$ are independent, and that $\frac{w}{\|w\|}$ is uniformly distributed on $\mathbb{S}^{n-2}$, the unit sphere of $v_1^{\perp}$.

Proof of Lemma 3.2.2. We work conditionnally on v_1. Let O be an orthogonal transformation of the hyperplane $v_1^{\perp}$ (such that $Ov_1 = v_1$). Since the GOE distribution is invariant by rotation and A and H are independent, $\widetilde{B} := O^{\top} A O + \xi O^{\top} H O$ has he same distribution as $B = A + \xi H$.

Note that $Ow' = v_1 + Ow$ is an eigenvector of $\widetilde{B}$ for the eigenvalue λ_1. Since the distribution of the matrix of eigenvectors $(v_2, \ldots, v_n)$ is the Haar measure on the orthogonal

group $\mathcal{O}_{n-1}\left(v_1^\perp\right)$, denoted by $d\mathcal{H}$, the distribution of w is also invariant by rotation in the orthogonal complement of v_1. Furthermore, for any f, g bounded continuous functions and $O \in \mathcal{O}_{n-1}\left(v_1^\perp\right)$,

$$\begin{aligned}\mathbb{E}\left[f(\|w\|)g\left(\frac{w}{\|w\|}\right)\right] &= \mathbb{E}\left[f(\|w\|)g\left(\frac{Ow}{\|Ow\|}\right)\right] = \mathbb{E}\left[f(\|w\|)\int_{\mathcal{O}_{n-1}\left(v_1^\perp\right)} d\mathcal{H}(O)g\left(\frac{Ow}{\|Ow\|}\right)\right] \\ &= \mathbb{E}\left[f(\|w\|)\int_{\mathbb{S}^{n-2}} \frac{g(u)du}{\mathrm{Vol}\left(\mathbb{S}^{n-2}\right)}\right] = \mathbb{E}\left[f(\|w\|)\right]\mathbb{E}\left[g\left(\frac{w}{\|w\|}\right)\right].\end{aligned}$$

This completes the proof of Lemma 3.2.2. □

We can now show the main result of this section, Proposition 3.1.1.

Proof of Proposition 3.1.1. Recall the decomposition $w' = v_1 + w$ with $w = \sum_{i=2}^n \theta_i v_i$. According to Lemma 3.2.2, conditioned to v_1, $\frac{w}{\|w\|}$ is uniformly distributed on $\mathbb{S}^{n-2}$, the unit sphere of $v_1^\perp$. We now state a classical result about sampling uniform vectors on a sphere:

Lemma 3.2.3. *Let E be $p-$dimensional Euclidean space, endowed with an orthogonal basis $\mathcal{B} = (e_1, \ldots, e_p)$. Let u be a random vector uniformly distributed on the unit sphere $\mathbb{S}^{p-1}$ of E. Then, in basis $\mathcal{B}$, u has the same distribution as*

$$\left(\frac{Z_1}{\sqrt{\sum_{i=1}^p Z_i^2}}, \ldots, \frac{Z_p}{\sqrt{\sum_{i=1}^p Z_i^2}}\right),$$

where $Z_1, \ldots, Z_p$ are i.i.d. standard normal random variables.

We refer e.g. to [OVW16], Lemma 10.1, for the proof of this result. In our context, this proves that the joint distribution of the coordinates $w_2, \ldots, w_n$ of w along $v_2, \ldots, v_n$ is always that of a normalized standard Gaussian vector (on $\mathbb{R}^{n-1}$). This joint probability does not dependent on v_1. Hence, there exist $Z_2, \ldots, Z_n$ standard Gaussian independent variables, independent from v_1 (and from $\|w\|$ by Lemma 3.2.2), such that:

$$w' = v_1 + \frac{\|w\|}{\left(\sum_{i=2}^n Z_i^2\right)^{1/2}} \sum_{i=2}^n Z_i v_i.$$

Let Z_1 be another standard Gaussian variable, independent from everything else. Then

$$w' = \left(1 - \frac{\|w\| Z_1}{\left(\sum_{i=2}^n Z_i^2\right)^{1/2}}\right) v_1 + \frac{\|w\|}{\left(\sum_{i=2}^n Z_i^2\right)^{1/2}} \sum_{i=1}^n Z_i v_i.$$

Let $Z = \sum_{i=1}^n Z_i v_i$, which is a standard Gaussian vector. Since the distribution of Z is invariant by permutation of the $(Z_i)_{1 \le i \le n}$, Z and v_1 are independent. We have

$$\begin{aligned} v_1' &= \frac{w'}{\|w'\|} = \frac{w'}{\sqrt{1 + \|w\|^2}} \\ &= \frac{1}{\sqrt{1 + \|w\|^2}}\left(1 - \frac{\|w\| Z_1}{\left(\sum_{i=2}^n Z_i^2\right)^{1/2}}\right) v_1 + \frac{\|w\|\|Z\|}{\sqrt{1 + \|w\|^2}\left(\sum_{i=2}^n Z_i^2\right)^{1/2}} \frac{Z}{\|Z\|}. \end{aligned}$$

Taking

$$\mathbf{S} = \frac{\|w\|\|Z\|}{\left(\sum_{i=2}^n Z_i^2\right)^{1/2} - \|w\| Z_1},$$

we get

$$v_1' = \frac{1}{\sqrt{1+\|w\|^2}} \left(1 - \frac{\|w\| Z_1}{\left(\sum_{i=2}^n Z_i^2\right)^{1/2}}\right) \left(v_1 + \mathbf{S}\frac{Z}{\|Z\|}\right). \tag{3.12}$$

Proposition 3.2.1 together with Lemma 3.2.1 yield

$$\|w\|^2 = \|w' - v_1\|^2 = (1 + o_{\mathbb{P}}(1)) \cdot \xi^2 \sum_{i=2}^n \frac{\langle H v_i, v_1\rangle^2}{(\lambda_1 - \lambda_i)^2} \asymp \xi^2 n^{1/3},$$

the last quantity being $o(1)$ under assumption (3.6). With the previous computation, equation (3.12) becomes

$$v_1' = (1 + o_{\mathbb{P}}(1)) \left(v_1 + \mathbf{S}\frac{Z}{\|Z\|}\right),$$

with $\mathbf{S} = (1 + o_{\mathbb{P}}(1))\|w\| \asymp \xi n^{1/6}$. □

3.3. Definition and analysis of a toy model

Now that we have established a expansion of v_1' with respect to v_1, our main question boils down to the study of the effect of a random Gaussian perturbation of a Gaussian vector in terms of rank of its coordinates: if these ranks are preserved, the permutation that aligns these two vectors will be $\hat{\Pi} = \Pi^\star = I_n$. Otherwise we want to understand the error made between $\hat{\Pi}$ and $\Pi^\star = I_n$.

3.3.1. Definitions and notations

We refer to Section 3.1.2 for the definition of the toy model $\mathsf{J}(n, s)$. Recall that we want to compute, when $(X, Y) \sim \mathsf{J}(n, s)$, the probability

$$p(n, s) := \mathbb{P}\left(r_1(X) = r_1(Y)\right).$$

In this section, we denote by E the probability density function of a standard Gaussian variable, and F its cumulative distribution function. Namely

$$E(u) := \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \quad \text{and} \quad F(u) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^u e^{-z^2/2} dz.$$

We hereafter elaborate on the link between this toy model and our first matrix model (3.2) in Section 3.2. Since v_1 is uniformly distributed on the unit sphere, we have the equality in distribution $v_1 = \frac{X}{\|X\|}$ where X is a standard Gaussian vector of size n, independent of Z by Proposition 3.1.1. We write

$$\begin{aligned} v_1 &= \frac{X}{\|X\|}, \\ v_1' &= (1 + o_{\mathbb{P}}(1)) \left(\frac{X}{\|X\|} + \mathbf{S}\frac{Z}{\|Z\|}\right). \end{aligned}$$

Note that for all $\lambda > 0$, $r_1(\lambda T) = r_1(T)$, hence

$$r_1(v_1) = r_1(X), \quad r_1(v_1') = r_1\left(X + \mathbf{s}Z\right), \tag{3.13}$$

where

$$\mathbf{s} = \frac{\mathbf{S}\|X\|}{\|Z\|} \asymp \xi n^{1/6},$$

where we used the law of large numbers ($\|X\|/\|Z\| \to 1$ p.s.) as well as Proposition 3.1.1 in the last expansion. Equation (3.13) shows that this toy model is relevant for our initial problem, up to the fact that the noise term $\mathbf{s}$ is random in the matrix model (though we know its order of magnitude to be $\asymp \xi n^{1/6}$).

Remark 3.3.1. *The intuition for the zero-one law for $p(n,s)$ is as follows. If we sort the n coordinates of X on the real axis, all coordinates being typically perturbed by a factor s, it seems natural to compare s with the typical gap between two coordinates of order $1/n$ to decide whether the rank of the first coordinate of X is preserved in Y.*

Let us show that this intuition is rigorously verified. For every couple (x,y) of real numbers, define

$$\mathcal{N}^+_{n,s}(x,y) := |\{1 \le i \le n,\ X_i > x, Y_i < y\}|,$$
$$\mathcal{N}^-_{n,s}(x,y) := |\{1 \le i \le n,\ X_i < x, Y_i > y\}|.$$

In the following, we omit all dependencies in n and s, using the notations $\mathcal{N}^+$ and $\mathcal{N}^-$. The corresponding regions are shown on Figure 3.3. We will also need the following probabilities

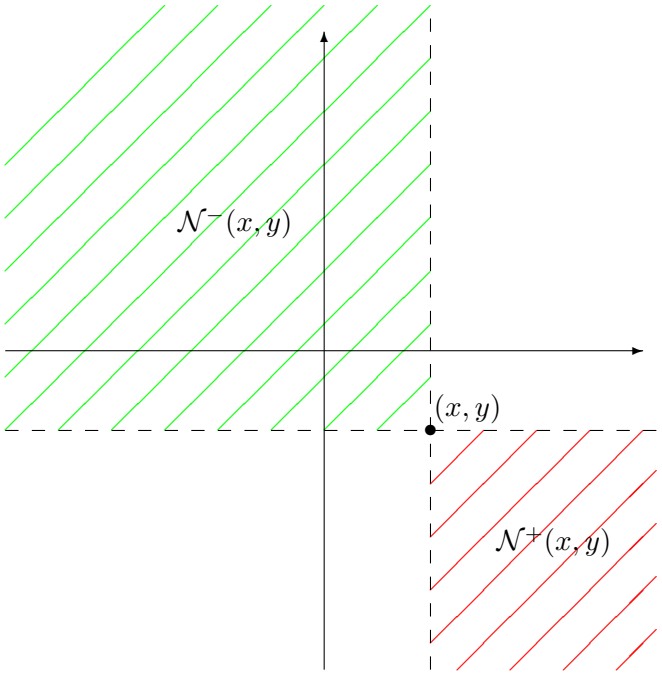


Figure 3.3 – *Areas corresponding to $\mathcal{N}^+(x,y)$ and $\mathcal{N}^-(x,y)$.*

$$S^+(x,y) := \mathbb{P}\left(X_1 > x, Y_1 < y\right), \text{ and}$$
$$S^-(x,y) := \mathbb{P}\left(X_1 < x, Y_1 > y\right) = S^+(-x,-y).$$

In terms of distribution, the random vector

$$\left(\mathcal{N}^+(x,y), \mathcal{N}^-(x,y), n-1-\mathcal{N}^+(x,y)-\mathcal{N}^+(x,y)\right)$$

follows a multinomial distribution of parameters

$$\left(n-1, S^+(x,y), S^-(x,y), 1-S^+(x,y)-S^-(x,y)\right).$$

In order to have $r_1(X) = r_1(Y)$, there must be the same number of points on the two domains

on Figure 3.3, for $x = X_1$ and $y = Y_1$. We then have the following expression of $p(n, s)$:

$$\begin{aligned} p(n,s) &= \mathbb{E}\left[\mathbb{P}\left(\mathcal{N}^+(X_1,Y_1) = \mathcal{N}^-(X_1,Y_1)\right)\right] \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{P}(dx,dy)\mathbb{P}(\mathcal{N}^+(x,y) = \mathcal{N}^-(x,y)) \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} E(x)E(z)\phi_{x,z}(n,s)\,dx\,dz, \end{aligned}$$

with

$$\phi_{x,z}(n,s) := \sum_{k=0}^{\lfloor (n-1)/2 \rfloor} \binom{n-1}{k}\binom{n-1-k}{k} \left(S^+_{x,z}\right)^k \left(S^-_{x,z}\right)^k \left(1 - S^+_{x,z} - S^-_{x,z}\right)^{n-1-2k}, \quad (3.14)$$

using the notations $S^+_{x,z} = S^+(x, x+sz)$ and $S^-_{x,z} = S^-(x, x+sz)$. A simple computation shows that

$$\begin{aligned} S^+(x,x+sz) &= \int_x^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \left(\int_{-\infty}^{z+\frac{x-u}{s}} \frac{1}{\sqrt{2\pi}} e^{-v^2/2}\,dv \right) du \\ &= \int_x^{+\infty} E(u)\,F\left(z - \frac{u-x}{s}\right) du, \quad (3.15) \\ &= s\int_0^{+\infty} E(x+vs)\,F(z-v)\,dv. \quad (3.16) \end{aligned}$$

We have the classical integration result

$$\int_{-\infty}^{z} F(u)du = zF(z) + E(z). \quad (3.17)$$

From (3.15), (3.16) and (3.17) we derive the following easy lemma:

Lemma 3.3.1. *For all x and z,*

$$\begin{aligned} S^+(x,x+sz) &\underset{s\to 0}{=} s\left[E(x)\left(zF(z) + E(z)\right)\right] + o(s), \\ S^+(x,x+sz) &\underset{s\to\infty}{\longrightarrow} F(z)\left(1 - F(x)\right), \\ S^-(x,x+sz) &\underset{s\to 0}{=} s\left[E(x)\left(-z + zF(z) + E(z)\right)\right] + o(s), \\ S^-(x,x+sz) &\underset{s\to\infty}{\longrightarrow} F(x)\left(1 - F(z)\right). \end{aligned}$$

Moreover, both $s \mapsto S^+(x, x+sz)$ and $s \mapsto S^-(x, x+sz)$ are increasing.

3.3.2. Zero-one law for $p(n, s)$

In this Section we give a proof of Proposition 3.1.2.

Proof of Proposition 3.1.2. ***First case*** *(i).* If $s = o(1/n)$, we have the following inequality

$$p(n,s) \geq \int_{\mathbb{R}}\int_{\mathbb{R}} dxdzE(x)E(z)\mathbb{P}\left(\mathcal{N}^+(x,x+sz) = \mathcal{N}^-(x,x+sz) = 0\right). \quad (3.18)$$

According to Lemma 3.3.1, for all $x, z \in \mathbb{R}$

$$\begin{aligned} \mathbb{P}\left(\mathcal{N}^+(x,x+sz) = \mathcal{N}^-(x,x+sz) = 0\right) &= \left(1 - S^+(x,x+sz) - S^-(x,x+sz)\right)^{n-1} \\ &\sim \exp\left(-nsE(x)\left[z(2F(z)-1) + 2E(z)\right]\right) \end{aligned}$$

$$\underset{n\to\infty}{\longrightarrow} 1,$$

By applying the dominated convergence theorem in (3.18), we conclude that $p(n,s) \to 1$.

Second case (ii)**.** If $sn \to \infty$, recall that

$$p(n,s) = \int_{\mathbb{R}}\int_{\mathbb{R}} dxdzE(x)E(z)\phi_{x,z}(n,s), \tag{3.19}$$

with $\phi_{x,z}$ defined in equation (3.14). In the rest of the proof, we fix x and z two real numbers. Letting

$$b(n,s,k) := \binom{n-1}{k}\left(S^+_{x,z}\right)^k\left(1-S^+_{x,z}\right)^{n-1-k}$$

and

$$M(n,s) := \max_{0\le k\le n-1} b(n,s,k).$$

Note that by Lemma 3.3.1, there exists $C = C(x,z) < 1$ such that for n large enough, $S^+_{x,z} < C < 1$. Moreover, combining this Lemma with assumption (ii) gives that $nS^+_{x,z} \to \infty$. It is also known that $M(n,s) = b(n,s,\lfloor nS^+_{x,z}\rfloor)$ and a classical computation shows that in this case (see e.g. [Bol01], formula 1.5):

$$\begin{aligned} M(n,s) &= \binom{n-1}{\lfloor nS^+_{x,z}\rfloor}\left(S^+_{x,z}\right)^{\lfloor nS^+_{x,z}\rfloor}\left(1-S^+_{x,z}\right)^{n-1-\lfloor nS^+_{x,z}\rfloor} \\ &\sim \frac{1}{\sqrt{2\pi nt(1-t)}}t^{-(n-1)t}(1-t)^{-(n-1)(1-t)}\left(S^+_{x,z}\right)^{(n-1)t}\left(1-S^+_{x,z}\right)^{(n-1)(1-t)} \\ &= \left(nS^+_{x,z}\right)^{-1/2}(1+O(1)) \to 0. \end{aligned}$$

where $t := \frac{\lfloor nS^+_{x,z}\rfloor}{n-1} \sim S^+_{x,z}$. Working with equation (3.14), we obtain the following control

$$\begin{aligned} \phi_{x,z}(n,s) &\le M(n,s)\times \sum_{k=0}^{\lfloor (n-1)/2\rfloor}\binom{n-1-k}{k}\left(S^-_{x,z}\right)^k\frac{\left(1-S^+_{x,z}-S^-_{x,z}\right)^{n-1-2k}}{\left(1-S^+_{x,z}\right)^{n-1-k}} \\ &\overset{(a)}{=} M(n,s)\times\frac{\left(1-S^+_{x,z}\right)\left(1-\left(\frac{-S^+_{x,z}}{1-S^-_{x,z}}\right)^n\right)}{1+S^-_{x,z}-S^+_{x,z}} \\ &\overset{(b)}{=} M(n,s)\times O(1) \underset{n\to\infty}{\longrightarrow} 0. \end{aligned}$$

We used in (b) the fact that $S^+_{x,z}+S^-_{x,z}$ is increasing in s, and that given x and z, for all $s>0$, by Lemma 3.3.1,

$$S^+_{x,z}+S^-_{x,z} < F(x)\left(1-F(z)\right)+F(z)\left(1-F(x)\right) < 1.$$

We used in (a) the following combinatorial result:

Lemma 3.3.2. *For all* $\alpha > 0$*,*

$$\sum_{k=0}^{\lfloor (n-1)/2\rfloor}\binom{n-1-k}{k}\alpha^k = \frac{1}{\sqrt{1+4\alpha}}\left[\left(\frac{1+\sqrt{1+4\alpha}}{2}\right)^n-\left(\frac{1-\sqrt{1+4\alpha}}{2}\right)^n\right]. \tag{3.20}$$

We refer to Appendix 3.B.1 for a proof of this result. To obtain (a) from Lemma 3.3.2, we apply (3.20) to $\alpha = \frac{S^-_{x,z}\left(1-S^+_{x,z}\right)}{\left(1-S^+_{x,z}-S^-_{x,z}\right)^2}$, with $\sqrt{1+4\alpha} = \frac{1-S^+_{x,z}+S^-_{x,z}}{1-S^+_{x,z}-S^-_{x,z}}$. Some simple simplifications

then give the claimed result. The dominated convergence theorem in (3.19) shows that $p(n,s) \to 0$ and ends the proof. □

Remark 3.3.2. *The above computations also imply the existence of a non-degenerate limit of $p(n,s)$ in the critical case where $sn \to c > 0$: in this case, previous discussions as well as Lemma 3.3.1 show that the joint distribution of $(\mathcal{N}^+(x, x+sz), \mathcal{N}^-(x, x+sz))$ is asymptotically*

$$\mathsf{Poi}(c\,[E(x)\,(zF(z)+E(z))]) \otimes \mathsf{Poi}(c\,[E(x)\,(-z+zF(z)+E(z))]).$$

Therefore, $p(n,s)$ has a non-degenerate limit given by

$$\int_{\mathbb{R}}\int_{\mathbb{R}} E(x)E(z)\cdot \mathbf{G}\left(c\,[E(x)\,(zF(z)+E(z))]\,, c\,[E(x)\,(-z+zF(z)+E(z))]\right)\,dx\,dz, \quad (3.21)$$

where

$$\mathbf{G}(a,b) := \mathbb{P}(\mathsf{Poi}(a) = \mathsf{Poi}(b)) = e^{-(a+b)}\sum_{k\geq 0}\frac{a^k b^k}{(k!)^2}. \quad (3.22)$$

3.4. Analysis of the EIG1 method for matrix alignment

By now, we come back to our initial problem, which is the analysis of EIG1 method. Recall that for any estimator $\hat{\Pi}$ of $\Pi^\star$, its overlap is defined as follows

$$\mathrm{ov}(\hat{\Pi}, \Pi^\star) := \frac{1}{n}\sum_{u=1}^{n} \mathbb{1}_{\hat{\Pi}(u)=\Pi^\star(u)}.$$

The aim of this section is to show how Propositions 3.1.1 and 3.1.2 can be assembled to show the main result of our study, namely Theorem 3.1.

Proof of Theorem 3.1. ***First case*** *(i).* Assuming $\xi = o(n^{-7/6-\epsilon})$ for some $\epsilon > 0$, then in particular condition (3.6) holds. Proposition 3.1.1 as well as equation (3.13) in Section 3.3 enable to identify v_1 and v_1' with the following vectors:

$$v_1 \sim X, \quad v_1' \sim X + \mathbf{s}Z, \quad (3.23)$$

where X and Z are two independent Gaussian vectors from the toy model, and where $\mathbf{s} \asymp \xi n^{1/6}$ w.h.p. Recall that we work under the assumptions $\Pi^\star = I_n$ and $\langle v_1, v_1'\rangle > 0$. In this case, we expect Π_+ to be very close to I_n.

We will use the notations of Section 3.3 hereafter. Let's take $f \in \mathcal{F}$ such that w.h.p., $\xi n^{1/6} f(n)^{-1} \leq \mathbf{s} \leq \xi n^{1/6} f(n)$. We have for all $1 \leq i \leq n$,

$$\begin{aligned}
\mathbb{P}\left(\Pi_+(i) = \Pi^\star(u)\right) &= \mathbb{P}\left(\Pi_+(1) = \Pi^\star(1)\right) \\
&= \mathbb{E}\left[\iint dxdzE(x)E(z)\phi_{x,z}\left(n,\mathbf{s}\right)\mathbb{1}_{\xi n^{1/6}f(n)^{-1}\leq \mathbf{s}\leq \xi n^{1/6}f(n)}\right] + o(1) \\
&= \iint dxdzE(x)E(z)\mathbb{E}\left[\phi_{x,z}\left(n,\mathbf{s}\right)\mathbb{1}_{\xi n^{1/6}f(n)^{-1}\leq \mathbf{s}\leq \xi n^{1/6}f(n)}\right] + o(1).
\end{aligned}$$

When conditioning on the event $\mathcal{A}$ where $\xi n^{1/6} f(n)^{-1} \leq \mathbf{s} \leq \xi n^{1/6} f(n)$, we know that $\mathbf{s}n \to 0$ by condition (i) and for all x, z, $\mathbb{E}\left[\phi_{x,z}\left(n,\mathbf{s}\right) \mid \mathcal{A}\right] \to 1$ as shown in Section 3.3. Since $\mathcal{A}$ occurs w.h.p. we have

$$\mathbb{E}\left[\phi_{x,z}\left(n,\mathbf{s}\right)\mathbb{1}_{\mathcal{A}}\right] \longrightarrow 1,$$

which implies with the dominated convergence theorem that

$$\mathbb{E}\left[\mathrm{ov}(\Pi_+, \Pi^\star)\right] \underset{n\to\infty}{\longrightarrow} 1 \tag{3.24}$$

and thus

$$\mathrm{ov}(\Pi_+, \Pi^\star) \overset{L^1}{\to} 1.$$

We now check that w.h.p., Π_+ is preferred to Π_- in the EIG1 method:

Lemma 3.4.1. *In the case (i) of Theorem 3.1, if $\langle v_1, v_1' \rangle > 0$, we have w.h.p.*

$$\langle A, \Pi_+ B \Pi_+^\top \rangle > \langle A, \Pi_- B \Pi_-^\top \rangle,$$

in other words Algorithm EIG1 returns w.h.p. $\hat{\Pi} = \Pi_+$.

This Lemma is proved in Appendix 3.B.3 and implies, together with (3.24), that

$$\begin{aligned}
\mathbb{E}\left[\mathrm{ov}(\hat{\Pi}, \Pi^\star)\right] &\geq \mathbb{E}\left[\mathrm{ov}(\hat{\Pi}, \Pi^\star)\mathbb{1}_{\hat{\Pi}=\Pi_+}\right] = \mathbb{E}\left[\mathrm{ov}(\Pi_+, \Pi^\star)\mathbb{1}_{\hat{\Pi}=\Pi_+}\right] \\
&= \mathbb{E}\left[\mathrm{ov}(\Pi_+, \Pi^\star)\right] - \mathbb{E}\left[\mathrm{ov}(\Pi_+, \Pi^\star)\mathbb{1}_{\hat{\Pi}=\Pi_-}\right] \\
&= 1 - o(1).
\end{aligned}$$

and thus

$$\mathrm{ov}(\hat{\Pi}, \Pi^\star) \underset{n\to\infty}{\overset{L^1}{\longrightarrow}} 1. \tag{3.25}$$

Second case (ii)**.** If condition (3.6) is verified then the identification (3.23) still holds and the proof of case (i) adapts well. However, if (3.6) is not verified, we can still make a link with the toy model studied in Section 3.3. Let's use a simple coupling argument: if $\xi = \omega(n^{-1/2-\alpha})$ for some $\alpha \geq 0$, let's take $\xi_1, \xi_2 > 0$ such that

$$\xi^2 = \xi_1^2 + \xi_2^2$$

and

$$n^{-7/6+\epsilon} \ll \xi_1 \ll n^{-1/2-\alpha},$$

fixing for instance $\xi_1 = n^{-1}$. We will use the notation $\widetilde{v}_1$, now viewed as the leading eigenvector of the matrix

$$\widetilde{B} = A + \xi_1 H + \xi_2 \widetilde{H},$$

where $\widetilde{H}$ is an independent copy of H. This has no consequence in terms of distribution : $(A, \widetilde{B})$ is still drawn under model (3.2). Let's denote v_1' the leading eigenvector of $B_1 = A + \xi_1 H$, chosen so that $\langle v_1, v_1' \rangle > 0$. It is clear that condition (3.6) holds for ξ_1. We have the following result, based on the invariance by rotation of the GOE distribution:

Lemma 3.4.2. *We still have the following equality in distribution:*

$$\left(r_1(v_1), r_1(\widetilde{v}_1)\right) \overset{(d)}{=} \left(r_1(X), r_1(X + \mathbf{s}Z)\right),$$

where X, Z are two standard Gaussian vectors from the toy model, with w.h.p.

$$\mathbf{s} \geq \mathbf{s^1} \asymp \xi_1 n^{1/6}.$$

We refer to Appendix 3.B.2 for a proof. Since w.h.p. $\mathbf{s} \geq \mathbf{s^1}$ and $\mathbf{s^1} n \asymp \xi_1 n^{7/6} \to \infty$, we have for all $1 \leq i \leq n$,

$$\mathbb{P}\left(\Pi_+(i) = \Pi^\star(u)\right) = \mathbb{P}\left(\Pi_+(1) = \Pi^\star(1)\right)$$

$$= \mathbb{E}\left[\iint dxdzE(x)E(z)\phi_{x,z}(n,\mathbf{s})\mathbb{1}_{\mathbf{s}n\to\infty}\right] + o(1)$$
$$= \iint dxdzE(x)E(z)\mathbb{E}\left[\phi_{x,z}(n,\mathbf{s})\mathbb{1}_{\mathbf{s}n\to\infty}\right] + o(1).$$

With the same arguments as in the case (i), we show that $\phi_{x,z}(n,\mathbf{s})\mathbb{1}_{\mathbf{s}n\to\infty} \xrightarrow{L^1} 0$, which implies

$$\mathbb{E}\left[\mathrm{ov}(\Pi_+,\Pi^\star)\right] \underset{n\to\infty}{\longrightarrow} 0,$$

hence $\mathrm{ov}(\Pi_+,\Pi^\star) \underset{n\to\infty}{\xrightarrow{L^1}} 0$. The last step is to verify that the overlap achieved by Π_- does not outperform that of Π_+. We prove the following Lemma in Appendix 3.B.4:

Lemma 3.4.3. *In the case (ii), if $\langle v_1, v_1'\rangle > 0$, we also have*

$$\mathrm{ov}(\Pi_-,\Pi^\star) \underset{n\to\infty}{\xrightarrow{L^1}} 0.$$

Lemma 3.4.3 then gives

$$\mathbb{E}\left[\mathrm{ov}(\hat{\Pi},\Pi^\star)\right] \leq \mathbb{E}\left[\mathrm{ov}(\Pi_+,\Pi^\star)\right] + \mathbb{E}\left[\mathrm{ov}(\Pi_-,\Pi^\star)\right] \underset{n\to\infty}{\longrightarrow} 0,$$

and thus

$$\mathrm{ov}(\hat{\Pi},\Pi^\star) \underset{n\to\infty}{\xrightarrow{L^1}} 0. \tag{3.26}$$

Of course, the convergences in (3.25) and (3.26) also hold in probability, by Markov's inequality. □

Appendix of Chapter 3

3.A. Additional proofs for Section 3.2

Throughout the proofs, all variables denoted by C_i with $i = 1, 2, \dots$ are unspecified, independent, positive constants.

3.A.1. Proof of Proposition 3.2.1

Proof of Proposition 3.2.1. Let us establish a first inequality: since the GOE distribution is invariant by rotation (see e.g. [AGZ09]), the random variables $\langle Hv_j, v_i\rangle$ are zero-mean Gaussian, with variance $1/n$ of $i \neq j$ and $2/n$ if $i = j$. Hence, w.h.p.

$$\sup_{1\leq i,j\leq n} |\langle Hv_j, v_i\rangle| \leq C_1\sqrt{\frac{\log n}{n}}. \tag{3.27}$$

We will use the following short-hand notation for $1 \leq i, j \leq n$:

$$m_{i,j} := \langle Hv_j, v_i\rangle,$$

The defining eigenvector equations projected on vectors v_i write

$$\begin{cases} \theta_i &= \dfrac{\xi}{\lambda_1' - \lambda_i} \sum_{j=1}^n \theta_j m_{i,j}, \\ \lambda_1' - \lambda_1 &= \xi \sum_{j=1}^n \theta_j m_{1,j}. \end{cases} \tag{3.28}$$

In order to approximate the θ_i variables, we define the following iterative scheme:

$$\begin{cases} \theta_i^k &= \dfrac{\xi}{\lambda_1^{k-1} - \lambda_i} \sum_{j=1}^n \theta_j^{k-1} m_{i,j}, \\ \lambda_1^k - \lambda_1 &= \xi \sum_{j=1}^n \theta_j^{k-1} m_{1,j}, \end{cases} \tag{3.29}$$

with initial conditions $\left(\theta_i^0\right)_{2\leq i\leq n} = 0$ and $\lambda_1^0 = \lambda_1$, and setting $\theta_1^k = 1$ for all k. For $k \geq 1$, define

$$\Delta_k := \sum_{i\geq 2} \left|\theta_i^k - \theta_i^{k-1}\right|,$$

and for $k \geq 0$,

$$S_k := \sum_{i\geq 1} \left|\theta_i^k\right|.$$

Recall that under assumption (3.6), there exists $\alpha > 0$ such that $\xi = o\left(n^{-1/2-\alpha}\right)$. We define ϵ as follows:

$$\epsilon = \epsilon(n) = \sqrt{\xi n^{1/2+\alpha}}.$$

The idea is to show that the sequence $\{\Delta_k\}_{k\geq 1}$ decreases geometrically with k at rate ϵ. More specifically, we show the following result:

Lemma 3.A.1. *With the same notations and under the assumption (3.6) of Proposition 3.2.1, one has w.h.p.*

(*i*) $\forall k \geq 1,\ \Delta_k \leq \Delta_1 \epsilon^{k-1}$,

(*ii*) $\forall k \geq 0, \forall\, 2 \leq i \leq n,\ \left|\lambda_1^k - \lambda_i\right| \geq \frac{1}{2}\left|\lambda_1 - \lambda_i\right| \left(1 - \epsilon - \ldots - \epsilon^{k-1}\right)$,

(*iii*) $\forall k \geq 0,\ S_k \leq 1 + (1 + \ldots + \epsilon^{k-1})\Delta_1$,

(*iv*) $\sum_{i=2}^n \left|\theta_i - \theta_i^1\right|^2 = o\left(\sum_{i=2}^n \left|\theta_i^1\right|^2\right)$.

This Lemma is proved in the next section. Equation (*iv*) of Lemma 3.A.1 yields

$$\begin{aligned} w' &= v_1 + \sum_{i=2}^n \theta_i^1 v_i + \sum_{i=2}^n \left(\theta_i - \theta_i^1\right) v_i \\ &= v_1 + \xi \sum_{i=2}^n \frac{\langle H v_i, v_1\rangle}{\lambda_1 - \lambda_i} v_i + o_{\mathbb{P}}\left(\xi \sum_{i=2}^n \frac{\langle H v_i, v_1\rangle}{\lambda_1 - \lambda_i} v_i\right). \end{aligned}$$

□

3.A.2. Proof of Lemma 3.A.1

Proof of Lemma 3.A.1. In this proof we will use the same notations as defined in the proof of Proposition 3.2.1, and we make the assumption (3.6). We now state three technical lemmas controlling some statistics of eigenvalues in the GOE which are useful hereafter.

Lemma 3.A.2. *W.h.p., for all $\delta > 0$,*

$$\sum_{j=2}^n \frac{1}{\lambda_1 - \lambda_j} \leq O\left(n^{1+\delta}\right). \tag{3.30}$$

Lemma 3.A.3. *We have*

$$\sum_{j=2}^n \frac{1}{\left(\lambda_1 - \lambda_j\right)^2} \asymp n^{4/3}. \tag{3.31}$$

Lemma 3.A.4. *For any $C > 0$, w.h.p.*

$$\lambda_1 - \lambda_2 \geq n^{-2/3} \left(\log n\right)^{-C \log\log n}. \tag{3.32}$$

Proofs of these three Lemmas can be found in the next sections. We will work under the event (that occurs w.h.p.) on which the equations (3.30), (3.31), (3.32), (3.10) and (3.27) are satisfied. We show the following inequalities:

(*i*) $\forall k \geq 1,\ \Delta_k \leq \Delta_1 \epsilon^{k-1}$,

(*ii*) $\forall k \geq 0, \forall\, 2 \leq i \leq n,\ \left|\lambda_1^k - \lambda_i\right| \geq \frac{1}{2}\left|\lambda_1 - \lambda_i\right| \left(1 - \epsilon - \ldots - \epsilon^{k-1}\right)$,

(*iii*) $\forall k \geq 0,\ S_k \leq 1 + (1 + \ldots + \epsilon^{k-1})\Delta_1$.

Recall that ϵ is given by

$$\epsilon = \epsilon(n) = \sqrt{\xi n^{1/2+\alpha}}.$$

We will denote by $f_i(n)$, with i an integer, functions as defined in Lemma 3.A.3. All the following inequality will be valid for n large enough (uniformly in i and in k).

Step 1: propagation of the first equation. Let $k \geq 3$. We work by induction, assuming that (i), (ii) and (iii) are verified until $k-1$.

$$\begin{aligned}
\left|\theta_i^k - \theta_i^{k-1}\right| &\leq \left|\frac{\xi}{\lambda_1^{k-1} - \lambda_i} \sum_{j=2}^{n} \left(\theta_j^{k-1} - \theta_j^{k-2}\right) m_{i,j}\right| + \left|\frac{\xi\left(\lambda_1^{k-2} - \lambda_1^{k-1}\right)}{\left(\lambda_1^{k-1} - \lambda_i\right)\left(\lambda_1^{k-2} - \lambda_i\right)} \sum_{j=1}^{n} \theta_j^{k-2} m_{i,j}\right| \\
&\leq \frac{\xi}{\left|\lambda_1^{k-1} - \lambda_i\right|} C_1 \sqrt{\frac{\log n}{n}} \Delta_{k-1} + \xi C_1 \sqrt{\frac{\log n}{n}} S_{k-2} \frac{\left|\lambda_1^{k-2} - \lambda_1^{k-1}\right|}{\left|\lambda_1^{k-1} - \lambda_i\right|\left|\lambda_1^{k-2} - \lambda_i\right|} \\
&\overset{(a)}{\leq} \xi \frac{3}{|\lambda_1 - \lambda_i|} C_1 \sqrt{\frac{\log n}{n}} \Delta_{k-1} + \xi C_1 \sqrt{\frac{\log n}{n}} S_{k-2} \frac{9\left|\lambda_1^{k-2} - \lambda_1^{k-1}\right|}{|\lambda_1 - \lambda_i|^2} \\
&\overset{(b)}{\leq} \xi \frac{3}{|\lambda_1 - \lambda_i|} C_1 \sqrt{\frac{\log n}{n}} \Delta_{k-1} + \xi C_1 \sqrt{\frac{\log n}{n}} 2 \frac{9\left|\lambda_1^{k-2} - \lambda_1^{k-1}\right|}{|\lambda_1 - \lambda_i|^2}.
\end{aligned}$$

We applied (ii) to $k-1, k-2$ in (a) and (iii) to $k-2$ in (b). Note that

$$\left|\lambda_1^{k-2} - \lambda_1^{k-1}\right| = \left|\xi \sum_{j=1}^{n} \left(\theta_j^{k-2} - \theta_j^{k-3}\right) m_{i,j}\right| \leq \xi C_1 \sqrt{\frac{\log n}{n}} \Delta_{k-2},$$

which yields the inequality:

$$\left|\theta_i^k - \theta_i^{k-1}\right| \leq \frac{\xi}{|\lambda_1 - \lambda_i|} f_1(n) n^{-1/2} \Delta_{k-1} + \frac{\xi^2}{|\lambda_1 - \lambda_i|^2} f_2(n) n^{-1} \Delta_{k-2}.$$

We choose δ such that $0 < \delta < \alpha$ (where α is fixed by (3.6)), and we sum from $i = 2$ to n:

$$\begin{aligned}
\Delta_k &\leq \xi f_1(n) n^{1/2+\delta} \Delta_{k-1} + \xi^2 f_3(n) n^{1/3} \Delta_{k-2} \\
&\overset{(a)}{\leq} o(\epsilon)\epsilon^{k-2}\Delta_1 + o(\epsilon^2)\epsilon^{k-3}\Delta_1 \\
&\leq \epsilon^{k-1}\Delta_1.
\end{aligned}$$

We used $\xi f_1(n) n^{1/2+\delta} = o(\epsilon)$, $\xi^2 f_3(n) n^{1/3} = o(\epsilon^2)$ and we applied (i) to $k-1$ and $k-2$ in (a).

Step 2: propagation of the second equation. Let $k \geq 2$, and $0 < \delta < \alpha$. We work by induction, assuming that (i), (ii) and (iii) are verified until $k-1$.

$$\begin{aligned}
\left|\lambda_1^k - \lambda_1^{k-1}\right| &\leq \xi f_1(n) n^{-1/2} \Delta_{k-1} \\
&\overset{(a)}{\leq} \xi f_1(n) n^{-1/2} \epsilon^{k-2} \Delta_1 \\
&\leq n^{-2/3} (\log n)^{-C \log\log n} \epsilon^{k-2} \Delta_1 \\
&\leq \frac{\lambda_1 - \lambda_2}{2} \epsilon^{k-2} \Delta_1
\end{aligned}$$

$$\leq \frac{\lambda_1 - \lambda_i}{2}\epsilon^{k-2}\Delta_1.$$

We applied (i) to $k-1$ in (a). Note that

$$\Delta_1 = \sum_{j=2}^{n} \frac{\xi}{\lambda_1 - \lambda_i} |m_{i,1}| \leq \xi f_1(n) n^{1/2+\delta} \leq o(\epsilon).$$

Applying (ii) to $k-1$, we get

$$\begin{aligned}
\left|\lambda_1^k - \lambda_i\right| &\geq \left|\lambda_1 - \lambda_1^{k-1}\right| - \left|\lambda_1^k - \lambda_1^{k-1}\right| \\
&\geq \frac{\lambda_1 - \lambda_i}{2}\left(1 - \epsilon - \ldots - \epsilon^{k-2}\right) - \frac{\lambda_1 - \lambda_i}{2}\epsilon^{k-1} \\
&\geq \frac{\lambda_1 - \lambda_i}{2}\left(1 - \epsilon - \ldots - \epsilon^{k-1}\right).
\end{aligned}$$

Step 3: propagation of the third equation. Let $k \geq 1$. Here again, we work by induction, assuming that (i), (ii) and (iii) are verified until $k-1$.

$$\begin{aligned}
S_k &= 1 + \sum_{j=2}^{n} \left|\theta_j^k\right| \\
&\leq 1 + \Delta_k + S_{k-1} - 1 \\
&\overset{(a)}{\leq} \epsilon^{k-1}\Delta_1 + 1 + \left(1 + \ldots + \epsilon^{k-2}\right)\Delta_1 \\
&\leq 1 + \left(1 + \epsilon + \ldots + \epsilon^{k-1}\right)\Delta_1.
\end{aligned}$$

We applied (i) to k and (iii) to $k-1$ in (a).

Step 4: Proof of (i) ***for*** $k = 1, 2$***,*** (ii) ***for*** $k = 0, 1$ ***and*** (iii) ***for*** $k = 0, 1$**.** The equation (i) for $k = 1$ is obvious. For $k = 2$:

$$\left|\theta_i^2 - \theta_i^1\right| \leq \left|\frac{\xi}{\lambda_1^1 - \lambda_i}\sum_{j=2}^{n}\left(\theta_j^1 - \theta_j^0\right)m_{i,j}\right| + \left|\frac{\xi\left(\lambda_1^0 - \lambda_1^1\right)}{\left(\lambda_1^1 - \lambda_i\right)\left(\lambda_1^0 - \lambda_i\right)}\sum_{j=1}^{n}\theta_j^0 m_{i,j}\right|.$$

We have

$$\begin{aligned}
\left|\lambda_1^1 - \lambda_i\right| &\geq \left|\lambda_1 - \lambda_i\right| - \left|\lambda_1 - \lambda_1^1\right| \geq \left|\lambda_1 - \lambda_i\right| - \xi\left|m_{1,1}\right| \\
&\geq \left|\lambda_1 - \lambda_i\right| - \frac{1}{2}\left|\lambda_1 - \lambda_2\right| \geq \frac{1}{2}\left|\lambda_1 - \lambda_i\right|,
\end{aligned}$$

which shows (ii) for $k = 0, 1$. Thus, for $0 < \delta < \alpha$:

$$\left|\theta_i^2 - \theta_i^1\right| \leq \frac{2\xi}{\lambda_1 - \lambda_i}C_1\sqrt{\frac{\log n}{n}}\Delta_1 + \frac{4\xi}{\left(\lambda_1 - \lambda_i\right)^2}\xi\left|m_{1,1}\right|\left|m_{i,1}\right|,$$

and

$$\begin{aligned}
\Delta_2 &\leq \xi f_1(n) n^{1/2+\delta}\Delta_1 + 4\xi\left|m_{1,1}\right|\sum_{i=2}^{n}\frac{\xi\left|m_{i,1}\right|}{\left(\lambda_1 - \lambda_i\right)^2} \\
&\leq \xi f_1(n) n^{1/2+\delta}\Delta_1 + 4\xi f_4(n) n^{-1/2} n^{2/3}\sum_{i=2}^{n}\frac{\xi\left|m_{i,1}\right|}{\left(\lambda_1 - \lambda_i\right)}
\end{aligned}$$

$$\leq \xi f_1(n) n^{1/2+\delta} \Delta_1 + 4\xi f_4(n) n^{1/6} \Delta_1$$
$$\leq \epsilon \Delta_1.$$

The proof of (iii) for $k = 0, 1$ is obvious.

Step 5: Proof of equation (iv)**.** Let $k \geq 2$ and $2 \leq i \leq n$. In the same way as in Step 1, we have

$$\left|\theta_i^k - \theta_i^{k-1}\right| \leq \frac{2\xi C_1}{\lambda_1 - \lambda_i} \sqrt{\frac{\log n}{n}} \epsilon^{k-2} \Delta_1 + \frac{8\xi^2 C_1^2}{(\lambda_1 - \lambda_i)^2} \frac{\log n}{n} \epsilon^{(k-3)_+} \Delta_1.$$

In the right-hand term, the ratio of the second term on the first one is smaller that

$$\frac{4\xi C_1}{\lambda_1 - \lambda_i} \sqrt{\frac{\log n}{n}} \epsilon^{-1} \leq \xi n^{1/6} f(n) \epsilon^{-1} \leq \epsilon \to 0,$$

using Lemma 3.A.4, with $f \in \mathcal{F}$. It follows that for n big enough (uniformly in k and i) one has

$$\left|\theta_i^k - \theta_i^{k-1}\right| \leq \frac{\xi f(n)}{\lambda_1 - \lambda_i} n^{-1/2} \epsilon^{k-2} \Delta_1. \tag{3.33}$$

Equation (3.33) shows that the scheme (3.29) converges, and that the limits are indeed the solutions $\theta_1 = 1, \theta_2, \ldots, \theta_n$ of the fixed-point equations. By a simple summation of (3.33) over $k \geq 2$, applying Lemma 3.A.2 and inequality (3.27) we have

$$\left|\theta_i - \theta_i^1\right| \leq \frac{2\xi f(n)}{\lambda_1 - \lambda_i} n^{-1/2} \Delta_1 \leq \frac{2\xi^2 f(n)}{\lambda_1 - \lambda_i} n^{\delta},$$

where $\delta > 0$ is a positive quantity of Lemma 3.A.2 specified later. Using Lemma 3.A.3 one has the following control

$$\sum_{i=2}^{n} \left|\theta_i - \theta_i^1\right|^2 \leq 4\xi^4 n^{2\delta} f(n) n^{4/3}.$$

Moreover, Lemma 3.2.1 shows that

$$\sum_{i=2}^{n} \left|\theta_i^1\right|^2 \asymp \xi^2 n^{1/3} \geq g(n)^{-1} \xi^2 n^{1/3},$$

where g is another function in $\mathcal{F}$. This yields

$$\sum_{i=2}^{n} \left|\theta_i - \theta_i^1\right|^2 \leq \sum_{i=2}^{n} \left|\theta_i^1\right|^2 4\xi^2 n^{2\delta+1} f(n) g(n).$$

The proof is completed by taking $\delta = \alpha/2$ and applying (3.6). □

3.A.3. Proof of Lemma 3.A.4

Proof of Lemma 3.A.4. This lemma provides a control of the spectral gap $\lambda_1 - \lambda_2$. Given a good rescaling (in $n^{2/3}$), the asymptotic joint law of the eigenvalues in the edge has been investigated in a great amount of research work, for Gaussian ensembles, and for more general Wigner matrices. The GOE case has been mostly studied by Tracy, Widom, and Forrester among many others; in [For93] and [TW98], the convergence of the joint distribution of the first k eigenvalues towards a density distribution is established:

Proposition 3.A.1 ([For93], [TW98]). *For a given $k \geq 1$, and all $s_1, \ldots, s_k$ real numbers,*

$$\mathbb{P}\left(n^{2/3}(\lambda_1 - 2) \leq s_1, \ldots, n^{2/3}(\lambda_k - 2) \leq s_k\right) \underset{n\to\infty}{\longrightarrow} \mathcal{F}_{1,k}(s_1, \ldots, s_k), \tag{3.34}$$

where the $\mathcal{F}_{1,k}$ are continuous and can be expressed as solutions of non linear PDEs. Thus the re-scaled spectral gap $n^{2/3}(\lambda_1 - \lambda_2)$ has a limit probability density law supported by $\mathbb{R}_+$, which implies that

$$\mathbb{P}\left(n^{2/3}(\lambda_1 - \lambda_2) \geq (\log n)^{-C\log\log n}\right) \underset{n\to\infty}{\longrightarrow} 1.$$

Of course, the choice of the function $n \mapsto (\log n)^{-C\log\log n}$ is here arbitrary and the result is also true for any function tending to 0. □

3.A.4. Proof of Lemma 3.A.3

Proof of Lemma 3.A.3. This result needs an understanding of the behavior of the spectral gaps of matrix A, in the bulk and in the edges (left and right). The eigenvalues in the *edge* correspond to indices i such that $i = o(n)$ (left) or $i = n - o(n)$ (right). Eigenvalues in the *bulk* are the remaining eigenvalues. For this, we use a result of rigidity of eigenvalues, due to L. Erdös et al. [EYY10], which consists in a control of the probability of the gap between the eigenvales of A and the typical eigenvalues γ_j of the semi-circle law, defined as follows

$$\forall i \in \{1, \ldots, n\}, \ \frac{1}{2\pi}\int_{-2}^{\gamma_j} \sqrt{4 - x^2}dx = 1 - \frac{j}{n}. \tag{3.35}$$

Proposition 3.A.2 ([EYY10]). *For some positive constants $C_5 > 0$ and $C_6 > 0$, for n large enough,*

$$\mathbb{P}\left(\exists j \in \{1, \ldots, n\} \mid |\lambda_j - \gamma_j| \geq (\log n)^{C_5\log\log n}\left(\min(j, n+1-j)\right)^{-1/3} n^{-2/3}\right) \leq C_5 \exp\left(-(\log n)^{C_6\log\log n}\right). \tag{3.36}$$

Remark 3.A.1. *Another similar result that goes in the same direction for the* GOE *is already known: it has been shown by O'Rourke in [O'R10] that the variables $\lambda_i - \gamma_i$ behave as Gaussian variables when $n \to \infty$. However, the rigidity result in* (3.36) *obtained in [EYY10] can apply in more general models. This quantitative probabilistic statement was not previously known even for the* GOE *case.*

Remark 3.A.2. *Let us note that one of the assumptions made in [EYY10] is that variances of each column sum to 1, which is not directly the case in our model* (3.2)*. Nevertheless, one may use (3.36) for the re-scaled matrix $\tilde{A} := A\left(1 + \frac{1}{n}\right)^{-1/2}$, then easily check that there is a possible step back to A: $|\lambda_j - \gamma_j| \leq \left|\lambda_j\left(1 + \frac{1}{n}\right)^{-1/2} - \gamma_j\right| + n^{-1} + o(n^{-1})$, and $n^{-1} + o(n^{-1}) \leq 2\left(\min(j, n+1-j)\right)^{-1/3} n^{-2/3}$ for n big enough. Tolerating a slight increase of the constant C_5, the result (3.36) is thus valid in the* GOE*.*

Let us now compute an asymptotic expansion of γ_j in the right edge, which is for $j = o(n)$. Define

$$G(x) := \frac{1}{2\pi}\int_{-2}^{x} \sqrt{4 - t^2}dt = \frac{x\sqrt{4 - x^2} + 4\arcsin(x/2)}{4\pi} + \frac{1}{2}, \tag{3.37}$$

for all $x \in [-2, 2]$. We have $\gamma_j = G^{-1}(1 - j/n) = -G^{-1}(j/n)$, observing that the integrand in (3.37) is an even function. We get the following expansion when $x \to -2$,

$$G(x) \underset{x\to-2}{=} \frac{2(x+2)^{3/2}}{3\pi} + o\left((x+2)^{3/2}\right)$$

which implies that

$$G^{-1}(y) \underset{y\to 0}{=} -2 + \left(\frac{3\pi y}{2}\right)^{2/3} + o\left(y^{2/3}\right),$$

hence

$$\gamma_j \underset{j/n\to 0}{=} 2 - \left(\frac{3\pi j}{2n}\right)^{2/3} + o\left((j/n)^{2/3}\right). \tag{3.38}$$

Remark 3.A.3. *One can observe the coherence of this result that arises naturally in [O'R10] as the expectation of the eigenvalues in the edge.*

Let $\epsilon > 0$, to be specified later. To establish our result we will split the variables j in three sets:

$$\begin{aligned}
A_1 &:= \left\{2 \le j \le (\log n)^{(C_5+1)\log\log n}\right\} \text{ (a small part of the right edge)},\\
A_2 &:= \left\{(\log n)^{(C_5+1)\log\log n} < j \le n^{1-\epsilon}\right\} \text{ (a larger part of the right edge)},\\
A_3 &:= \left\{n^{1-\epsilon} < j \le n\right\} \text{ (everything else)}.
\end{aligned}$$

We show that the sum over A_1 is the major contribution in (3.31). The split in the right edge in A_1 and A_2 is driven by the error term of (3.36): this term is small compared to γ_j if and only if $(\log n)^{C_5 \log\log n} = o(j)$.

Step 1: estimation of the sum over A_1. According to (3.36) and Lemma 3.A.4, w.h.p.

$$n^{-4/3}(\log n)^{-C_6\log\log n} \le (\lambda_1 - \lambda_2)^2 \le C_7 n^{-4/3}(\log n)^{C_6\log\log n},$$

where C_6, C_7 are positive constants. Hence, w.h.p.

$$\begin{aligned}
\frac{n^{4/3}}{C_7(\log n)^{C_6\log\log n}} &\le \sum_{j\in A_1}\frac{1}{(\lambda_1-\lambda_j)^2}\\
&\le \sum_{j\in A_1}\frac{1}{(\lambda_1-\lambda_2)^2}\\
&\le n^{4/3}(\log n)^{(C_5+C_6+1)\log\log n}.
\end{aligned}$$

Step 2: estimation of the sum over A_2. Let us show that the sum over A_2 is asymptotically small compared to the sum over A_1: using (3.36) and (3.38), we know that there exists $C_8 > 0$ such that for all $j \in A_2$, w.h.p.

$$\lambda_j = 2 - C_8\left(\frac{j}{n}\right)^{2/3} + o\left((j/n)^{2/3}\right),$$

and we know furthermore (se e.g. [AGZ09]) that w.h.p.

$$\lambda_1 = 2 + o\left((j/n)^{2/3}\right), \forall j \in A_2 \tag{3.39}$$

hence w.h.p.

$$\begin{aligned}
\sum_{j\in A_2}\frac{1}{(\lambda_1-\lambda_j)^2} &= n^{4/3}\sum_{j\in A_2}\frac{1}{C_9 j^{4/3}(1+o(1))}\\
&= n^{4/3}(1+o(1))\sum_{j\in A_2}\frac{1}{C_9 j^{4/3}} = o\left(n^{4/3}\right),
\end{aligned}$$

using in the last line the fact that the Riemann's series $\sum j^{-4/3}$ converges.

Step 3: estimation of the sum under A_3. With the previous results (3.36), (3.38) and (3.39), assuming that $\epsilon < 1$, we get w.h.p.

$$\lambda_1 - \lambda_{n^{1-\epsilon}} = C_8 n^{-2\epsilon/3} + O\left(n^{-2\epsilon/3}\right),$$

which gives w.h.p. the following control

$$\begin{aligned}\sum_{j\in A_3} \frac{1}{(\lambda_1-\lambda_j)^2} &\leq \left(n - n^{1-\epsilon}\right) \frac{1}{(\lambda_1 - \lambda_{n^{1-\epsilon}})^2} \\ &= \left(n - n^{1-\epsilon}\right) \frac{n^{4\epsilon/3}}{C_9(1+o(1))} = O\left(n^{1+4\epsilon/3}\right) = o\left(n^{4/3}\right),\end{aligned}$$

as long as $\epsilon < 1/4$. Taking such a ϵ, these three controls end the proof. □

3.A.5. Proof of Lemma 3.A.2

Proof of Lemma 3.A.2. We follow the same steps as in the proof of Lemma 3.A.3. Let's take $\delta > 0$. We split the j variables in three sets:

$$\begin{aligned}A_1 &:= \left\{2 \leq j \leq n^{1/3}\right\}, \\ A_2 &:= \left\{n^{1/3} < j \leq n^{1-\delta}\right\}, \\ A_3 &:= \left\{n^{1-\delta} < j \leq n\right\}.\end{aligned}$$

We use Lemma 3.A.4 to obtain the following control w.h.p.

$$\sum_{j\in A_1} \frac{1}{\lambda_1 - \lambda_j} \leq n^{1/3} n^{2/3} (\log n)^{C_5 \log\log n} = O(n^{1+\delta}).$$

Similarly, for A_2

$$\begin{aligned}\sum_{j\in A_2} \frac{1}{\lambda_1 - \lambda_j} &\leq \sum_{j\in A_2} \frac{1}{o(n^{-2/3}) + C_8(j/n)^{2/3} + O\left((\log n)^{C_5 \log\log n} n^{-2/3} j^{-1/3}\right)} \\ &= n^{2/3} \sum_{j\in A_2} \frac{1}{o(1) + C_8 j^{2/3}} \leq C_{10} n^{2/3} n^{(1-\delta)/3} \leq O(n^{1+\delta}).\end{aligned}$$

Finally, using Cauchy–Schwarz inequality

$$\sum_{j\in A_3} \frac{1}{\lambda_1 - \lambda_j} \leq \sqrt{n} \left(\sum_{j\in A_3} \frac{1}{(\lambda_1-\lambda_j)^2}\right)^{1/2} \leq \sqrt{n} O(n^{1/2+2\delta/3}) = O(n^{1+\delta}).$$

□

3.A.6. Proof of Lemma 3.2.1

Proof of Lemma 3.2.1. We show that w.h.p.

$$\sum_{i=2}^{n} \frac{\langle H v_i, v_1\rangle^2}{(\lambda_1-\lambda_i)^2} - \frac{1}{n}\sum_{i=2}^{n} \frac{1}{(\lambda_1-\lambda_i)^2} = o\left(\frac{1}{n}\sum_{i=2}^{n} \frac{1}{(\lambda_1-\lambda_i)^2}\right) \tag{3.40}$$

Let us recall that H is drawn according to the GOE, hence its law is invariant by rotation. This implies that the $\langle Hv_i, v_1\rangle$ are independent variables with variance $1/n$, independent of $\lambda_1, \dots, \lambda_n$. Define

$$M_n := \sum_{i=2}^{n} \frac{\langle Hv_i, v_1\rangle^2 - 1/n}{(\lambda_1 - \lambda_i)^2}.$$

Computing the second moment of M_n, we get

$$\mathbb{E}\left[M_n^2 | \lambda_1, \dots, \lambda_n\right] = \mathrm{Var}(M_n | \lambda_1, \dots, \lambda_n) = \frac{1}{n^4} \sum_{i=2}^{n} \frac{2}{(\lambda_1 - \lambda_i)^4}.$$

Adapting the proof of Lemma 3.A.3, following the same steps, one can also show that w.h.p.

$$\sum_{i=2}^{n} \frac{1}{(\lambda_1 - \lambda_i)^4} \asymp n^{8/3}. \tag{3.41}$$

Let $\epsilon = \epsilon(n) > 0$ to be specified later. By Markov's inequality

$$\begin{aligned}\mathbb{P}\left(|M_n| \geq \frac{\epsilon}{n} \sum_{i=2}^{n} \frac{1}{(\lambda_1 - \lambda_i)^2} | \lambda_1, \dots, \lambda_n\right) &\leq \frac{n^2}{\epsilon^2} \frac{\mathbb{E}\left[M_n^2 | \lambda_1, \dots, \lambda_n\right]}{\left(\sum_{i=2}^{n} \frac{1}{(\lambda_1 - \lambda_i)^2}\right)^2} \\ &\asymp \frac{1}{\epsilon^2 n^2},\end{aligned}$$

by Lemma 3.A.3 and equation (3.41). Taking e.g. $\epsilon(n) = n^{-1/2}$ concludes the proof. □

3.B. Additional proofs for Sections 3.3 & 3.4

3.B.1. Proof of Lemma 3.3.2

Proof of Lemma 3.3.2. We fix $\alpha > 0$ and we want to prove

$$\sum_{k=0}^{\lfloor (n-1)/2 \rfloor} \binom{n-1-k}{k} \alpha^k = \frac{1}{\sqrt{1+4\alpha}} \left[\left(\frac{1+\sqrt{1+4\alpha}}{2}\right)^n - \left(\frac{1-\sqrt{1+4\alpha}}{2}\right)^n\right]. \tag{3.42}$$

We denote in the following $\phi_+ := \frac{1+\sqrt{1+4\alpha}}{2}$ and $\phi_- := \frac{1-\sqrt{1+4\alpha}}{2}$, and for all $n \geq 1$:

$$u_n = u_n(\alpha) := \sum_{k=0}^{\lfloor (n-1)/2 \rfloor} \binom{n-1-k}{k} \alpha^k.$$

We clearly have $u_n(\alpha) \leq (1+\alpha)^n$. For all $t > 0$ small enough (e.g. $t < \frac{1}{1+\alpha}$), define

$$f(t) := \sum_{n=1}^{\infty} u_n t^n.$$

On one hand,

$$\frac{t}{1-t-\alpha t^2} = t \sum_{m=0}^{\infty} (t + \alpha t^2)^m = \sum_{m=0}^{\infty} \sum_{l=0}^{m} \binom{m}{l} \alpha^l t^{l+m+1}$$

$$= \sum_{n=1}^{\infty} \left(\sum_{\substack{0 \le l \le m \\ l+m=n-1}} \binom{m}{l} \alpha^l \right) t^n = \sum_{n=1}^{\infty} u_n t^n = f(t).$$

On the other hand,

$$\frac{t}{1-t-\alpha t^2} = \frac{t}{(1-\phi_- t)(1-\phi_+ t)} = \frac{1}{\phi_+ - \phi_-} \left(\frac{1}{1-\phi_+ t} - \frac{1}{1-\phi_- t} \right)$$
$$= \frac{1}{\sqrt{1+4\alpha}} \sum_{n=1}^{\infty} \left(\phi_+^n - \phi_-^n \right) t^n.$$

This proves (3.42). □

3.B.2. Proof of Lemma 3.4.2

Proof of Lemma 3.4.2. Let us represent the situation in the plane spanned by v_1 and v_1', as shown on Figure 3.4. Since $\widetilde{v}_1$ is taken such that $\langle v_1, \widetilde{v}_1 \rangle > 0$ and ξ_1 satisfies (3.6), we have

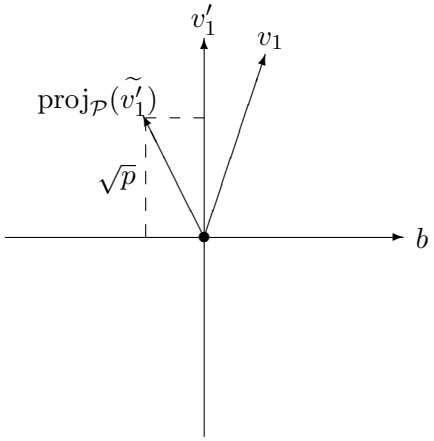


Figure 3.4 – *Orthogonal projection of* $\widetilde{v}_1$ *on* $\mathcal{P} := \mathrm{span}(v_1', v_1)$.

$\langle \widetilde{v}_1, v_1' \rangle > 0$ for n large enough by Proposition 3.1.1. Let $p := \langle \widetilde{v}_1, v_1' \rangle^2$ and $\widetilde{w} := \widetilde{v}_1 - \sqrt{p} v_1' \in (v_1')^{\perp}$. By invariance by rotation we can obtain that $\frac{\widetilde{w}}{\|\widetilde{w}\|} = \frac{\widetilde{w}}{\sqrt{1-p}}$ is uniformly distributed on the unit sphere $\mathbb{S}^{n-2}$ of $(v_1')^{\perp}$, and independent of p, v_1 and v_1'. Hence

$$\langle b, \widetilde{v}_1 \rangle = \langle b, \widetilde{w} \rangle \overset{(d)}{=} \sqrt{1-p} \cdot \frac{\widetilde{Z}_1}{\sqrt{\sum_{i=1}^{n-1} \left(\widetilde{Z}_i \right)^2}},$$

where the $\widetilde{Z}_i$ are independent Gaussian standard variables, independent from everything else. According to Section 3.2 we know that $1 - \langle v_1, v_1' \rangle \asymp \xi_1^2 n^{1/3}$ and thus $\langle v_1, b \rangle \asymp \xi_1 n^{1/6}$. This yields, for n large enough, w.h.p,

$$\begin{aligned} 0 < \langle \widetilde{v}_1, v_1 \rangle &\le \sqrt{p} \langle v_1, v_1' \rangle + \sqrt{\frac{1-p}{n}} \widetilde{Z}_1 \xi_1 n^{1/6} f(n) \\ &\le \sqrt{p} \langle v_1, v_1' \rangle + \sqrt{1-p} n^{-4/3} g(n) \\ &\le \max\left(\sqrt{p}, \sqrt{1-p} \right) \langle v_1, v_1' \rangle \\ &\le \langle v_1, v_1' \rangle, \end{aligned}$$

where f and g are two functions as defined in Lemma 3.A.3. From this point one can still make the link with the toy model, as done in the beginning of Section 3.3. By invariance by rotation, letting $t := \widetilde{v}_1 - \langle \widetilde{v}_1, v_1 \rangle v_1$, we know that $\|t\|$ and $\frac{t}{\|t\|}$ are independent, and that $\frac{t}{\|t\|}$ is uniformly distributed on the unit sphere in $v_1^\perp$. We have the following equality in distribution:

$$(r_1(v_1), r_1(\widetilde{v}_1)) \overset{(d)}{=} (r_1(X), r_1(X + \mathbf{s}Z)),$$

with w.h.p.

$$\mathbf{s} \geq \mathbf{s^1} = \frac{\|w\|\|X\|}{\left(\sum_{i=2}^n Z_i^2\right)^{1/2}\left(1 - \frac{\|w\|Z_1}{\left(\sum_{i=2}^n Z_i^2\right)^{1/2}}\right)} \asymp \xi_1 n^{1/6},$$

where the X_i, Z_i and w are defined in Section 3.3, for $\xi = \xi_1$. □

3.B.3. Proof of Lemma 3.4.1

Proof of Lemma 3.4.1. Recall that we work in the case (i) ($\xi = o(n^{-7/6-\epsilon})$ for some $\epsilon > 0$), with $\langle v_1, v_1' \rangle > 0$ and $\Pi^\star = I_n$. We want to show that w.h.p.

$$\langle A, \Pi_+ B \Pi_+^\top \rangle > \langle A, \Pi_- B \Pi_-^\top \rangle. \tag{3.43}$$

Define

$$\mathcal{G} := \{u, \Pi_+(u) = \Pi^\star(u) = i\}.$$

and

$$\mathcal{A} := \left\{\xi n^{1/6} f(n)^{-1} \leq \mathbf{s} \leq \xi n^{1/6} f(n)\right\},$$

with $f \in \mathcal{F}$ such that $\mathbb{P}(\mathcal{A}) \to 1$. For n large enough, on the event $\mathcal{A}$, we have $0 \leq \mathbf{s}n \leq n^{-\epsilon} f(n)$. Hence, retaking the proof of Proposition 3.1.2, we have

$$\begin{aligned}\phi_{x,z}(n, \mathbf{s}) &\geq \mathbb{P}\left(\mathcal{N}^+(x, x + \mathbf{s}z) = \mathcal{N}^-(x, x + \mathbf{s}z) = 0\right) \\ &\sim \exp\left(-\mathbf{s}nE(x)\left[z(2F(z) - 1) + 2E(z)\right]\right) = 1 - O(n^{-\epsilon} f(n)).\end{aligned}$$

Thus, with dominated convergence, for n large enough,

$$\mathbb{P}(\Pi_+(u) = \Pi^\star(u) | \mathcal{A}) = \iint dx dz E(x) E(z) \mathbb{E}\left[\phi_{x,z}(n, \mathbf{s}) | \mathcal{A}\right] \geq 1 - O(n^{-\epsilon} f(n)). \tag{3.44}$$

We use Markov's inequality with (3.44) to show that $\mathbb{P}\left(|\mathcal{G}| \leq n - n^{1-\epsilon/2} \mid \mathcal{A}\right) \leq O\left(n^{-\epsilon/2} f(n)\right)$, hence w.h.p.

$$|\mathcal{G}| \geq n - n^{1-\epsilon/2}. \tag{3.45}$$

Splitting the sum

$$\langle A, \Pi_+ B \Pi_+^\top \rangle = \sum_{u,v} A_{u,v} B_{\Pi_+(u), \Pi_+(v)} = \sum_{(u,v) \in \mathcal{G}^2} A_{u,v} B_{u,v} + \sum_{(i,j) \notin \mathcal{G}^2} A_{u,v} B_{\Pi_+(u), \Pi_+(v)},$$

one has, w.h.p.,

$$\begin{aligned}\langle A, \Pi_+ B \Pi_+^\top \rangle = &\sum_{(i,j) \in \mathcal{G}^2} A_{u,v}^2 + \sum_{\substack{(i,j) \notin \mathcal{G}^2 \\ (\Pi_+(u), \Pi_+(v)) \neq (v,u)}} A_{u,v} A_{\Pi_+(u), \Pi_+(v)} \\ &+ \sum_{\substack{(i,j) \notin \mathcal{G}^2 \\ (\Pi_+(u), \Pi_+(v)) = (v,u)}} A_{u,v}^2 + \xi \sum_{1 \leq i,j \leq n} A_{u,v} H_{\Pi_+(u), \Pi_+(v)}\end{aligned}$$

$$\geq C_1 \frac{|\mathcal{G}|^2}{n} - C_2 \left(n^2 - |\mathcal{G}|^2\right) \frac{\log n}{n} - C_2 \xi n^2 \frac{\log n}{n}.$$

We applied the law of large numbers for the first sum, lower-bounded the third sum by zero, and the classical inequality $\max_{u,v} \{A_{u,v}, H_{u,v}\} \leq C_2 \frac{\log n}{n}$ (which holds w.h.p.) for the two others.
Inequality (3.45) and condition (i) lead to, w.h.p.

$$\langle A, \Pi_+ B \Pi_+^\top \rangle \geq C_1 n - 2C_1 n^{1-\epsilon/2} - 2C_2 n^{1-\epsilon/2} \log n - C_2 n^{-1/6-\epsilon} \log n \geq C_3 n.$$

On the other hand, since by definition $\Pi_-(i) = \Pi_+(n+1-i)$, w.h.p.,

$$\begin{aligned}\langle A, \Pi_- B \Pi_-^\top \rangle = \sum_{(u,v)\in\mathcal{G}^2} A_{u,v} B_{n+1-u,n+1-v} + \sum_{(u,v)\notin\mathcal{G}^2} A_{u,v} B_{\Pi_-(u),\Pi_-(v)} \\ \leq O(\log n) + \frac{|\mathcal{G}|^2}{n} o(1) + C_2 \left(n^2 - |\mathcal{G}|^2\right) \frac{\log n}{n}.\end{aligned}$$

For the first sum, we used the law of large numbers: the variables $A_{u,v}$ and $B_{n+1-u,n+1-v}$ are independent in all cases but at most $n+1$, and this part of the sum is bounded by $O(\log n)$. We used the same control on Gaussian variables as above.
This gives

$$\left(\langle A, \Pi_- B \Pi_-^\top \rangle\right)_+ = o_{\mathbb{P}}(n),$$

where $(x)_+ := \max(0, x)$, which proves (3.43). □

3.B.4. Proof of Lemma 3.4.3

Proof of Lemma 3.4.3. Recall that we work in the case (ii) ($\xi = \omega(n^{-7/6+\epsilon})$ for some $\epsilon > 0$), with $\langle v_1, v_1' \rangle > 0$ and $\Pi^\star = I_n$. We want to show that the aligning permutation between v_1 and $-v_1'$ has a very bad overlap. Considering the pair $(X, -Y)$ where $(X, Y) \sim \mathsf{J}(n, s)$, one can adapt the proof of Proposition 3.1.2, with the new definitions

$$\begin{aligned}\widetilde{S}^+(x, y) &:= \mathbb{P}\left(X_1 > x, -Y_1 < -y\right), \text{ and} \\ \widetilde{S}^-(x, y) &:= \mathbb{P}\left(X_1 < x, -Y_1 > -y\right).\end{aligned}$$

The analysis is even easier since for all x, z, there exist two constants c, C such that

$$0 < c \leq \widetilde{S}^+(x, x + sz),\ \widetilde{S}^-(x, x + sz) \leq C < 1.$$

It is then easy to check that the proof of Proposition 3.1.2, case (ii) adapts well. □

Chapter 4

Alignment of sparse Erdős-Rényi graphs: information-theoretic results

In this chapter, we study fundamental limits of graph alignment: in the correlated Erdős-Rényi model, we prove an impossibility result for partial recovery in the sparse regime, with constant average degree and correlation, as well as a general bound on the maximal reachable overlap. This bound is tight in the noiseless case (the graph isomorphism problem) and we conjecture that it is still tight with noise. The proof of this negative result relies on a careful application of the probabilistic method to build automorphisms between tree components of a subcritical Erdős-Rényi graph.

This chapter is based on the paper *Impossibility of partial recovery in the graph alignment problem* [GML21b], published at *COLT 2021*, which is a joint work with M. Lelarge and L. Massoulié.

4.1. Introduction

As we have seen in Section 1.3.4 of the introduction, a vast majority of previous works focus on the exact (resp. quasi-exact) alignment, which is known to be feasible in the dense case, when $nqs \geq \log n$ (resp. $nqs \to \infty$). On the computational side, many algorithms are proposed for (quasi-)exact alignment; however, none of these succeed in the sparse setting with constant correlation and average degree $\lambda > 0$, i.e. with $q = \lambda/n$. It is thus natural and interesting to tackle the challenging question of partial alignment in the sparse setting.

4.1.1. A colored view on the correlated Erdős-Rényi model

Let us recall the definition of the *correlated Erdős-Rényi model* (already introduced in (1.10)) in the sparse case: in this chapter, we represent the graphs $(G, G') \sim \mathsf{G}(n, \lambda/n, s)$ with respectively blue and red edges, and with the same set of nodes $[n]$. For each edge, the colors are samples independently:

- with probability $\lambda s/n$ to get two-colored edges;
- with probability $\lambda(1-s)/n$ to get a blue (monochromatic) edge;
- with probability $\lambda(1-s)/n$ to get a red (monochromatic) edge;
- with probability $1 - \lambda(2-s)/n$ to get a non-edge,

where $\lambda > 0$ and $s \in [0, 1]$ are fixed parameters and n is large. In this model, G and G' are both sparse $\mathsf{G}(n, \lambda/n)$ graphs. For large values of n, the fraction of edges of one graph that are shared with the other is of order s (see Figure 4.1).

We then relabel the vertices of the red graph G' with a uniform independent permutation $\pi^\star \in \mathcal{S}_n$, and we observe G and $H := G'^{\pi^\star}$, see Figure 4.2. Upon observing G and H, the goal

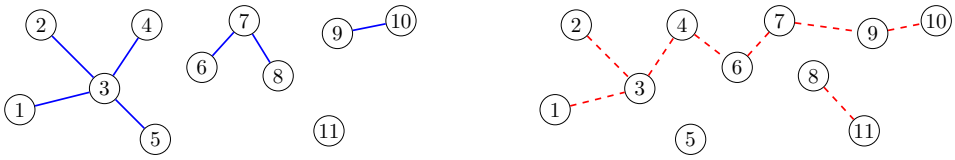


Figure 4.1 – *A realization of* (G, G') *from the correlated Erdős-Rényi model, with* $n = 11$, $\lambda = 1.9$, *and* $s = 0.7$. *For the sake of readability, red edges are always dashed.*

is to recover (or, reconstruct) partially the latent vertex correspondence $\pi^\star$ with probability converging to 1 as $n \to \infty$.

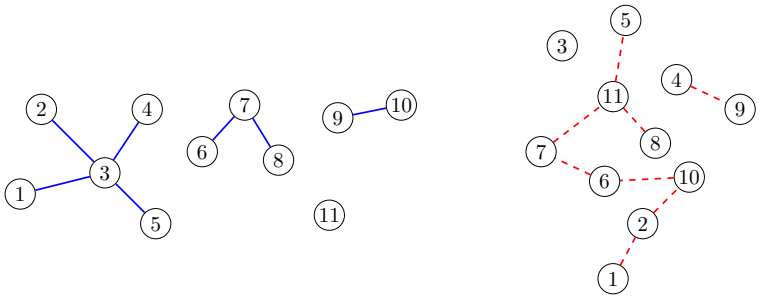


Figure 4.2 – *The pair* (G, H) *corresponding to* (G, G') *of Figure 4.1, after relabeling* G' *with the permutation* $\pi^\star = (6)(1\ 5\ 3\ 11\ 9\ 2\ 8\ 4\ 7\ 10)$.

4.1.2. Partial alignment in the sparse regime

First note that since we are in the sparse regime, then even without noise, i.e. with $s = 1$, there is no way to be able to map the $\Theta(n)$ isolated vertices[1] in G and H better than chance. Hence, we rather focus on the partial alignment problem where we ask for the best possible fraction of matched vertices between G and H. More formally, an *estimator* $\hat{\pi}$ (of $\pi^\star$) is a $\mathcal{S}_n$-valued measurable function of (G, H).

Note that in this problem, the graphs could very well be unlabeled in the first place. We could assign the labels uniformly, the only interesting information being the correspondence between vertices. Hence, any estimator $\hat{\pi}$ must satisfy the *equivariance* property, in the sense that for all $\sigma \in \mathcal{S}_n$,

$$\hat{\pi}(G^\sigma, H) = \hat{\pi}(G, H) \circ \sigma^{-1} . \tag{4.1}$$

Remark 4.1.1. *Note that unsurprisingly, the maximum a posteriori estimator* $\hat{\pi}_{\mathrm{MAP}}$, *which is the permutation solving the maximization problem* (1.5), *satisfies* (4.1).

Another – though more cumbersome – approach to enforce some notion of equivariance (and put aside some trivial estimators such as $\hat{\pi} = \mathrm{id}$*) would be to redefine the overlap as follows:*

$$\mathrm{ov}(\hat{\pi}(G, H), \pi^\star) := \frac{1}{n \cdot n!} \sum_{\sigma \in \mathcal{S}_n} \sum_{u=1}^{n} \mathbb{1}_{\hat{\pi}(G^\sigma, H)(u) = \pi^\star \circ \sigma^{-1}(u)} .$$

With this definition, it is ensured that for any $\sigma \in \mathcal{S}_n$,

$$\mathrm{ov}(\hat{\pi}(G^\sigma, H), \pi^\star) = \mathrm{ov}(\hat{\pi}(G, H), \pi^\star \circ \sigma) .$$

We recall that partial alignment consists in finding a estimator $\hat{\pi}$ of $\pi^\star$ satisfying $\mathrm{ov}(\hat{\pi}, \pi^\star) > \alpha n$ with high probability, for some $\alpha > 0$. Let us start by stating a conjecture[2]:

[1] We refer to Theorem 1.1 of the introduction for a proof of this result.

[2] At the time this manuscript is being completed, this conjecture and a more general form are proved in

Conjecture.

(*i*) *If* $\lambda s \leq 1$*, partial reconstruction is impossible, i.e. for any* $\alpha > 0$*, for all estimator* $\hat{\pi}$*,*

$$\mathbb{P}\left(\mathrm{ov}(\hat{\pi}, \pi^\star) > \alpha n\right) \underset{n\to\infty}{\longrightarrow} 0.$$

(*ii*) *If* $\lambda s > 1$*, partial reconstruction is possible (feasible), i.e. there exists* $\alpha > 0$ *and an estimator* $\hat{\pi}$ *such that*

$$\mathbb{P}\left(\mathrm{ov}(\hat{\pi}, \pi^\star) > \alpha n\right) \underset{n\to\infty}{\longrightarrow} 1.$$

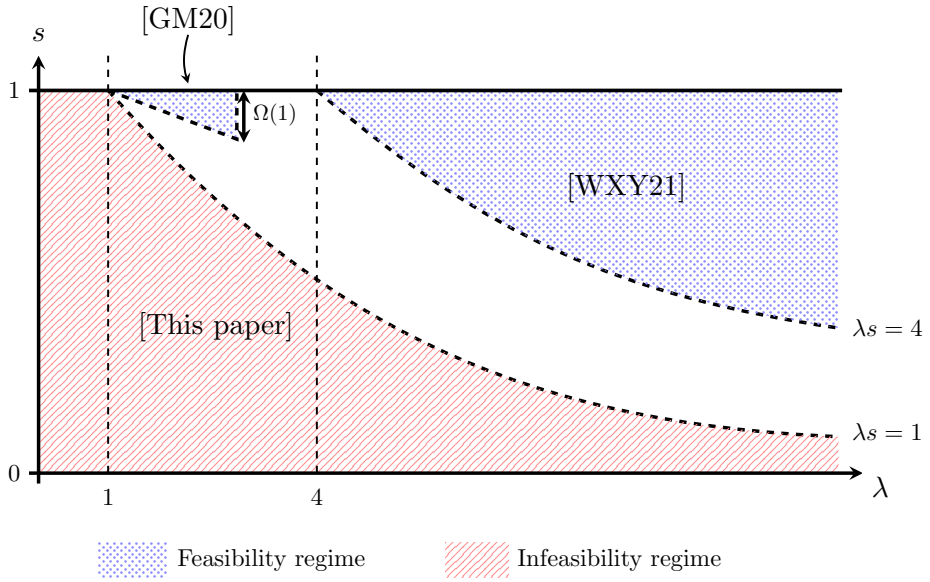


Figure 4.3 – *Diagram of the* (λ, s) *regions where partial reconstruction is known*[3] *to be impossible (resp. possible), in the sparse regime where* λ, s *are fixed constants.*

Main result The main result of the chapter is as follows:

Theorem 4.1. *For* $\lambda > 0$ *and* $s \in [0, 1]$*, we have for any* $\alpha > 0$*, for any estimator* $\hat{\pi}$*:*

$$\mathbb{P}\left(\mathrm{ov}(\hat{\pi}, \pi^\star) > (c(\lambda s) + \alpha)n\right) \underset{n\to\infty}{\longrightarrow} 0, \tag{4.2}$$

where $c(\mu)$ *is the greatest non-negative solution to the equation* $e^{-\mu x} = 1 - x$*.*

Note that a well-known result (see e.g. [Bol01]) is that $c(\mu)$ is the typical fraction of nodes in the largest component of an Erdős-Rényi graph with average degree μ, and that $c(\mu) = 0$ if $\mu \leq 1$, and $c(\mu) \in (0, 1)$ whenever $\mu > 1$. Hence, Theorem 4.1 implies that partial reconstruction is impossible for $\lambda s \leq 1$. Moreover, if $\lambda s > 1$, any estimator can reach an overlap of at most $c(\lambda s)n + o(n)$. Note that $c(\lambda s)$ is the typical fraction of nodes in the largest component of the intersection graph.

Related work in the sparse regime In this chapter, we work in the regime where $\lambda > 0$ and $s \in [0, 1]$ are fixed constants. Our results prove part (i) of the conjecture, which had not been previously studied, and give an upper bound on the maximal reachable overlap in case (ii).

[DD22]. The result of Theorem 4.1 however does not lose of its appeal, since it gives an upper bound on the best reachable overlap.

[3] at the time of this contribution.

Most relevant related results[4] for our conjecture [GM20] which proves[5] that partial recovery is possible in polynomial time in a region $\mathcal{R} := \{(\lambda, s);\ \lambda \in [1, \lambda_0)$ and $s \in (s^*(\lambda), 1]\}$ for some function $s^*(\lambda) < 1$, so that interestingly the case $\lambda > \lambda_0$ is left open, nevertheless much in step with (ii). Previous results from [HM20] showed that partial reconstruction was feasible for $\lambda s > C$, with an unspecified constant $C > 20$. The work [WXY21] significantly improves these results, narrowing down the gap for (ii): when translated with our notations, it is shown that partial alignment is possible (theoretically) if $\lambda s \geq 4 + \varepsilon$. In addition, an impossibility condition of the form $nqs \leq 1 - \varepsilon$ is also established, but in a denser case, where $nq/s = \omega(\log^2 n)$. Note that this last impossibility result does not cover our regime, where both the mean degree nq and the correlation parameter s are of order 1.

These results are summed up in a diagram in Figure 4.3. In particular, our bound is tight and our conjecture is almost solved for the case $s = 1$, with a remaining gap $[\lambda_0, 4]$ being still open.

For the impossibility part, [WXY21] works with the mutual information $I(\pi^\star; G, G')$, closely related to the minimum mean squared error. They are able to derive an upper bound on the expectation of $\mathrm{ov}(\hat{\pi}, \pi^\star)$, for any estimator, which happens to be $o(1)$ when the mean degree in the parent graph of G and G' is at least of order $\log^2 n$, but not when λ, s are of order 1. In our result, we do not work directly with the mutual information, but we are considering the posterior distribution of $\pi^\star$: in simple words, we show that under the assumption $\lambda s \leq 1$ the posterior distribution puts equal weights on permutations that overlap only on a vanishing fraction of points. This is done by building ad hoc permutations with the probabilistic method.

In this work, we derive information-theoretic results: our proof is not related to a particular algorithm. The search for efficient algorithms in this field is a very active field of research: we refer once again to Section 1.3.4. Unfortunately, all proposed algorithms are not known to give a positive fraction of overlap in the regime $\lambda s \geq 1$, hence leaving the question of the tightness of our bound open. New light will be shed on this question in Chapter 6.

4.2. Main results and global intuition

4.2.1. Some definitions

Let us first recall that for two permutations $\sigma, \sigma' \in \mathcal{S}_n$ we denote by $\mathrm{ov}(\sigma, \sigma')$ the number of points on which $\sigma = \sigma'$, namely

$$\mathrm{ov}(\sigma, \sigma') := \sum_{u=1}^{n} \mathbb{1}_{\sigma(u)=\sigma'(u)}\,.$$

Through all the chapter, we will implicitly consider that every graph G of size n has the canonical vertex set $[n]$. We will denote by $E(G)$ its edge set and $e(G)$ its number of edges.

For any pair of graphs (G, G'), both labeled on $[n]$, we denote by $G \vee G'$ (resp. by $G \wedge G'$) the union graph (resp. intersection graph) of G and G', that is the graph with same node set and edge set $E(G) \cup E(G')$ (resp. $E(G) \cap E(G')$). The symmetric difference of G and G', denoted by $G \triangle G'$, is the subgraph made of edges of $G \vee G'$ that are not in $G \wedge G'$.

In the case where edges are colored, say edges of G (resp. G') are blue (resp. red), these definitions extend to ensure colour preservation: note e.g. that in this case $G \wedge G'$ is simply the subgraph of $G \vee G'$ consisting of two-colored edges (see Figure 4.4).

When the pair (G, G') is drawn under the correlated Erdős-Rényi model, for all $u, v \in [n]$, we write $u \longleftrightarrow v$ (resp. $u \longleftrightarrow v$) if u and v are connected in G, that is the edge is either

[4] at the time of this contribution.

[5] see Chapter 5.

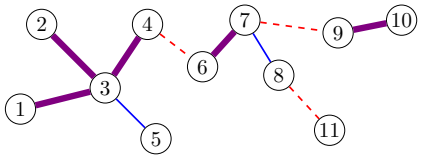


Figure 4.4 – *The graph $G \vee G'$ with (G, G') of Figure 4.1. For the sake of readability, the two-colored edges of $G \wedge G'$ are always drawn thick and purple.*

blue or two-colored (resp. in G', either red or two-colored).

For G a graph with vertex set $[n]$ and $\sigma \in \mathcal{S}_n$, we denote by G^σ the *relabeling of G with σ*, which is the graph with same vertex set $[n]$ and edges $\{\sigma(u), \sigma(v)\}$ for all $\{u, v\} \in E(G)$.

Finally we recall the definition of $c(\mu)$: for all $\mu > 0$, $c(\mu)$ is the greatest non-negative solution to the equation $e^{-\mu x} = 1 - x$. We also recall the fact that for $\mu \leq 1$, $c(\mu) = 0$.

4.2.2. General intuition on the main result

Let us describe the general intuition for our result : recall that we are given (G, H) drawn under the correlated Erdős-Rényi model with planted relabeling $\pi^\star$. The idea of the argument for impossibility is to show that, there are w.h.p. lots of permutations that have the same weight for the posterior distribution of $\pi^\star$ given G, H, and that are far apart. In other words, an informal statement is as follows :

(Informal Statement). *We want to show that there exists lots of relabelings G^{σ_i} of G such that:*

(*i*) *There is no way of deciding (statistically) whether the two graphs we observe are (G, G') or some (G^{σ_i}, G').*

(*ii*) *These relabelings are far apart from each other on small components of $G \wedge G'$.*

Let us give a formal version of the previous intuition. First note that for any labeled graphs g, g' on $[n]$:

$$\mathbb{P}(G = g, G' = g') = \left(\frac{\lambda s}{n}\right)^{e(g \wedge g')} \left(\frac{\lambda(1-s)}{n}\right)^{e(g \triangle g')} \left(1 - \frac{\lambda(2-s)}{n}\right)^{\binom{n}{2} - e(g \vee g')}.$$

Since

$$e(g \vee g') = e(g) + e(g') - e(g \wedge g') \quad \text{and} \quad e(g \triangle g') = e(g \vee g') - e(g \wedge g'),$$

$\mathbb{P}(G = g, G' = g')$ is uniquely determined by $e(g), e(g')$ and $e(g \wedge g')$. In particular, the dependence of the joint distribution in $e(g \wedge g')$ is given by:

$$\mathbb{P}(G = g, G' = g') \propto \left(\frac{s(n - \lambda(2-s))}{\lambda(1-s)^2}\right)^{e(g \wedge g')}. \tag{4.3}$$

In view of (4.3), preserving the posterior distribution by relabeling a graph G is simply preserving the number of edges of their intersection graph. We now have a formal rephrasing for our conditions (*i*) and (*ii*) above: we encapsulate them in a theorem, which will constitute the bulk of this chapter.

Theorem 4.2. *Fix an integer $p > 0$. Consider (G, G') drawn under the correlated Erdős-Rényi model $\mathsf{G}(n, \lambda/n, s)$. Then, with high probability, there exists $\{\sigma_i\}_{i \in [p]}$ – that depend on the intersection graph $G \wedge G'$ – such that*

(*i*) $\forall i \in [p],\ e\left(G^{\sigma_i} \wedge G'\right) = e\left(G \wedge G'\right),$

(*ii*) $\forall i, j \in [p],\ i \neq j \implies \mathrm{ov}(\sigma_i, \sigma_j) \leq c(\lambda s)n + o(n)$, *where the* $o(n)$ *is independent of* $i, j \in [p]$.

Let us now explain how Theorem 4.2 implies our impossibility result via a simple pigeonhole principle.

Proof of Theorem 4.1. Let us take $\alpha > 0$. We want to control the probability that the overlap between an estimator $\hat{\pi}$ and $\pi^\star$ is greater than $\alpha n + c(\lambda s)n$. Fix $\varepsilon > 0$, and take p large enough so that

$$\alpha \varepsilon p > 2.$$

First note that point (i) together with (4.3) gives that the joint probability of $(G, G', \pi^\star)$ is is equal to that of $(G^{\sigma_i}, G', \pi^\star)$, for all $i \in [p]$. Thus, for all estimator $\hat{\pi}$ depending on $G, H = G'^{\pi^\star}$, one has

$$\forall i \in [p],\ \mathrm{ov}\left(\hat{\pi}(G^{\sigma_i}, H), \pi^\star\right) \overset{(d)}{=} \mathrm{ov}\left(\hat{\pi}(G, H), \pi^\star\right), \tag{4.4}$$

and by (4.1), we also have

$$\forall i \in [p],\ \mathrm{ov}\left(\hat{\pi}(G^{\sigma_i}, H), \pi^\star\right) = \mathrm{ov}\left(\hat{\pi}(G, H), \pi^\star \circ \sigma_i\right). \tag{4.5}$$

Let

$$X := \sum_{i \in [p]} \mathbb{1}_{\mathrm{ov}(\hat{\pi}, \pi^\star \circ \sigma_i) > (c(\lambda s) + \alpha)n}$$

Note that because of point (ii), all $\mathrm{ov}(\pi^\star \circ \sigma_i, \pi^\star \circ \sigma_j)$ are at most $c(\lambda s)n + o(n)$ for $i \neq j \in [p]$. Thus, there are at least $X \times (\alpha - o(1))n$ distinct points among the node set $[n]$. This gives that one necessarily has

$$X \leq \frac{1}{\alpha - o(1)}. \tag{4.6}$$

Then, taking the expectation and considering the event on which the set $\{\sigma_i\}_{i \in [p]}$ of Theorem 4.2 exists – which happens with probability $1 - o(1)$ – gives

$$\begin{aligned} \mathbb{E}[X] &\geq \sum_{i=1}^{p} \mathbb{P}\left(\mathrm{ov}(\hat{\pi}, \pi^\star \circ \sigma_i) > (c(\lambda s) + \alpha)n\right) - p \times o(1) \\ &= p \times \mathbb{P}\left(\mathrm{ov}(\hat{\pi}, \pi^\star) > (c(\lambda s) + \alpha)n\right) - o(1). \end{aligned}$$

Hence,

$$\mathbb{P}\left(\mathrm{ov}(\hat{\pi}, \pi^\star) > (c(\lambda s) + \alpha)n\right) \leq \frac{1}{p(\alpha - o(1))} + o(1). \tag{4.7}$$

For n large enough, the right-hand side of the last term is less that $\frac{1}{p(\alpha/2)}$, which is less than ε. This proves as desired that for all $\alpha > 0$

$$\mathbb{P}\left(\mathrm{ov}(\hat{\pi}, \pi^\star) > (c(\lambda s) + \alpha)n\right) \underset{n \to \infty}{\longrightarrow} 0. \tag{4.8}$$

□

We are now left to understand how to build ad hoc permutations verifying points (i) and (ii) of Theorem 4.2. In order to build these permutations, we are going to relabel the vertices on small tree components of the intersection graph $G \wedge G'$. As a first step, we hereafter check that they indeed nearly cover the whole graph, when letting aside the giant component.

4.2.3. Vertices on small tree components

We briefly recall the definition of the simple Erdős-Rényi model $\mathsf{G}(n,p)$: it consist in drawing a (single) graph with node set $[n]$ in which every edge is independently present with probability p. Let us begin with a classical result:

Lemma 4.2.1 ([Bol01], Corollary 5.8, Theorem 6.11)**.** *Let $G \sim \mathsf{G}(n,\mu/n)$ with $\mu > 0$, and $a_n \to \infty$. Then, with high probability, G has a giant component of order $c(\mu)n + o(n)$ and outside the giant component, at least $(1-c(\mu))n - a_n$ vertices are on tree components.*

We need here a slight adaptation of this result, showing that $(1-c(\mu))n - o(n)$ vertices are in fact on *small* tree components.

Lemma 4.2.2. *Let $G \sim \mathsf{G}(n,\mu/n)$ with $\mu > 0$, and $K(n) \to \infty$. Then with high probability, $(1-c(\mu))n - o(n)$ vertices are on tree components of size at most $K(n)$.*

Proof. Assume without loss of generality that $K(n) = o(\log n)$. Let $T_>$ be the number of vertices that are on tree components of size $\geq K(n)$. Taking $a_n = o(n)$ in Lemma 4.2.1, it remains to show that w.h.p., $T_> = o(n)$. This is done easily by bounding very roughly the first moment. Another classical result (see e.g. [JLR00], Theorem 5.4) is that with probability $1-o(1)$, all tree components are of size $O(\log n)$, which gives

$$\begin{aligned}\frac{\mathbb{E}\left[T_>\right]}{n} &\leq o(1) + \sum_{k=K(n)}^{O(\log n)} \frac{1}{n} \cdot k \cdot \binom{n}{k} k^{k-2} \left(\frac{\mu}{n}\right)^{k-1} \left(1-\frac{\mu}{n}\right)^{k(n-k)+\binom{k}{2}-k+1} \\ &\leq o(1) + (1+o(1)) \sum_{k=K(n)}^{O(\log n)} \frac{e^k}{k} \mu^{k-1} e^{-k\mu},\end{aligned}$$

using $\binom{n}{k} \leq \left(\frac{en}{k}\right)^k$ together with Cayley's formula[6] and the fact that for all indices $K(n) \leq k \leq O(\log n)$ in the sum, $k^2 \leq o(n)$ (uniformly). Now, the series in the right hand term has general terms which is $O\left(e^{-k(\mu - \log\mu + 1)}\right)$, and since $\mu - \log\mu + 1 > 0$ the series converges, which implies that $\mathbb{E}\left[T_>/\, n\right] = o(1)$. The proof is concluded by Markov's inequality. □

Since in our model $G \wedge G'$ is an Erdős-Rényi graph of parameters $(n, \lambda s/n)$, the previous results ensures that all but a vanishing part of the $(1-c(\mu))n$ vertices outside the giant component are on small (i.e. $\leq K(n)$) tree components of the intersection graph. For the rest of the chapter, we will take

$$K(n) = \lfloor \sqrt{\log n} \rfloor.$$

This first step suggests to build the permutations (relabelings) only by looking at $G \wedge G'$. Hence, we will first consider the random generation of the intersection graph, then create some permutations σ_i, and finally reveal the monochromatic edges.

The generating process is as follows: since almost all $(1-c(\mu))n$ vertices are on small trees in $G \wedge G'$, we can prove that each small tree up to isomorphism will have a number of occurrences in the intersection graph of order n (this is claimed more precisely in Lemma 4.3.1). Permuting iteratively these isomorphic trees, we may derange them quite a lot, and each time differently.

In order to prove Theorem 4.2, we use the *probabilistic method*[7]: we give in the next section a simple detailed stochastic method to build p permutation candidates, and we will next prove that these permutations satisfy conditions (i) and (ii) with positive probability, hence proving the desired existence.

[6] Cayley's formula states that the number of trees on k labeled vertices is k^{k-2}.

[7] The main interest of this widely used method (see [AS16]) is to be non-constructive. Indeed, as detailed in the next Sections, explicitly giving the p permutations considered in Theorem 4.2 is very cumbersome, because of the extra double edges that may appear (see Section 4.3.3).

4.3. Building automorphisms of $G \wedge G'$ tree-wise

Through all this section, we work conditionally on the intersection graph $G \wedge G'$ (that is the two-colored edges).

4.3.1. Mathematical formalization

Recall that we fix $K := K(n) = \lfloor \sqrt{\log n} \rfloor$. For all $k \in [K]$, we will denote by $\mathbb{T}_k$ the set of *unlabeled* trees of size k. $\mathbb{T}_k$ can also be viewed as the set of equivalence classes of labeled trees of size k for the isomorphism relation. Note that $\mathbb{T}_k$ is finite and that we can roughly upper bound its size by the number of *labeled* trees of size k which equals k^{k-2}, by Cayley's formula[8].

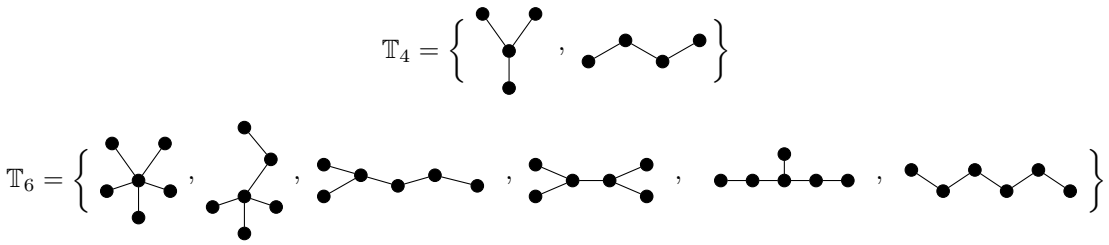


Figure 4.5 – *Explicit composition of* $\mathbb{T}_4$ *(of size* 2*) and* $\mathbb{T}_6$ *(of size* 6*).*

For a given tree $\mathbf{T} \in \mathbb{T}_k$, we will denote by $X_{\mathbf{T}}$ the number of distinct connected components of $G \wedge G'$ that are isomorphic to $\mathbf{T}$, $H_{\mathbf{T}} := \{T_1, T_2, \dots, T_{X_{\mathbf{T}}}\}$ the set of the corresponding labeled subgraphs of $G \wedge G'$, and $V(H_{\mathbf{T}})$ the set of vertices of $[n]$ that belong to one of the trees in $H_{\mathbf{T}}$.

Our global finite recursion will be done on the finite set

$$\mathbb{T} := \bigcup_{k=1}^{K} \mathbb{T}_k = \{\mathbf{T}_1, \mathbf{T}_2, \dots, \mathbf{T}_M\}, \tag{4.9}$$

which we assume to have been ordered increasingly according to tree sizes, for convenience. The global permutation σ is built block-wise by composing permutations $\sigma_{\mathbf{T}}$ for $\mathbf{T} \in \mathbb{T}$ such that each $\sigma_{\mathbf{T}}$ only acts on vertices of $H_{\mathbf{T}}$.

More precisely, for a fixed $\mathbf{T} \in \mathbb{T}$, $\sigma_{\mathbf{T}}$ will consists in permuting the vertices tree by tree, so $\sigma_{\mathbf{T}}$ will be determined by a tree permutation $\Sigma_{\mathbf{T}}$ of size $X_{\mathbf{T}}$. Assume that for all trees $T_1, \dots, T_{X_{\mathbf{T}}}$ isomorphic to $\mathbf{T}$ in $G \wedge G'$, we fix some isomorphisms $\psi_1, \dots, \psi_{X_{\mathbf{T}}}$ such that $T_i \underset{\psi_i}{\widehat{=}} \mathbf{T}$ for all $i \in [X_{\mathbf{T}}]$. More generally we will denote $\mathfrak{i}(u)$ the index of the tree that $u \in V(H_{\mathbf{T}})$ belongs to (when there is no ambiguity on $\mathbf{T}$), and $u \simeq u'$ when two vertices of $G \wedge G'$ are sent onto the same point of $\mathbf{T}$ by these isomorphisms. Then, the natural definition of the node permutation $\sigma_{\mathbf{T}}$ according to $\Sigma_{\mathbf{T}}$ and these isomorphisms is given by

$$\sigma_{\mathbf{T}} : u \mapsto \begin{cases} \psi^{-1}_{\Sigma_{\mathbf{T}}(\mathfrak{i}(u))} \circ \psi_{\mathfrak{i}(u)}(u) \ \ (\in T_{\Sigma_{\mathbf{T}}(\mathfrak{i}(u))}) & \text{if } u \in V(H_{\mathbf{T}}), \\ u & \text{if } u \notin V(H_{\mathbf{T}}). \end{cases} \tag{4.10}$$

Note that by definition, $V(H_{\mathbf{T}})$ is stable by $\sigma_{\mathbf{T}}$, and $\sigma_{\mathbf{T}}$ fixes all nodes in $[n] \setminus V(H_{\mathbf{T}})$. Recall that M denotes the total size of $\mathbb{T}$ as defined in (4.9). The recursive construction is as follows

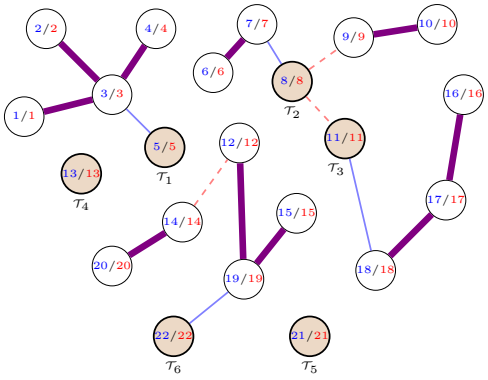
[8] This upper bound is far from being optimal, but is enough for our use.

:

Algorithm 4.1: Recursive construction of σ

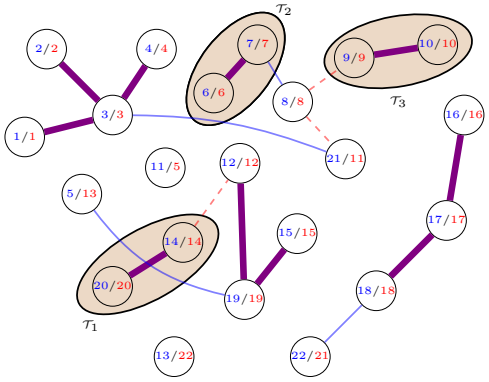
```
Initialize σ_0 ← id;
for i = 1 to M do
    Consider T = T_i and draw uniformly at random the tree permutation Σ_T ∈ S_{X_T},
      independently from the past;
    Consider σ_T the node permutation associated with Σ_T by (4.10);
    σ_i ← σ_T ∘ σ_{i−1};
end
return σ = σ_M
```

Note that at the end of the procedure, σ fixes all points that are either on the giant component of the intersection graph, or on a component that is not a tree a size $\leq K(n)$. Figure 4.6 gives an example of this random recursive construction (for convenience, $\lambda s < 1$, the true labels are in red, whereas blue labels enables to keep track of the relabeling recursively built on the blue graph).



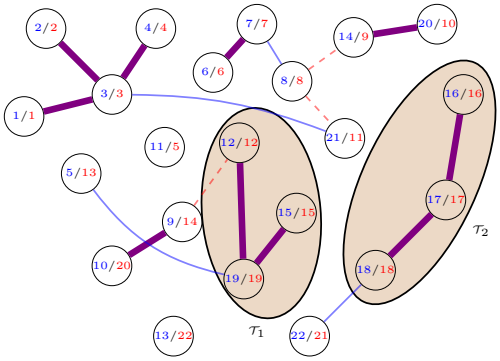
$\Sigma_{\mathbf{T}_1} = (2)(1\ 3\ 5\ 6\ 4)$

$\sigma_{\mathbf{T}_1} = (8)(5\ 11\ 21\ 22\ 13)$



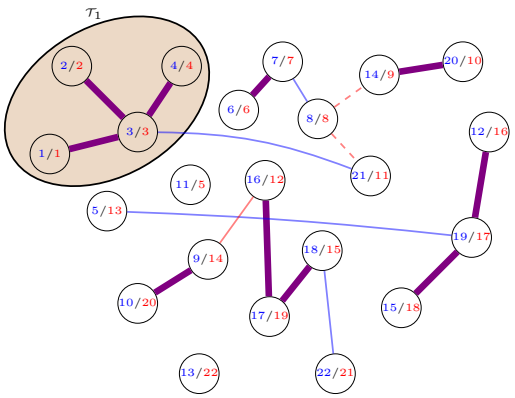
$\Sigma_{\mathbf{T}_2} = (2)(1\ 3)$

$\sigma_{\mathbf{T}_2} = (6)(7)(9\ 14)(10\ 20)$



$\Sigma_{\mathbf{T}_3} = (1\ 2)$

$\sigma_{\mathbf{T}_3} = (15\ 18)(19\ 17)(12\ 16)$



$\Sigma_{\mathbf{T}_4} = (1)$

$\sigma_{\mathbf{T}_4} = (1)(2)(3)(4)$

Figure 4.6 – *Example of recursive (tree-wise) generation of a permutation with Algorithm 4.1.*

Through the analysis we will need the following control on $X_{\mathbf{T}}$ for $\mathbf{T} \in \mathbb{T}$:

Lemma 4.3.1. *Recall that $K(n) = \lfloor \sqrt{\log n} \rfloor$. For all $k \in [K(n)]$, define $f(k) := \frac{(\lambda s)^{k-1} e^{-\lambda s k}}{k!}$. Then, with high probability (on the intersection graph),*

$$\forall k \in [K(n)],\ \forall \mathbf{T} \in \mathbb{T}_k,\ X_{\mathbf{T}} \geq n(1 - o(1)) f(k). \tag{4.11}$$

The proof of this result is deferred to Appendix 4.B.1.

Remark 4.3.1. *Note that since $\lambda s e^{-\lambda s} < 1$, $k \mapsto f(k)$ is decreasing with k. Moreover, for $K(n) \leq \sqrt{\log n}$, we have that for any $t > 0$,*

$$f(K(n)) \geq \exp\left(-C\sqrt{\log n} \log\log n\right) \gg n^{-t}.$$

4.3.2. Ensuring that the permutations are 'far apart'

We check in this section that Algorithm 4.1 generates permutations that will verify condition (ii) of Theorem 4.2, w.h.p. Let $\sigma_1, \dots, \sigma_p$ be generated independently with Algorithm 4.1. We then have the following results:

Lemma 4.3.2. *With high probability, for all $i \neq j \in [p]$,*

$$\mathrm{ov}(\sigma_i, \sigma_j) = c(\lambda s)n + o(n).$$

This lemma is proved in Appendix 4.B.2. In the sequel we will denote by V_∞ the set of vertices that are on the giant component of $G \wedge G'$ (if there is one), and by $V_>$ the vertices of $[n] \setminus V_\infty$ that are *not* on tree components of size $\leq K(n)$. Finally we set $V_{\infty,>} := V_\infty \cup V_>$. Define

$$\mathcal{S}_{in} := \binom{[n] \setminus V_{\infty,>}}{2}, \quad \mathcal{S}_{out} := \binom{[n]}{2} \setminus \left(\binom{V_{\infty,>}}{2} \cap \binom{[n] \setminus V_{\infty,>}}{2} \right), \quad \mathcal{S} := \mathcal{S}_{in} \cup \mathcal{S}_{out}. \tag{4.12}$$

$\mathcal{S}_{in}$ is the set of edges that have both endpoints outside $V_{\infty,>}$, whereas edges of $\mathcal{S}_{out}$ have exactly one endpoint in $V_{\infty,>}$. We say that an edge $(u, v) \in \binom{[n]}{2}$ is a *common fixed edge* of permutations $\sigma_1, \dots, \sigma_r$ if

$$\{\sigma_1(u), \sigma_1(v)\} = \dots = \{\sigma_r(u), \sigma_r(v)\}.$$

For all subset of edges $\mathcal{W} \subseteq \binom{[n]}{2}$, we define

$$F(\mathcal{W}, \sigma_1, \dots, \sigma_r) := \sum_{e \in \mathcal{W}} \mathbb{1}_{e \text{ is a common fixed edge of } \sigma_1, \dots, \sigma_r}. \tag{4.13}$$

We now state a result – which proof is deferred to 4.B.3 – that will be useful in next section.

Lemma 4.3.3. *With high probability, we have, for any $t > 0$,*

- *for any $i_1 \neq i_2 \in [p]$,*
$$F(\mathcal{S}, \sigma_{i_1}, \sigma_{i_2}) \leq n^{1+t}, \tag{4.14}$$
- *for any $i_1, i_2, i_3 \in [p]$ pairwise distinct,*
$$F(\mathcal{S}, \sigma_{i_1}, \sigma_{i_2}, \sigma_{i_3}) \leq n^t, \tag{4.15}$$
- *for any $r \geq 4$, $i_1, \dots, i_r \in [p]$ pairwise distinct,*
$$F(\mathcal{S}, \sigma_{i_1}, \dots, \sigma_{i_r}) = 0. \tag{4.16}$$

4.3.3. Emergence of extra double edges

In the example of Figure 4.6, we can see that the number of two-colored edges in the relabeled union graph $G^{\sigma_i} \vee G'$ is constant through time. This property is fundamental for point (i) of Theorem 4.2. However, depending on the random $\sigma_{\mathbf{T}_i}$ drawn through the process – we recall that they are drawn independently from the monochromatic edges, that are not revealed yet – we may see extra two-colored edges appear (extra double edges hereafter). Figure 4.7 shows a case in which there is an emergence of an extra double edge in the process.

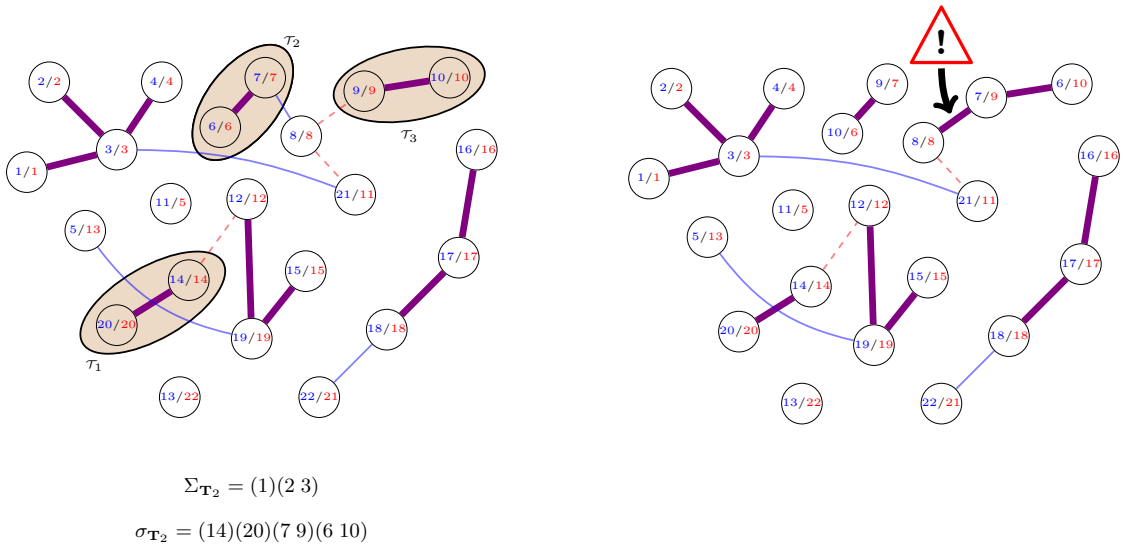


Figure 4.7 – *Example of the emergence of an extra double edge in Algorithm 4.1.*

Note that the number of two-coloured edges can only be greater or equal to $e(G \wedge G')$ through this process, since by definition we are preserving edges of the intersection graph.

The last part of our work is to prove that there is a positive probability that applying independently Algorithm 4.1 p times gives p permutations that do not present extra double edges, before using the probabilistic method. This step will require a Poisson approximation, described hereafter.

4.4. Poisson approximation, proof of Theorem 4.2

In this section we introduce n' to be the number of vertices that the permutations actually act on:

$$n' := |[n] \setminus V_{\infty,>}| \sim (1 - c(\lambda s))n \text{ w.h.p.} \tag{4.17}$$

4.4.1. Poisson approximation for extra double edges

In the sequel, we will assume that we fix a set $\{\sigma_i\}_{i\in[p]}$ of p permutations of $[n']$, verifying :

$$\text{for all } t > 0, \text{for all } m \neq m' \in [p], F(\mathcal{S}, \sigma_m, \sigma_{m'}) \leq n^{1+t}. \tag{H1}$$

$$\text{for all } t > 0, \text{for all } m_1, m_2, m_3 \in [p] \text{ pairwise distinct }, F(\mathcal{S}, \sigma_{m_1}, \sigma_{m_2}, \sigma_{m_3}) \leq n^t. \tag{H2}$$

$$\text{There are no common fixed edge of any } r\text{-tuple in } \{\sigma_i\}_{i\in[p]}. \tag{H3}$$

We will work under the event $\mathcal{E}_{\mathcal{S}}$ on which $n' \sim (1 - c(\lambda s))n$ and $|\mathcal{S}| \sim \binom{n'}{2} \sim n'^2/2 = (1 - c(\lambda s))^2 n^2/2$. It is easy (see e.g. [Bol01]) to show that $\mathcal{E}_{\mathcal{S}}$ is satisfied w.h.p. As explained before, some extra double edges (e.d.e. hereafter) may appear when revealing the non double edges of $\mathcal{S}$ (that is, blue and red edges that are not between vertices of $V_{\infty,>}$). Note that for

every edge we have

$$\begin{aligned}\mathbb{P}\left(u \longleftrightarrow v \,|\, (u,v) \notin E(G \wedge G')\right) &= \mathbb{P}\left(u \longleftrightarrow v \,|\, (u,v) \notin E(G \wedge G')\right) \\ &= \frac{\mathbb{P}\left(u \longleftrightarrow v, (u,v) \notin E(G \wedge G')\right)}{\mathbb{P}\left((u,v) \notin E(G \wedge G')\right)} \\ &= \frac{\lambda(1-s)/n}{1-\lambda s/n} \sim \frac{\lambda(1-s)}{n}.\end{aligned}$$

For any permutation σ, define the number of created e.d.e. by the relabeling of G by σ as follows:

$$\Delta(\sigma) := \sum_{\{u,v\} \in \mathcal{S}} \mathbb{1}_{u \longleftrightarrow v} \mathbb{1}_{\sigma(u) \longleftrightarrow \sigma(v)}. \tag{4.18}$$

We now present the key result for our analysis, with the notation $n^{\underline{k}}$ for the *falling factorial*

$$n^{\underline{k}} := n(n-1)\cdots(n-k+1).$$

Theorem 4.3 (Asymptotic Poisson behavior of $\{\Delta(\sigma_i)\}_{i\in[p]}$)**.** *Assume that* $\{\sigma_i\}_{i\in[p]}$ *verify* (H1), (H2) *and* (H3). *Then, for all* $\ell_1, \ell_2, \ldots, \ell_p \geq 0$,

$$\mathbb{E}\left[\Delta(\sigma_1)^{\underline{\ell_1}} \Delta(\sigma_2)^{\underline{\ell_2}} \cdots \Delta(\sigma_p)^{\underline{\ell_p}} \,\middle|\, G \wedge G', \mathcal{E}_{\mathcal{S}}\right] \underset{n\to\infty}{\longrightarrow} \left(\frac{\lambda^2(1-s)^2(1-c(\lambda s))^2}{2}\right)^{\ell_1+\ell_2+\ldots+\ell_p}. \tag{4.19}$$

In other words, conditionally to graph $G \wedge G'$ *and event* $\mathcal{E}_{\mathcal{S}}$, *the random variables* $\{\Delta(\sigma_i)\}_{i\in[p]}$ *are asymptotically distributed as independent Poisson variables of parameter* $\frac{\lambda^2(1-s)^2(1-c(\lambda s))^2}{2}$.

The proof of Theorem 4.3, based on a fine control of terms of unusually high contribution, is deferred to Appendix 4.A.

4.4.2. Proof of Theorem 4.2

Proof. The proof is quite straightforward now. Fixing $p > 0$, Lemma 4.3.3 gives that (H1), (H2) and (H3) are verified w.h.p. by some $\sigma_1, \ldots, \sigma_p$ generated independently with Algorithm 4.1. Then, the probability (on the remaining monochrome edges) that the p permutations given satisfy conditions (i) and (ii) of Theorem 4.2 is equivalent to

$$\begin{aligned}(1-o(1)) \times \mathbb{P}\left(\mathsf{Poi}\left(\frac{\lambda^2(1-s)^2(1-c(\lambda s))^2}{2}\right) = 0\right)^p \\ = (1-o(1)) \exp\left(-p\frac{\lambda^2(1-s)^2(1-c(\lambda s))^2}{2}\right) > 0,\end{aligned}$$

which gives the existence with high probability of a set a permutations of size p satisfying conditions (i) and (ii) of Theorem 4.2. □

Appendix of Chapter 4

4.A. Proof of Theorem 4.3

Proof of Theorem 4.3. Let $\ell_1, \ell_2, \dots, \ell_p$ be non negative integers. Recall that conditioned to $G \wedge G'$, each edge of $\mathcal{S}$ is independently blue (resp. red) with probability

$$q = q(\lambda, s, n) := \frac{\lambda(1-s)}{n - \lambda s}.$$

Now, let us explain why convergence (4.19) holds. First recall that for a given $\ell \geq 0$, $\mathbb{E}\left[\Delta(\sigma)^{\underline{\ell}}\right]$ is nothing else but the expected number of (ordered) $p-$tuples of edges $\{u, v\} \in \mathcal{S}$ such that $\mathbb{1}_{u \longleftrightarrow v} \mathbb{1}_{\sigma(u) \longleftrightarrow \sigma(v)} = 1$. Using the notation $\sum^*$ for summation of ordered tuples of edges in $\mathcal{S}$ as well as linearity of expectation, we get:

$$\mathbb{E}\left[\Delta(\sigma_1)^{\underline{\ell_1}} \Delta(\sigma_2)^{\underline{\ell_2}} \cdots \Delta(\sigma_p)^{\underline{\ell_p}}\right] = \sum^*_{\substack{\{u_1^{(1)}, v_1^{(1)}\}, \\ \{u_2^{(1)}, v_2^{(1)}\}, \\ \dots, \\ \{u_{\ell_1}^{(1)}, v_{\ell_1}^{(1)}\}}} \sum^*_{\substack{\{u_1^{(2)}, v_1^{(2)}\}, \\ \{u_2^{(2)}, v_2^{(2)}\}, \\ \dots, \\ \{u_{\ell_2}^{(2)}, v_{\ell_2}^{(2)}\}}} \cdots \sum^*_{\substack{\{u_1^{(p)}, v_1^{(p)}\}, \\ \{u_2^{(p)}, v_2^{(p)}\}, \\ \dots, \\ \{u_{\ell_p}^{(p)}, v_{\ell_p}^{(p)}\}}} \mathbb{E}\left[\prod_{m=1}^{p} \prod_{j=1}^{\ell_m} \mathbb{1}_{u_j^{(m)} \longleftrightarrow v_j^{(m)}} \mathbb{1}_{\sigma_m(u_j^{(m)}) \longleftrightarrow \sigma_m(v_j^{(m)})}\right] \tag{4.20}$$

First observe that the total number of terms N in the previous sum is

$$N := |\mathcal{S}|^{\underline{\ell_1}} \times |\mathcal{S}|^{\underline{\ell_2}} \times \cdots |\mathcal{S}|^{\underline{\ell_p}} \sim \left(\frac{(1 - c(\lambda s))^2 n^2}{2}\right)^{\ell_1 + \dots + \ell_p},$$

since $|\mathcal{S}| \sim \frac{(1-c(\lambda s))^2 n^2}{2}$ on event $\mathcal{E}_{\mathcal{S}}$.

Lower bound. Observe that the N terms in the sum of eq. (4.20) are made in general of $2(\ell_1 + \dots + \ell_p)$ indicator variables, not necessarily distinct. For most of the terms however, all involved edges are distinct, thus independent, and their contribution to the sum is $q^{2(\ell_1 + \dots + \ell_p)}$.

Whenever a pair of blue (resp. red) indicators are equal, at least one term may be canceled, so the contribution to the expectation is higher than $q^{2(\ell_1 + \dots + \ell_p)}$.

Whenever a pair of edges that appear in a blue/red pair of indicators are equal, the product of the indicators is necessarily 0 (indeed, an edge in $\mathcal{S}$ cannot be two-colored). These terms, where at least one equality of the form $\{u_j^{(m)}, v_j^{(m)}\} = \{\sigma_{m'}(u_{j'}^{(m')}), \sigma_{m'}(v_{j'}^{(m')})\}$ occurs, cover the case where the contribution is strictly less that $q^{2(\ell_1 + \dots + \ell_p)}$ (it is 0). There are at most

$$\binom{\ell_1 + \dots + \ell_p}{2} \left(\frac{n^2}{2}\right)^{\ell_1 + \dots + \ell_p - 1}$$

such terms. Thus

$$\mathbb{E}\left[\Delta(\sigma_1)^{\underline{\ell_1}}\Delta(\sigma_2)^{\underline{\ell_2}}\cdots\Delta(\sigma_p)^{\underline{\ell_p}}\right] \geq \left(N - \binom{\ell_1+\ldots+\ell_p}{2}\left(\frac{n^2}{2}\right)^{\ell_1+\ldots+\ell_p-1}\right) \times q^{2(\ell_1+\ldots+\ell_p)}$$
$$\sim \left(\frac{(1-c(\lambda s))^2 n^2}{2}\right)^{\ell_1+\ldots\ell_p} \times \left(\frac{\lambda(1-s)}{n}\right)^{2(\ell_1+\ldots+\ell_p)}$$
$$\xrightarrow[n\to\infty]{} \left(\frac{\lambda^2(1-s)^2(1-c(\lambda s))^2}{2}\right)^{\ell_1+\ell_2+\ldots+\ell_p}.$$

Upper bound. The terms that we now want to study are the terms for which the contribution is greater than $q^{2(\ell_1+\ldots+\ell_p)}$. Looking closely at the general product in (4.20), an unusual high contribution is the consequence of three possible type of constraints:

(*i*) constraints of the form $\{u_j^{(m)}, v_j^{(m)}\} = \{u_{j'}^{(m')}, v_{j'}^{(m')}\}$: note that since the sums are made of ordered tuples, this equality may happen only for pairs such that $m \neq m'$. Moreover, transitivity of equality implies that a constraint implying some $\{u_j^{(m)}, v_j^{(m)}\}$ may happen at most once for each $m' \in [p], m' \neq m$ (otherwise we would have a relationship of the form $\{u_{j'}^{(m')}, v_{j'}^{(m')}\} = \{u_{k'}^{(m')}, v_{k'}^{(m')}\}$, which is impossible).

(*ii*) constraints of the form $\{\sigma_m(u_j^{(m)}), \sigma_m(v_j^{(m)})\} = \{\sigma_{m'}(u_{j'}^{(m')}), \sigma_{m'}(v_{j'}^{(m')})\}$. For the same reasons as in case (*i*), a constraint implying some $\{\sigma_m(u_j^{(m)}), \sigma_m(v_j^{(m)})\}$ may happen at most once for each $m' \in [p], m' \neq m$.

(*iii*) the last case is made of intersection of cases (*i*) and (*ii*), i.e. edges satisfying both constraints $\{u_j^{(m)}, v_j^{(m)}\} = \{u_{j'}^{(m')}, v_{j'}^{(m')}\}$ and $\{\sigma_m(u_j^{(m)}), \sigma_m(v_j^{(m)})\} = \{\sigma_{m'}(u_{j'}^{(m')}), \sigma_{m'}(v_{j'}^{(m')})\}$. This implies in particular that $\{u_j^{(m)}, v_j^{(m)}\}$ is an common fixed edge for σ_m and $\sigma_{m'}$. By assumption (H3), note that there cannot be a connected path of constraints of the form (*iii*) of length greater or equal to 3.

Let us now represent these constraints with a dependency graph. Each vertex a the graph represent one edge $\{u_j^{(m)}, v_j^{(m)}\}$ of the sum, that we will align column-wise according to $m \in [p]$. We put a plain (resp. dashed) edge between two nodes if they are enforced by constraint (*i*) but not (*iii*) (resp. (*ii*) but not (*iii*)). Finally we draw a thick plain edge between two nodes if they are enforced by constraint (*iii*).

In view of discussion in points $(i) - (ii) - (iii)$, this dependency graph must be p-partite. Moreover, the subgraph made of plain thick or plain edges (resp. plain thick of dashed edges) only consists in a union of disjoint paths. The thick plain subgraph is only made of isolated edges and paths fo size 3. Finally, transitivity of the equality relationship enables to draw any path in any order: we shall take the left to right order by convention (no backtracking).

We denote by k_1 (resp. k_2) the number of plain (resp. dashed) edges. We also denote track k_3 the number of thick plain isolated edges, and k_4 the number of thick plain isolated paths of length 2. Figure 4.8 gives an example of such a dependency graph.

In order to upper bound the contribution due to large terms, we must understand both the expectation of the product of indicators in (4.20) (this only depends on (k_1, k_2, k_3, k_4)), as well as the number of possible (labeled) dependency graphs with a given (k_1, k_2, k_3, k_4).

First, all plain (resp. dashed) dependency edge makes 1 (resp. 1) indicators disappear in the expectation (for any event $\mathcal{A}, \mathbb{1}_{\mathcal{A}}^2 = \mathbb{1}_{\mathcal{A}}$). In the same way, all thick plain isolated edge (resp. thick plain isolated path of length 2) makes 2 (resp. 4) indicators disappear the expectation for a given case with given (k_1, k_2, k_3, k_4) is

$$q^{2(\ell_1+\ldots+\ell_p)-(k_1+k_2+2k_3+4k_4)} \leq C_1 n^{-2(\ell_1+\ldots+\ell_p)+(k_1+k_2+2k_3+4k_4)} \tag{4.21}$$

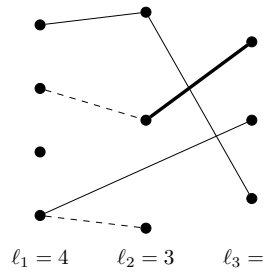


Figure 4.8 – *Example of a dependency graph, with* $(k_1, k_2, k_3, k_4) = (3, 2, 1, 0)$.

where C_1 is a constant depending on $\ell_1, \ldots, \ell_p$,

Second, an upper bound for the number of possible (labeled) dependency graphs with a given (k_1, k_2, k_3, k_4) can be established as follows. First, we have $k_1 + k_2 + k_3 + 2k_4$ equalities, leaving at most $\ell_1 + \ldots + \ell_p - (k_1 + k_2 + k_3 + 2k_4)$ degrees of freedom in the choices of the edges. Moreover, we force k_3 of these edges to be common fixed edges between two (distinct) permutations, and k_4 of them to be common fixed edges between three (pairwise distinct) permutations. In view of hypotheses (H1) and (H2), the number of possible (labeled) dependency graphs with a given (k_1, k_2, k_3, k_4) is at most

$$\begin{aligned}\binom{k_1 + k_2 + k_3 + k_4}{k_3 + k_4} |\mathcal{S}|^{\ell_1+\ldots+\ell_p-(k_1+k_2+k_3+2k_4)-k_3-k_4} \times (n^{1+t})^{k_3} \times n^{tk_4} \\ \leq C_2 n^{2(\ell_1+\ldots+\ell_p)-2(k_1+k_2)-(3-t)k_3-(6-t)k_4}, \qquad (4.22)\end{aligned}$$

where C_2 is a constant depending on $\ell_1, \ldots, \ell_p$.

Hence, in view of (4.21) and (4.22), the total contribution of higher terms is upper bounded by

$$\begin{aligned}&\sum_{s=1}^{\ell_1+\ldots+\ell_p} \sum_{k_1+k_2+k_3+2k_4=s} C_1 C_2 n^{-2(\ell_1+\ldots+\ell_p)+(k_1+k_2+2k_3+4k_4)} n^{2(\ell_1+\ldots+\ell_p)-2(k_1+k_2)-(3-t)k_3-(6-t)k_4} \\ &\qquad \leq C_1 C_2 \sum_{s=1}^{\ell_1+\ldots+\ell_p} \sum_{k_1+k_2+k_3+2k_4=s} n^{-k_1} n^{-k_2} n^{-(1-t)k_3} n^{-(2-t)k_4} \\ &\qquad\qquad \leq C_1 C_2 \times (\ell_1 + \ldots + \ell_p) \times (\ell_1 + \ldots + \ell_p)^{4(\ell_1+\ldots+\ell_p)} \times n^{-(1-t)} \underset{n\to\infty}{\longrightarrow} 0.\end{aligned}$$

This last convergence concludes the proof. □

4.B. Proofs of Lemmas

4.B.1. Proof of Lemma 4.3.1

Proof. For the control of $X_{\mathbf{T}}$ we follow classical computations made in [Bol01] to establish asymptotic behavior of $X_{\mathbf{T}}$. For our purpose, we only need the two first moments. Assume that $\mathbf{T}$ is of size $k = k(\mathbf{T}) \leq K$, and that its automorphism group has $a = a(\mathbf{T})$ elements. Then, letting $\mu = \lambda s$,

$$\mathbb{E}[X_{\mathbf{T}}] = \binom{n}{k} \times \frac{k!}{a} \times \left(\frac{\mu}{n}\right)^{k-1} \left(1 - \frac{\mu}{n}\right)^{k(n-k)+\binom{k}{2}-k+1}.$$

Indeed, we have $\binom{n}{k}$ choices for the nodes, then $\frac{k!}{a}$ ways of putting the edges. Using $\binom{n}{k} \sim \frac{n^k}{k!}$ and $\left(1-\frac{\mu}{n}\right)^{-k^2+\binom{k}{2}-k+1} \sim 1$ as soon as $k = o(\sqrt{n})$, we get

$$\mathbb{E}\left[X_{\mathbf{T}}\right] \sim n\mu^{k-1}e^{-\mu k}/a.$$

We now compute $\mathbb{E}\left[X_{\mathbf{T}}(X_{\mathbf{T}}-1)\right]$ by classically counting the number of ordered pairs of distinct isolated tree components of $G \wedge G'$ isomorphic to $\mathbf{T}$. This number is then multiplied by the probability of observing these two distinct isolated components. This gives

$$\mathbb{E}\left[X_{\mathbf{T}}(X_{\mathbf{T}}-1)\right] = \binom{n}{k}\binom{n-k}{k} \times \left(\frac{k!}{a}\right)^2 \times \left(\frac{\mu}{n}\right)^{2(k-1)} \left(1-\frac{\mu}{n}\right)^{2\left(k(n-2k)+\binom{k}{2}-k+1\right)} \left(1-\frac{\mu}{n}\right)^{k^2}.$$

Here again, $k = o(\sqrt{n})$ gives that

$$\mathbb{E}\left[X_{\mathbf{T}}(X_{\mathbf{T}}-1)\right] \sim n^2\mu^{2(k-1)}e^{-2\mu k}/a^2.$$

Denoting $\alpha = \alpha(\mathbf{T}) := n\mu^{k-1}e^{-\mu k}/a(\mathbf{T})$, these computations give that $\mathbb{E}\left[X_{\mathbf{T}}\right] \sim \mathrm{Var}\left(X_{\mathbf{T}}\right) \sim \alpha(\mathbf{T})$ when $n \to \infty$, uniformly in $k \leq K(n)$ as soon as $K(n) = o(\sqrt{n})$. Let us fix $\varepsilon = \varepsilon(n) > 0$ small enough. Applying Chebyshev's inequality together with the union bound gives

$$\begin{aligned}
\mathbb{P}\left(\exists (k,\mathbf{T}) \in [K(n)] \times \mathbb{T}, X_{\mathbf{T}} \leq (1-\varepsilon)\alpha(\mathbf{T})\right) &\leq \sum_{k=1}^{K(n)} \sum_{\mathbf{T}\in\mathbb{T}_k} \mathbb{P}\left(X_{\mathbf{T}} - \mathbb{E}\left[X_{\mathbf{T}}\right] \leq (1-\varepsilon)\alpha(\mathbf{T}) - \mathbb{E}\left[X_{\mathbf{T}}\right]\right) \\
&\overset{(a)}{\leq} \sum_{k=1}^{K(n)} \sum_{\mathbf{T}\in\mathbb{T}_k} \frac{\mathrm{Var}\left(X_{\mathbf{T}}\right)}{\left((1-\varepsilon)\alpha(\mathbf{T}) - \mathbb{E}\left[X_{\mathbf{T}}\right]\right)^2} \\
&\overset{(b)}{\leq} (1+o(1)) \sum_{k=1}^{K(n)} \sum_{\mathbf{T}\in\mathbb{T}_k} \frac{1}{\varepsilon^2\alpha(\mathbf{T})} \\
&\overset{(c)}{\leq} (1+o(1)) \sum_{k=1}^{K(n)} \sum_{\mathbf{T}\in\mathbb{T}_k} \frac{1}{\varepsilon^2 n f(k)} \\
&\overset{(d)}{\leq} (1+o(1))K(n)^{K(n)} \frac{1}{\varepsilon^2 n f(K(n))},
\end{aligned}$$

where

$$f(k) := \frac{\mu^{k-1}e^{-\mu k}}{k!}. \tag{4.23}$$

We used in (a) that all $(1-\varepsilon)\alpha(\mathbf{T}) - \mathbb{E}\left[X_{\mathbf{T}}\right]$ are negative for n large enough, in (b) uniformity in $k \leq K(n)$, in (c) the lower bound $nf(k)$ for $\alpha(T)$, and finally in (d) that $k \mapsto f(k)$ is decreasing since $\mu e^{-\mu} < 1$.

Taking now e.g. $\varepsilon = n^{-1/4}$, the last fact to check to establish the Lemma is that $K^K/f(K) = o(n^{1/2})$ when $K = K(n) = \log^{1/2}(n)$:

$$\begin{aligned}
K^K/f(K) &= K^K K!(1/\mu)^{K-1}e^{\mu K} \\
&\leq \exp\left(2K\log K + (\log(1/\mu)+\mu)K\right) \\
&= \exp\left(\log^{1/2}(n)\log\log n + (\log(1/\mu)+\mu)\log^{1/2}(n)\right) = o(n^{1/2}).
\end{aligned}$$

□

4.B.2. Proof of Lemma 4.3.2

Proof. Denote $T_\infty := |V_\infty|$ and $T_> := |V_>|$. First notice that for any permutations σ_i, σ_j with $i \neq j$ generated with Algorithm 4.1, we have the following equality:

$$\mathrm{ov}(\sigma_i, \sigma_j) = T_\infty + T_> + \sum_{k=1}^{K(n)} \sum_{\mathbf{T} \in \mathbb{T}_k} k \cdot \mathrm{ov}(\Sigma_{\mathbf{T}}^{(i)}, \Sigma_{\mathbf{T}}^{(j)}), \tag{4.24}$$

where $\Sigma_{\mathbf{T}}^{(i)}$ (resp. $\Sigma_{\mathbf{T}}^{(j)}$) is the tree permutation associated with $\mathbf{T}$ in σ_i (resp. in σ_j). We know that $T_\infty = c(\lambda s)n + o(n)$ w.h.p. and by Lemma 4.2.2, $T_> = o(n)$ w.h.p.

Define

$$\mathrm{ov}'(\sigma_i, \sigma_j) := \sum_{k=1}^{K(n)} \sum_{\mathbf{T} \in \mathbb{T}_k} k \cdot \mathrm{ov}(\Sigma_{\mathbf{T}}^{(i)}, \Sigma_{\mathbf{T}}^{(j)}), \tag{4.25}$$

the second term in (4.24). We dominate $\mathrm{ov}'(\sigma_i, \sigma_j)$ as follows:

Lemma 4.B.1. *If $X = \mathrm{ov}(\Sigma_{\mathbf{T}}^{(i)}, \Sigma_{\mathbf{T}}^{(j)})$,, then for all $t \in \mathbb{R}$,*

$$\mathbb{E}\left[e^{tX}\right] \leq \exp(e^t). \tag{4.26}$$

Proof.

$$\mathbb{E}\left[e^{tX}\right] = \sum_{m \geq 0} e^{tm} \mathbb{P}(X \geq m).$$

Noting that $\mathbb{P}(X \geq m) \leq \mathbb{E}\left[\binom{X}{m}\right]$ and that

$$\begin{aligned}\mathbb{E}\left[\binom{X}{m}\right] &= \frac{1}{m!}\mathbb{E}\left[X(X-1)\dots(X-m+1)\right] \\ &= \frac{1}{m!}k(k-1)\dots(k-m+1)\frac{(k-m)!}{k!} = \frac{1}{m!}\end{aligned}$$

gives

$$\mathbb{E}\left[e^{tX}\right] \leq \sum_{m \geq 0} \frac{e^{tm}}{m!} \leq \exp(e^t).$$

□

Using independence of the X variables, Equation (4.26) of Lemma 4.B.1 give that for all $t \in \mathbb{R}$,

$$\mathbb{E}\left[e^{t \cdot \mathrm{ov}'(\sigma_i, \sigma_j)}\right] \leq \prod_{k=1}^{K(n)} \prod_{\mathbf{T} \in \mathbb{T}_k} \exp(e^{tk}) \leq \exp\left(e^{tK(n)} K(n)^{K(n)+1}\right). \tag{4.27}$$

Now, we use the classical Chernoff bound, for positive t,

$$\begin{aligned}\mathbb{P}\left(\mathrm{ov}'(\sigma_i, \sigma_j) \geq n^\alpha\right) &\leq \exp\left(-tn^\alpha + e^{tK(n)} K(n)^{K(n)+1}\right) \\ &\leq \exp\left(-\frac{n^\alpha}{K(n)}\left[\log\left(\frac{n^{1-\alpha}}{K(n)^{K(n)+2}}\right) - 1\right]\right),\end{aligned}$$

taking $t = \frac{1}{K(n)} \log\left(\frac{n^\alpha}{K(n)^{K(n)+2}}\right)$. The right hand side tend to 0 for any $\alpha \in (0,1)$, and a

simple use of the union bound ends the proof. □

4.B.3. Proof of Lemma 4.3.3

Proof. Fix $t > 0$. We use a standard first moment method. We will use the results of Lemmas 4.2.2 and 4.3.1, conditioning on the event $\mathcal{A}$ where the corresponding results hold. Since $\mathbb{P}(\mathcal{A}) = 1 - o(1)$, this conditioning is legitimate for our purpose.

Step 1. Let us first control the term $F(\mathcal{S}_{out}, \sigma_{i_1}, \dots, \sigma_{i_r})$: edges of $\mathcal{S}_{out}$ are made of exactly one vertex in $V_{\infty,>}$. There are at most n^2 such edges, and the probability for a given edge of $\mathcal{S}_{out}$ being a common fixed edge of $\sigma_{i_1}, \dots, \sigma_{i_r}$ is $\frac{1}{X_{\mathbf{T}}^{r-1}}$, which can be upper-bounded on $\mathcal{A}$ by $(nf(K(n)))^{1-r} \leq n^{1-r+t/2}$ by Remark 4.3.1.

Edges of $\mathcal{S}_{out}$ thus have a contribution in $\mathbb{E}\left[F(\sigma_{i_1}, \dots, \sigma_{i_r})|\mathcal{A}\right]$ of at most $n^{3-r+t/2}$.

Step 2. In the edges appearing in $F(\sigma_{i_1}, \dots, \sigma_{i_r})$, we consider three cases:

(i) edges of ***Intra***: these are edges made with two vertices in the same tree $T \widehat{=} \mathbf{T} \in \mathbb{T}$. On event $\mathcal{A}$, there are at most

$$\sum_{k=1}^{K(n)} \sum_{\mathbf{T} \in \mathbb{T}_k} X_{\mathbf{T}} k^2 \leq nK(n)$$

such edges. The probability for a given edge of ***Intra*** made of vertices of $\mathbf{T} \in \mathbb{T}$ being a common fixed edge of $\sigma_{i_1}, \dots, \sigma_{i_r}$ is $\frac{1}{X_{\mathbf{T}}^{r-1}}$, which can be upper-bounded by $(nf(K(n)))^{1-r} \leq n^{1-r+t/2}$. Edges of ***Intra*** thus have a contribution in $\mathbb{E}\left[F(\sigma_{i_1}, \dots, \sigma_{i_r})|\mathcal{A}\right]$ of at most $n^{2-r+t/2}$.

(ii) edges of $\boldsymbol{Inter}_1$: these are edges made with two vertices u, v in different trees $T \neq T'$ (but that may be $\sim$ to the same $\mathbf{T} \in \mathbb{T}$), and verifying $u \not\simeq v$. There are at most n^2 such edges. Since $u \not\simeq v$, there are only one possibility to map two edges of $\boldsymbol{Inter}_1$. The probability for a given edge of $\boldsymbol{Inter}_1$ made of vertices of $T \widehat{=} \mathbf{T}, T' \widehat{=} \mathbf{T}'$ being a common fixed edge is $\frac{1}{(X_{\mathbf{T}}(X_{\mathbf{T}}-1))^{r-1}}$, and edges of $\boldsymbol{Inter}_1$ thus have a contribution in the expectation of at most $n^{4-2r+t/2}$.

(iii) edges of $\boldsymbol{Inter}_2$: these are edges similar to case (ii), except that their endpoints belong necessarily to isomorphic trees, and verifying $u \simeq v$. There are at most n^2 such edges. Since $u \simeq v$, there are two ways to map two edges of $\boldsymbol{Inter}_2$. The probability for a given edge of $\boldsymbol{Inter}_2$ made of vertices of $T, T' \widehat{=} \mathbf{T}$ being a common fixed edge is time $\left(\frac{2}{X_{\mathbf{T}}(X_{\mathbf{T}}-1)}\right)^{r-1}$, and edges of $\boldsymbol{Inter}_2$ thus have a contribution in the expectation of at most $n^{4-2r+t/2}$.

Step 3. The first two steps show that $\mathbb{E}\left[F(\sigma_{i_1}, \dots, \sigma_{i_r})|\mathcal{A}\right] \leq Cn^{3-r+t/2}$ for all $t > 0$. Summing over all possible r-tuples of permutations, Markov inequality yields

$$\begin{aligned}\mathbb{P}\left(\exists r \geq 4,\, \exists \sigma_{i_1}, \dots, \sigma_{i_r} \text{ pairwise distinct},\, F(\mathcal{S}, \sigma_{i_1}, \dots, \sigma_{i_r}) \geq 1\right) &\leq o(1) + \sum_{r=4}^{\infty} p^r C n^{3-r+t/2} \\ &\leq Cp^4 n^{t/2-1} \to 0,\end{aligned}$$

for t small enough, and

$$\mathbb{P}\left(\exists \sigma_{i_1}, \sigma_{i_2}, \sigma_{i_3} \text{ pairwise distinct},\, F(\mathcal{S}, \sigma_{i_1}, \sigma_{i_2}, \sigma_{i_3}) \geq n^t\right) \leq o(1) + p^3 \times Cn^{-t/2} \to 0,$$

and

$$\mathbb{P}\left(\exists \sigma_{i_1} \neq \sigma_{i_2}, F(\mathcal{S}, \sigma_{i_1}, \sigma_{i_2}) \geq n^{1+t}\right) \leq o(1) + p^2 \times Cn^{-t/2} \to 0.$$

□

Chapter 5

From tree matching to sparse graph alignment

In this chapter, we consider alignment of sparse graphs, for which we introduce the *Neighborhood Tree Matching Algorithm* (`NTMA`), based on a measure of similarity between trees. For correlated Erdős-Rényi random graphs, we prove that the algorithm returns – in polynomial time – a positive fraction of correctly matched vertices, and a vanishing fraction of mismatches. This result holds with average degree of the graphs in $O(1)$ and correlation parameter s that can be bounded away from 1, conditions under which random graph alignment is particularly challenging. As a byproduct of the analysis we introduce a matching metric between trees and characterize it for several models of correlated random trees. These results may be of independent interest, yielding for instance efficient tests for determining whether two random trees are correlated or independent[1].

This chapter is based on the paper *From tree matching to sparse graph alignment* [GM20], published at *COLT 2020*, a joint work with L. Massoulié.

5.1. Introduction

As seen in the introduction (Section 1.3.4), previously existing methods for Erdős-Rényi graph alignment only succeed in a dense regime where the mean degree of the graphs is $\Omega(\log n)$. When the mean degree is constant, several phenomena occur – degrees do not concentrate any more and the graph looses its connectivity (see Theorem 1.1), among other things – and make the performance of standard dense methods collapse.

We recall in particular that results from [CK17, CKMP18] show that in the sparse regime, there is no hope of recovering $\pi^\star$ exactly or almost exactly, or in other words, of perfectly re-aligning G and H. Nevertheless, their result does not rule out the possibility of partially recovering the unknown permutation $\pi^\star$. For the applications mentioned earlier in Section 1.3.1, it is at the same time natural to assume that the graphs involved are sparse, and potentially useful to recover only a fraction of the unknown matches $(u, \pi^\star(u))$.

This motivates the present work, whose goal is to show that *partial alignment of sparse correlated graphs is feasible*, and to introduce a polynomial-time algorithm for producing such partial alignments.

We do not recall here the definition of the correlated Erdős-Rényi model, already introduced in the introduction (see (1.10)), and specified in Section 4.1.1 of Chapter 4 in the sparse case. We only recall that the parameters of $\mathsf{G}(n, \lambda/n, s)$ are the number of nodes n, the mean degree $\lambda > 0$ and the correlation parameter $s \in [0, 1]$. The vertices of the second graph G' are relabeled with a uniform independent permutation $\pi^\star \in \mathcal{S}_n$, and we observe G and $H := G'^{\pi^\star}$.

Notations Let us recall a few notations. For an undirected graph G, denote by $V(G)$ its set of vertices, $E(G)$ (resp. $\overrightarrow{E}(G) := \{(u, v), \{u, v\} \in E(G)\}$) its set of non-oriented (resp.

[1] This related problem first appeared in this contribution, and will be the focus of Chapter 6.

oriented) edges. We use the notations $u \longleftrightarrow v$ if $\{u, v\} \in E(G)$ and $u \to v$ if $(u, v) \in \overrightarrow{E}(G)$. The usual graph distance in G will be denoted δ_G. For $u \in V(G)$, let $\mathcal{N}_G(u)$ denote the neighborhood – the set of neighbors – of v in G, and $\deg_G(v)$ its degree.

For $d \geq 1$ we also define $\mathcal{B}_G(v, d)$ the set of vertices at (graph) distance at most d from v, and $\mathcal{S}_G(v, d) := \mathcal{B}_G(v, d) \setminus \mathcal{B}_G(v, d-1)$ the set of vertices at distance exactly d from v.

For a rooted tree t, we let $\rho(t)$ denote its root node. For any $u \in V(T) \setminus \{\rho(t)\}$, we let $\pi_t(u)$ denote the parent of node u in T. For $d \geq 1$, we note $\mathcal{B}_d(t) = \mathcal{B}_t(\rho(t), d)$ and $\mathcal{L}_d(t) = \mathcal{S}_t(\rho(t), d)$.

We omit the dependencies in G or t of these notations when there is no ambiguity.

Objectives and main result Our main result is the proposal of the so-called *Neighborhood Tree Matching Algorithm* (`NTMA` hereafter) together with the following

Theorem 5.1. *For some $\lambda_0 > 1$, for all $\lambda \in (1, \lambda_0]$, there exists $s^*(\lambda) < 1$ such that, provided $s \in (s^*(\lambda), 1]$, for $(G, H) \sim \mathsf{G}(n, \lambda/n, s)$. `NTMA` returns a matching $\mathcal{S} = \mathcal{S}(G, H)$ verifying the following properties with high probability:*

$$|\mathcal{S} \cap \{(u, \pi^\star(u)),\ u \in [n]\}| = \Omega(n) \quad \textit{and} \quad |\mathcal{S} \setminus \{(u, \pi^\star(u)),\ u \in [n]\}| = o(n)\,. \tag{5.1}$$

In words, our algorithm returns a set of node alignments which contains a negligible fraction of mismatches, and $\Omega(n)$ good matches, that is performs *one-sided partial alignment* (see 1.4.4). This result covers values of λ arbitrarily close to 1, and thus applies to very sparse graphs. For $\lambda s < 1$, Erdős-Rényi graphs in our correlated model have connected components of size at most logarithmic in n, and we saw earlier on in Chapter 4 that there is no hope to recover a positive fraction of correct matches. This result can be interpreted as follows. For partial graph alignment of sparse Erdős-Rényi correlated random graphs, there is an "easy phase" that includes the parameter range $\{(\lambda, s) : \lambda \in (1, \lambda_0],\ s \in (s^*(\lambda), 1]\}$.

Organization of the chapter The description of the Neighborhood Tree Matching Algorithm and the proof strategy for establishing Theorem 5.1 are given in Section 5.3. Our algorithm relies essentially on a tree matching operation. To pave the way for Section 5.3, we introduce in Section 5.2 a notion of matching weight between trees that is key for our algorithm, and can be computed efficiently in a recursive manner. We further obtain probabilistic guarantees on the matching weights between random trees drawn according to some (correlated) Galton-Watson branching processes. These are instrumental in the proof of Theorem 5.1, but also of independent interest. Indeed we introduce in Section 5.2 a natural hypothesis testing problem on pairs of random trees, for which we obtain a successful test based on computation of tree matching weights. This last problem will next be the main focus of Chapter 6.

5.2. Tree matching

In this section, we introduce the matching weight between rooted trees and the related matching rate. We then establish high probability bounds on the latter for (correlated) Galton-Watson random trees. We also give an application to a hypothesis testing problem of correlation detection in trees.

5.2.1. Matching weight of two rooted trees

For any pair of rooted trees (τ, t), we say that a mapping $g : V(\tau) \to V(t)$ is *tree-preserving* if

- $f(\rho(\tau)) = \rho(t)$ (the root of τ is sent onto the root of t), and
- $\forall\, u \in V(\tau) \setminus \{\rho(\tau)\},\ f(\pi_\tau(u)) = \pi_t(f(u))$ (the parent of u is matched on the match of its parent).

For any $d \geq 0$, let $\mathcal{A}_d$ denote the collection of rooted trees whose leaves are all of depth d. Given two rooted trees t and t' of depth at most d, let $\{t \cap t'\}$ denote the collection of trees $\tau \in \mathcal{A}_d$ such that there exist tree-preserving injective embeddings $f : V(\tau) \to V(t)$, $f' : V(\tau) \to V(t')$. The *matching weight of t and t' at depth d*, as introduced in Section 1.4.4, is defined as follows:

$$\mathcal{W}_d(t, t') := \sup_{\tau \in \{t \cap t'\}} |\mathcal{L}_d(\tau)| \,, \tag{5.2}$$

i.e. the size of the largest common subtree of t and t', measured in terms of the number of leaves at depth d.

Remark 5.2.1. *Note that by definition* (5.2),

$$\mathcal{W}_0(t, t') = 1 \quad \textit{and} \quad \mathcal{W}_1(t, t') = \max\left(\deg_t(\rho(t)), \deg_{t'}(\rho(t'))\right) .$$

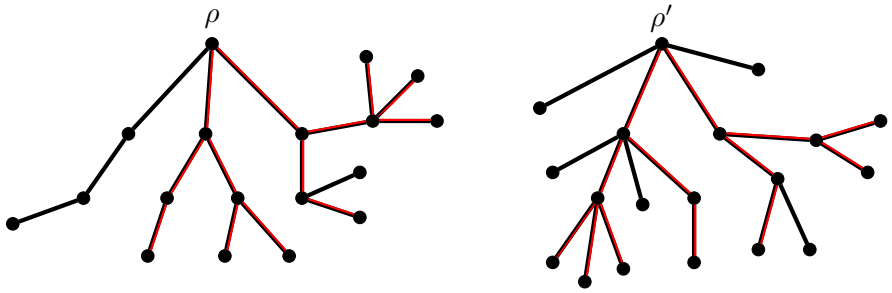


Figure 5.1 – *Example of two trees t, t' with $\mathcal{W}_3(t, t') = 7$, where an optimal $t \in \mathcal{A}_3$ is drawn in red.*

Before going any further, we need to recall and introduce a few new notations. For a rooted tree t of depth at most d, $u \in V(t)$, t_u is the downstream subtree of t re-rooted at u. More generally[2] $u, v \in V(t)$ such that $v \to u$, $t_{u \leftarrow v}$ denotes the subtree of t re-rooted at u where edge $\{u, v\}$ has been removed, that is the subtree pointed by the oriented edge $v \to u$.

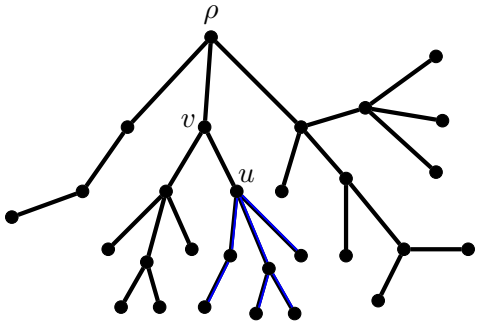


Figure 5.2 – *An example of a tree t and its corresponding $t_{u \leftarrow v}$ highlighted in blue.*

For a given pair of trees t and t' of depth at most D, for pairs of vertices $(u, u'), (v, v') \in V(t) \times V(t')$ such that $v \to u$, $v' \to u'$, $t_{u \leftarrow v}$ and $t'_{u' \leftarrow v'}$ are of depth at most d, the *matching weight of edges $v \to u$ and $v' \to u'$* is then defined as:

$$\mathcal{W}_d(u \leftarrow v, u' \leftarrow v') := \sup_{\tau \in \{t_{u \leftarrow v} \cap t'_{u' \leftarrow v'}\}} |\mathcal{L}_d(\tau)| \,. \tag{5.3}$$

Remark 5.2.2. *Note that by definition* (5.3),

$$\mathcal{W}_0(u \leftarrow v, u' \leftarrow v') = 1 \quad \textit{and} \quad \mathcal{W}_1(u \leftarrow v, u' \leftarrow v') = \max\left(\deg_t(u), \deg_{t'}(u')\right) - 1 \,.$$

[2]Note that in tree t, $t_u = t_{u \leftarrow \rho(t)}$.

5.2.2. Recursive computation of $\mathcal{W}_d$

From definition (5.3), doing a first step conditioning, i.e. distinguishing on the matching of pairs of nodes at depth 1 in both trees, gives the following:

$$\mathcal{W}_d(u \leftarrow v, u' \leftarrow v') = \sup_{\mathfrak{m} \in \mathcal{M}(\mathcal{N}_t(u)\setminus\{v\}, \mathcal{N}_{t'}(u')\setminus\{v'\})} \sum_{(w,w')\in\mathfrak{m}} \mathcal{W}_{d-1}(w \leftarrow u, w' \leftarrow u'), \qquad (5.4)$$

where for all sets U, V, $\mathcal{M}(U,V)$ is the set of all partial matchings between U and V, that is one-to-one mappings $\mathfrak{m} : U_0 \subseteq U \to V$. In the same way, for two trees t, t' of depth at most d, we have

$$\mathcal{W}_d(t,t') = \sup_{\mathfrak{m} \in \mathcal{M}(\mathcal{N}_t(\rho(t)), \mathcal{N}_{t'}(\rho(t')))} \sum_{(u,u')\in\mathfrak{m}} \mathcal{W}_{d-1}(u \leftarrow \rho(t), u' \leftarrow \rho(t')). \qquad (5.5)$$

These recursive formulae (5.5) and (5.4) show that matching weights at depth d can be obtained by computing weights at depth $d-1$ and solving a linear assignment problem (LAP) [Kuh55], and yield the following simple recursive algorithm to compute matching weights at depth d.

Algorithm 5.1: $\mathcal{W}_d(u \leftarrow v, u' \leftarrow v')$

```
if d = 0 then
    return 1;
else
    U ← N_t(u) \ {v} ;
    V ← N_{t'}(u') \ {v'} ;
    for (w, w') ∈ E × F do
        Compute W_{d-1}(w ← u, w' ← u');
    end
    Solve the LAP problem W* := sup_{m∈M(U,V)} Σ_{(w,w')∈m} W_{d-1}(w ← u, w' ← u');
    return W*;
end
```

Remark 5.2.3. *It is easy to show that computing the matching weight $\mathcal{W}_d(t,t')$ with the recursive algorithm 5.1 takes $O\left(d_{\max}^{2d}\right)$ time, where $d_{\max}$ is the maximal degree in t and t', which is not polynomial in d.*

However, we can do better using dynamic programming, namely storing for all $k \in [d]$ the weights $\mathcal{W}_k(e,e')$ in a array of size the number of pairs (e,e') where e and e' are two oriented edges in t,t' (there are $4 \times |t| \times |t'|$ such pairs). Each time we increase k and update the array, we solve one LAP for each pair (e,e'), e.g. with the Hungarian algorithm that running in cubic time complexity [Kuh55]. The size of the – small – matrix on which the LAP is done does not exceed $d_{\max} \times d_{\max}$, hence updating the array from k to $k+1$ is done in $O(|t| \times |t'| \times d_{\max}^3)$ steps. This gives a time complexity of $O\left(d \times |t| \times |t'| \times d_{\max}^3\right)$, which is better in general[3]*.*

5.2.3. Matching rate of random trees

For each $d \geq 0$, let us consider a pair of random trees (T_d, T'_d) sampled according to some distribution μ_d (we will further introduce models in the sequel). The *matching rate of the family of distributions* $\{\mu_d\}_{d\geq 0}$ is defined as follows

$$\gamma(\{\mu_d\}_{d\geq 0}) :=$$

[3]Note that however, if $|t|, |t'| = \Theta(n^\alpha)$ and $d_{\max} = O(\log n)$, which will be the case later in Section 5.3, then for small values of d and large values of n, the recursive algorithm 5.1 is faster.

$$\inf\left\{\gamma : \exists\, m, c, d_0 > 0,\ \forall x \geq 0,\ \forall d \geq d_0,\ \mu_d(\{(t,t') : \mathcal{W}_d(t,t') \geq mx\gamma^d\}) \leq e^{-(x-c)_+}\right\}. \tag{5.6}$$

This important quantity (5.6) captures the asymptotic geometric growth rate of matching weights of random trees drawn under μ_d. A simpler alternative definition could have been

$$\widetilde{\gamma}(\{\mu_d\}_{d\geq 0}) := \inf\left\{\gamma : \mu_d(\{(t,t') : \mathcal{W}_d(t,t') \geq \gamma^d\}) \underset{d\to\infty}{\longrightarrow} 0\right\}.$$

However, definition (5.6) better suits our purpose.

Remark 5.2.4. *By definition, note that for any $\gamma > \gamma(\{\mu_d\}_{d\geq 0})$, $\mu_d\left(\mathcal{W}_d(t,t') \geq \gamma^d\right)$ converges to 0 very fast, like $O\left(\exp\left(-c(\gamma)^d\right)\right)$ with $c(\gamma) > 1$, so that $\widetilde{\gamma}(\{\mu_d\}_{d\geq 0}) \leq \gamma(\{\mu_d\}_{d\geq 0})$.*

5.2.4. Models of random trees

We now introduce[4] three models of random trees that are relevant to sparse graph alignment.

Galton-Watson trees with Poisson offspring The *Galton-Watson tree with offspring* $\mathsf{Poi}(\mu)$ *up to depth* d, denoted by $\mathsf{GW}_d^{(\mu)}$, is defined recursively as follows. First, the distribution $\mathsf{GW}_0^{(\mu)}$ is a Dirac at the trivial tree, containing only the root. Then, for $d \geq 1$, sample a number $Z \sim \mathsf{Poi}(\mu)$ of independent $\mathsf{GW}_{d-1}^{(\lambda)}$ trees, and attach each of them as c children of the root, to form a tree of depth at most d.

Independent model $\mathbb{P}_d^{(\lambda)}$ Under the independent model $\mathbb{P}_d^{(\lambda)}$, t and t' are two independent $\mathsf{GW}_d^{(\lambda)}$, where $\lambda > 0$ is the mean number of children in the graph.

Tree augmentation For $\lambda > 0$ and $s \in [0,1]$, a (random) $(\lambda, s)-$*augmentation* of a given tree $\tau = (V, E)$, denoted by $\mathsf{Aug}_d^{(\lambda,s)}(\tau)$, is defined as follows. First, to each node u in V of depth $< d$, we attach a number Z_u^+ of additional children, where the Z_u^+ are i.i.d. of distribution $\mathsf{Poi}(\lambda(1-s))$. Let V^+ be the set of these additional children. To each $v \in V^+$ at depth d_v, we attach another random tree of distribution $\mathsf{GW}_{d-d_v}^{(\lambda)}$, independently of everything else.

Correlated shifted model $\mathbb{P}_d^{(\lambda,s,\delta)}$ In the correlated shifted model $\mathbb{P}_d^{(\lambda,s,\delta)}$, the tree T is rooted at ρ and T' is rooted at ρ', and ρ' is also a node of T, at distance δ from its root ρ. The two trees are generated as follows. First, all nodes u in T on the path from ρ to the parent of ρ' in T have, besides their child leading to ρ', extra $Z_u^+ \sim \mathsf{Poi}(\lambda)$ children in T, and all extra child v at depth d_v has an additional offspring in T sampled from $\mathsf{GW}_{d-d_v}^{(\lambda)}$. Then, sample an *intersection tree* $\tau^\star \sim \mathsf{GW}_{d-\delta}^{(\lambda s)}$ starting from ρ'. Independently, we finish the construction of T (resp. of T') with a $(\lambda, s)-$augmentation of $\tau^\star$ of depth $d-\delta$ (resp. of depth d). See Figure 5.3 for an illustration. We denote $(T, T') \sim \mathbb{P}_d^{(\lambda,s,\delta)}$.

Correlated model $\mathbb{P}_d^{(\lambda,s)}$ It is the previous model with $\delta = 0$, so that the two correlated trees T and T' have same root ρ. We denote $(T, T') \sim \mathbb{P}_d^{(\lambda,s)}$. In other words, the correlated model $\mathbb{P}_d^{(\lambda,s)}$ is built as follows: starting from an *intersection tree* $\tau^\star \sim \mathsf{GW}_d^{(\lambda s)}$, and T and T' are obtained as two independent $(\lambda, s)-$augmentations of $\tau^\star$. We denote $(T, T') \sim \mathbb{P}_d^{(\lambda,s)}$.

In all these models, the labels of the trees T and T' are always forgotten, or randomly uniformly re-sampled. We however still distinguish the roots af the two trees. It can easily be verified that the marginals of T and T' are the same under $\mathbb{P}_d^{(\lambda)}$ and $\mathbb{P}_d^{(\lambda,s)}$, namely $\mathsf{GW}_d^{(\lambda)}$. The parameters are λ, the mean number of children of a node, and the correlation s.

[4] Some of them are already mentioned in the introduction, see Section 1.4.

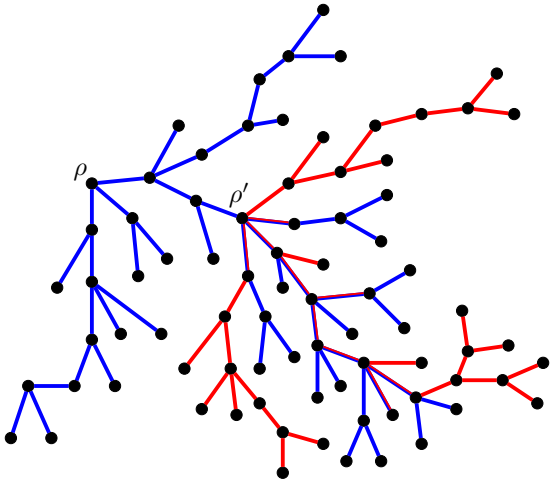


Figure 5.3 – *Random trees T (blue) and T' (red) from model $\mathbb{P}_d^{(\lambda,s,\delta)}$ with $\delta = 3$.*

We now move to the analysis of matching rates for these models, which as explained before are crucial quantities which can help discriminate between the independent and the correlated setting.

5.2.5. Matching rate of independent and correlated Galton-Watson trees

Proposition 5.2.1. *Let $\lambda > 1$ and $s \in [0,1]$ such that $\lambda s > 1$. Then, letting $\gamma(\lambda, s) := \gamma(\{\mathbb{P}_d^{(\lambda,s)}\}_{d\geq 0})$, we have:*

$$\gamma(\lambda, s) \geq \lambda s.$$

Proof. Let $\tau^\star$ be the intersection tree between T and T'. Branching process theory implies that $(\lambda s)^{-d}\big|\mathcal{L}_d(\tau^\star)\big|$ converges almost surely to a random variable Z as $d \to \infty$, such that $\mathbb{P}(Z > 0) = 1 - p_{\text{ext}}$, with p_{ext} the extinction probability of the branching tree $\tau^\star$. Since $p_{\text{ext}} < 1$ when $\lambda s > 1$, and for every small enough $\varepsilon > 0$,

$$\lim_{d\to\infty} \mathbb{P}_d^{(\lambda,s)} \left(\mathcal{W}_d(T,T') \geq (\lambda s(1-\varepsilon))^d\right) \geq 1 - p_{\text{ext}} > 0\,,$$

the result follows. □

Theorem 5.2. *Let $\gamma(\lambda) := \gamma(\{\mathbb{P}_d^{(\lambda)}\}_{d\geq 0})$. There exists $\lambda_0 > 1$ such that for all $\lambda \in (1, \lambda_0]$, we have*

$$\gamma(\lambda) < \lambda. \tag{5.7}$$

Evaluations of $\gamma(\lambda)$ by simulations, confirming and illustrating Theorem 5.2, are provided in Appendix 5.A.1.

Outline of proof of Theorem 5.2. The full proof of Theorem 5.2 is detailed in Appendix 5.B.1, but we here give the key steps. We introduce some notations. First, for a tree t of depth at most d, let $r_d(t)$ denote the tree obtained by iteratively pruning leaves of depth strictly less than d. When computing $\mathcal{W}_d(t,t')$, the only informative subtrees are precisely in $r_d(t)$ and in $r_d(t')$, one of these being empty if t or t' doesn't survive up to depth d. In the rest of the chapter, we define T_d the random variable $r_d(T)$ where T is conditioned to survive up to depth d.

Consider $(T,T') \sim \mathbb{P}_d^{(\lambda)}$. We let $\mathcal{E}_d$ (respectively, $\mathcal{E}'_d$) denote the event that tree T (respectively, T') becomes extinct before d generations, i.e. $\mathcal{L}_d(T) = \varnothing$ (respectively, $\mathcal{L}_d(T') = \varnothing$). We let $p_d = \mathbb{P}(\mathcal{E}_d) = \mathbb{P}(\mathcal{E}'_d)$. It is well known that it satisfies the recursion

$$p_0 = 0,\ p_d = e^{-\lambda(1-p_{d-1})}.$$

We now state a lemma on the structure of T_d.

Lemma 5.2.1. *For any $\lambda > 1$, T_d can be constructed by first sampling the number of children D of the root $\rho(T)$ according to distribution*

$$\mathbb{P}(D = k) = \mathbb{1}_{k>0} \frac{\mathbb{P}(\mathsf{Poi}(\lambda(1-p_{d-1})) = k)}{\mathbb{P}(\mathsf{Poi}(\lambda(1-p_{d-1})) > 0)} =: q_{d,k},$$

and then attaching D independent copies of T_{d-1} to the D children of $\rho(T)$.

Proof of Lemma 5.2.1. For a tree t, we identify t to $(t_1, \dots, t_k)$ the tuple of offsprings of its k children. Write, defining D the number of children of $\rho(T)$, fixing $k \geq 1$, $t_1, \dots, t_k \in \mathcal{A}_{d-1}$, and letting $S = i_1 < \ldots < i_k$ run over all k subsets of $[\ell]$:

$$\begin{aligned}
\mathbb{P}(T_d = (t_1, \dots, t_k)) &= \sum_{\ell \geq 0} \mathbb{P}(T_d = (t_1, \dots, t_k), D = \ell) \\
&= \sum_{\ell \geq k} \sum_{S} \mathbb{P}\left(D = \ell, r_d(T_{i_j}) = t_j, j \in [k], r_d(T^v) = \varnothing, v \notin S \middle| \overline{\mathcal{E}}_d\right) \\
&= \frac{1}{1-p_d} \sum_{\ell \geq k} \binom{\ell}{k} e^{-\lambda} \frac{\lambda^\ell}{\ell!} p_{d-1}^{\ell-k} \prod_{j=1}^{k} \mathbb{P}(T_{d-1} = t_j)(1-p_{d-1}) \\
&= \frac{1}{1-p_d} \frac{(\lambda(1-p_{d-1}))^k}{k!} \prod_{j=1}^{k} \mathbb{P}(T_{d-1} = t^j) \sum_{\ell \geq k} e^{-\lambda} \frac{(\lambda p_{d-1})^{\ell-k}}{(\ell-k)!} \\
&= \frac{e^{-\lambda(1-p_{d-1})}}{1-p_d} \frac{(\lambda(1-p_{d-1}))^k}{k!} \prod_{j=1}^{k} \mathbb{P}(T_{d-1} = t_j)\,.
\end{aligned}$$

The conclusion follows by noting that $1 - p_d = 1 - e^{-\lambda(1-p_{d-1})}$. □

Assume $\varepsilon = \lambda - 1$ to be small enough. Fix $r \in (0,1)$, let $\gamma = 1 + r\varepsilon$. We first show using exponential moments that there exist $m, c > 0$ and $d_0 > 0$ such that for all $x > 0$

$$\mathbb{P}\left(\mathcal{W}_{d_0}\left(T_{d_0}, T'_{d_0}\right) \geq mx\right) \leq e^{-x+c}.$$

Then we define the random variables

$$X_d := \gamma^{-(d-d_0)} m^{-1} \mathcal{W}_d\left(T_d, T'_d\right).$$

Then, considering the number D of children of the root in T_d (resp. D' in T'_d), using the previous lemma, one can establish, for all $x > 0$, a recursive formula of the following form

$$\mathbb{P}\left(X_d \geq x\right) \leq \sum_{k,\ell \geq 1} q_{d,k} q_{d,\ell} \mathbb{P}\left(\exists \mathfrak{m} \in \mathcal{M}\left([k], [\ell]\right), \sum_{(i,u) \in \mathfrak{m}} X_{d-1,i,u} \geq \gamma x\right),$$

where the $X_{d-1,i,u}$ are i.i.d. copies of X_{d-1}. The union bound yields

$$\mathbb{P}\left(X_d \geq x\right) \leq \sum_{k,\ell \geq 1} q_{d,k} q_{d,\ell} \min\left(1, (k \vee \ell)^{\underline{k \wedge \ell}} \times \mathbb{P}\left(\sum_{i=1}^{k \wedge \ell} X_{d-1,i,u} \geq \gamma x\right)\right),$$

where $m^{\underline{p}} := m(m-1)\dots(m-p+1) = \frac{m!}{(m-p)!}$. This inequality enables, with a few more technical steps (see 5.B.1), to propagate recursively the inequality

$$\mathbb{P}\left(X_d \geq x\right) \leq e^{-(x-c)_+}.$$

5.2.6. Implications for a hypothesis testing problem

Let a pair of trees (T, T') be distributed according to $\mathbb{P}_d^{(\lambda)}$ under the null hypothesis $\mathcal{H}_0$, and according to $\mathbb{P}_d^{(\lambda,s)}$ under the alternative hypothesis $\mathcal{H}_1$. They are thus independent under $\mathcal{H}_0$, and correlated under $\mathcal{H}_1$. Consider the following test:

$$\text{Decide } \mathcal{H}_0 \text{ if } \mathcal{W}_d(T, T') < \gamma^d, \ \mathcal{H}_1 \text{ otherwise.}$$

Assume that $\gamma(\lambda) < \gamma < \lambda s$. Then in view of Remark 5.2.4 and Theorem 5.2 one has for some $c(\gamma) > 1$:

$$\mathbb{P}\left(\text{decide } \mathcal{H}_1 \middle| \mathcal{H}_0\right) = O(e^{-c(\gamma)^d}),$$

thus a super-exponential decay of the probability of false positive (first type error). Conversely, in view of Proposition 5.2.1, noting $\tau^\star$ the intersection tree under $\mathcal{H}_1$, one has

$$\mathbb{P}\left(\text{decide } \mathcal{H}_0 \middle| \mathcal{H}_1, \text{non-extinction of } \tau^\star\right) = o_d(1).$$

The false negative probability of this test thus also goes to zero, provided the intersection tree survives. As we will see in next section, this hypothesis testing problem on a pair of random trees is related to our original graph alignment problem much as the so-called tree reconstruction problem, reviewed in [MP03], is related to community detection in sparse random graphs (see e.g. [BLM15]). This fundamental correspondence is studied in detail in Chapter 6.

5.2.7. Matching rate of correlated shifted trees

Theorem 5.3. *Let* $\gamma(\lambda, s, \delta) := \gamma(\{\mathbb{P}_d^{(\lambda,s,\delta)}\}_{d\geq 0})$. *There exists* $\lambda_0 > 1$ *such that for all* $\lambda \in (1, \lambda_0]$ *we have*

$$\sup_{\delta \geq 1} \gamma(\lambda, s, \delta) < \lambda. \tag{5.8}$$

Evaluations of $\gamma(\lambda, s, \delta)$ by simulations, confirming and illustrating Theorem 5.3, are provided in Appendix 5.A.1.

Outline of proof of Theorem 5.3. The full proof of Theorem 5.3 is detailed in Appendix 5.B.2, but we here give the key steps. The proof will again be by induction on d, the initial step being established with the same argument as in the proof of Theorem 5.2. The difference $\varepsilon = \lambda - 1$ is assumed to be small enough. We fix $r \in (0, 1)$, and we let $\gamma = 1 + r\varepsilon'$. We now work with the random variables

$$X'_d := \gamma^{-(d-d_0)} m^{-1} \mathcal{W}_d\left(T_d, T'_d\right),$$

conditionally on the event that the path from ρ to ρ' survives down to depth d in T. Then, considering D the number of children of ρ in T_d, D' the number of children of ρ' in T'_d that are in the intersection tree $T_d \cap T'_d$, and D'' the number of children of ρ' in $T'_d \setminus T_d$, we establish for all $x > 0$ a recursive formula of the following form

$$\mathbb{P}\left(X'_d \geq x\right) \leq \sum_{k,\ell \geq 1} \mathbb{P}\left(D' + D'' = k, D = \ell\right) \min\left(1, (k \vee \ell)^{\frac{k \wedge \ell}{}} \mathbb{P}\left(X'_{d-1} + \sum_{i=1}^{k\wedge\ell - 1} X_{d-1,i,u} \geq \gamma x\right)\right),$$

where the $X_{d-1,i,u}$ are i.i.d. copies of X_{d-1} as defined in the proof of Theorem 5.2. Again, with a few more technical steps (see 5.B.2), we are able to propagate recursively the inequality

$$\mathbb{P}\left(X'_d \geq x\right) \leq e^{-(x-c)_+}.$$

5.3. Sparse graph alignment by matching trees

We now describe our main algorithm and its theoretical guarantees. For simplicity we assume that the underlying permutation $\pi^\star$ is the identity.

5.3.1. Neighborhood Tree Matching Algorithm (NTMA), main result

The main intuition for the NTMA algorithm is as follows. In order to distinguish matched pairs of nodes (u, u'), we consider their neighborhoods at a certain depth d, that are close to Galton-Watson trees. In the case where the two vertices are actual matches, the largest common subtree measured in terms of children at depth (exactly) d is w.h.p. of size $\geq (\lambda s)^d$. However, when the two nodes u and u' are sufficiently distant in the union graph aligned with the ground truth, $G \cap G'$, previous study of matching rates shows that the growth rate of largest common subtree will be $< \lambda s$. The natural idea is thus to apply the test comparing $\mathcal{W}_d(\mathcal{B}_G(u,d), \mathcal{B}_H(u',d))$ to γ^d for some well-chosen γ to decide whether u is matched to u'.

However, as the reader may have noticed, testing $\mathcal{W}_d(\mathcal{B}_G(u,d), \mathcal{B}_H(u',d)) > \gamma^d$ is not enough, because two-hop neighbors in $G \cap G'$ would dramatically increase the number of incorrectly matched pairs, making the performance collapse. To fix this, we use the *dangling trees trick*: instead of just looking at their neighborhoods, we look for the downstream trees from two distinct neighbors $v \neq w$ of u, and $v' \neq w'$ of u'. The trick is now to compare both $\mathcal{W}_{d-1}(v \leftarrow u, v' \leftarrow u')$ and $\mathcal{W}_{d-1}(w \leftarrow u, w' \leftarrow u')$ to γ^{d-1}. This way, even if $u \neq u'$ and u and u' are close by, the pairs of rooted trees that can be considered will lead to one of the four cases considered and illustrated on Figure 5.4, that are settled in the proof of Theorem 5.5.

Our algorithm is as follows, where matching tree weights $\mathcal{W}_{d-1}(j \leftarrow i, v \leftarrow u)$ are defined in (5.3):

Algorithm 5.2: Neighborhood Tree Matching Algorithm for sparse graph alignment

Input: Two graphs G and H of size n, average degree λ, depth d, parameter γ.
Output: A set of pairs $\mathcal{S} \subset V(G) \times V(H)$.
$\mathcal{S} \leftarrow \varnothing$
for $(u, u') \in V(G) \times V(H)$ **do**
  if $\mathcal{B}_G(u,d)$ *and* $\mathcal{B}_H(u',d)$ *contain no cycle, and* $\exists v \neq w \in \mathcal{N}_G(u), \exists v' \neq w' \in \mathcal{N}_H(u')$ *such that* $\mathcal{W}_{d-1}(v \leftarrow u, v' \leftarrow u') > \gamma^{d-1}$ *and* $\mathcal{W}_{d-1}(w \leftarrow u, w' \leftarrow u') > \gamma^{d-1}$ **then**
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{(u, u')\}$
  end
end
return $\mathcal{S}$

Remark 5.3.1. *For $d = \lfloor c \log n \rfloor$, in view of Remark 5.2.3, with high probability the complexity of NTMA is*

$$O\left(|V(G)|\,|V(H)|\,(\log n)^2 n^{2c\log\lambda} d_{\max}^2\right) + O\left(|E(G)|\,|E(H)|\,(\log n) d_{\max}^3\right),$$

where $d_{\max}$ is the maximum degree in G and H. In the context of Theorems 5.4 and 5.5 the complexity is then $O\left((\log n)^4 n^{5/2}\right)$.

The two results to follow will readily imply Theorem 5.1.

Theorem 5.4. *Let $(G, H) \sim \mathsf{G}(n, \lambda/n, s)$ be $s-$correlated Erdős-Rényi graphs such that $\lambda s > 1$. Let $d = \lfloor c \log n \rfloor$ with $c \log(\lambda(2-s)) < 1/2$. Then for $\gamma \in (1, \lambda s)$, with high*

probability, if $\mathcal{S}$ denotes the matching returned by **NTMA**,

$$\frac{1}{n}\sum_{u\in[n]} \mathbb{1}_{\{(u,u)\in\mathcal{S}\}} = \Omega(1). \tag{5.9}$$

In other words, a non vanishing fraction of nodes is correctly recovered by **NTMA** *(Algorithm 5.2).*

Theorem 5.5. *Let $(G,H)\sim \mathsf{G}(n,\lambda/n,s)$ be two $s-$correlated Erdős-Rényi graphs. Assume that $\gamma_0(\lambda) := \max\left(\gamma(\lambda), \sup_{\delta\geq 1}\gamma(\lambda,s,\delta)\right) < \lambda s$, and that $d=\lfloor c\log n\rfloor$ with $c\log\lambda < 1/4$. Then for $\gamma\in(\gamma_0(\lambda),\lambda s)$, with high probability,*

$$\mathrm{err}(n) := \frac{1}{n}\sum_{u=1}^{n} \mathbb{1}_{\{\exists u'\neq u,\ (u,u')\in\mathcal{S}\}} = o(1)\,, \tag{5.10}$$

i.e. only at most a vanishing fraction of nodes are incorrectly matched by **NTMA** *(Algorithm 5.2).*

Remark 5.3.2. *The set $\mathcal{S}$ returned by the* **NTMA** *is not necessarily a matching. Let $\mathcal{S}'$ be obtained by removing all pairs (i,u) of $\mathcal{S}$ such that i or u appears at least twice. Theorems 5.4 and 5.5 guarantee that $\mathcal{S}'$ still contains a non-vanishing number of correct matches and a vanishing number of incorrect matches. Theorem 5.1 easily follows. Simulations of* **NTMA-2**, *a simple variant of of* **NTMA**, *are reported in Appendix 5.A.2. These confirm our theory, as the algorithm returns many good matches and few mismatches.*

5.3.2. Proof of Theorems 5.4 and 5.5

We start by stating Lemmas, adapted from [Mas14] and [BLM15] and proven in Appendix 5.C, that are instrumental in the proofs of Theorems 5.4 and 5.5.

Lemma 5.3.1 (Control of the sizes of the neighborhoods)**.** *Let $G\sim\mathsf{G}(n,\lambda/n)$, $d=\lfloor c\log n\rfloor$ with $c\log\lambda<1$. For all $\gamma>0$, there is a constant $C=C(\gamma)>0$ such that with probability $1-O\left(n^{-\gamma}\right)$, for all $u\in[n]$, $t\in[d]$:*

$$|\mathcal{S}_G(u,t)| \leq C(\log n)\lambda^t. \tag{5.11}$$

Lemma 5.3.2 (Cycles in the neighborhoods in an ER graph)**.** *Let $G\sim\mathsf{G}(n,\lambda/n)$, $d=\lfloor c\log n\rfloor$ with $c\log\lambda<1/2$. There exists $\varepsilon>0$ such that for any vertex $u\in[n]$, one has*

$$\mathbb{P}\left(\mathcal{B}_G(u,d) \text{ contains a cycle}\right) = O\left(n^{-\varepsilon}\right). \tag{5.12}$$

Lemma 5.3.3 (Two logarithmic neighborhoods are typically size-independent)**.** *Let $G\sim\mathsf{G}(n,\lambda/n)$ with $\lambda>1$, $d=\lfloor c\log n\rfloor$ with $c\log\lambda<1/2$. Then there exists $\varepsilon>0$ such that for any fixed nodes $u\neq v$, the variation distance between the joint law of the neighborhoods $\mathcal{L}\left((\mathcal{S}_G(u,t),\mathcal{S}_G(v,t))_{t\leq d}\right)$ and the product law $\mathcal{L}\left((\mathcal{S}_G(u,t))_{t\leq d}\right)\otimes\mathcal{L}\left((\mathcal{S}_G(v,t))_{t\leq d}\right)$ tends to 0 as $O\left(n^{-\varepsilon}\right)$ for some $\varepsilon>0$ when $n\to\infty$.*

Lemma 5.3.4 (Coupling the $|\mathcal{S}_G(i,t)|$ with a Galton-Watson process)**.** *Let $G\sim\mathsf{G}(n,\lambda/n)$, $d=\lfloor c\log n\rfloor$ with $c\log\lambda<1/2$. For a fixed $u\in[n]$, the variation distance between the law of $(|\mathcal{S}_G(u,t)|)_{t\leq d}$ and the law of $(Z_t)_{t\leq d}$ where $(Z_t)_t$ is a Galton-Watson process of offspring distribution $\mathsf{Poi}(\lambda)$ tends to 0 as $O\left(n^{-\varepsilon}\right)$ when $n\to\infty$.*

Proof of Theorems 5.4 and 5.5

Proof of Theorem 5.4. Define the joint graph $G_\cup = G\cup G'$. We recall that we assume that $\pi^\star=\mathrm{id}$, without loss of generality, hence $H=G'$. For $u\in[n]$, let M_u denote the event

that the algorithm matches u in G with u in H, i.e. on which $\mathcal{B}_G(u,d)$ and $\mathcal{B}_H(u,d)$ contain no cycle, and $\exists v \neq w \in \mathcal{N}_G(u), \exists v' \neq w' \in \mathcal{N}_H(u)$ such that $\mathcal{W}_{d-1}(v \leftarrow u, v' \leftarrow u) > \gamma^{d-1}$ and $\mathcal{W}_{d-1}(w \leftarrow u, w' \leftarrow u) > \gamma^{d-1}$. Denote by $C_{\cup,u,d}$ the event that there is no cycle in $\mathcal{B}_{G_\cup}(u,d)$.

With the same arguments as in the proof of Lemma 5.3.4, the two neighborhoods $\mathcal{B}_G(u,d)$ and $\mathcal{B}_H(u,d)$ can be coupled with trees distributed as $\mathbb{P}_d^{(\lambda,s)}$ of Section 5.2. However, we will instead consider the intersection graph $G_\cap = G \cap H$. Obviously, $G_\cap \sim \mathsf{G}(n, \lambda s/n)$. By Lemma 5.3.4, the random variables $|\mathcal{S}_{G_\cap}(u,t)|$ can be coupled with a Galton-Watson process with offspring distribution $\mathsf{Poi}(\lambda s)$ up to depth $t = d$. Let P_u denote the event that this coupling succeeds. Since $\lambda s > 1$, there is a probability $2\alpha > 0$ that the first generation has at least two children whose offsprings survive up to depth $d-1$. Note S this event. On event S, the matching given by the identity on the intersection tree implies the existence of two neighbors $v \neq w \in \mathcal{N}_G(u)$ and $v' \neq w' \in \mathcal{N}_H(u)$ such that with high probability $\mathcal{W}_{d-1}(v \leftarrow u, v' \leftarrow u) > \gamma^{d-1}$ and $\mathcal{W}_{d-1}(w \leftarrow u, w' \leftarrow u) > \gamma^{d-1}$, by standard martingale arguments, as in Proposition 5.2.1. This gives the lower bound for $\mathbb{P}(M_u)$:

$$\mathbb{P}(M_u) \ \geq \mathbb{P}\left(C_{\cup,u,d} \cap P_u \cap S\right) \ \geq 2\alpha - o(1) > \alpha > 0.$$

It is easy to see that $G_\cup \sim \mathsf{G}(n, \lambda(2-s)/n)$. For $u \neq v \in [n]$, define $I_{u,v}$ the event on which the two neighborhoods of u and v in $G_\cup$ coincide with their independent couplings up to depth d. By lemma 5.3.3, $\mathbb{P}(I_{u,v}) = 1 - o(1)$. Then for $0 < \varepsilon < \alpha$ Markov's inequality yields

$$\mathbb{P}\left(\frac{1}{n}\sum_{u\in[n]} \mathbb{1}_{\{(u,u)\in\mathcal{S}\}} < \alpha - \varepsilon\right) \leq \mathbb{P}\left(\sum_{u\in[n]} \left(\mathbb{P}(M_u) - \mathbb{1}_{M_u}\right) > \varepsilon n\right) \tag{5.13}$$

$$\leq \frac{1}{n^2\varepsilon^2}\left(n\mathrm{Var}\left(\mathbb{1}_{M_1}\right) + n(n-1)\mathrm{Cov}\left(\mathbb{1}_{M_1}, \mathbb{1}_{M_2}\right)\right) \tag{5.14}$$

$$\leq \frac{\mathrm{Var}\left(\mathbb{1}_{M_1}\right)}{n\varepsilon^2} + \frac{1 - \mathbb{P}\left(I_{1,2}\right)}{\varepsilon^2} \to 0. \tag{5.15}$$

□

Proof of Theorem 5.5. Define

$$d_{\max} := \max\left(\max_u \deg_G(u), \max_{u'} \deg_H(u')\right).$$

We use the same notations as in the former proof: $G_\cup = G \cup H$ and $G_\cap = G \cap H$. Fix $u \in [n]$. In the rest of the proof we work conditionally to the event $C_{\cup,u,2d}$ that $\mathcal{B}_{G_\cup}(u,2d)$ has no cycle. Since $c\log\lambda < 1/4$, $\mathbb{P}\left(C_{\cup,u,2d}\right) = 1 - o(1)$ by Lemma 5.3.2.

Fix another vertex $u' \neq u$. The $d-$neighborhoods $\mathcal{B}_G(u,d)$ and $\mathcal{B}_H(u',d)$ have offspring distribution stochastically dominated by $\mathsf{Bin}(n, \lambda/n)$, which is also dominated by $\mathsf{Poi}(\lambda')$ as soon as $\lambda' = \lambda + O(1/n)$ (see e.g. [KM09]). We can choose λ' such that $\gamma > \gamma(\lambda', 0)$ still holds: indeed, by a standard coupling argument, one can see that $\gamma : \lambda \mapsto \gamma(\lambda)$ is increasing. We now build two dominating (in the usual edge presence sense) tree-like $d-$neighborhoods of i and u with the following construction.

- First, if the two neighborhoods do not intersect, we simply sample two independent trees from model $\mathbb{P}_d^{(\lambda')}$ rooted in u and in u'.
- If the two neighborhoods intersect, condition to the event that α is the contact point in the path $\mathfrak{p}_\cup$ (unique by conditioning on $C_{\cup,u,2d}$) from u to u' in the joint graph. Then there is a path of edges of G (say, blue) from u to α, then a path of edges of H (say, red) from α to u'. Next, complete this construction: along $\mathfrak{p}_\cup$, propagate the blue path from

α towards u' with probability s on each edge, stopping at the first time when one red edge is not selected. Do the symmetrical construction to propagate the red path from α towards u. Finally, to each double-colored vertex w, attach independent realizations of model $\mathbb{P}_{d(w)}^{(\lambda',s)}$ of adapted depth, and to each single-colored vertex z, attach independent realizations of model $\mathbb{P}_{d(z)}^{(\lambda')}$ of adapted depth.

Note that these constructions lead to at most one path $\mathfrak{p}_\cup$ between u and v' in $\mathcal{B}_G(u,d) \cup \mathcal{B}_H(u',d)$, so a fortiori in $\mathcal{B}_G(u,d) \cap \mathcal{B}_H(u',d)$. Denote by $\mathfrak{p}_\cap$ this hypothetical path (cf. Figure 5.4). We then distinguish between several cases.

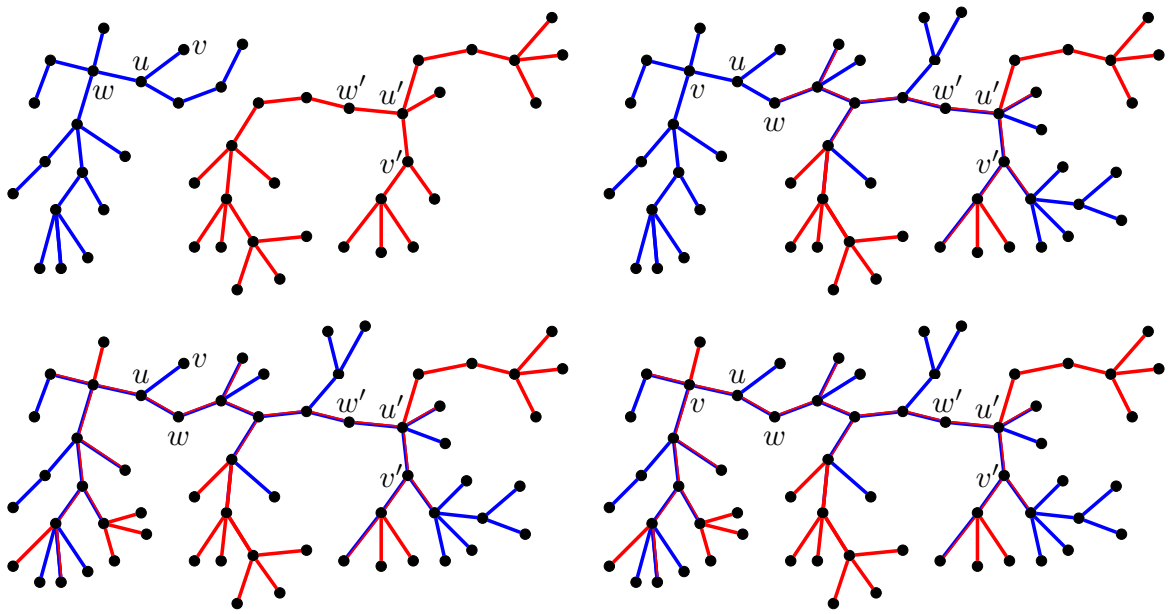


Figure 5.4 – *Possible realizations of $\mathcal{B}_G(u,d)$ (blue) and $\mathcal{B}_H(u',d)$ (red), with distinct cases (i) (top left), (ii) (top right), (iii.a) (bottom left) and (iii.b) (bottom right).*

***Case** (i): $\delta_{G_\cup}(u,u') > 2d$* (Figure 5.4, top left), i.e. $\mathcal{B}_G(u,d) \cap \mathcal{B}_H(u',d) = \varnothing$. The construction gives a coupling with two independent trees from model $\mathbb{P}_d^{(\lambda)}$. By assumption $\gamma(\lambda) < \lambda s$, the probability that there exist v in $\mathcal{N}_G(u)$ and v' in $\mathcal{N}_H(u')$ such that $\mathcal{W}_{d-1}(v \leftarrow u, v' \leftarrow u') > \gamma^{d-1}$ is upper bounded by $O\left(d_{\max}^2 \exp\left(-n^\varepsilon\right)\right)$, following Remark 5.2.4. Hence u is matched to u' with at most this probability.

Case** (ii): $\delta_{G_\cup}(u,u') \leq 2d$ **but** $\mathfrak{p}_\cap$ **does not exist (see Figure 5.4, top right). Take $v \neq w$ two neighbors of u and $v' \neq w'$ two neighbors of u'. Then (at least) one of these vertices is not on $\mathfrak{p}_\cup$ (e.g. vertex v on Figure 5.4): the downstream tree from this vertex is independent from every other neighborhood in the other graph. They can be coupled with model $\mathbb{P}_d^{(\lambda)}$, and the same bound as in case (i) holds.

Now assume that $\mathfrak{p}_\cap$ exists, and let $v \neq w$ two neighbors of u and $v \neq v'$ two neighbors of u'.

***Case** (iii.a):* at least one of the edges $\{u,v\},\{u,w\},\{u',v'\},\{u',w'\}$ is not in $G_\cap$ (e.g. edge (u,v) on Figure 5.4, bottom left): again, the same argument applies.

***Case** (iii.b):* Edges $\{u,v\},\{u,w\},\{u',v'\},\{u',w'\}$ are all in $G_\cap$ (see Figure 5.4, bottom right). Then one pair of vertices (say (w,w') as on Figure 5.4) can be on $\mathfrak{p}_\cap$ and bring a high $\mathcal{W}_{d-1}(w \leftarrow u, w' \leftarrow u') > \gamma^{d-1}$ matching weight, if their descendants spread over a great part

of the intersection. In that case, since v and v' can't be on $\mathfrak{p}_\cap$, the associated downstream trees are independent, and again $\mathcal{W}_{d-1}(j \leftarrow i, v \leftarrow u) < \gamma^{d-1}$ with high probability.

The remaining subcase to be considered is that of matches (v, w') and (w, v'), with w, w' on $\mathfrak{p}_\cap$. All trees involved are then correlated. However, the coupling construction induces a coupling of the two pairs of $(d-1)-$neighborhoods (from (v, w') and from (w, v'), see Figure 5.4) with two pairs of trees from model $GW(\lambda', s, \delta)$ where $\delta = |\mathfrak{p}_\cap|$. We assume in the Theorem that $\gamma(\lambda, s, \delta) < \lambda s$ so by Theorem 5.3, the probability that $\mathcal{W}_{d-1}(v \leftarrow u, w' \leftarrow u') > \gamma^{d-1}$ and $\mathcal{W}_{d-1}(w \leftarrow u, v' \leftarrow u') > \gamma^{d-1}$ is upper bounded by $O\left(\exp\left(-n^{\varepsilon}\right)\right)$.

Thus, for u fixed, one has

$$\mathbb{P}\left(\exists u' \neq u,\ (u, u') \in \mathcal{S}\right) \leq 1 - \mathbb{P}\left(C_{\cup,u,2d}\right) + n \times \mathbb{P}\left(C_{\cup,u,2d}\right) \times d_{\max}^2 \times O\left(\exp\left(-n^{\varepsilon}\right)\right) = o(1).$$

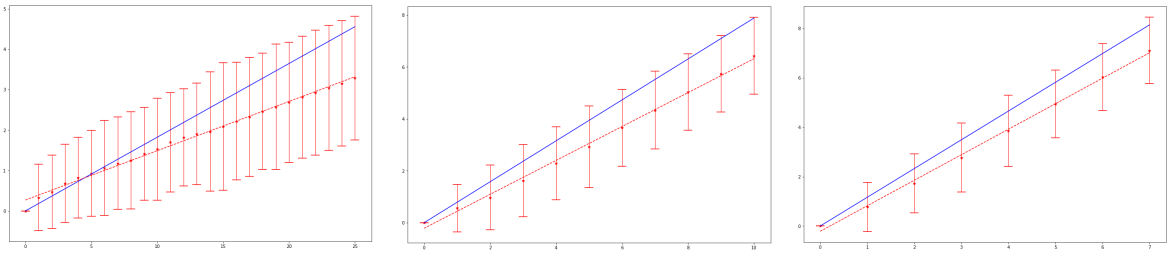
The theorem then follows by appealing to Markov's inequality. □

Appendix of Chapter 5

5.A. Numerical experiments

5.A.1. Simulations for tree matching

We here present some simulations of matching rates $\gamma(\lambda)$ (Figure 5.5) and $\gamma(\lambda, s, \delta)$ for $s = 1$ (Figure 5.6) in order to illustrate Theorems 5.2 and 5.3 and the final conjecture. For these simulations, error bars correspond to one standard deviation.

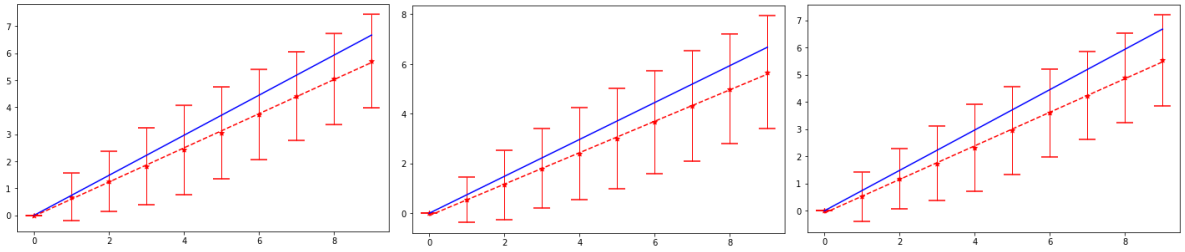


(a) – $\lambda = 1.2$, $\log \lambda \sim 0.18$. *Red dashed slope* ~ 0.12

(b) – $\lambda = 2.2$, $\log \lambda \sim 0.79$. *Red dashed slope* ~ 0.65

(c) – $\lambda = 3.2$, $\log \lambda \sim 1.16$. *Red dashed slope* ~ 1.03

Figure 5.5 – *Comparison of* $d \log \lambda$ *(blue) and* $\log \mathcal{W}_d(T, T')$ *(red) for* $\mathcal{W}_d(T, T') \sim \mathbb{P}_d^{(\lambda)}$ *conditioned to survive (*100 *iterations)*



(a) – $\lambda = 2.1, \log \lambda \sim 0.74, \delta = 1$. *Red dashed slope* ~ 0.63

(b) – $\lambda = 2.1, \log \lambda \sim 0.74, \delta = 2$. *Red dashed slope* ~ 0.63

(c) – $\lambda = 2.1, \log \lambda \sim 0.74, \delta = 5$. *Red dashed slope* ~ 0.62

Figure 5.6 – *Comparison of* $d \log \lambda$ *(blue) and* $\log \mathcal{W}_d(T, T')$ *(red) for* $\mathcal{W}_d(T, T') \sim \mathbb{P}_d^{(\lambda,s,\delta)}$ *with* $s = 1$*, conditioned to survive (*50 *iterations)*

5.A.2. Simulations for a simple variant algorithm of NTMA

We here present some simulations of simple variant algorithm of NTMA, NTMA-2, which happens to be more efficient in practice. The Algorithm NTMA-2 is as follows.

Algorithm 5.3: NTMA-2

Input: Two graphs G and H of size n, average degree λ, depth d, parameter γ.
Output: A set of pairs $\mathcal{S} \subset V(G) \times V(H)$.
$\mathcal{S} \leftarrow \varnothing$
for $(u, u') \in V(G) \times V(H)$ **do**
  if $\mathcal{B}_G(u, d)$ *and* $\mathcal{B}_H(u', d)$ *contain no cycle, and if, denoting*

$$\mathcal{W}_d(u, v') := \mathbb{1}_{\mathcal{B}_G(u,d) \text{ and } \mathcal{B}_H(v',d) \text{ contain no cycle}} \mathcal{W}_d(\mathcal{B}_G(u, d), \mathcal{B}_H(v', d)),$$

  one has $\mathcal{W}_d(u, u') > \gamma^d$, $\mathcal{W}_d(u, u') = \max_v \mathcal{W}_d(v, u')$ *and* $\mathcal{W}_d(u, u') = \max_{v'} \mathcal{W}_d(u, v')$ **then**
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{(u, v')\}$
  end
end
for $(u, u') \neq (v, v') \in \mathcal{S}$ **do**
  if $u = v$ **then**
    $\mathcal{S} \leftarrow \mathcal{S} \setminus \{(u, y), y \in V(H)\}$
  end
  if $u' = v'$ **then**
    $\mathcal{S} \leftarrow \mathcal{S} \setminus \{(x, u'), x \in V(G)\}$
  end
end
return $\mathcal{S}$

This algorithm only selects rows and columns weight maximums and match the corresponding pairs. The last part ensures that $\mathcal{S}$ is a matching. For these simulations, error bars correspond to a confidence interval for the mean value of scores. In Figures 5.7 and 5.8 we compare the scores of NTMA-2 for $s = 0.95$ with the isomorphism case $s = 1.0$, for different values of n. We illustrate the fact that nearly no vertex is mismatched, whereas a non-negligible fraction of nodes is indeed recovered. In Figure 5.9, we compare the scores of NTMA-2 for fixed n but varying s, illustrating the existence of a 'critical' parameter $s^*(\lambda)$.

5.B. Detailed proofs for Section 5.2

5.B.1. Proof of Theorem 5.2

Proof of Theorem 5.2. We first state an easy corollary:

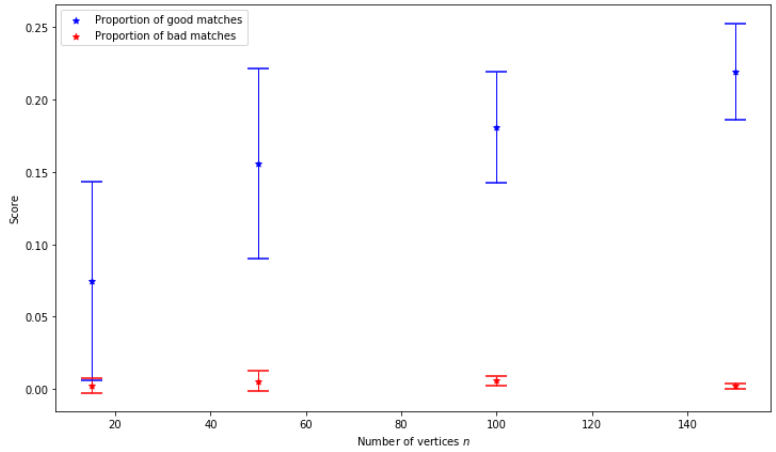
Corollary 5.B.1. *For any $d \geq 1$, the random variable $X = |\mathcal{L}_d(T_d)|$ is such that $\mathbb{E}\left[e^{\theta X}\right] < \infty$ for all $\theta > 0$.*

Proof. This is easily seen by induction, based on the structure of T_d given in Lemma 5.2.1. □

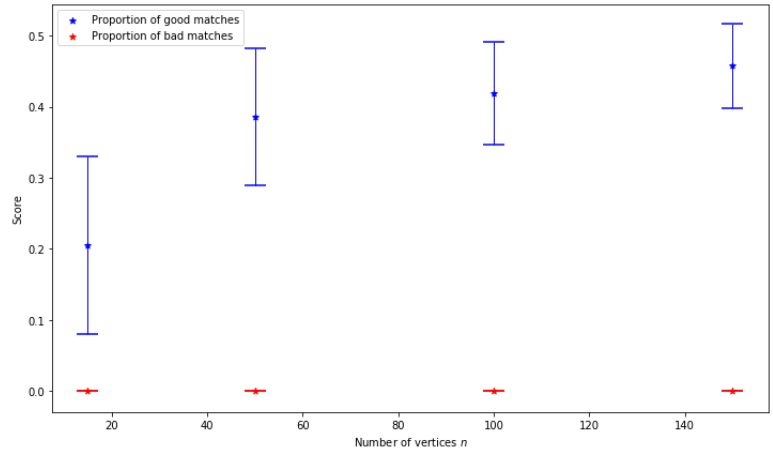
Recall that we let $\mathcal{E}_d$ (respectively, $\mathcal{E}'_d$) denote the event that tree T (respectively, T') becomes extinct before d generations, i.e. $\mathcal{L}_d(T) = \varnothing$ (respectively, $\mathcal{L}_d(T') = \varnothing$). We let $p_d = \mathbb{P}(\mathcal{E}_d)$. It is well known that it satisfies the recursion

$$p_0 = 0,\ p_d = e^{-\lambda(1-p_{d-1})},$$

and converges monotonically to the smallest root in $[0, 1]$ of $x = e^{-\lambda(1-x)}$. This root, that we



(a) – $s = 0.95$



(b) – *Isomorphism case,* $s = 1.0$

Figure 5.7 – *Mean score of* `NTMA-2` *for* $\lambda = 2.1$, $d = 5$ *(25 iterations per value of* n*)*

denote p_e, is the probability of ultimate extinction. For small enough $\varepsilon = \lambda - 1$, it holds that

$$p_e = 1 - 2\varepsilon + O(\varepsilon^2),$$

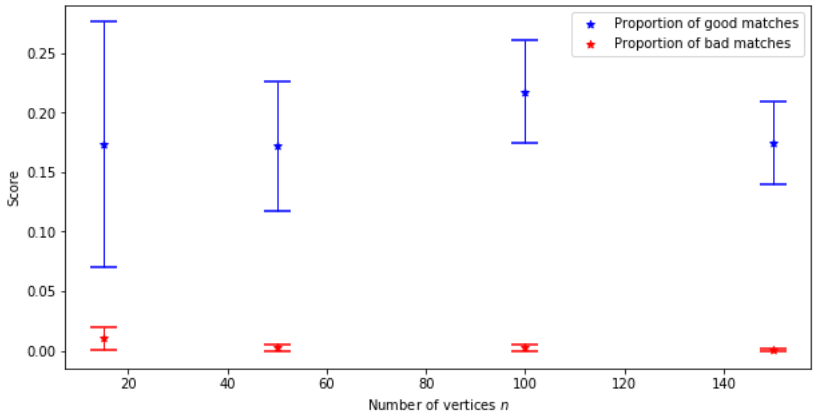
as can be seen by analysis of the fixed point equation satisfied by p_e. Let then d_0 be such that for all $d \geq d_0$, $p_d = 1 - 2\varepsilon + O(\varepsilon^2)$. Clearly, on the event $\mathcal{E}_d \cup \mathcal{E}'_d$, the set of matchings $M(T, T')$ is empty, so that $\mathcal{W}_d(T, T') = 0$. Recall that we define T_d the random variable $r_d(T)$ where T is conditioned to survive up to depth d.
Now fix $r \in (0, 1)$. We shall prove that for sufficiently small $\varepsilon > 0$, letting $\gamma = 1 + r\varepsilon$, there exists some constants $c, m, d_0 > 0$ such that for all $x > 0$, all $d \geq d_0$, one has

$$\mathbb{P}\left(\mathcal{W}_d(T_d, T'_d) \geq \gamma^{d-d_0} mx\right) \leq e^{-(x-c)_+}. \tag{5.16}$$

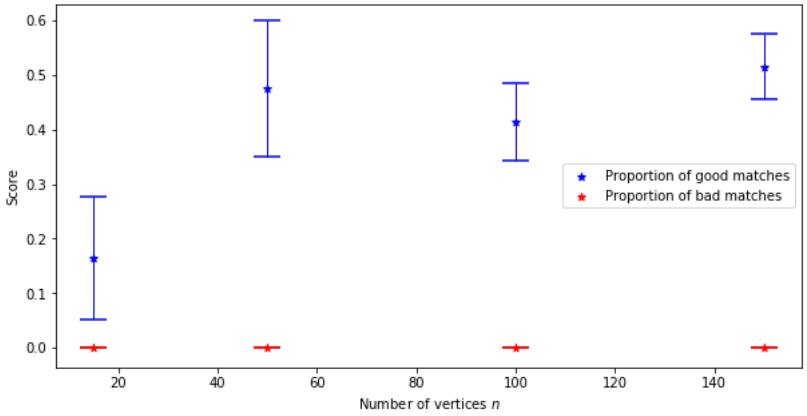
We proceed by induction over $d - d_0$. To initialize the induction, notice that one obviously has $\mathcal{W}_{d_0}(T_{d_0}, T'_{d_0}) \leq |\mathcal{L}_{d_0}(T_{d_0})| =: X$. By Corollary 5.B.1, for all $m, x, \theta > 0$, one has:

$$\mathbb{P}\left(\mathcal{W}_{d_0}(T_{d_0}, T'_{d_0}) > mx\right) \leq \mathbb{P}(X > mx) \leq \mathbb{E}e^{\theta X} e^{-\theta mx}.$$

Let now $\theta = 1/m$. By taking m sufficiently large, from dominated convergence we can make



(a) – $s = 0.95$



(b) – *Isomorphism case,* $s = 1.0$

Figure 5.8 – *Mean score of* **NTMA-2** *for* $\lambda = 3.1$, $d = 4$ *(25 iterations per value of* n*)*

$\mathbb{E}e^{(1/m)X}$ as close to 1 as we like. Choose for instance m such that $\mathbb{E}e^{(1/m)X} \leq 2$. Then

$$\mathbb{P}(\mathcal{W}_{d_0}(T_{d_0}, T'_{d_0}) > mx) \leq 2e^{-x} \leq e^{-x+c}.$$

for any $c \geq \ln(2)$. Hence, for sufficiently large m, we can initialize the induction at $d = d_0$ with any $c \geq \ln(2)$.

Recall we set $\gamma = 1 + r\varepsilon$. Define the random variables

$$X_d := \gamma^{-(d-d_0)} m^{-1} \mathcal{W}_d \left(T_d, T'_d\right).$$

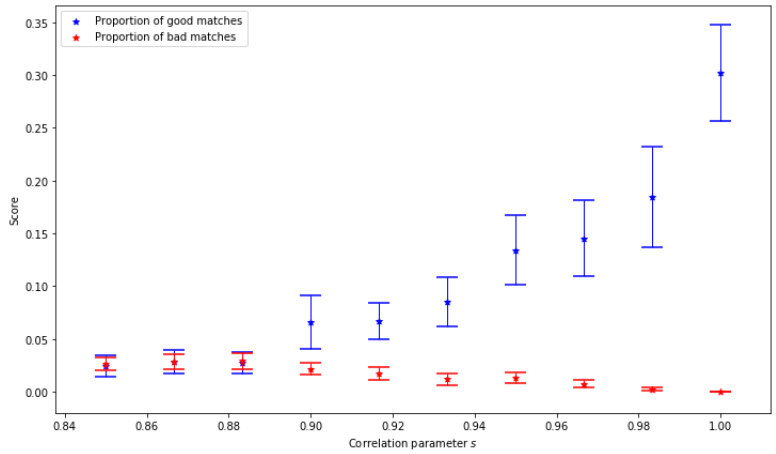
Let D (resp. D') denote the number of children of the root in T_d (resp. D' in T'_d). Given D and D', noting $T_d = (T_{d-1,1}, \ldots, T_{d-1,D})$ and $T'_d = \left(T'_{d-1,1}, \ldots, T'_{d-1,D'}\right)$, we have that

$$\mathcal{W}_d(T_d, T'_d) = \sup_{\mathfrak{m} \in \mathcal{M}([D],[D'])} \sum_{(i,u)\in\mathfrak{m}} \mathcal{W}_{d-1}(T_{d-1,i}, T'_{d-1,u}),$$

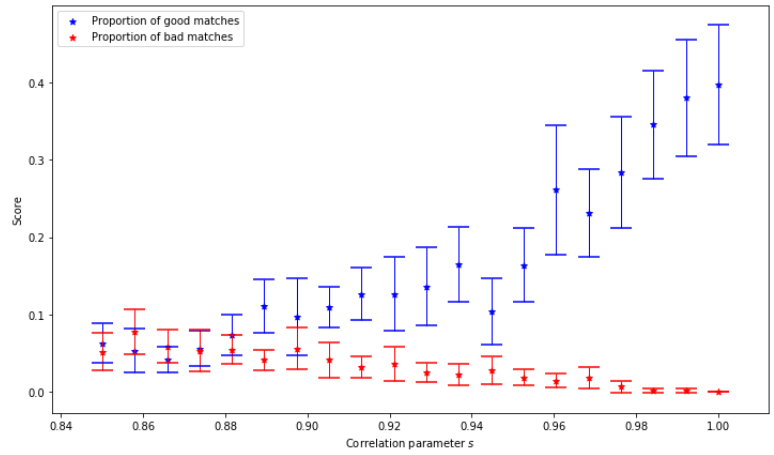
where $\mathcal{M}([D],[D'])$ denotes the set of all $(D \vee D')^{\underline{D \wedge D'}}$ maximal injective mappings between $\mathcal{E}_0 \subseteq [D]$ and $[D']$. Let

$$X_{d-1,i,u} := \gamma^{-(d-1-d_0)} m^{-1} \mathcal{W}_{d-1}(T_{d-1,i}, T'_{d-1,u}).$$

Note that conditional on D and D', for each matching $\mathfrak{m} \in \mathcal{M}([D],[D'])$, the variables



(a) – $s = 0.95$



(b) – *Isomorphism case,* $s = 1.0$

Figure 5.9 – *Mean score of* `NTMA-2` *with different values of* s *(25 iterations per value of* n*)*

$(X_{d-1,i,u})_{(i,u)\in\mathfrak{m}}$ are i.i.d. with the same distribution as X_{d-1}. The induction hypothesis states that each $X_{d-1,i,u}$ is less, for the strong stochastic ordering of comparison of cumulative distribution functions, than c plus an exponential random variable with parameter 1. With an easy union bound, we can derive the following bounds:

$$\mathbb{P}\left(X_d > x\right) \leq \sum_{1\leq k\leq \ell<\infty} \mathbb{P}\left(D \wedge D' = k, D \vee D' = \ell\right) \min\left(1, \ell^k \mathbb{P}\left(\mathcal{E}_1 + \ldots + \mathcal{E}_k > \gamma x - kc\right)\right), \tag{5.17}$$

where $\mathcal{E}_1, \ldots, \mathcal{E}_k$ are independent exponential random variables of parameter 1. Lemma 5.2.1 states that

$$\mathbb{P}(D = k) = e^{-\lambda(1-p_{d-1})} \frac{\lambda^k (1-p_{d-1})^k}{k!\,(1-p_d)} =: q_{d,k}.$$

We can increase d_0 such that for some constant $\kappa > 0$, for all $d \geq d_0$:

$$q_{d,1} \leq 1 - \varepsilon + \kappa\varepsilon^2, \quad q_{d,k} \leq \frac{(3\varepsilon)^{k-1}}{k!}, \ k \geq 2.$$

Note that for $x \leq c$, there is nothing to prove in (5.16), since a probability is always upper-bounded by 1. We thus only need to consider the case $x > c$. We conclude the proof of this

Theorem by appealing to the following technical Lemma, proved later on in Appendix 5.B.3:

Lemma 5.B.1. *Let $\kappa, C > 0$ and $r \in (0,1)$ be given constants. Then there exists $c > 0$ large enough and $\varepsilon_0 > 0$ such that, for all $\varepsilon \in (0, \varepsilon_0)$, letting $\gamma = 1 + r\varepsilon$, $q_1 = 1 - \varepsilon + \kappa\varepsilon^2$, $q_k = (C\varepsilon)^{k-1}/k!$ for $k \geq 2$, one has*

$$\forall x > c, \quad \sum_{k,\ell \geq 1} q_k q_l \min\left(1, (k \vee \ell)^{\frac{k \wedge \ell}{}} \mathbb{P}\left(\mathcal{E}_1 + \ldots + \mathcal{E}_{k\wedge\ell} > \gamma x - (k \wedge \ell)c\right)\right) \leq e^{-(x-c)}, \quad (5.18)$$

where the $\mathcal{E}_i$ are independent exponential random variables of parameter 1.

Its assumptions are indeed verified here with $C = 3$, so (5.16) can be propagated by using this Lemma in (5.17), and the conclusion of Theorem 5.2 follows.

□

5.B.2. Proof of Theorem 5.3

Proof of Theorem 5.3. We assume that $\lambda = 1 + \varepsilon$. We fix $r \in (0,1)$, and we let $\gamma = 1 + r\varepsilon$ for some fixed $r \in (0,1)$. We work with trees such that $(T, T') \sim \mathbb{P}_d^{(\lambda,s,\delta)}$. If we assume that the path from ρ to ρ' does not survive down to depth d in T, then this path is no more present in T_d, and the two trees T_d and T'_d can be coupled with two trees $\widetilde{T}_d$ and $\widetilde{T}'_d$ where $(\widetilde{T}, \widetilde{T}') \sim \mathbb{P}_d^{(\lambda)}$, and we are in the case of Theorem 5.2.

In the following proof, we will thus condition to the event $S_{\rho,d}$ that the path from ρ to ρ' survives down to depth d in T. Recall that the tree T_d (resp. T_d) is obtained, conditionally on the fact that T (resp. in T') survives down to depth d, by suppressing nodes at depth greater than d in T (resp. in T'), and then pruning alternatively leaves of depth strictly less than d. As in the proof of Theorem 5.2, we shall establish that for sufficiently small $\varepsilon > 0$, there exist constants $c, m, d_0 > 0$ such that for all $x > 0$, all $d \geq d_0$, one has

$$\mathbb{P}\left(\mathcal{W}_d(T_d, T'_d) \geq \gamma^{d-d_0} m x \middle| S_{\rho,d}\right) \leq e^{-(x-c)^+}. \quad (5.19)$$

Define the random variables

$$X'_d := \gamma^{-(d-d_0)} m^{-1} \mathcal{W}_d\left(T_{d+\delta}, T'_d\right),$$

conditional on $S_{\rho,d}$. The proof will again be by induction on d, the initial step being established with the same argument as in the proof of Theorem 5.2. Note that this argument does not depend on δ.

Denote by D the number of children of ρ in T_d, D' the number of children of ρ' in T'_d that are in the intersection tree $T_d \cap T'_d$, and D'' the number of children of ρ' in $T'_d \setminus T_d$. By branching property, note that these three variables are independent.

Recall that p_d denotes the probability that a Galton-Watson tree with offspring $\mathsf{Poi}(\lambda)$ becomes extinct before d generations. Then, conditionally on $S_{\rho,d}$, the random variables D, D' and D'' have the following distributions:

$$D \sim 1 + \mathsf{Poi}\left(\lambda\left(1 - p_{d-1}\right)\right), \quad D' \sim \mathsf{Poi}\left(\lambda s\left(1 - p_{d-1}\right)\right),$$
$$D'' \sim \mathsf{Poi}\left(\lambda(1-s)\left(1 - p_{d-1}\right)\right), \text{ conditionally on } D' + D'' > 0.$$

We show an illustration on Figure 5.10.

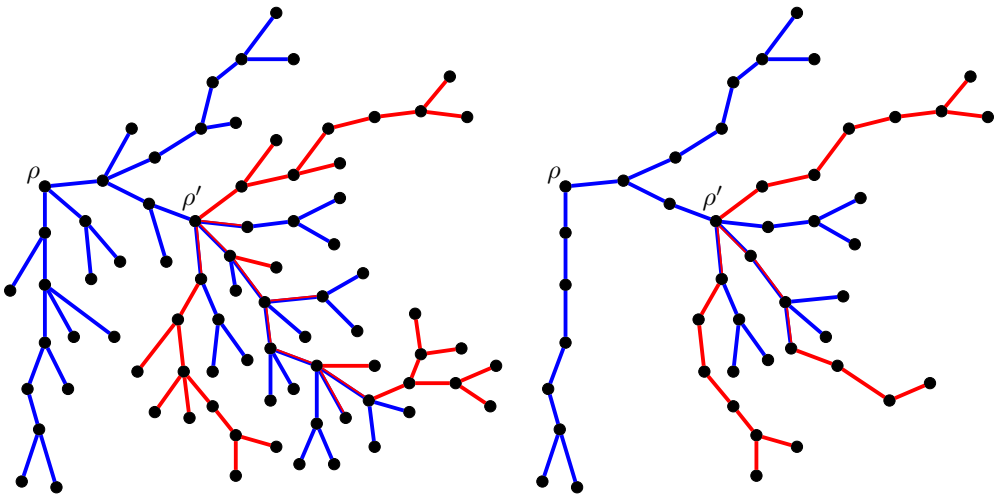


Figure 5.10 – *Random trees T (blue) and T' (red) from Figure 5.3 (left), and the results T_d and T'_d after applying r_d (right). In this example, $\delta = 3$ and $d = 6$, $D = 2$, $D' = 2$ and $D'' = 1$.*

We condition on the values ℓ, k', k'' taken by D, D', D''. The number of maximal one-to-one mappings between the children of ρ in T_d and those of ρ' in T'_d is given by $[(k' + k'') \vee \ell]^{\underline{(k'+k'')\wedge\ell}}$, and each of them is of size $\ell \wedge (k' + k'')$. Note here again that for a fixed matching between the children of ρ and ρ', the weights of the matched subtrees are independent. We distinguish between several cases (to help understand these cases, the reader could keep Figure 5.10 in mind):

- For a child u of ρ that is not on the path to ρ', the corresponding subtrees are independent so that the corresponding weight is distributed as $\mathcal{W}_{d-1}(\widetilde{T}_{d-1}, \widetilde{T}'_{d-1})$ in the independent model $\mathbb{P}_d^{(\lambda)}$.
- If the child of ρ on the path to ρ' is matched with a child of ρ' that is not in the intersection tree, again the corresponding weight is similarly distributed.
- Finally, if the child u of ρ leading to ρ' is matched to a child u' of ρ' in the intersection tree, setting the new root at $\widetilde{\rho} := u$ in T and at $\widetilde{\rho}' := u'$ in T', the corresponding weight has the same distribution as $\mathcal{W}_{d-1}(T_{d-1}, T'_{d-1})$ in the model $\mathbb{P}_{d-1}^{(\lambda,s,\delta)}$, still conditioned to $S_{\widetilde{\rho},d-1}$. Indeed, there is a path from $\widetilde{\rho}$ to $\widetilde{\rho}'$, and the corresponding Poisson distributions are conserved.

The induction hypothesis for case 3, together with Theorem 5.2 for cases 1 and 2, therefore give us:

$$\mathbb{P}\left(X'_d \geq x\right) \leq \sum_{k,\ell} \mathbb{P}\left(D' + D'' = k, D = \ell\right) \min\left(1, (k \vee \ell)^{\underline{k\wedge\ell}}\, \mathbb{P}\left(\mathcal{E}_1 + \ldots + \mathcal{E}_{k\wedge\ell} > \gamma x - (k \wedge \ell)c\right)\right),$$

where the $\mathcal{E}_i$ are independent exponential random variables of parameter 1. Assume, as in the proof of Theorem 5.2, that d_0 is chosen such that for all $d \geq d_0$,

$$p_d^{\lambda} = 1 - 2\varepsilon + O(\varepsilon^2).$$

With simple computations, we can then ensure that for some $\kappa > 0$, noting $q_{d,\cdot}$ the distribution of D, one has

$$q_{d,1} \leq 1 - \varepsilon + \kappa\varepsilon^2, \quad q_{d,k} \leq \frac{(3\varepsilon)^{k-1}}{(k-1)!} \leq \frac{(6\varepsilon)^{k-1}}{k!},\ k \geq 2,$$

where we used $k \le 2^{k-1}$ in the last step. By independence of D' and D'', $D' + D''$ follows a $\mathsf{Poi}(\lambda(1-p_{d-1}))$ distribution, conditional on being positive. Noting $q'_{d,\cdot}$ this distribution, we have, as in the previous proof,

$$q'_{d,1} \le 1 - \varepsilon + \kappa\varepsilon^2, \quad q'_{d,k} \le \frac{(3\varepsilon)^{k-1}}{k!},\ k \ge 2.$$

We can then invoke Lemma 5.B.1 to conclude. Note that every control in the proof is made uniformly on $\delta \ge 1$. □

5.B.3. Proof of Lemma 5.B.1

Proof of Lemma 5.B.1. . Let

$$S_1 := e^{x-c} q_1^2 e^{-(\gamma x - c)_+} + 4 q_1 q_2 e^{-(\gamma x - c)_+}, \quad S_2 := 2e^{x-c} q_1 \sum_{\ell \ge 3} q_\ell \min\left(1, \ell e^{-(\gamma x - c)_+}\right),$$
$$S_3 := 2e^{x-c} \sum_{2 \le k \le \ell} q_k q_\ell \min\left(1, \ell^{\underline{k}}\, \mathbb{P}\left(\mathcal{E}_1 + \ldots + \mathcal{E}_k > \gamma x - kc\right)\right).$$

Our goal is to show that for a suitable choice of c, for all $x > c$, $S_1 + S_2 + S_3 \le 1$. One has

$$S_1 \le e^{-r\varepsilon x}\left((1 - \varepsilon + \kappa\varepsilon^2)^2 + 2C\varepsilon\right) \le e^{-r\varepsilon x}(1 + 2C\varepsilon), \tag{5.20}$$

and

$$S_2 \le 2e^{-r\varepsilon x}(1 - \varepsilon + \kappa\varepsilon^2) \sum_{\ell \ge 3} \frac{(C\varepsilon)^{\ell-1}}{(\ell-1)!} \le 2e^{-r\varepsilon x}\left(e^{C\varepsilon} - 1 - (C\varepsilon)\right) \le 2e^{-r\varepsilon x} C^2 \varepsilon^2. \tag{5.21}$$

We let k_0 be such that $\gamma x \in [k_0 c, (k_0+1)c)$. We then upper-bound S_3 by $A + B$ where

$$A = 2e^{x-c} \sum_{k=2}^{k_0} q_k \sum_{\ell \ge k} q_\ell \frac{\ell!}{(\ell-k)!}\, \mathbb{P}\left(\mathcal{E}_1 + \ldots + \mathcal{E}_k > \gamma x - kc\right), \tag{5.22}$$

$$B = 2e^{x-c} \sum_{k \ge (k_0+1)\vee 2} \sum_{\ell \ge k} q_k q_\ell. \tag{5.23}$$

One readily has

$$B \le 2e^{-r\varepsilon x} e^{\gamma x - c} \sum_{k \ge (k_0+1)\vee 2} \frac{(C\varepsilon)^{k-1}}{k!} \sum_{\ell \ge k} \frac{(C\varepsilon)^{\ell-1}}{\ell!} \tag{5.24}$$

$$\le 2e^{-r\varepsilon x} e^{\gamma x - c} \sum_{k \ge (k_0+1)\vee 2} \frac{(C\varepsilon)^{2(k-1)}}{k!} \tag{5.25}$$

$$\le 2e^{-r\varepsilon x} e^{k_0 c} (C\varepsilon)^{2((k_0+1)\vee 2 - 1)} \tag{5.26}$$

$$\le 2e^{-r\varepsilon x} \left(C\varepsilon e^c\right)^2, \tag{5.27}$$

where in the last steps we assumed that $C\varepsilon e^c < 1$, so that

$$e^{k_0 c}(C\varepsilon)^{2((k_0+1)\vee 2 - 1)} \le (C\varepsilon e^c)^{2((k_0+1)\vee 2 - 1)} \le (C\varepsilon e^c)^2.$$

Note that for $y \geq 0$, $\mathbb{P}(\mathcal{E}_1 + \ldots + \mathcal{E}_k > y) = \mathbb{P}(\mathsf{Poi}(y) < k) = e^{-y} \sum_{j=0}^{k-1} y^j/j!$. Write then

$$\begin{aligned} A &\leq 2e^{x-c} \sum_{k=2}^{k_0} \frac{(C\varepsilon)^{2(k-1)}}{k!} \sum_{j=0}^{k-1} e^{-\gamma x + kc} \frac{(\gamma x)^j}{j!} \\ &\leq 2e^{-r\varepsilon x} \sum_{k=2}^{k_0} \frac{(C^2 e^c)^k}{k!} \sum_{j=0}^{k-1} \frac{\left(\gamma x \varepsilon^2\right)^j}{j!} \varepsilon^{2(k-1-j)} \\ &\leq 2e^{-r\varepsilon x} \sum_{k=2}^{k_0} \frac{(C^2 e^c)^k}{k!} \left[\varepsilon^{2(k-1)} + e^{\gamma x \varepsilon^2} - 1\right] \\ &\leq 2e^{-r\varepsilon x} \left[\varepsilon^{-2} \left(e^{\varepsilon^2 C^2 e^c} - 1 - \varepsilon^2 C^2 e^c\right) + e^{C^2 e^c} \left(e^{\gamma x \varepsilon^2} - 1\right)\right] \end{aligned} \tag{5.28}$$

Summing the upper bounds (5.20)-(5.28), the desired property will then hold if for all $x > c$, one has:

$$e^{-r\varepsilon x} \left[1 + 2C\varepsilon + 2C^2\varepsilon^2 + 2[C\varepsilon e^c]^2 + 2\varepsilon^{-2} \left(e^{\varepsilon^2 C^2 e^c} - 1 - \varepsilon^2 C^2 e^c\right)\right] + 2e^{-r\varepsilon x} e^{C^2 e^c} \left(e^{\gamma x \varepsilon^2} - 1\right) \leq 1. \tag{5.29}$$

The first term is, for any fixed c, and for sufficiently small ε, upper bounded by

$$e^{-r\varepsilon x} \left(1 + (2C+1)\varepsilon\right).$$

We now distinguish three cases for x.

Case 1: $x \in [c, 1/\sqrt{\varepsilon}]$. The second term is then $O(\varepsilon\sqrt{\varepsilon})$. Provided $rc > 2C + 1$, since $e^{-r\varepsilon x} \leq e^{-r\varepsilon c} = 1 - r\varepsilon c + O(\varepsilon^2)$, the left-hand side of (5.29) is then upper-bounded by $1 - (rc - 2C - 1)\varepsilon + O(\varepsilon\sqrt{\varepsilon})$, and is thus less than 1.

Case 2: $x \in [1/\sqrt{\varepsilon}, 1/\varepsilon]$. Since $e^{-r\varepsilon x} \leq e^{-r\sqrt{\varepsilon}} = 1 - \Omega(\sqrt{\varepsilon})$, and $e^{\gamma x \varepsilon^2} - 1 \leq e^{\gamma\varepsilon} - 1 = O(\varepsilon)$, the left-hand side of (5.29) is upper-bounded by $1 - \Omega(\sqrt{\varepsilon})$ and is thus less than 1.

Case 3: $x \geq 1/\varepsilon$. The first term is then bounded by $e^{-r}(1 + (2C+1)\varepsilon)$, which is less than $1 - \Omega(1)$ for ε small enough. Letting $y = \varepsilon x$, the second term reads

$$e^{-ry}[e^{\varepsilon\gamma y} - 1](2e^{C^2 e^c}).$$

For small ε, this function is maximized for $y = 1/r + O(\varepsilon)$, at which point it evaluates to $O(\varepsilon)$. Thus the left-hand side of (5.29) is upper-bounded by $1 - \Omega(1)$ in that range.

We have thus shown that for any $r > 0$, provided $c > (2C+1)/r$, then for all sufficiently small ε, the desired property holds with $\gamma = 1 + r\varepsilon$. □

5.C. Detailed proofs for Section 5.3

The following proofs are adapted from the previous work of [Mas14] and [BLM15].

5.C.1. Proof of Lemma 5.3.1

Proof of Lemma 5.3.1. Fix $K > 0$ to be specified later and $\gamma > 0$. Fix $u \in [n]$, and define

$$T := \inf\left\{t \leq d, |\mathcal{S}_G(u,t)| \geq K \log n\right\}.$$

If $T = \infty$, there is nothing to prove. Given $|\mathcal{S}_G(u, T-1)|$,

$$|\mathcal{S}_G(u,T)| \sim \mathsf{Bin}(n - |\mathcal{S}_G(u,0)| - \ldots - |\mathcal{S}_G(u,T-1)|\,, 1-(1-\lambda/n)^{|\mathcal{S}_G(u,T-1)|}).$$

Thus

$$|\mathcal{S}_G(u,T)| \overset{\text{sto.}}{\leq} \mathsf{Bin}\left(n, \lambda K(\log n)/n\right).$$

Using Bennett's inequality, for $K' > \lambda K$:

$$\mathbb{P}\left(|\mathcal{S}_G(u,T)| \geq K' \log n\right) \leq \exp\left[-\lambda K h\left(\frac{K'-\lambda K}{\lambda K}\right) \log n\right],$$

with $h(x) = (1+x)\log(1+x) - x$. This probability is $\leq n^{-2-\gamma}$ if K' is large enough to verify $\lambda K h\left(\frac{K'-\lambda K}{\lambda K}\right) > \gamma + 2$. With a simple use of the union bound, one gets that $|\mathcal{S}_G(u,T)| \in [K \log n, K' \log n]$ for all $u \in [n]$ with probability $1 - O(n^{-1-\gamma})$.

Fix $\varepsilon > 0$ to be specified later. We then check by induction that with high probability, for all $T \leq t \leq d$,

$$|\mathcal{S}_G(u,t)| \in \left[K(\lambda/2)^{t-T}(\log n) \prod_{s=T}^{t}\left(1-\varepsilon\,(\lambda/2)^{-(s-T)/2}\right), K'\lambda^{t-T}(\log n)\prod_{s=T}^{t}\left(1+\varepsilon\lambda^{-(s-T)/2}\right)\right]. \tag{5.30}$$

The case $t = T$ is proved here above. We will next use the inequality

$$\lambda x/(2n) \leq \lambda x/n - \lambda^2 x^2/(2n^2) \leq 1-(1-\lambda/n)^x \leq \lambda x/n. \tag{5.31}$$

that holds as soon as $\lambda x/n < 1$.
Assuming (5.30) holds up to t, inequality (5.31) holds for $x = |\mathcal{S}_G(u,t)|$ for n large enough, since $|\mathcal{S}_G(u,t)| < n/\lambda$ for $c \log \lambda < 1$. Thus for n large enough $\mathbb{E}\,|\mathcal{S}_G(u,t+1)|$ lies in the interval

$$\left[K\,(\lambda/2)^{t-T}(\log n) \underbrace{\prod_{s=T}^{t}\left(1-\varepsilon\,(\lambda/2)^{-(s-T)/2}\right)}_{=1-O(\varepsilon)}, \lambda K'\lambda^{t-T}(\log n)\prod_{s=T}^{t}\left(1+\varepsilon\lambda^{-(s-T)/2}\right)\right],$$

with $\hat{\varepsilon} > 0$ to be specified later, Bennett's inequality writes

$$\mathbb{P}\left(\left|\,|\mathcal{S}_G(u,t+1)| - \mathbb{E}\,|\mathcal{S}_G(u,t+1)|\,\right| \geq \hat{\varepsilon}\mathbb{E}\,|\mathcal{S}_G(u,t+1)|\right) \leq 2\exp\left[-(\lambda/2)^{t-T+1}\log n\,(1-O(\varepsilon))\,h(\hat{\varepsilon})\right],$$

which is $\leq n^{-2-\gamma}$ if $K\,(\lambda/2)^{t+1-T}\,h(\hat{\varepsilon}) > 2+\gamma$. Since for $u \to 0$, $h(u) = u^2/2 + o(u^2)$, it suffices to take $\hat{\varepsilon} = \varepsilon\,(\lambda/2)^{-(t+1-T)/2}$ with ε small enough and K large enough such that $K\varepsilon > 2+\gamma$. Thus (5.30) holds for $t+1$ with probability $1 - O(n^{-2-\gamma})$.

All this ensures that the desired inequality (5.11) holds for all $u \in [n]$, $t \in [d]$ with probability $1 - O(n^{-\gamma})$. □

5.C.2. Proof of Lemma 5.3.2

Proof of Lemma 5.3.2. Fix $u \in [n]$. Define

$$k^* := \inf\{t \leq d,\ \mathcal{B}_G(u,t) \text{ contains a cycle}\}.$$

Note that $k^* \geq 2$, and that if $k^* = \infty$ then $\mathcal{B}_G(u,d)$ does not contain any cycle. Now assume that $k^* < \infty$. For any $k \geq 2$, $k^* = k$ if and only if there are two vertices of $\mathcal{S}_G(u,k-1)$ that are connected, or if there is a vertex of $\mathcal{S}_G(u,k)$ connected to two vertices of $\mathcal{S}_G(u,k-1)$. On the event

$$\mathcal{A} := \bigcap_{t \leq d} \left\{ |\mathcal{S}_G(u,t)| < C(\log n)\lambda^t \right\},$$

this happens with probability at most

$$|\mathcal{S}_G(u,k-1)|^2 \times \frac{\lambda}{n} + |\mathcal{S}_G(u,k)| \times |\mathcal{S}_G(u,k-1)|^2 \times \frac{\lambda^2}{n^2} \leq C^2 \frac{(\log n)^2 \lambda^{2k}}{n} + C^3 \frac{(\log n)^3 \lambda^{3k}}{n^2}.$$

Taking $\varepsilon > 0$ such that $c \log \lambda \leq 1/2 - \varepsilon$, choosing C such that $\mathbb{P}(\mathcal{A}) = 1 - O\left(n^{-2\varepsilon}\right)$ with Lemma 5.3.1, the probability that $\mathcal{B}_G(u,d)$ contains a cycle is less than

$$\begin{aligned}
\mathbb{P}(k^* < \infty) &\leq \mathbb{P}\left(\bar{\mathcal{A}}\right) + \sum_{k=2}^{d} \mathbb{P}(k^* = k \,|\, \mathcal{A}) \\
&\leq O\left(n^{-2\varepsilon}\right) + O\left(\frac{(\log n)^2 \lambda^{2d}}{n}\right) + O\left(\frac{(\log n)^3 \lambda^{3d}}{n^2}\right) \\
&\leq O\left(n^{-2\varepsilon}\right) + O\left((\log n)^2 n^{-2\varepsilon}\right) + O\left((\log n)^3 n^{-3\varepsilon}\right) \leq O(n^{-\varepsilon}).
\end{aligned}$$

□

5.C.3. Proof of Lemma 5.3.3

Proof of Lemma 5.3.3. For fixed $u \neq v \in [n]$, let $\left(\tilde{\mathcal{S}}(u,t)\right)_{t \leq d}$ and $\left(\tilde{\mathcal{S}}(v,t)\right)_{t \leq d}$ denote two independent realizations of the neighborhoods (i.e. with independent underlying Bernoulli variables). We then construct recursively a coupling $(\mathcal{S}(u,t), \mathcal{S}(v,t))_{t \leq k}$:

- For $k = 1$, take $\mathcal{S}(u,t)$ to be a set of vertices uniformly chosen among sets of $[n]$ of size $\left|\tilde{\mathcal{S}}(u,0)\right|$. Independently, take $\mathcal{S}(v,t)$ to be a set of vertices uniformly chosen among sets of $[n]$ of size $\left|\tilde{\mathcal{S}}(v,0)\right|$.

- Now if $k > 1$, construct $\mathcal{S}(u,k)$ as follows: select a subset of $[n] \setminus \left(\bigcup_{s \leq k-1} \mathcal{S}(u,s) \right)$ of size $\left|\tilde{\mathcal{S}}(u,k)\right|$ uniformly at random. Then we construct independently $\mathcal{S}(v,k)$ taking a uniform subset of $[n] \setminus \left(\bigcup_{s \leq k-1} \mathcal{S}(v,s) \right)$ of size $\left|\tilde{\mathcal{S}}(v,k)\right|$.

This coupling is well defined, and coincides with the independent setting up to step k as long as the sets $\bigcup_{s \leq k} \mathcal{S}(u,s)$ and $\bigcup_{s \leq k} \mathcal{S}(v,s)$ do not intersect. On the event

$$\mathcal{A} := \bigcap_{t \leq d} \left\{ |\mathcal{S}(u,t)|, |\mathcal{S}(v,t)| < C(\log n)\lambda^t \right\},$$

one has

$$\mathbb{E}\left[\left| \bigcup_{k \leq d} \mathcal{S}(u,s) \cap \bigcup_{k \leq d} \mathcal{S}(v,s) \right| \right] \leq \mathbb{E}\left[\sum_{k=1}^{d} \mathsf{Bin}\left(C(\log n)\lambda^k, \frac{\sum_{t=1}^{k} C(\log n)\lambda^t}{n - \sum_{t=1}^{k} C(\log n)\lambda^t} \right) \right]$$

$$\leq C^2(\log n)^2 \left(\frac{\lambda}{\lambda-1}\right) \sum_{k=1}^{d} \frac{\lambda^{2k}}{n - \frac{\lambda}{\lambda-1} C(\log n)\lambda^k}$$
$$\leq O\left((\log n)^2 \lambda^{2d}/n\right)$$

if $(\log n)\lambda^d = o(n)$, which is the case if $c \log \lambda < 1$. The expectation is upper-bounded by $O\left((\log n)^2\lambda^{2d}/n\right) = O\left((\log n)^2 n^{-2\varepsilon}\right)$ if $c \log \lambda \leq 1/2 - \varepsilon$.

With Lemma 5.3.1, choosing C such that $\mathbb{P}(\mathcal{A}) = 1 - O\left(n^{-2\varepsilon}\right)$, we get

$$\mathrm{d_{TV}}\left(\mathcal{L}\left((\mathcal{S}_G(u,t), \mathcal{S}_G(v,t))_{t\leq d}\right), \mathcal{L}\left((\mathcal{S}_G(u,t))_{t\leq d}\right) \otimes \mathcal{L}\left((\mathcal{S}_G(v,t))_{t\leq d}\right)\right)$$
$$\leq O((\log n)^2 n^{-2\varepsilon}) + \mathbb{P}\left(\bar{\mathcal{A}}\right) \leq O(n^{-\varepsilon}).$$

□

5.C.4. Proof of Lemma 5.3.4

Proof. We work here conditionally on

$$\mathcal{A} := \bigcap_{t\leq d} \left\{ |\mathcal{S}_G(u,t)| < C(\log n)\lambda^t \right\}.$$

Let's define a Galton-Watson process as follows: set $Z_0 = 1$, and for $t > 0$, $\mathcal{L}(Z_t|\mathcal{G}_{t-1}) = \mathsf{Poi}(\lambda Z_{t-1})$, where $\mathcal{G}_t = \sigma(Z_s, s \leq t)$. Fix $t > 0$. Conditionally on $\mathcal{F}_{t-1} := \sigma(|\mathcal{S}_G(u,s)|, s \leq t-1)$, define a random variable W_t with distribution $\mathsf{Poi}(\lambda |\mathcal{S}_G(u,t-1)|)$. Note that

$$\mathcal{L}\left(|\mathcal{S}_G(u,t)| \,\middle|\, \mathcal{F}_{u-1}\right) = \mathsf{Bin}(n - |\mathcal{S}_G(u,0)| - \ldots - |\mathcal{S}_G(u,t-1)|,\ 1 - (1-\lambda/n)^{|\mathcal{S}_G(u,t-1)|}).$$

The Stein-Chen method (see e.g. [BC05]) enables to bound $\mathrm{d_{TV}}(\mathsf{Bin}(n,\lambda/n), \mathsf{Poi}(\lambda))$ by $\min(1,\lambda^{-1})\lambda^2/n \leq \lambda/n$. We also use the classical bound $\mathrm{d_{TV}}(\mathsf{Poi}(\lambda), \mathsf{Poi}(\lambda')) \leq |\lambda - \lambda'|$ together with inequality (5.31) (which holds for n large enough since $c \log \lambda < 1$) to obtain that conditionally on $\mathcal{F}_{t-1}$:

$$\mathrm{d_{TV}}(|\mathcal{S}_G(u,t)|, W_t) \leq n^{-1}(n - |\mathcal{S}_G(u,0)| - \ldots - |\mathcal{S}_G(u,t-1)|) \frac{\lambda |\mathcal{S}_G(u,t-1)|}{n}$$
$$+ \left|(n - |\mathcal{S}_G(u,0)| - \ldots - |\mathcal{S}_G(u,t-1)|)\left(1 - (1-\lambda/n)^{|\mathcal{S}_G(u,t-1)|}\right) - \lambda |\mathcal{S}_G(u,t-1)|\right|$$
$$\leq \frac{\lambda |\mathcal{S}_G(u,t-1)|}{n} + \lambda |\mathcal{S}_G(u,t-1)| - (n - |\mathcal{S}_G(u,0)| - \ldots - |\mathcal{S}_G(u,t-1)|) \frac{\lambda |\mathcal{S}_G(u,t-1)|}{n}$$
$$+ \frac{\lambda^2 |\mathcal{S}_G(u,t-1)|^2}{2n}.$$

Now, for $\varepsilon > 0$ such that $c \log \lambda \leq 1/2 - \varepsilon$, on the event $\mathcal{A}$, all variables $|\mathcal{S}_G(u,s)|$ are bounded by $C(\log n)n^{1/2-\varepsilon}$. This leads to

$$\mathrm{d_{TV}}(|\mathcal{S}_G(u,t)|, W_t) \leq O\left((\log n)n^{-1/2-\varepsilon}\right) + O\left((\log n)^3 n^{-2\varepsilon}\right) + O\left((\log n)^2 n^{-2\varepsilon}\right)$$
$$= O\left((\log n)^3 n^{-2\varepsilon}\right).$$

This proves by induction that the total variation distance between $(|\mathcal{S}_G(u,t)|)_{t\leq d}$ and $(Z_t)_{t\leq d}$ is bounded by $O\left((\log n)^4 n^{-2\varepsilon}\right) = O\left(n^{-\varepsilon}\right)$, taking C large enough in Lemma 5.3.1 so that $\mathbb{P}(\mathcal{A}) \geq 1 - O\left(n^{-2\varepsilon}\right)$. □

Chapter 6

Detecting correlation in trees

Following the way paved in Chapter 5, motivated by alignment of correlated sparse random graphs, we are now studying more in detail the hypothesis testing problem of deciding whether or not two random trees are correlated. We obtain conditions under which this task is impossible or feasible.

We propose `MPAlign`, a message-passing algorithm for graph alignment inspired by the tree correlation detection problem. We prove `MPAlign` to succeed in polynomial time at partial alignment whenever tree detection is feasible. As a result, our analysis of correlation detection in trees reveals new ranges of parameters for which partial alignment of sparse random graphs is feasible in polynomial time.

We then conjecture that graph alignment is not feasible in polynomial time when the associated tree detection problem is impossible. If true, this conjecture together with our sufficient conditions on tree detection impossibility would imply the existence of a hard phase for graph alignment, i.e. a parameter range where alignment cannot be performed in polynomial time even though it is known to be feasible in non-polynomial time.

This chapter is based on the paper *Correlation detection in trees for partial graph alignment* [GML21a] (submitted), a joint work with M. Lelarge and L. Massoulié. A short version of this work, *Correlation Detection in Trees for Planted Graph Alignment*, [GML22] is published at *ITCS 2021*.

6.1. Introduction

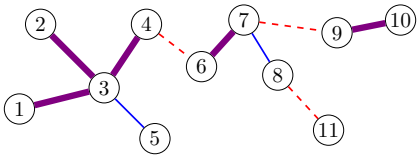
We refer to Section 1.3 for a presentation of the graph alignment, so as not to repeat ourselves.

As done in Chapter 5, we do not recall here the definition of the correlated Erdős-Rényi model, already introduced in the introduction (see (1.10)), and specified in Section 4.1.1 of Chapter 4 in the sparse case. We only recall that the parameters of $\mathsf{G}(n, \lambda/n, s)$ are the number of nodes n, the mean degree $\lambda > 0$ and the correlation parameter $s \in [0, 1]$. The vertices of the second graph G' are relabeled with a uniform independent permutation $\pi^\star \in \mathcal{S}_n$, and we observe G and $H := G'^{\pi^\star}$.

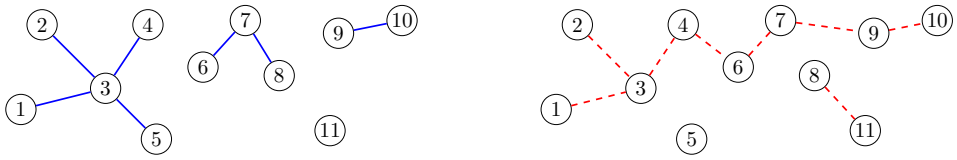
The previous model is used to study planted graph alignment – the mean-case version of graph alignment – consisting in finding an estimator $\hat{\pi}$ of the planted solution $\pi^\star$ upon observing G and H. As stated earlier, for any subset $\mathcal{C} \subset [n]$, the performance of any one-to-one estimator $\hat{\pi} : \mathcal{C} \to [n]$ is now assessed through $\mathrm{ov}(\pi^\star, \hat{\pi})$, its *overlap* with the unknown permutation $\pi^\star$, defined as

$$\mathrm{ov}(\pi^\star, \hat{\pi}) := \frac{1}{n} \sum_{u \in \mathcal{C}} \mathbb{1}_{\hat{\pi}(u) = \pi^\star(u)}. \tag{6.1}$$

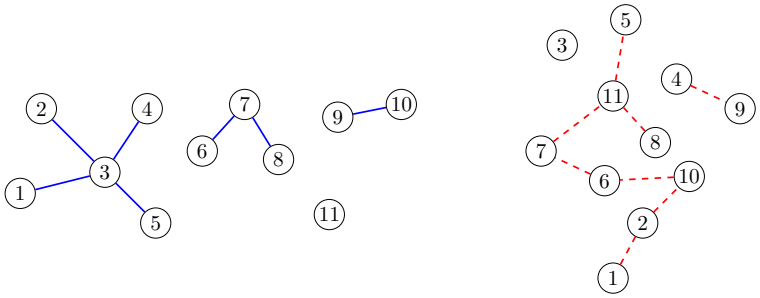
Note that the estimator $\hat{\pi}$ may not be in $\mathcal{S}_n$, and only consists in a partial matching. The



(a) – *Union graph* (G, G')



(b) – *Graphs* G, G' *in separate views*



(c) – *Graphs* G, H

Figure 6.1 – *A sample from model* $\mathsf{G}(n, \lambda/n, s)$ *with* $n = 11$, $\lambda = 1.9$, $s = 0.7$ *(for the sake of readability, the two-colored edges are drawn thick and purple).*

error fraction of $\hat{\pi}$ with the unknown permutation $\pi^\star$ is defined as

$$\mathrm{err}(\pi^\star, \hat{\pi}) := \frac{1}{n} \sum_{u \in \mathcal{C}} \mathbb{1}_{\hat{\pi}(u) \neq \pi^\star(u)} = \frac{|\mathcal{C}|}{n} - \mathrm{ov}(\pi^\star, \hat{\pi}). \tag{6.2}$$

We recall that sequence of injective estimators $\{\hat{\pi}_n\}_n$ – omitting the dependence in n – is said to achieve

- *Exact recovery* if $\mathbb{P}(\hat{\pi} = \pi^\star) \underset{n\to\infty}{\longrightarrow} 1$,
- *Almost exact recovery* if $\mathbb{P}(\mathrm{ov}(\pi^\star, \hat{\pi}) = 1 - o(1)) \underset{n\to\infty}{\longrightarrow} 1$,
- *Partial recovery* if there exists some $\varepsilon > 0$ such that $\mathbb{P}(\mathrm{ov}(\pi^\star, \hat{\pi}) > \varepsilon) \underset{n\to\infty}{\longrightarrow} 1$,
- *One-sided partial recovery* if it achieves partial recovery and $\mathbb{P}(\mathrm{err}(\pi^\star, \hat{\pi}) = o(1)) \underset{n\to\infty}{\longrightarrow} 1$.

Remark 6.1.1. *One-sided partial recovery is by definition at least as hard as partial recovery. As already stated in the introduction, from an application standpoint it is more appealing than partial recovery: indeed, it may be of little use to know one has a permutation with 30% of correctly matched nodes if one does not have a clue about which pairs are correctly matched. Our proposed algorithm will achieve one-sided partial recovery under suitable conditions.*

Phase diagram In the studied sparse regime where the graphs have constant mean degree λ, it is known [CK17, CKMP18, GML21b] that the presence of $\Omega(n)$ isolated vertices in the underlying intersection graph of G and H makes exact and almost exact recovery impossible. The main questions consist then in determining the phase diagram of the model $\mathsf{G}(n, \lambda/n, s)$

for partial alignment (or recovery), which we here recall the definition. We are interested in the range of parameters (λ, s) for which, in the large n limit:

- Any sequence of estimators fails to achieve partial recovery for any $\varepsilon > 0$. We refer to the corresponding range as the *impossible phase*;
- There is a sequence of estimators $\hat{\pi}$ achieving partial recovery (not necessarily one-sided) with some $\varepsilon > 0$, which we refer to as the *IT-feasible phase*;
- There is a sequence of estimators $\hat{\pi}$ *that can be computed in polynomial-time* achieving partial recovery with some $\varepsilon > 0$ (and sometimes even more, achieving also one-sided partial recovery): the *easy phase*.

An interesting perspective on this problem is provided by research on community detection, or graph clustering, for random graphs drawn according to the stochastic block model. In that setup, above the so-called Kesten-Stigum threshold, polynomial-time algorithms for clustering are known [BLM18, KMM$^+$13, MNS16], and the consensus among researchers in the field is that no polynomial-time algorithms exist below that threshold. Yet, there is a range of parameters with non-empty interior below the Kesten-Stigum threshold for which exponential-time algorithms are known to succeed at clustering [BMNN16]. In other words, for graph clustering, it is believed that there is a non-empty *hard phase*, consisting of the set difference between the IT-feasible phase and the polynomial-time feasible phase.

The picture available to date[1] for partial graph alignment is as follows. Work presented in Chapter 4 [GML21b] shows that the IT-impossible phase includes the range of parameters $\{(\lambda, s) : \lambda s \leq 1\}$, and Wu et al. [WXY21] have established that the IT-feasible phase includes the range of parameters $\{(\lambda, s) : \lambda s > 4\}$ (condition $\lambda s > C$ for some large C had previously been established in [HM20]). For the easy phase, we established in Chapter 5 [GM20] that it includes the range of parameters $\{(\lambda, s) : \lambda \in [1, \lambda_0], s \in [s(\lambda), 1]\}$ for some parameter $\lambda_0 > 1$ and some function $s(\lambda) : (1, \lambda_0] \to [0, 1]$. The `NTMA` algorithm proposed in Chapter 5 based on tree matching weights achieves in this regime one-sided partial recovery. Figure 6.2 depicts a phase diagram describing these prior results together with the new results of this chapter.

Problem description and main contributions This partial picture leaves open the question of whether, similarly to the case of graph clustering, graph alignment features a hard phase or not. The contribution of the present work can be summarized in three points:

(1) We investigate a fundamental statistical problem, which to the best of our knowledge had not been previously studied: hypothesis testing for correlation detection in trees. We study the regimes in which the optimal test on trees succeeds or fails in the setting when the trees are correlated Galton-Watson trees (see Theorem 6.1);

(2) For this detection problem on trees, the computation of the likelihood ratio can be made recursively on the depth, which yields an optimal message-passing algorithm for this task running in polynomial-time in the number of nodes;

(3) We remark that the previous detection problem on trees arises naturally from a local point of view in the related problem of one-sided partial recovery for graph alignment. In light of the previous analysis we then draw conclusions for our initial problem on graphs and doing so we precise the phase diagram shown in Figure 6.2, extending the regime for which one-sided partial alignment is provably feasible in polynomial time, and exhibiting the presence of a conjectured hard phase (see Theorem 6.2).

[1] at the time of this contribution.
[2] at the time of this contribution.

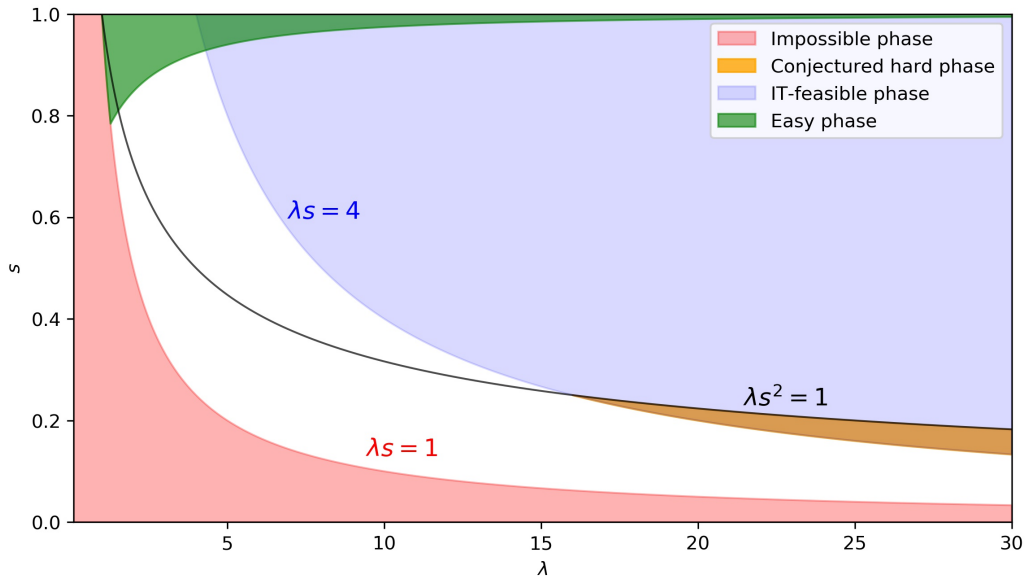


Figure 6.2 – *Diagram of the (λ, s) regions where partial recovery is known*[2] *to be IT-impossible ([GML21b]), IT-feasible ([WXY21]), or easy ([GM20] and this chapter). In the orange region, though partial graph alignment is IT-feasible, one-sided detectability is impossible in the tree correlation detection problem, and partial graph alignment is conjectured to be hard (this chapter).*

Our approach to point (3) follows the way paved in Chapter 5. It essentially relies on an algorithm which lets $\hat{\pi}(u) = u'$ for u such that the local structure of graph G in the neighborhood of node u is 'close' to the local structure of graph H in the neighborhood of node u'. As exploited in Chapter 5, the neighborhoods to distance d of two nodes u, u' in G and H, provided that $u' = \pi^\star(u)$, are asymptotically distributed as correlated Galton-Watson branching trees (denoted $\mathbb{P}_d^{(\lambda,s)}$ hereafter). On the other hand, for pairs of nodes (u, u') taken at random in $[n]$, the joint neighborhoods of nodes u and u' in G and H respectively, up to depth d, are asymptotically distributed as a pair of independent Galton-Watson branching trees (distribution denoted $\mathbb{P}_d^{(\lambda)}$).

Thus a fundamental step in our approach is to determine the efficiency of tests for deciding whether a pair of branching trees is drawn from either a product distribution, or a correlated distribution. [GM20] relied on tests based on a so-called *tree matching weight* to measure the similarity between two trees. In the present work we are instead interested in studying the existence of *one-sided tests*, which are tests asymptotically guarantying a vanishing type I error and a non vanishing power. According to the Neyman-Pearson Lemma, optimal one-sided tests are based on the likelihood ratio L_d of the distributions under the distinct hypotheses $\mathbb{P}_d^{(\lambda,s)}$ and $\mathbb{P}_d^{(\lambda)}$ (trees correlated or not)[3]. The mathematical formalization of point (1) here above is the following

Theorem 6.1 (Correlation detection in trees)**.** *Let*

$$\mathrm{KL}_d := \mathrm{KL}(\mathbb{P}_d^{(\lambda,s)} \| \mathbb{P}_d^{(\lambda)}) = \mathbb{E}_{1,d}\left[\log(L_d)\right].$$

Then the following propositions are equivalent:

(*i*) *There exists a one-sided test for deciding $\mathbb{P}_d^{(\lambda)}$ versus $\mathbb{P}_d^{(\lambda,s)}$,*

[3]This guarantees that whenever the test based on tree matching weight in Chapter 5 [GM20] succeeds, the optimal test studied in this chapter also succeeds. On this point, Theorem 6.4 (see Section 6.4) extends the sufficient conditions established in Chapter 5 for partial alignment (for small λ and s close to 1).

(*ii*) $\lim_{d\to\infty} \mathrm{KL}_d = +\infty$ *and* $\lambda s > 1$,

(*iii*) *There exists* $(a_d)_d$ *such that* $a_d \to \infty$, $\mathbb{P}_d^{(\lambda)}(L_d > a_d) \to 0$ *and* $\liminf_d \mathbb{P}_d^{(\lambda,s)}(L_d > a_d) > 0$.

(*iv*) *The martingale* $(L_d)_d$ *(w.r.t.* $\mathbb{P}_\infty^{(\lambda)}$*) is not uniformly integrable.*

(*v*) $\lambda s > 1$ *and* $\mathbb{P}_\infty^{(\lambda,s)}\left(\liminf_{d\to\infty}(\lambda s)^{-d}\log L_d > 0\right) \geq 1 - p_{\mathrm{ext}}(\lambda s)$, *where* $p_{\mathrm{ext}}(\lambda s)$ *is the probability that a Galton-Watson tree with offspring distribution* $\mathsf{Poi}(\lambda s)$ *gets extinct.*

Remark 6.1.2. *This Theorem gives general necessary and sufficient conditions for the existence of a one-sided test in the tree correlation detection problem. Several more explicit conditions in terms of* λ *and* s *will be obtained throughout the chapter which guarantee that the equivalent conditions of Theorem 6.1 either fail or hold. Condition* (v) *will be used in the design of the algorithm in Section 6.7, choosing an appropriate threshold that will guarantee for the method to output both a substantial part of the underlying permutation and a vanishing number of mismatches.*

The link between the problem on trees and sparse graph alignment is given in the following

Theorem 6.2 (Consequences for one-sided partial graph alignment)**.** *For given* (λ, s)*, if one-sided correlation detection is feasible, i.e. any of the conditions in Theorem 6.1 holds, then one-sided partial alignment in the correlated Erdős-Rényi model* $\mathsf{G}(n, \lambda/n, s)$ *is achieved in polynomial time by our algorithm* `MPAlign` *(Algorithm 6.1 in Section 6.7).*

Conjecture. *We conjecture that if one-sided correlation detection in trees fails, i.e. none of the equivalent conditions in Theorem 6.1 holds, then no polynomial-time algorithm achieves partial recovery. In view of Theorem 6.6 of Section 6.6, which guarantees existence of a non-empty parameter region where one-sided tree detection fails while partial graph alignment can be done in non-polynomial time, our conjecture would imply the* hard phase *to be non-empty.*

Chapter organization The outline of the chapter is as follows. We recall some notations and model of random trees and the in Section 6.2. The derivation of the likelihood ratio between the relevant distributions is done in Section 6.3, where points (*iii*) and (*iv*) of Theorem 6.1 are proved (see 6.3.3). In Section 6.4, points (*ii*) and (*v*) of Theorem 6.1 are proved (see Section 6.4.1) and a first sufficient condition for one-sided tree detectability (Theorem 6.4) is obtained by analyzing Kullback-Leibler divergences: this condition is of the same kind as the one following from [GM20] in Chapter 5, however with a more direct derivation as well as a more explicit condition. Using a different approach, a second sufficient condition – that of Theorem 6.5 – is established in Section 6.5 by analyzing the number of automorphisms of Galton-Watson trees.

Next, we prove in Section 6.6 another condition (see Theorem 6.6) for the failure of one-sided detectability, hence showing that the conjectured hard phase is non-empty. The precise message-passing method for aligning graphs is introduced in Section 6.7, and guarantees on its output are established as well as the proof of Theorem 6.2.

Appendix 6.A is dedicated to numerical experiments as well as the description of the algorithm used in practice (`MPAlign2`). Some additional proofs are deferred to Appendix 6.B.

6.2. Notations and problem statement

6.2.1. Notations

In this first part we briefly introduce – or recall – some basic definitions that are used throughout the chapter.

Finite sets, permutations. For all $n > 0$, we define $[n] := \{1, 2, \dots, n\}$. For any finite set $\mathcal{X}$, we denote by $|\mathcal{X}|$ its cardinal. $\mathcal{S}_\mathcal{X}$ is the set of permutations on $\mathcal{X}$. We also denote $\mathcal{S}_k = \mathcal{S}_{[k]}$ for brevity, and we will often identify $\mathcal{S}_k$ to $\mathcal{S}_\mathcal{X}$ whenever $|\mathcal{X}| = k$. For any $0 \leq k \leq \ell$, we will write $\mathcal{S}(k, \ell)$ (resp. $\mathcal{S}(A, B)$) for the set of injective mappings from $[k]$ to $[\ell]$ (resp. between finite sets A and B). By convention, $|\mathcal{S}(0, \ell)| = 1$.

Graphs. In a graph $G = (V, E)$ – with node set V and edge set E – we denote by $d_G(u)$ the degree of node u in G and $\mathcal{N}_{G,d}(u)$ (resp. $\mathcal{S}_{G,d}(u)$) the set of vertices at distance $\leq d$ (resp. exactly d) from node u in G, $\mathcal{S}_{G,d}(i)$. The *neighborhood* of a node $u \in V$ is $\mathcal{N}_G(u) := \mathcal{N}_{G,1}(u)$, i.e. the set of all vertices that are connected to u by an edge in G.

Labeled rooted trees. A *labeled rooted tree* $t = (V, E)$ is an undirected graph with node set V and edge set E with no cycle. The *root* of t is a given distinguished node $\rho \in V$, and the *depth* of a node is defined as its distance to the root ρ. The depth of tree t is given as the maximum depth of all nodes in t. Each node u at depth $d \geq 1$ has a unique *parent* in t, which can be defined as the unique node at depth $d - 1$ on the path from u to the root ρ. Similarly, the *children* of a node u of depth d are all the neighbors of u at depth $d + 1$.

For any $u \in V$, we denote by t_u the subtree of t rooted at node u, and $c_t(u)$ the number of children of u in t – or simply $c(u)$ where there is no ambiguity. Finally we define $\mathcal{V}_d(t)$ (resp. $\mathcal{L}_d(t)$) to be the set of nodes of t at depth less than or equal to d (resp. exactly d).

Canonical labeling. A *labeled rooted tree* can be canonically labeled by ordering nodes' children, giving the following labels. First, the label of the root node is set to the empty list $\varnothing$. Then, recursively, the label of a node u is a list $\{m, k\}$ where m is the label of its parent node, and k is the rank of u among the children of its parent.

We denote by $\mathcal{Y}_d$ the collection of such canonically labeled rooted trees of depth no larger than d. Obviously, $\mathcal{Y}_0$ contains a single element, namely the rooted tree with only one node – its root. Each tree t in $\mathcal{Y}_d$ can be represented with a unique *ordered list* $(t_1, \dots, t_{c(\rho)})$ where each t_u is the subtree of t rooted at the $u-$th child of the root, and thus belongs to $\mathcal{Y}_{d-1}$. When $c(\rho) = 0$, the previous ordered list is empty.

Tree subsampling. For $s \in (0, 1)$, a $s-$*subsampling* of a tree t is obtained by conserving every edge independently with probability s, and outputting the connected component of the root (which is still a tree). The nodes in the resulting tree inherit a canonical labeling from their order in the original tree.

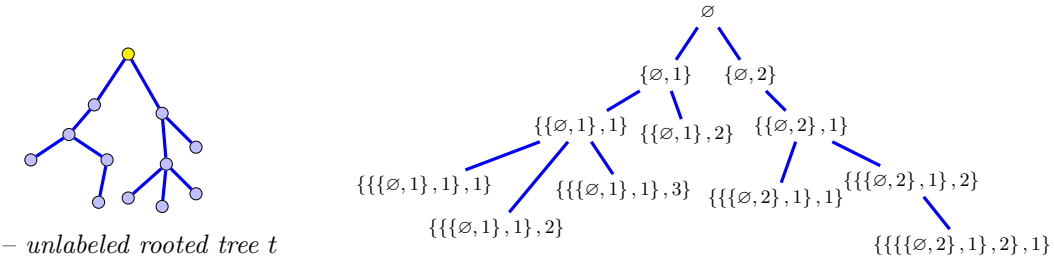
Relabelings of trees. A *relabeling* $r(t)$ of a tree $t \in \mathcal{Y}_d$ is recursively identified as a permutation $\sigma \in \mathcal{S}_{c(\rho)}$ of the children of the root node, together with relabelings $r_u(t_u)$ of its subtrees, resulting in tree

$$r(t) = \left(r_{\sigma(1)}(t_{\sigma(1)}), \dots, r_{\sigma(c(\rho))}(t_{\sigma(c(\rho))})\right).$$

A *random uniform relabeling* $r(t)$ of a (un-)labeled tree t of depth at most d is defined as follows. Associate independently to each node i of t a permutation σ_i of its children, uniformly distributed in $\mathcal{S}_{c(i)}$. The relabeling is then defined by induction on the depth of nodes: the new label $r(\rho)$ of the root is $\varnothing$, and recursively, if the label of u is $\{m, k\}$ and v is the parent of u, we assign to u the new label

$$r(u) := \{r(v), \sigma_v(k)\}.$$

An important and easily verified property is that, for a given labeled tree $t \in \mathcal{Y}_d$, $r(t)$ is indeed uniformly distributed on the set of all possible relabelings of t.



(a) – *unlabeled rooted tree* t

(b) – *a random uniform relabeling of* t

Figure 6.3 – *A rooted tree* $t \in \mathcal{Y}_d$ *with* $n = 4$ *(the root is highlighted in yellow).*

Automorphisms of labeled trees. Some of the relabelings of a labeled tree t may be *indistinguishable* from t, that is, equal to t as labeled trees. These relabelings are called *automorphisms of* t, and their set is denoted by $\mathrm{Aut}(t)$.

Injective mappings between labeled trees. For two labeled trees $\tau, t \in \mathcal{Y}_d$, the set of *injective mappings* from τ to t, denoted $\mathcal{S}(\tau, t)$, is the set of injective mappings from the labels of vertices of τ to the labels of vertices of t that preserve the rooted tree structure, in the sense that any $\sigma \in \mathcal{S}(\tau, t)$ must verify

$$\sigma(\varnothing) = \varnothing \quad \text{and} \quad \sigma(\{p, k\}) = \{\sigma(p), j\} \text{ for some } j.$$

Note that $\mathcal{S}(\tau, t)$ is not empty if and only if τ is, up to some relabeling, a subtree of t.

Probability. For the sake of readability, we will denote by π_μ the Poisson distribution of parameter μ, namely for all $k \geq 0$, $\pi_\mu(k) := e^{-\mu} \frac{\mu^k}{k!}$.

6.2.2. Models of random trees, hypothesis testing

We recall hereafter two models of random trees of Chapter 5.

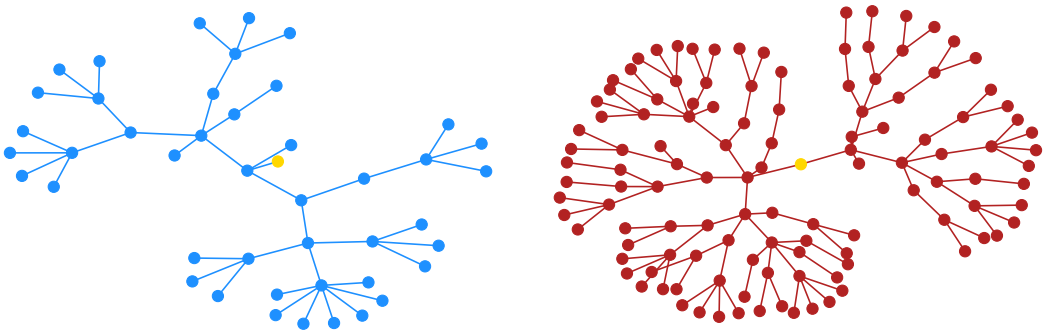
Independent model $\mathbb{P}_d^{(\lambda)}$ Under the independent model $\mathbb{P}_d^{(\lambda)}$, t and t' are two independent $\mathsf{GW}_d^{(\lambda)}$, where $\lambda > 0$ is the mean number of children in the graph.

Tree augmentation For $\lambda > 0$ and $s \in [0, 1]$, a (random) $(\lambda, s)-$*augmentation* of a given tree $\tau = (V, E)$, denoted by $\mathsf{Aug}_d^{(\lambda,s)}(\tau)$, is defined as follows. First, to each node u in V of depth $< d$, we attach a number Z_u^+ of additional children, where the Z_u^+ are i.i.d. of distribution $\mathsf{Poi}(\lambda(1-s))$. Let V^+ be the set of these additional children. To each $v \in V^+$ at depth d_v, we attach another random tree of distribution $\mathsf{GW}_{d-d_v}^{(\lambda)}$, independently of everything else.

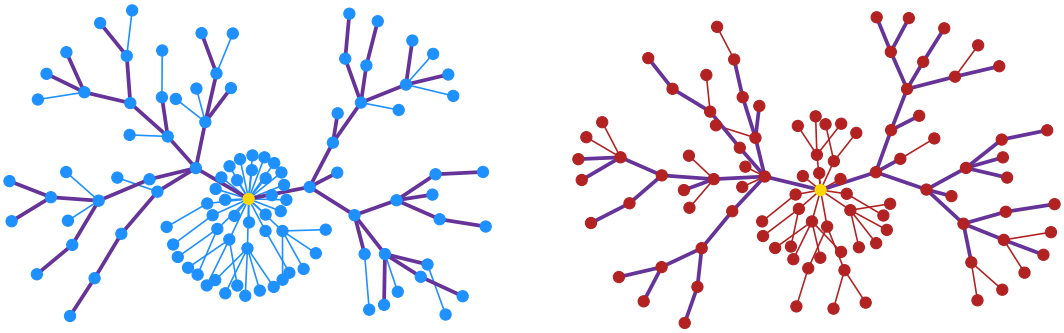
Correlated model $\mathbb{P}_d^{(\lambda,s)}$ The correlated model $\mathbb{P}_d^{(\lambda,s)}$ is built as follows: starting from an *intersection tree* $\tau^\star \sim \mathsf{GW}_d^{(\lambda s)}$, and T and T' are obtained as two independent $(\lambda, s)-$augmentations of $\tau^\star$. We denote $(T, T') \sim \mathbb{P}_d^{(\lambda,s)}$.

In all these models, the labels of the trees T and T' are then uniformly resampled at random by the procedure described above. It can easily be verified that T and T' are marginally both $\mathsf{GW}_d^{(\lambda)}$ under $\mathbb{P}_d^{(\lambda)}$ and $\mathbb{P}_d^{(\lambda,s)}$, namely. The parameters are λ, the mean number of children of a node, and the correlation s.

Hypothesis testing, one-sided test As mentioned earlier, we observe *finite* trees in practice. A property that we will use implicitly in the sequel is that for $T, T' \sim \mathbb{P}_d^{(\lambda)}$ (resp. $\sim \mathbb{P}_d^{(\lambda,s)}$) and $d' < d$, then $p_{d'}(T), p_{d'}(T') \sim \mathbb{P}_{d'}^{(\lambda)}$ (resp. $\sim \mathbb{P}_{d'}^{(\lambda,s)}$).



(a) – *Samples T, T' from $\mathbb{P}_d^{(\lambda)}$.*



(b) – *Samples T, T' from $\mathbb{P}_d^{(\lambda,s)}$. The common subtree τ is drawn thick and purple.*

Figure 6.4 – *Samples from models $\mathbb{P}_d^{(\lambda)}$ and $\mathbb{P}_d^{(\lambda,s)}$, with $\lambda = 1.8$, $s = 0.8$, and $d = 5$. The root node is highlighted in yellow. Labels are not shown.*

The hypothesis testing considered in this study can be formalized as follows: given the observation of a pair of trees (t, t') in $\mathcal{Y}_d \times \mathcal{Y}_d$, we want to test

$$\mathcal{H}_0 = \text{"}t, t' \text{ are realizations under } \mathbb{P}_d^{(\lambda)}\text{"} \quad \text{versus} \quad \mathcal{H}_1 = \text{"}t, t' \text{ are realizations under } \mathbb{P}_d^{(\lambda,s)}\text{"}. \tag{6.3}$$

More specifically, we are interested in being able to ensure the existence of a (asymptotic) *one-sided test*, that is a test $\mathcal{T}_n : \mathcal{Y}_d \times \mathcal{Y}_d \to \{0, 1\}$ such that $\mathcal{T}_n$ chooses hypothesis $\mathcal{H}_0$ under $\mathbb{P}_d^{(\lambda)}$ with probability $1 - o(1)$, and chooses $\mathcal{H}_1$ with some positive probability uniformly bounded away from 0 under $\mathbb{P}_d^{(\lambda,s)}$, guaranteeing a vanishing type I error and a non vanishing power.

Remark 6.2.1. *We here motivate one-sided tests once again. In statistical detection problems, the commonly considered tasks are that of*

- strong detection, *i.e. tests $\mathcal{T}_d$ that verify*

$$\lim_{n\to\infty} \left[\mathbb{P}_d^{(\lambda)} \left(\mathcal{T}_d(T, T') = 1 \right) + \mathbb{P}_d^{(\lambda,s)} \left(\mathcal{T}_d(T, T') = 0 \right) \right] = 0,$$

- weak detection, *i.e. tests $\mathcal{T}_d$ that verify*

$$\lim_{n\to\infty} \left[\mathbb{P}_d^{(\lambda)} \left(\mathcal{T}_d(T, T') = 1 \right) + \mathbb{P}_d^{(\lambda,s)} \left(\mathcal{T}_d(T, T') = 0 \right) \right] < 1.$$

In other words, strong detection corresponds to discriminate w.h.p. exactly the hypotheses, whereas weak detection corresponds to strictly outperforming random guess. We recall hereafter why neither strong detection nor weak detection are relevant for our problem.

First, because of the event that the intersection tree does not survive, which is of positive probability under $\mathbb{P}_d^{(\lambda,s)}$: we always have $\mathbb{P}_d^{(\lambda,s)}(t,t') \geq C \cdot \mathbb{P}_d^{(\lambda)}(t,t')$, with

$$C := \frac{\pi_{\lambda s}(0)\pi_{\lambda(1-s)}(c)\pi_{\lambda(1-s)}(c')}{\pi_\lambda(c)\pi_\lambda(c')},$$

where c (resp. c') is the degree of the root in t (resp t'). This implies that $\mathbb{P}_d^{(\lambda)}$ is always absolutely continuous w.r.t. $\mathbb{P}_d^{(\lambda,s)}$, hence strong detection can never be achieved.

Second, weak detection is always achievable as soon as $s > 0$: with the same notations as here above, the distribution of $c - c'$ is always centered but has different variance under $\mathbb{P}_d^{(\lambda)}$ and under $\mathbb{P}_d^{(\lambda,s)}$, hence these two distributions can be weakly distinguished, without any further assumption than $s > 0$. Since we know by [GML21b] that partial graph alignment is not feasible for $\lambda s \leq 1$, we conclude that weak detection in tree detection is not a relevant task either for graph alignment.

6.2.3. Warm-up discussion: the isomorphic case ($s = 1$)

In this section, we discuss the graph alignment problem in the case where $s = 1$ in the correlated Erdős-Rényi model (1.10), namely when the graphs G and H are isomorphic, $\pi^\star$ being one of the graph isomorphisms between G and H. We then ask the question: *what is the best fraction of nodes that can be recovered with high probability?*

The answer to the above question comes with the following easy remark: the joint distribution of (G, H) is invariant by any relabeling of G according to some $\sigma \in \mathrm{Aut}(G)$, where $\mathrm{Aut}(G)$ denotes the automorphism group of G. The set of nodes that can be aligned w.h.p. is hence

$$\mathcal{I}(G) := \{u \in V(G),\ \forall\, \sigma \in \mathrm{Aut}(G), \sigma(u) = u\}\,. \tag{6.4}$$

In other words, $\mathcal{I}(G)$ is the set of vertices of G invariant under any automorphism.

Let us denote $\mathcal{C}_1(G)$ the largest connected component of G (the *giant component*), and $\overline{\mathcal{C}_1(G)}$ the subgraph made of all the smaller components. It is clear that

$$\mathrm{Aut}(G) = \mathrm{Aut}(\mathcal{C}_1(G)) \times \mathrm{Aut}(\overline{\mathcal{C}_1(G)}).$$

Recent work [GML21b] shows that $\mathcal{I}(G) \cap \overline{\mathcal{C}_1(G)}$ contains at most a vanishing fraction of the points: it is not hard to see indeed that smaller components mainly consist in isolated trees, which are proved to have many copies in the graph when n gets large, yielding some automorphisms that swap almost all vertices in $\overline{\mathcal{C}_1(G)}$. Hence, for our purpose, the main part of $\mathcal{I}(G)$ comes from the study of $\mathrm{Aut}(\mathcal{C}_1(G))$ and $\mathcal{I}(\mathcal{C}_1(G))$.

When $G \sim \mathsf{G}(n, q)$, these sets have been thoroughly studied by Łuczak in [Luc88]. Vertices of the giant component that are not invariant under automorphism are mainly (i.e. up to $o(n)$ errors) vertices that do not belong to the *2-core*[4] of G, denoted by $\mathcal{C}^{(2)}(G)$.

Simple structures appearing in $\mathcal{C}_1(G) \setminus \mathcal{I}(G)$ are leaves (degree one nodes) v, w with common a neighbor u in $\mathcal{C}_1(G)$. [Luc88] upper-bounds the size of $\mathcal{C}_1(G) \setminus \mathcal{I}(G)$ by the number of (generalizations) of such structures, thus obtaining the following

Theorem 6.3 ([Luc88], Theorems 3 and 4)**.** *Let $G \sim \mathsf{G}(n, q)$ with $q = \lambda/n$. Let $(K_n)_n$ be a sequence such that $K_n \to \infty$. There exists $\lambda_0 > 0$ such that if $\lambda > \lambda_0$, then with high probability,*

$$\left|\mathcal{C}^{(2)}(G)\right| - \left|\mathcal{I}(\mathcal{C}^{(2)}(G))\right| \leq K_n, \quad \textit{and} \quad |\mathcal{C}_1(G)| - |\mathcal{I}(\mathcal{C}_1(G))| \leq \lambda(\lambda+5)e^{-2\lambda}n. \tag{6.5}$$

Equation (6.5) of Theorem 6.3 states that for λ large enough, almost all vertices of the 2-core of G are invariant, whereas at most a fraction $\lambda(\lambda + 5)e^{-2\lambda}$ of the nodes are in the

[4]The *2-core* of a graph is defined as the maximal subgraph of minimal degree at least 2.

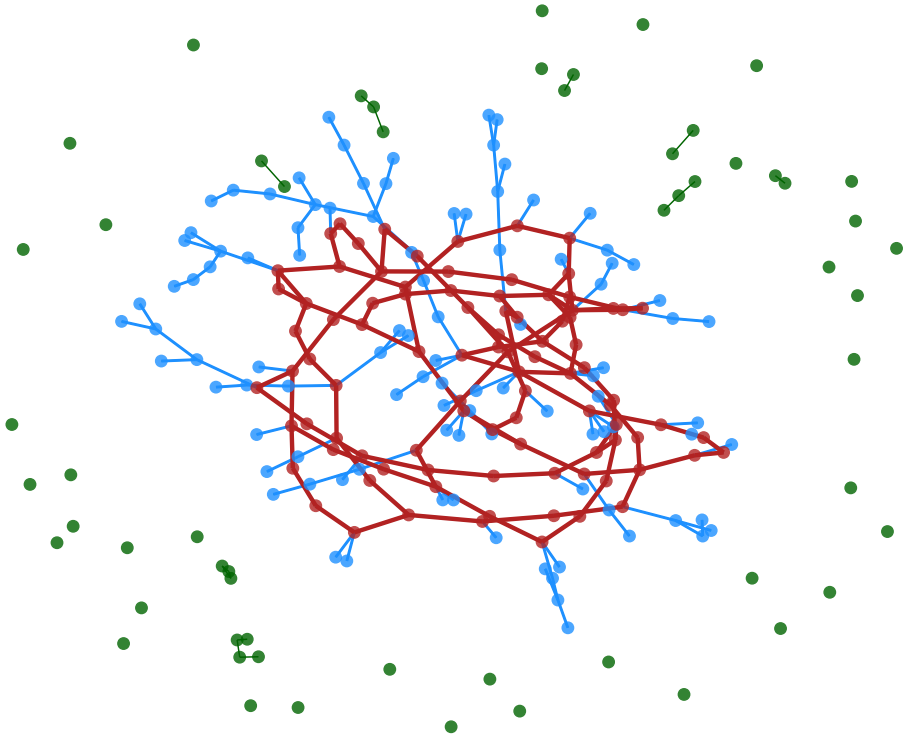


Figure 6.5 – *Sample G from model $\mathsf{G}(n, \lambda/n)$, with $\lambda = 2$ and $n = 250$. Vertices of $\overline{\mathcal{C}_1(G)}$ (resp. of $\mathcal{C}_1(G) \setminus \mathcal{C}^{(2)}(G)$, $\mathcal{C}^{(2)}(G)$) are drawn in green (resp. blue, red).*

giant component and not in $\mathcal{I}(G)$. In this case, with high probability, any isomorphism $\hat{\pi}$ between G and H will achieve partial recovery and will satisfy

$$\mathrm{ov}\,(\hat{\pi}, \pi^\star) \geq 1 - p_{\mathrm{ext}}(\lambda) - \lambda(\lambda+5)e^{-2\lambda},$$

where $p_{\mathrm{ext}}(\lambda)$ is defined as the probability that a Galton-Watson tree of offspring $\mathsf{Poi}(\lambda)$ survives.

However, finding efficiently such an isomorphism $\hat{\pi}$ is known to be challenging in the general case (see e.g. [AK02]): hence, whether there exists a polynomial-time algorithm achieving this optimal bound remains an open question[5].

6.3. Derivation of the likelihood ratio

For $t, t' \in \mathcal{Y}_d$, we introduce the likelihood ratio

$$L_d(t, t') := \frac{\mathbb{P}_d^{(\lambda,s)}(t, t')}{\mathbb{P}_d^{(\lambda)}(t, t')}. \tag{6.6}$$

6.3.1. Recursive computation

In this section, our aim is to obtain a recursive representation of the likelihood ratio L_d. First note that for two trees $t = (t_1, \ldots, t_c)$, $t' = (t'_1, \ldots, t'_{c'})$ both in $\mathcal{Y}_d$, we have

$$\mathbb{P}_d^{(\lambda)}(t, t') = \mathsf{GW}_d^{(\lambda)}(t) \times \mathsf{GW}_d^{(\lambda)}(t'), \tag{6.7}$$

[5]We can however cite a famous result of Bollobás ([Bol01], Theorem 9.9) showing that in the dense case $np \geq \Theta(\log n)$, the vertices of every $G \sim \mathsf{G}(n,p)$ graph are uniquely determined by their distance sequences, and the automorphism group of G is w.h.p. trivial.

and that conditioned to c, $\mathsf{GW}_d^{(\lambda)}(t)$ satisfies the recursion

$$\mathsf{GW}_d^{(\lambda)}(t) = \pi_\lambda(c) \prod_{u\in[c]} \mathsf{GW}_{d-1}^{(\lambda)}(t_u). \tag{6.8}$$

In the construction of t, t' under $\mathcal{H}_1$, partitioning on the permutations $\sigma \in \mathcal{S}_c, \sigma' \in \mathcal{S}_{c'}$ used to shuffle the children of the root nodes of t, t', as well as on the number k of children of the root in $\tau^\star$, we have the following

$$\begin{aligned}\mathbb{P}_d^{(\lambda,s)}(t,t') = \sum_{k=0}^{c\wedge c'} &\pi_{\lambda s}(k)\pi_{\lambda(1-s)}(c-k)\pi_{\lambda(1-s)}(c'-k) \\ &\times \sum_{\sigma\in\mathcal{S}_c,\sigma'\in\mathcal{S}_{c'}} \frac{1}{c!\times c'!}\left(\prod_{u=1}^{k}\mathbb{P}_{1,n-1}(t_{\sigma(u)},t'_{\sigma'(u)})\right) \\ &\times\left(\prod_{u=k+1}^{d}\mathsf{GW}_{d-1}^{(\lambda)}(t_{\sigma(u)})\right)\times\left(\prod_{i=k+1}^{d'}\mathsf{GW}_{d-1}^{(\lambda)}(t'_{\sigma'(u)})\right).\end{aligned}$$

This together with Equations (6.7), (6.8) readily implies the following recursive formula for the likelihood ratio L_d:

$$L_d(t,t') = \sum_{k=0}^{c\wedge c'} \frac{\pi_{\lambda s}(k)\pi_{\lambda(1-s)}(c-k)\pi_{\lambda(1-s)}(c'-k)}{\pi_\lambda(c)\pi_\lambda(c')\times c!\times c'!} \sum_{\sigma\in\mathcal{S}_c,\sigma'\in\mathcal{S}_{c'}}\prod_{u=1}^{k} L_{d-1}(t_{\sigma(u)},t'_{\sigma'(u)}). \tag{6.9}$$

In this expression, by convention the empty product equals 1. We will use in the sequel the following shorthand notation

$$\begin{aligned}\psi(k,c,c') &:= \frac{\pi_{\lambda s}(k)\pi_{\lambda(1-s)}(c-k)\pi_{\lambda(1-s)}(c'-k)}{\pi_\lambda(c)\pi_\lambda(c')} \times \frac{(c-k)!\times(c-k')!}{c!\times c'!} \\ &= e^{\lambda s}\times\frac{s^k(1-s)^{c+c'-2k}}{\lambda^k k!},\end{aligned}$$

which enables an alternative, more compact recursive expression:

$$L_d(t,t') = \sum_{k=0}^{c\wedge c'}\psi(k,c,c') \sum_{\substack{\sigma\in\mathcal{S}(k,c)\\ \sigma'\in\mathcal{S}(k,c')}}\prod_{u=1}^{k} L_{d-1}(t_{\sigma(u)},t'_{\sigma'(u)}), \tag{6.10}$$

where we recall that $\mathcal{S}(k,\ell)$ denotes the set of injective mappings from $[k]$ to $[\ell]$ and that by convention $|\mathcal{S}(0,\ell)| = 1$.

Remark 6.3.1. *The above expression* (6.10) *will be useful for efficient computations of the likelihood ratio in Algorithm 6.1 in Section 6.7, through message-passing.*

6.3.2. Explicit computation

We now use the recursive expression (6.10) to prove by induction on d the following explicit formula for L_d.

Lemma 6.3.1. *With the previous notations, we have*

$$L_d(t,t') = \sum_{\tau\in\mathcal{Y}_d}\sum_{\substack{\sigma\in\mathcal{S}(\tau,t)\\ \sigma'\in\mathcal{S}(\tau,t')}}\prod_{u\in\mathcal{V}_{d-1}(\tau)}\psi\left(c_\tau(u),c_t(\sigma(u)),c_{t'}(\sigma'(u))\right). \tag{6.11}$$

Proof of Lemma 6.3.1. We prove this result by recursion on d. An empty product being set to 1, there is nothing to prove in the case $d = 0$. Let us first establish formula (6.11) for $d = 1$. In that case, the depth 1 trees t, t' are identified by the degrees c, c' of their root node. Since $\mathcal{Y}_0$ is a singleton, L_0 is identically 1, and from (6.9) we have that

$$L_1(t, t') = \sum_{k=0}^{c\wedge c'} \frac{\pi_{\lambda s}(k)\pi_{\lambda(1-s)}(c-k)\pi_{\lambda(1-s)}(c'-k)}{\pi_\lambda(c)\pi_\lambda(c')}. \tag{6.12}$$

On the other hand, in evaluating expression (6.11), we only need consider trees τ in $\mathcal{Y}_1$ with root degree $k \leq c \wedge c'$, since for larger k, one of the two sets $\mathcal{S}(\tau, t)$ or $\mathcal{S}(\tau, t')$ is empty. For such k, we have $|\mathcal{S}(\tau, t)| = c!/(c-k)!$. The right-hand term in (6.11) thus writes

$$\sum_{k=0}^{c\wedge c'} \frac{c! \times c'!}{(c-k)! \times (c'-k)!} \psi(k, c, c'),$$

which gives precisely (6.12).

Assume that (6.11) has been established up to some $n - 1 \geq 1$. Expressing L_d in terms of L_{d-1} based on (6.9), and replacing in there the expression of L_{d-1} by (6.11) , we get

$$L_d(t, t') = \sum_{k=0}^{c\wedge c'} \frac{\psi(k, c, c')}{(c-k)!(c'-k)!}$$
$$\times \sum_{\sigma\in\mathcal{S}_c, \sigma'\in\mathcal{S}_{c'}} \prod_{u=1}^{k} \left[\sum_{\substack{\tau_u\in\mathcal{Y}_{d-1} \\ \sigma'_u\in\mathcal{S}(\tau_u, t'_{\sigma(u)})}} \sum_{\sigma_u\in\mathcal{S}(\tau_u, t_{\sigma(u)})} \prod_{v\in\mathcal{V}_{d-1}(\tau_u)} \psi\left(c_{\tau_u}(v), c_{t_{\sigma(u)}}(\sigma_u(v)), c_{t'_{\sigma'(u)}}(\sigma'_u(v))\right) \right].$$

Note that the product term in the above expression depends on the permutations σ, σ' only through their restriction to $[k]$: for given such restrictions there are $(c-k)! \times (c'-k)!$ corresponding pairs of permutations σ, σ'.

Moreover, there is a bijective mapping between an integer $k \in \{0, \ldots, c \wedge c'\}$, pairs of injections $\sigma : [k] \to [c]$, $\sigma' : [k] \to [c']$, k trees $\tau_1, \ldots, \tau_k \in \mathcal{Y}_{d-1}$, injections $\sigma_u \in \mathcal{S}(\tau_u, t_{\sigma(u)})$ and $\sigma'_u \in \mathcal{S}(\tau_u, t'_{\sigma'(u)})$ for all $u \in [k]$ and a tree $\tau \in \mathcal{Y}_d$ together with a pair of injections $\sigma, \sigma' \in \mathcal{S}(\tau, t) \times \mathcal{S}(\tau, t')$. This establishes formula (6.11) at step d.

□

6.3.3. Martingale properties and the objective of one-sided test

In this part, we assume that we observe T, T' drawn under one of the two models $\mathbb{P}_\infty^{(\lambda)}$ or $\mathbb{P}_\infty^{(\lambda,s)}$. For $d \geq 0$, let $\mathcal{F}_d := \sigma(p_d(T), p_d(T'))$ be the sigma-field of the two trees T, T' observed down to depth d. We then have

Lemma 6.3.2. *The sequence* $\{L_d := L_d(p_d(T), p_d(T'))\}_{d\geq 0}$ *is a* $\mathcal{F}_d$*-martingale under* $\mathbb{P}_\infty^{(\lambda)}$.

The above martingale property follows from general considerations of likelihood ratios. It is however informative to derive it by calculus, which we now do.

Proof of Lemma 6.3.2. There are several ways to see that $\{L_d\}_{d\geq 0}$ is a $\mathcal{F}_d$-martingale under $\mathbb{P}_\infty^{(\lambda)}$, depending on the formula used to write L_{d+1} in terms of L_d. We here choose to use the developed expression (6.11), enabling simple computations:

$$L_{d+1} = \sum_{\tau\in\mathcal{Y}_{d+1}} \sum_{\substack{\sigma\in\mathcal{S}(\tau,T)\\ \sigma'\in\mathcal{S}(\tau,T')}} \prod_{u\in\mathcal{V}_d(\tau)} \psi\left(c_\tau(u), c_T(\sigma(u)), c_{T'}(\sigma'(u))\right)$$
$$= \sum_{\chi\in\mathcal{Y}_d} \sum_{\substack{\sigma\in\mathcal{S}(\chi,p_d(T))\\ \sigma'\in\mathcal{S}(\chi,p_d(T'))}} \prod_{i\in\mathcal{V}_{d-1}(\chi)} \psi\left(c_\chi(u), c_{p_d(T)}(\sigma(u)), c_{p_d(T')}(\sigma'(u))\right)$$
$$\times \prod_{u\in\mathcal{L}_d(\chi)} \sum_{k=0}^{c_T(\sigma(u))\wedge c_{T'}(\sigma'(u))} \frac{c_T(\sigma(u))!c_{T'}(\sigma'(u))!}{(c_T(\sigma(u))-k)!(c_{T'}(\sigma'(u))-k)!}\psi(k, c_T(\sigma(u)), c_{T'}(\sigma'(u))).$$

The last product is independent from $\mathcal{F}_d$. Moreover, under $\mathbb{P}_\infty^{(\lambda)}$, all terms in the last product are independent, the $c_T(u)$ and $c_{T'}(u)$ being independent $\mathsf{Poi}(\lambda)$ random variables. Since for any independent $\mathsf{Poi}(\lambda)$ random variables c, c', one has

$$\mathbb{E}\left[\sum_{k=0}^{c\wedge c'} \frac{\pi_{\lambda s}(k)\pi_{\lambda(1-s)}(c-k)\pi_{\lambda(1-s)}(c'-k)}{\pi_\lambda(c)\pi_\lambda(c')}\right] = 1,$$

taking the expectation conditionally to $\mathcal{F}_d$ entails the desired martingale property. □

We now consider the martingale almost sure limit L_∞, and define $\ell := \mathbb{E}_\infty^{(\lambda)}[L_\infty]$. Using the recursive formula (6.9) and conditioning on the root degrees c and c', it follows that ℓ verifies the following fixed point equation

$$\ell = \sum_{k\geq 0} \pi_{\lambda s}(k)\ell^k. \tag{6.13}$$

This is also (!) the fixed point equation for the extinction probability $p_{\text{ext}}(\lambda s)$ of a Galton-Watson branching process with offspring distribution $\mathsf{Poi}(\lambda s)$. For $\lambda s \leq 1$, the only solution of (6.13) is $\ell = 1$. For $\lambda s > 1$, the equation also admits a non-trivial solution $p_{\text{ext}}(\lambda s) \in (0,1)$.

Our goal is to find conditions on (λ, s) for which the martingale $\{L_d\}_{d\geq 0}$ is not uniformly integrable and *looses mass* at infinity, i.e. the conditions for which the martingale limit L_∞ has expectation $\mathbb{E}_\infty^{(\lambda)}[L_\infty] < 1$. By the previous calculation we know that if this holds, then necessarily $\mathbb{E}_\infty^{(\lambda)}[L_\infty] = p_{\text{ext}}(\lambda s) < 1$. Simulations of L_d displayed on Figure 6.7 seem to indicate that its transition to non-uniform integrability does not coincide with the condition $\lambda s > 1$. We shall obtain a theoretical confirmation of this fact with Theorem 6.6.

Our interest in conditions for non-uniform integrability stem from the following simple Lemma:

Lemma 6.3.3. *Assume that $\mathbb{E}_\infty^{(\lambda)}[L_\infty] < 1$. Then there exists a one-sided test.*

Proof of Lemma 6.3.3. Let us take $a > 0$ a continuity point of the law of L_∞ under $\mathbb{P}_\infty^{(\lambda)}$. We have

$$\lim_{d\to\infty} \mathbb{P}_\infty^{(\lambda)}(L_d > a) = \mathbb{P}_\infty^{(\lambda)}(L_\infty > a). \tag{6.14}$$

Moreover,

$$1 = \mathbb{E}_\infty^{(\lambda)}[L_d] = \mathbb{E}_\infty^{(\lambda)}[L_d \mathbb{1}_{L_d>a}] + \mathbb{E}_\infty^{(\lambda)}[L_d \mathbb{1}_{L_d\leq a}]$$
$$= \mathbb{P}_\infty^{(\lambda,s)}(L_d > a) + \mathbb{E}_\infty^{(\lambda)}[L_d \mathbb{1}_{L_d\leq a}].$$

The last equation implies, under the assumption $\mathbb{E}_\infty^{(\lambda)}[L_\infty] < 1$ (that is $\mathbb{E}_\infty^{(\lambda)}[L_\infty] = p_{\text{ext}}(\lambda s)$), that

$$\liminf_{d\to\infty} \mathbb{P}_\infty^{(\lambda,s)}(L_d > a) \geq 1 - \mathbb{E}_\infty^{(\lambda)}[L_\infty] = 1 - p_{\text{ext}}(\lambda s) > 0. \tag{6.15}$$

In view of (6.14) and (6.15), we can thus choose $a_d \to \infty$ such that:

$$\lim_{d\to\infty} \mathbb{P}_\infty^{(\lambda)}(L_d > a_d) = 0 \quad \text{and} \quad \liminf_{d\to\infty} \mathbb{P}_\infty^{(\lambda,s)}(L_d > a_d) \geq 1 - p_{\text{ext}}(\lambda s) > 0.$$

□

Proof of $(i) \iff (iii) \iff (iv)$ **in Theorem 6.1**

Proof. The previous proof shows first that $(i) \iff (iii)$ in Theorem 6.1 (applying Neyman-Pearson's Lemma and a diagonal extraction procedure) as well as $(iii) \iff (iv)$, since condition

$$\exists\, \varepsilon > 0,\ \forall a > 0,\ \liminf_{d\to\infty} \mathbb{P}_\infty^{(\lambda,s)}(L_d > a) \geq \varepsilon > 0 \tag{6.16}$$

is exactly the condition of non-uniform integrability of the martingale $(L_d)_d$ with respect to $\mathbb{P}_\infty^{(\lambda)}$. □

6.3.4. A Markov transition kernel on trees

In this section, we introduce a Markov transition semi-group on trees that arises naturally in our study. Indeed, the joint distribution of the pair of trees (T, T') under $\mathbb{P}_d^{(\lambda,s)}$ will be, up to relabeling, interpreted as the joint distribution of (X_0, X_r), where X_0 is the initial state of this Markov process, distributed according to its stationary distribution $\mathsf{GW}_d^{(\lambda)}$, and X_r is its state at time r. The time parameter r is in one-to-one correspondence with the correlation parameter s of our model, through the relation

$$r = -\log(s).$$

For $n > 0$, we define M_d the linear operator indexed on trees of $\mathcal{Y}_d$, defined as follows:

$$M_d(t, t') := \frac{\mathbb{P}_d^{(\lambda,s)}(t, t')}{\mathbb{P}_d^{(\lambda)}(t)}. \tag{6.17}$$

M_d is identified to the *transition kernel* of the Markov chain with transitions denoted $t \underset{\lambda,s}{\longrightarrow} t'$ where t' is obtained from t following the following three-step procedure:

1. Extracting τ, a $s-$subsampling of t;
2. Draw τ^+, an augmentation $\mathsf{Aug}_d^{(\lambda,s)}$ of τ;
3. Take t' to be a uniform relabeling of τ^+.

We next denote $M_d(s)$ this transition kernel to emphasize its dependence on s.

A remarkable property of this kernel is the following semi-group structure:

Proposition 6.3.1 (Consistency of kernels $M_d(s)$)**.** *Let $\lambda > 0$ and $s, s' \in [0, 1]$. Then, for all $n \geq 1$,*

$$M_d(s)M_d(s') = M_d(s')M_d(s) = M_d(ss'). \tag{6.18}$$

Proof. The proof consists in verifying that applying transitions $M_d(s)$ and $M_d(s')$ successively is equivalent in distribution to applying transition $M_d(ss')$. Let us first show that the unlabeled structures of the trees are equivalent in distribution. For $t \in \mathcal{Y}_d$, let us sample a sequence $t \underset{s}{\longrightarrow} \tilde{t} \underset{s'}{\longrightarrow} t'$ as follows.

For $t \in \mathcal{Y}_d$, let us apply a first transition $t \underset{s}{\longrightarrow} \tilde{t}$: we extract $\tilde{\tau}$, a $s-$subsampling of t. To each vertex u of $\tilde{\tau}$ we attach an independent number $\mathsf{Poi}(\lambda(1-s))$ of new children. The set

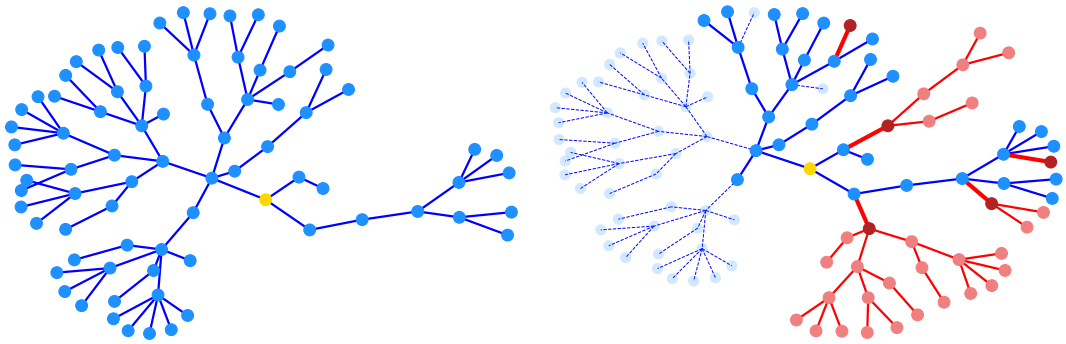


Figure 6.6 – *Example of a transition described hereabove, with* $\lambda = 1.85$, $s = 0.85$, *at depth* $d = 5$. *The original tree* t *is drawn on the left. On the right,* t' *is obtained as follows: first extracting a* $s-$*subsampling* τ *of* t *(dashed blue edges are deleted), and drawing a* $(\lambda, s)-$*augmentation of* τ *– first attaching new children to all vertices of* τ *(dark red nodes with thick edges), and attaching new Galton-Watson trees to these new children (light red nodes with standard edges). Labels are not shown.*

of these new vertices is denoted $\tilde{V}_+$. Then, to each vertex $u \in \tilde{V}_+$ we attach an independent tree $\tilde{t}_u$ with distribution GW_λ. We just sampled the unlabeled version of $\widetilde{t}$.

Let us now apply the second transition $\tilde{t} \underset{s'}{\longrightarrow} t'$. We sample t as follows:

1. First, we sample all vertices of $\tilde{\tau}$ in $\tilde{t}$, keeping them independently with probability s'. The obtained subtree is denoted by τ;
2. To any vertex u of τ, we keep each previous child vertex in $\tilde{V}_+$ independently with probability s', the set of children that are kept is denoted by V_+^1;
3. To any vertex u of τ, we attach an independent number $\mathsf{Poi}(\lambda(1-s'))$ of new children. The set of these new vertices are referred to as V_+^2.
4. To any vertex $u \in V_+^1$, we sample a transition $\tilde{t}_u \underset{s'}{\longrightarrow} t_u$, and attach t_u to node u.
5. To each vertex $v \in V_+^2$ we attach an independent tree t_v with distribution GW_λ.

Eventually we performed the following process: from the initial tree t, we extracted τ as a $ss'-$subsampling of t, and we attached to each vertex of τ some new children: the sum of two independent $\mathsf{Poi}(\lambda(1-s)s')$ (for children in V_+^1) and $\mathsf{Poi}(\lambda(1-s'))$ (for children in V_+^2), hence again of Poisson distribution with parameter $\lambda(1-s)s' + \lambda(1-s') = \lambda(1-ss')$. By steps 4. and 5., the trees attached to every vertex in $V_+ := V_+^1 \cup V_+^2$ are i.i.d. with distribution GW_λ, independent of t. Hence, the unlabeled version of t' can also be obtained from t with the transition $t \underset{ss'}{\longrightarrow} t'$.

Finally, the definition of the tree subsampling ensures that the composition of the two relabelings in the two steps gives indeed a uniform relabeling of t, which ends the proof. □

6.4. Conditions based on Kullback-Leibler divergences

In the sequel we shall denote

$$\mathrm{KL}_d := \mathrm{KL}(\mathbb{P}_d^{(\lambda,s)} \| \mathbb{P}_d^{(\lambda)}) = \mathbb{E}_d^{(\lambda,s)} \left[\log(L_d)\right]. \tag{6.19}$$

Note that by convexity of $\phi : x \to x\log(x)$, the martingale property of likelihood ratios L_d under $\mathbb{P}_d^{(\lambda)}$ and Jensen's inequality, the sequence KL_d is increasing with d and therefore admits a limit KL_∞ as $d \to \infty$.

6.4.1. Phase transition for KL_∞

Let us start with a simple proposition.

Proposition 6.4.1. *One has* $\mathrm{KL}_d \leq \mathrm{Ent}(\mathsf{GW}_d^{(\lambda s)})$.

Proof. Consider the Markov transition kernel K_d from $\mathcal{Y}_d^2$ to $\mathcal{Y}_d^2$ such that $K_d((\tau,\tau'),(t,t'))$ is the probability that independent $(\lambda,s)-$augmentations and relabelings of (τ,τ') to depth d produce the two trees (t,t').

Thus $\mathbb{P}_d^{(\lambda,s)}$ is the law obtained by applying kernel K_d to the distribution of (τ,τ), where $\tau \sim \mathsf{GW}_d^{(\lambda s)}$ whereas $\mathbb{P}_d^{(\lambda)}$ is the law obtained by applying kernel K_d to the distribution of two independent $\mathsf{GW}_d^{(\lambda s)}$ trees (τ,τ'). Standard monotonicity properties of Kullback-Leibler divergence then guarantee that KL_d is upper-bounded by $\mathrm{KL}(\mathcal{L}(\tau,\tau)\|\mathcal{L}(\tau,\tau'))$. This divergence reads

$$\sum_{\tau\in\mathcal{Y}_d} \mathsf{GW}_d^{(\lambda s)}(\tau)\log\left(\frac{\mathsf{GW}_d^{(\lambda s)}(\tau)}{\mathsf{GW}_d^{(\lambda s)}(\tau)^2}\right) = \mathrm{Ent}(\mathsf{GW}_d^{(\lambda s)}).$$

□

This readily implies the following

Corollary 6.4.1. *Assume* $\lambda s < 1$. *Then*

$$\mathrm{KL}_\infty = \lim_{d\to\infty} \mathrm{KL}_d \leq \frac{1}{1-\lambda s}\mathrm{Ent}(\pi_{\lambda s}) < +\infty. \tag{6.20}$$

Proof. Entropy $\mathrm{Ent}(\mathsf{GW}_d^{(\lambda s)})$ can be evaluated by the conditional entropy formula as

$$\mathrm{Ent}(\mathsf{GW}_d^{(\lambda s)}) = \mathrm{Ent}(\mathsf{GW}_{d-1}^{(\lambda s)}) + (\lambda s)^{d-1}\mathrm{Ent}(\pi_{\lambda s}).$$

The result follows from Proposition 6.4.1. □

We then have the following result:

Proposition 6.4.2. *Existence of one-sided tests holds if* $\lambda s > 1$ *and* $\mathrm{KL}_\infty = +\infty$, *whereas it fails if* $\mathrm{KL}_\infty < +\infty$.

Proof. Assume existence of one-sided tests. As previously mentioned, equivalently there exists $\varepsilon > 0$ such that

$$\forall a > 0,\ \liminf_{d\to\infty} \mathbb{P}_d^{(\lambda,s)}(L_d > a) \geq \varepsilon.$$

Fix $a > 0$, and define for $d \in \mathbb{N}$, $C_d := \{x \in \mathcal{Y}_d^2 : L_d(x) > a\}$. Write then, noting $\phi(u) := u\log(u)$:

$$\begin{aligned}
\mathrm{KL}_d &\geq \mathbb{P}_d^{(\lambda,s)}(C_d)\log(a) + \sum_{x\in\overline{C_d}} \mathbb{P}_d^{(\lambda,s)}(x)\log\frac{\mathbb{P}_d^{(\lambda,s)}(x)}{\mathbb{P}_d^{(\lambda)}(x)} \\
&\geq \mathbb{P}_d^{(\lambda,s)}(C_d)\log(a) + \mathbb{P}_d^{(\lambda)}(\overline{C_d}) \sum_{x\in\overline{C_d}} \frac{\mathbb{P}_d^{(\lambda)}(x)}{\mathbb{P}_d^{(\lambda)}(\overline{C_d})}\phi(L_d(x)) \\
&\overset{(a)}{\geq} \mathbb{P}_d^{(\lambda,s)}(C_d)\log(a) + \mathbb{P}_d^{(\lambda)}(\overline{C_d})\phi\left(\sum_{x\in\overline{C_d}} \frac{\mathbb{P}_d^{(\lambda)}(x)}{\mathbb{P}_d^{(\lambda)}(\overline{C_d})} L_d(x)\right) \\
&= \mathbb{P}_d^{(\lambda,s)}(C_d)\log(a) + \mathbb{P}_d^{(\lambda,s)}(\overline{C_d})\log\left(\mathbb{P}_d^{(\lambda,s)}(\overline{C_d})\right) - \mathbb{P}_d^{(\lambda,s)}(\overline{C_d})\log(\mathbb{P}_d^{(\lambda)}(\overline{C_d}))
\end{aligned}$$

$$\geq \mathbb{P}_d^{(\lambda,s)}(C_d)\log(a) + \inf_{u\in[0,1]} \phi(u) \geq \mathbb{P}_d^{(\lambda,s)}(C_d)\log(a) - e^{-1}\,.$$

We used convexity of ϕ in (a). It thus follows from characterization of one-sided testability that for all $a > 0$,

$$\mathrm{KL}_\infty \geq \varepsilon \log(a) - e^{-1},$$

and thus $\mathrm{KL}_\infty = +\infty$.

Conversely, assume $\lambda s > 1$ and $\mathrm{KL}_\infty = +\infty$. Let under $\mathbb{P}_\infty^{(\lambda,s)}$ define

$$w := \lim_{d\to\infty} |\mathcal{L}_d(\tau^\star)|(\lambda s)^{-d}\,.$$

On the event that $\tau^\star$ survives, which has strictly positive probability for $\lambda s > 1$, it holds that $w > 0$. In addition, we let $\pi^\star$, $(\pi')^\star$ denote the injections from $\tau^\star$ to T and T' respectively that result from uniform shuffling of the augmentations of $\tau^\star$.

Let d, m be two integers. One then has the lower bound:

$$\begin{aligned} L_{d+m}(T,T') &\geq \prod_{u\in\mathcal{V}_{d-1}(\tau^\star)} \psi(c_{\tau^\star}(u), c_T(\pi^\star(u)), c_{T'}((\pi')^\star(u)) \prod_{u\in\mathcal{L}_d(\tau^\star)} L_m(T_{\pi^\star(u)}, T'_{(\pi')^\star(u)}) \\ &\geq \prod_{u\in\mathcal{V}_{d-1}(\tau^\star)} \psi(c_{\tau^\star}(u), c_T(\pi^\star(u)), c_{T'}((\pi')^\star(u))) e^{|\mathcal{L}_d(\tau^\star)|[\mathbb{E}_m^{(\lambda,s)}[\log L_m] - o(1)]}. \end{aligned}$$

For d large, by the law of large numbers, the first product is with high probability lower-bounded by $e^{Cw(\lambda s)^d}$ for some fixed constant C. Choosing m of order 1 but sufficiently large, since by assumption $\lim_{m\to\infty} \mathbb{E}_m^{(\lambda,s)}[\log(L_m)] = +\infty$, we can ensure that the second factor is larger than $e^{C'w(\lambda s)^d}$ for some arbitrary C'. Taking C' large enough ensures that, on the event that $\tau^\star$ survives, $\lim_{d\to\infty} L_d = +\infty$ almost surely. This readily implies one-sided testability. □

Proof of $(i) \iff (ii) \iff (v)$ in Theorem 6.1

Proof. Proposition 6.4.2 gives the implication $(ii) \implies (i)$. Its proof further gives $(ii) \implies (v)$. The converse $(v) \implies (ii)$ is obvious. The second statement in Proposition 6.4.2 gives $(i) \implies \mathrm{KL}_\infty = +\infty$. To obtain that $(i) \implies (ii)$ and conclude, it thus only remains to show that $(i) \implies \lambda s > 1$.

As will be shown in Section 6.7, one-sided testability implies (polynomial-time) feasibility of partial graph alignment. However, [GML21b] established that partial alignment is not feasible when $\lambda s \leq 1$ (see Chapter 4). This establishes $(i) \implies \lambda s > 1$ as required. □

6.4.2. Applications

To apply condition (ii) of Theorem 6.1, let us first establish the following

Lemma 6.4.1. *For all $d \geq 1$, one has*

$$\mathrm{KL}_{d+1} \geq \lambda s \mathrm{KL}_d + \lambda s\,(\log(s/\lambda) + 1) + 2\lambda(1-s)\log(1-s). \tag{6.21}$$

Proof. Let c denote under $\mathbb{P}_d^{(\lambda,s)}$ the degree of $\tau^\star$'s root, and $c + \Delta$ (respectively $c + \Delta'$) the degree of the root nodes in T and T'. By the recursive formula for L_d, considering only the term for $k = c$ in the first summation as well as the injections $\sigma : [c] \to [c + \Delta]$, $\sigma' : [c] \to [c + \Delta']$ that correctly match the c children of $\tau^\star$'s root in T and T', of which there are exactly $c!$ pairs, one has:

$$L_d(T,T') \geq \psi(c, c+\Delta, c+\Delta') \times c! \times \prod_{u=1}^{c} L_{d-1}(T_u, T'_u)$$

$$\geq e^{\lambda s} \frac{s^c (1-s)^{\Delta+\Delta'}}{\lambda^c} \times \prod_{u=1}^{c} L_{d-1}(T_u, T'_u).$$

Taking logarithms and then expectations, since $\mathbb{E}_d^{(\lambda,s)}[c] = \lambda s$ and $\mathbb{E}_d^{(\lambda,s)}[\Delta] = \mathbb{E}_d^{(\lambda,s)}[\Delta'] = \lambda(1-s)$, the result follows. □

We then have

Corollary 6.4.2. *Assume that $\lambda s > 1$ and*

$$\mathrm{KL}_1 > \frac{1}{\lambda s - 1}\left[\lambda s(\log(\lambda/s) - 1) - 2\lambda(1-s)\log(1-s)\right]. \tag{6.22}$$

Then $\mathrm{KL}_\infty = +\infty$.

Proof. This follows from (6.21): indeed together with (6.22) it implies that for all $d \geq 1$,

$$\mathrm{KL}_{d+1} - \mathrm{KL}_1 \geq \lambda s(\mathrm{KL}_d - \mathrm{KL}_1),$$

hence KL_d diverges geometrically to infinity, provided we have $\mathrm{KL}_2 > \mathrm{KL}_1$. The latter property is established by writing

$$\mathrm{KL}_2 = \mathbb{E}_\infty^{(\lambda)}[\phi(L_2)] = \mathbb{E}_\infty^{(\lambda)}\left[\mathbb{E}_\infty^{(\lambda)}[\phi(L_2)|\mathcal{F}_1]\right]$$

where $\phi(x) = x\log(x)$ is strictly convex. Jensen's inequality thus guarantees $\mathrm{KL}_2 \geq \mathrm{KL}_1 = \mathbb{E}_\infty^{(\lambda)}[\phi(L_1)]$, with equality only if almost surely, $L_2 = L_1$. However this almost sure equality does not hold, hence the result. □

These results have the following consequence:

Theorem 6.4. *Assume that $\lambda \in (1, e)$. Let*

$$s^*(\lambda) := \sup\{s \in [0,1] : s(\log(\lambda/s) - 1) - 2(1-s)\log(1-s) \geq 0\}. \tag{6.23}$$

Then $s^*(\lambda) < 1$, and under the conditions

$$\lambda \in (1, e) \text{ and } s \in (s^*(\lambda), 1], \tag{6.24}$$

one-sided detectability holds.

Proof. The fact that $s^*(\lambda) < 1$ follows by continuity, since for $s = 1$ the function

$$s \to s(\log(\lambda/s) - 1) - 2(1-s)\log(1-s)$$

evaluates to $\log(\lambda) - 1$, which is negative by the assumption $\lambda < e$. By definition, for $s \in (s^*(\lambda), 1]$, the right-hand side of (6.22) is less than or equal to zero. Since $\mathrm{KL}_1 > 0$, the result is a consequence of Corollary 6.4.2 and Proposition 6.4.2. □

Remark 6.4.1. *A result similar to that of Theorem 6.4 follows from [GM20]. The present derivation is however more direct, and allows for more explicit upper bound $\lambda_0 = e$ on the range of values of λ considered, as well as characterization of the function $s^*(\lambda)$ involved.*

Condition (6.22) of Corollary 6.4.2 can also be used to identify conditions on s for one-sided testability for large values of λ, based on corresponding evaluations of KL_1. However, the resulting conditions do not appear as sharp as those obtained by the analysis of automorphisms of $\tau^\star$, that is the object of the next Section.

6.5. Number of automorphisms of Galton-Watson trees

In this Section, we show how counting automorphisms of Galton-Watson trees gives a sufficient condition for the existence of one-sided tests in the tree correlation detection problem, and provide evaluations of this number of automorphisms.

6.5.1. A lower bound on the likelihood ratio

Under $\mathbb{P}_\infty^{(\lambda,s)}$, recall that $\tau^\star$ is the *true* intersection tree used to perform correlated construction of T and T', and $\pi^\star$, $(\pi')^\star$ denote the injections from $\tau^\star$ to T and T' respectively that result from uniform shuffling of the augmentations of $\tau^\star$. Without loss of generality, we assume in this section that $\pi^\star$ and $(\pi')^\star$ are the identity map. We denote, for each $u \in \mathcal{V}_{d-1}(\tau^\star)$:

$$c_u := c_{\tau^\star}(u),\ \Delta_u := c_T(u) - c_{\tau^\star}(u),\ \Delta'_u := c_{T'}(u) - c_{\tau^\star}(u). \tag{6.25}$$

We now prove the following

Lemma 6.5.1. *Under $\mathbb{P}_d^{(\lambda,s)}$ we have the lower bound:*

$$L_d = L_d(T,T') \geq |\mathrm{Aut}(\tau^\star)| \prod_{u\in\mathcal{V}_{d-1}(\tau^\star)} \frac{s^{c_i}(1-s)^{\Delta_u+\Delta'_u}}{e^{-\lambda s}\lambda^{c_u}} \prod_{u\in\mathcal{L}_{d-1}(\tau^\star)} \binom{c_u+\Delta_u}{c_u}\binom{c_u+\Delta'_u}{c_u}, \tag{6.26}$$

where we recall that $\mathrm{Aut}(\tau^\star)$ denotes the set of tree automorphisms of $\tau^\star$.

Proof. In view of the developed expression (6.11), we can lower-bound $L_d(T,T')$ by writing

$$L_d(T,T') \geq \sum_{\substack{\tau\in\mathcal{Y}_d \\ \tau\equiv\tau^\star}} \sum_{\substack{\sigma\in\mathcal{S}(\tau,T) \\ \sigma'\in\mathcal{S}(\tau,T)}} \prod_{u\in\mathcal{V}_{d-1}(\tau)} \psi\left(c_\tau(u), c_T(\sigma(u)), c_{T'}(\sigma'(u))\right), \tag{6.27}$$

where $\equiv$ is used to denote equality up to some relabeling. Let us compute the right hand term in (6.27). Note that any tree $\tau \in \mathcal{Y}_d$ such that $\tau \equiv \tau^\star$ can be determined by a collection

$$\xi(\tau) := \left\{\xi_u(\tau) \in \mathcal{S}_{c_{\tau^\star}(u)}, u \in \mathcal{V}_{d-1}(\tau^\star)\right\},$$

giving the reordering of the children of each node of $\tau^\star$ at depth $d-1$. Moreover, the number of such permutations that produce this particular tree τ is precisely given by $|\mathrm{Aut}(\tau^\star)|$. Thus the number of trees in the summation (6.27) is precisely

$$|\{\tau \in \mathcal{Y}_d : \tau \equiv \tau^\star\}| = \frac{\prod_{u\in\mathcal{V}_{d-1}(\tau^\star)} c_{\tau^\star}(u)!}{|\mathrm{Aut}(\tau^\star)|}. \tag{6.28}$$

Note that for any tree $\tau \equiv \tau^\star$, we can construct

$$|\mathrm{Aut}(\tau^\star)|^2 \times \prod_{u\in\mathcal{L}_{d-1}(\tau^\star)} \binom{c_u+\Delta_u}{c_u}\binom{c_u+\Delta'_u}{c_u} \tag{6.29}$$

pairs of injections $(\sigma,\sigma') \in \mathcal{S}(\tau,T)\times\mathcal{S}(\tau,T')$. Indeed the factor $\binom{c_u+\Delta_u}{c_u}$ (respectively, $\binom{c_u+\Delta'_u}{c_u}$) denotes the number of subsets of the $c_u+\Delta_u$ children of u in t (respectively, of the $c_u+\Delta'_u$ children of u in t') that we can associate as children of u in the injection σ (respectively, σ'), the order in which they are considered being determined by the permutation ξ_u in ξ. We thus have the following lower bound, for any tree $\tau \equiv \tau^\star$:

$$\sum_{\substack{\sigma\in\mathcal{S}(\tau,T)\\ \sigma'\in\mathcal{S}(\tau,T)}}\prod_{u\in\mathcal{V}_{d-1}(\tau)}\psi\left(c_\tau(u),c_T(\sigma(u)),c_{T'}(\sigma'(u))\right)$$
$$\geq |\mathrm{Aut}(\tau^\star)|^2\prod_{u\in\mathcal{V}_{d-1}(\tau^\star)}\frac{s^{c_u}(1-s)^{\Delta_u+\Delta'_u}}{c_u!e^{-\lambda s}\lambda^{c_u}}\prod_{u\in\mathcal{L}_{d-1}(\tau^\star)}\binom{c_u+\Delta_u}{c_u}\binom{c_u+\Delta'_u}{c_u}. \tag{6.30}$$

Combined, (6.28) and (6.30) imply (6.26). □

We now turn to lower-bounding the number $|\mathrm{Aut}(\tau^\star)|$ of automorphisms for $\tau^\star\sim\mathsf{GW}_d^{(\lambda s)}$:

Proposition 6.5.1. *Let r be a sufficiently large constant (in particular, $r>1$). For $\tau^\star\sim\mathsf{GW}_d^{(r)}$, let us denote by w the almost sure limit:*

$$w:=\lim_{d\to\infty}\frac{1}{r^d}|\mathcal{L}_d(\tau^\star)|. \tag{6.31}$$

We place ourselves on the event on which $\tau^\star$ survives, which occurs with probability $1-p_{\mathrm{ext}}(r)>0$, and on which $w>0$. We let

$$K:=\frac{wr^d}{r-1}\cdot \tag{6.32}$$

We then have with high probability the lower bound

$$\log\left(\frac{|\mathrm{Aut}(\tau^\star)|}{\prod_{u\in\mathcal{V}_{d-1}(\tau^\star)}e^{-r}r^{c_{\tau^\star}(u)}}\right)\geq K(1-o_{\mathbb{P}}(1))\left[\frac{\log^{3/2}r}{3\sqrt{r}}+O_r\left(\frac{\log^{5/4}r}{\sqrt{r}}\right)\right]. \tag{6.33}$$

Note that hereabove, K is a high probability equivalent of $|\mathcal{V}_{d-1}(\tau^\star)|$. Proposition 6.5.1, proved in Appendix 6.B.1, could be of independent interest. We believe that a little more work could easily show that inequality (6.33) is exponentially tight, i.e. gives the right exponential order for the estimation of the number of automorphism of a Galton-Watson tree. We next show that Lemma 6.5.1 together with Proposition 6.5.1 yield a sufficient condition for the existence of one-sided test.

6.5.2. A sufficient condition for one-sided tests

We are now in a position to prove the following

Theorem 6.5. *There exists a constant r_0 such that if*

$$\lambda s>r_0\quad\text{and}\quad 1-s\leq\frac{1}{(3+\eta)}\sqrt{\frac{\log(\lambda s)}{\lambda^3 s}}, \tag{6.34}$$

for some $\eta>0$, then one-sided detectability of tree correlation holds.

Proof. The proof consists in showing that in this regime, L_d goes to $+\infty$ with positive probability under $\mathbb{P}_\infty^{(\lambda,s)}$. Throughout, X_μ will denote a Poisson random variable with parameter μ. In the lower bound (6.26) of Lemma 6.5.1, consider the factor

$$\prod_{u\in\mathcal{V}_{d-1}(\tau^\star)}\frac{s^{c_u}(1-s)^{\Delta_u+\Delta'_u}}{e^{-\lambda s}\lambda^{c_u}}\prod_{u\in\mathcal{L}_{d-1}(\tau^\star)}\binom{c_u+\Delta_u}{c_u}\binom{c_u+\Delta'_u}{c_u}.$$

Placing ourselves on the event on which $\tau^\star$ survives, reusing the notations w and K defined in equations (6.31) and (6.32), another appeal to the law of large numbers gives the following equivalents:

$$A := \log\left(\prod_{u\in\mathcal{V}_{d-1}} \frac{s^{c_u}(1-s)^{\Delta_u+\Delta'_u}}{e^{-\lambda s}\lambda^{c_u}}\right) \sim K\left(\lambda s(\log(s/\lambda)+1) + 2\lambda(1-s)\log(1-s)\right) \quad (6.35)$$

and

$$\begin{aligned} B &:= \log\left(\prod_{u\in\mathcal{L}_{d-1}(\tau^\star)} \binom{c_u+\Delta_u}{c_u}\binom{c_u+\Delta'_u}{c_u}\right) \\ &\sim w(\lambda s)^{n-1}\left(2\mathbb{E}\left[\log(X_\lambda!)\right] - 2\mathbb{E}\left[\log(X_{\lambda(1-s)}!)\right] - 2\mathbb{E}\left[\log(X_{\lambda s}!)\right]\right). \end{aligned} \quad (6.36)$$

Let us introduce the notations $r := \lambda s$, $\alpha := \lambda(1-s)$, such that $\lambda = \alpha + r$ and $s = \frac{r}{\alpha+r}$. We will identify equivalents of exponents of interest as $\alpha \to 0$ and $r \to \infty$. In this regime, (6.35) becomes

$$\begin{aligned} A &\sim K\left(-2r\log(1+\alpha/r) + 2\alpha\log\left(\frac{\alpha/r}{1+\alpha/r}\right) - r\log r + r\right) \\ &\sim K\left(-r\log r + r - 2\alpha\log r + 2\alpha\log\alpha + O(\alpha)\right) \end{aligned}$$

We have the classical estimate for large μ:

$$\mathbb{E}\left[\log(X_\mu!)\right] = \mu\log(\mu) - \mu + \frac{1}{2}\log(2\pi e\mu) + O\left(\frac{1}{\mu}\right), \quad (6.37)$$

Using (6.37) and noting that in this regime, $\mathbb{E}\left[\log(X_\alpha!)\right] = O(\alpha^2)$, (6.36) becomes

$$\begin{aligned} B &\sim 2wr^{n-1}\left((r+\alpha)\log(r+\alpha) - r - \alpha + \frac{1}{2}\log(2\pi e(r+\alpha)) - r\log(r) + r - \frac{1}{2}\log(2\pi er) - O(\alpha^2)\right) \\ &\sim 2wr^{n-1}\left(r\log(1+\alpha/r) + \alpha\log(1+\alpha/r) + \alpha\log r - \alpha + \frac{1}{2}\log(1+\alpha/r) + O(\alpha)\right) \\ &\sim 2wr^{n-1}\left(\alpha\log(r) + O(\alpha)\right). \end{aligned}$$

Combined, these approximations give:

$$\begin{aligned} A + B &\sim K\left(\left(1-\frac{1}{r}\right)\times 2\alpha\log(r) - r\log r + r - 2\alpha\log(r) + 2\alpha\log(\alpha) + O(\alpha)\right) \\ &\sim K\left(-r\log r + r + 2\alpha\log(\alpha) + O(\alpha)\right). \end{aligned} \quad (6.38)$$

Appealing to the strong law of large numbers gives

$$\begin{aligned} \log\left(\prod_{u\in\mathcal{V}_{d-1}(\tau^\star)} e^{-r}r^{c_{\tau^\star}(u)}\right) &= (1+o_{\mathbb{P}}(1))\left|\mathcal{V}_{d-1}(\tau^\star)\right|\mathbb{E}\left[-r + c_{\tau^\star}(\rho(\tau^\star))\log r\right] \\ &= (1+o_{\mathbb{P}}(1))K\left(-r + r\log(r)\right). \end{aligned} \quad (6.39)$$

Combining (6.38) and (6.39) with the results of Proposition 6.5.1 entails

$$\log L_d \geq K\left[r\log r - r + \frac{\log^{3/2}(r)}{3\sqrt{r}} + O\left(\frac{\log^{5/4} r}{\sqrt{r}}\right)\right] + K\left[-r\log r + r + 2\alpha\log(\alpha) + O(\alpha)\right]$$

$$= K\left[2\alpha\log\alpha + \frac{\log^{3/2}(r)}{3\sqrt{r}} + O\left(\frac{\log^{5/4} r}{\sqrt{r}}\right) + O(\alpha)\right].$$

Then, under assumption (6.34), we have $\alpha \le \frac{1}{3+\eta}\sqrt{\log(r)/r}$ so that, for sufficiently large r,

$$2\alpha\log\alpha + \frac{\log^{3/2}(r)}{3\sqrt{r}} > \Omega\left(\frac{\log^{3/2}(r)}{\sqrt{r}}\right).$$

It follows that on the event on which $\tau^\star$ survives, which happens with probability $1 - p_{\mathrm{ext}}(\lambda s) > 0$, under condition (6.34), L_d goes to $+\infty$ with d. Thus one-sided detectability holds. □

6.6. Impossibility of correlation detection: conjectured hard phase for partial graph alignment

In the present section we establish that, for $\lambda s^2 < 1$ and sufficiently large λ, $\mathrm{KL}_\infty < +\infty$ and hence, by Theorem 6.1, one-sided testability fails for our tree correlation problem. Since there exists a range of parameters (λ, s) for which partial alignment can be information-theoretically achieved while $\lambda s^2 < 1$ (it suffices to have $4 < \lambda s < s^{-1}$ in view of [WXY21]) we therefore conclude that the conjectured hard phase for partial graph alignment (see the conjecture at the end of Section 6.1) is non empty.

6.6.1. Mutual information formulation

Note that the Kullback-Leibler divergence KL_d also coincides with the mutual information between $T_d := p_d(T)$ and $T'_d := p_d(T')$ under $\mathbb{P}_\infty^{(\lambda,s)}$. To emphasize this interpretation we rewrite

$$\mathrm{KL}_d = I(T_d; T'_d).$$

Note that under $\mathbb{P}_\infty^{(\lambda,s)}$, conditionally on $\tau_d^\star := p_d(\tau^\star)$, T_d and T'_d are mutually independent, a property that we will depict with the dependence diagram

$$T_d \text{ —- } \tau_d^\star \text{ —- } T'_d.$$

By the data processing inequality, we thus have

$$\mathrm{KL}_d = I(T_d; T'_d) \le I(\tau_d^\star; T_d).$$

To establish that $\mathrm{KL}_\infty < \infty$, it therefore suffices to prove that $I(\tau_d^\star; T_d)$ is bounded, uniformly in d. Write then

$$\begin{aligned} I(\tau_d^\star, t_d) &= \mathbb{E}_d^{(\lambda,s)} \ln\left(\frac{\mathbb{P}_d^{(\lambda,s)}(\tau_d^\star, T_d)}{\mathbb{P}_d^{(\lambda,s)}(\tau_d^\star)\mathbb{P}_d^{(\lambda,s)}(T_d)}\right) \\ &\le \mathbb{E}_d^{(\lambda,s)}\left[\frac{\mathbb{P}_d^{(\lambda,s)}(\tau_d^\star, T_d)}{\mathbb{P}_d^{(\lambda,s)}(\tau_d^\star)\mathbb{P}_d^{(\lambda,s)}(T_d)} - 1\right] \le \mathbb{E}_d^{(\lambda,s)}\left[\frac{\mathbb{P}_d^{(\lambda,s)}(\tau_d^\star, T_d)}{\mathbb{P}_d^{(\lambda,s)}(\tau_d^\star)\mathbb{P}_d^{(\lambda,s)}(T_d)}\right]. \end{aligned}$$

We have established the bound

$$I(\tau_d^\star, T_d) \le V_d := \mathbb{E}_d^{(\lambda,s)}\left[\frac{\mathbb{P}_d^{(\lambda,s)}(\tau_d^\star | T_d)}{\mathbb{P}_d^{(\lambda,s)}(\tau_d^\star)}\right]. \tag{6.40}$$

6.6.2. Bounding the mutual information

Let us denote by c the degree of the root node in $\tau^\star$ and $c+\Delta$ the degree of the root node in T. Let us further write

$$\tau^\star = (\tau_1^\star, \dots, \tau_c^\star), \quad T = r(A(\tau_1^\star), \dots, A(\tau_c^\star), \theta_1, \dots, \theta_\Delta) = (T_1, \dots, T_{c+\Delta}),$$

where the $A(\tau_u^\star)$ are $(\lambda, s)-$augmentations, the θ_u are $\mathsf{GW}_{d-1}^{(\lambda)}$ trees, and r is a uniform relabeling. Observe that

$$\begin{aligned}
\mathbb{P}_d^{(\lambda,s)}(\tau^\star|T) &= \frac{\mathsf{GW}_d^{(\lambda s)}(\tau^\star)}{\mathsf{GW}_d^{(\lambda)}(T)} e^{-\lambda(1-s)} \frac{(\lambda(1-s))^\Delta}{\Delta!} \\
&\qquad \times \sum_{\sigma \in \mathcal{S}(c,c+\Delta)} \frac{\Delta!}{(c+\Delta)!} \prod_{u\in[c]} \mathbb{P}_{d-1}^{(\lambda,s)}(T_{\sigma(u)}|\tau_u^\star) \prod_{u=c+1}^{c+\Delta} \mathsf{GW}_{d-1}^{(\lambda)}(T_{\sigma(u)}) \\
&= \frac{e^{-\lambda s}(\lambda s)^c/c!}{e^{-\lambda}\lambda^{c+\Delta}/(c+\Delta)!} e^{-\lambda(1-s)} \frac{(\lambda(1-s))^\Delta}{\Delta!} \sum_{\sigma \in \mathcal{S}(c,c+\Delta)} \frac{\Delta!}{(c+\Delta)!} \prod_{u\in[c]} \mathbb{P}_{d-1}^{(\lambda,s)}(\tau_u^\star|T_{\sigma(u)}) \\
&= \frac{s^c(1-s)^\Delta}{c!} \sum_{\sigma \in \mathcal{S}(c,c+\Delta)} \prod_{u\in[c]} \mathbb{P}_{d-1}^{(\lambda,s)}(\tau_u^\star|T_{\sigma(u)}),
\end{aligned}$$

so that

$$\frac{\mathbb{P}_d^{(\lambda,s)}(\tau^\star|T)}{\mathbb{P}_d^{(\lambda,s)}(\tau^\star)} = \frac{s^c(1-s)^\Delta}{c!\pi_{\lambda s}(c)} \sum_{\sigma \in \mathcal{S}(c,c+\Delta)} \prod_{u\in[c]} \frac{\mathbb{P}_{d-1}^{(\lambda,s)}(\tau_u^\star|T_{\sigma(u)})}{\mathbb{P}_{d-1}^{(\lambda,s)}(\tau_u^\star)}.$$

Taking expectation entails the following formula for V_d defined in equation (6.40):

$$V_d = \sum_{c\geq 0}\sum_{\Delta\geq 0} \pi_{\lambda(1-s)}(\Delta) \frac{s^c(1-s)^\Delta}{c!} \sum_{\sigma\in\mathcal{S}(c,c+\Delta)} \mathbb{E}_{d-1}^{(\lambda,s)}\left[\prod_{i=1}^{c} \frac{\mathbb{P}_{d-1}^{(\lambda,s)}(\tau_i^\star|T_{\sigma(i)})}{\mathbb{P}_{d-1}^{(\lambda,s)}(\tau_i^\star)}\middle| c, \Delta\right]. \tag{6.41}$$

To evaluate the previous expression, we need to introduce the following notion of cycles.

Open paths, closed cycles For two integers $c, \Delta \geq 0$ and an injective mapping $\sigma \in \mathcal{S}(c, c+\Delta)$, a sequence $(i_1, \dots, i_\ell)$ of elements of $[c]$ is

- an *open path of* σ if

$$i_1 \notin \sigma([c]), \quad \forall k = 1, \dots, \ell-1,\ \sigma(i_k) = i_{k+1}, \quad \text{and } \sigma(i_\ell) \notin [c].$$

- a *closed cycle of* σ if

$$\forall k = 1, \dots, \ell-1,\ \sigma(i_k) = i_{k+1} \quad \text{and } \sigma(i_\ell) = i_1.$$

It is an easy fact to check that each injective mapping $\sigma \in \mathcal{S}(c, c+\Delta)$ can be factorized in disjoint open paths and closed cycles. Since each term i in the product in (6.41) only depends on the other terms j in its own open path (resp. closed cycle), the expectation term in (6.41) factorizes according to the path/cycle decomposition of σ.

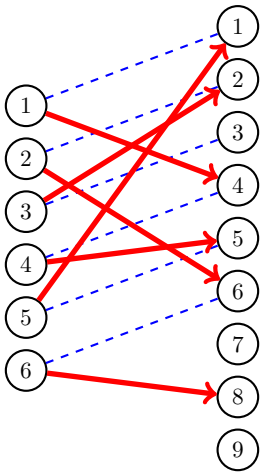


Figure 6.8 – *Representation of* $\sigma \in \mathcal{S}(c, c+\Delta)$ *with* $c = 6, \Delta = 3$, *and* $\sigma(1) = 4$, $\sigma(2) = 6$, $\sigma(3) = 2$, $\sigma(4) = 5$, $\sigma(5) = 1$, $\sigma(6) = 8$. *In this example,* $(1, 4, 5)$ *(resp.* $(3, 2, 6)$*) is an open path (resp. closed path) of* σ.

First consider an open path O_ℓ of σ of length ℓ, assumed without loss of generality to be given by $(1, \dots, \ell)$, so that $\sigma(1) = 2, \dots, \sigma(\ell-1) = \ell$, and $\sigma(\ell) = c+1$. The expectation of the corresponding factor reads:

$$\mathbb{E}_{d-1}^{(\lambda,s)}\left[\prod_{i\in O_\ell} \frac{\mathbb{P}_{d-1}^{(\lambda,s)}(\tau_i^\star|T_{\sigma(i)})}{\mathbb{P}_{d-1}^{(\lambda,s)}(\tau_i^\star)}\right] = \mathbb{E}_{d-1}^{(\lambda,s)}\left[\prod_{k=1}^{\ell-1} \frac{\mathbb{P}_{d-1}^{(\lambda,s)}(\tau_k^\star|A(\tau_{k+1}^\star))}{\mathbb{P}_{d-1}^{(\lambda,s)}(\tau_k^\star)} \times \frac{\mathbb{P}_{d-1}^{(\lambda,s)}(\tau_\ell^\star|\theta_1)}{\mathbb{P}_{d-1}^{(\lambda,s)}(\tau_\ell^\star)}\right].$$

Now integrated over θ_1, $\mathbb{P}_{d-1}^{(\lambda,s)}(\tau_\ell^\star|\theta_1)$ evaluates to $\mathbb{P}_{d-1}^{(\lambda,s)}(\tau_\ell^\star)$ and the last factor disappears. Integrating then successively with respect to $A(\tau_k^\star)$, $k = \ell, \ell-1, \dots, 2$, we obtain that the factors corresponding to open cycles evaluate to 1.

Consider next a closed cycle C_ℓ of σ of length ℓ. Assuming without loss of generality that $T_i = A(\tau_i^\star)$, the expectation reads

$$\mathbb{E}_{d-1}^{(\lambda,s)}\left[\prod_{i\in C_\ell} \frac{\mathbb{P}_{d-1}^{(\lambda,s)}(\tau_i^\star|T_{\sigma(i)})}{\mathbb{P}_{d-1}^{(\lambda,s)}(\tau_i^\star)}\right] = \mathbb{E}_{d-1}^{(\lambda,s)}\left[\prod_{k=1}^{\ell} \frac{\mathbb{P}_{d-1}^{(\lambda,s)}(\tau_k^\star|A(\tau_{(k+1) \bmod \ell}^\star))}{\mathbb{P}_{d-1}^{(\lambda,s)}(\tau_k^\star)}\right].$$

This reads, using for $t, \tau \in \mathcal{Y}_{d-1}$ the notations $p_{d-1}(t) := \mathsf{GW}_{d-1}^{(\lambda)}(t)$, $q_{d-1}(\tau|t) := \mathbb{P}_{d-1}^{(\lambda,s)}(\tau|t)$, $r_{d-1}(\tau) := \mathsf{GW}_{d-1}^{(\lambda s)}(\tau)$:

$$\sum_{\tau_1,t_1,\dots,\tau_\ell,t_\ell\in\mathcal{Y}_{d-1}} \prod_{i\in[\ell]} p_{d-1}(t_i) q_{d-1}(\tau_i|t_i) \times \frac{q_{d-1}(\tau_i|t_{(i+1) \bmod \ell})}{r_{d-1}(\tau_i)}. \tag{6.42}$$

Introduce the operator Ψ_{d-1}, indexed by trees in $\mathcal{Y}_{d-1}$:

$$\Psi_{d-1}(\tau_1, \tau_2) := \sum_{t\in\mathcal{Y}_{d-1}} p_{d-1}(t) \frac{q_{d-1}(\tau_1|t) q_{d-1}(\tau_2|t)}{\sqrt{r_{d-1}(\tau_2) r_{d-1}(\tau_2)}}. \tag{6.43}$$

M is symmetric and semi-definite positive, hence the operator is diagonalizable and its spectrum lies in $\mathbb{R}_+$. Note that the expectation in (6.42) coincides with the trace of matrix Ψ_{d-1}^ℓ.

It follows that [6]

$$\mathbb{E}_{d-1}^{(\lambda,s)}\left[\prod_{i\in C_\ell}\frac{\mathbb{P}_{d-1}^{(\lambda,s)}(\tau_i^\star|T_{\sigma(i)})}{\mathbb{P}_{d-1}^{(\lambda,s)}(\tau_i^\star)}\right]=\mathrm{Tr}(\Psi_{d-1}^\ell)\le\mathrm{Tr}(\Psi_{d-1})^\ell=V_{d-1}.$$

We now have the ingredients in place to prove the following

Lemma 6.6.1. *The quantity V_d verifies*

$$V_d\le f(V_{d-1}),\tag{6.44}$$

where

$$f(x)=\frac{1}{1-sx}\exp\left(\frac{\kappa(1-s)(x-1)}{1-sx}\right)\tag{6.45}$$

with $\kappa:=\lambda s^2$.

Proof. For given $c,\Delta\ge 0$ and an injection $\sigma\in\mathcal{S}(c,c+\Delta)$, let $F(\sigma)$ denote the number of elements $i\in[c]$ that belong to closed cycles of σ. From the previous evaluations (6.41) – (6.43) we already have obtained the bound

$$V_d\le\sum_{c,\Delta\ge0}\pi_{\lambda(1-s)}(\Delta)\frac{s^c(1-s)^\Delta}{c!}\sum_{\sigma\in\mathcal{S}(c,c+\Delta)}V_{d-1}^{F(\sigma)}.$$

To upper-bound this quantity, we use the facts that $V_{d-1}\ge 1$ and that $F(\sigma)\le|[c]\cap\sigma([c])|$. Then, for any $0\le k\le c$, there are $\binom{c}{k}\binom{\Delta}{c-k}$ ways to chose the set $\sigma([c])$ such that $|[c]\cap\sigma([c])|=k$, and $c!$ distinct injections σ with the same set $\sigma([c])$. Hence $V_d\le f(V_{d-1})$ with

$$\begin{aligned}
f(x)&:=\sum_{c,\Delta\ge0}e^{-\lambda(1-s)}\frac{s^c(\lambda(1-s)^2)^\Delta}{\Delta!}\sum_{k=0}^{c}\binom{c}{k}\binom{\Delta}{c-k}x^k\\
&=e^{-\lambda(1-s)}\sum_{k,\Delta\ge0}x^k\frac{(\lambda(1-s)^2)^\Delta}{\Delta!}\sum_{c\ge k}\binom{c}{k}\binom{\Delta}{c-k}s^c\\
&=e^{-\lambda(1-s)}\sum_{k,\Delta\ge0}x^k\frac{(\lambda(1-s)^2)^\Delta}{\Delta!}s^k\sum_{c=0}^{\Delta}\binom{c+k}{k}\binom{\Delta}{c}s^c\\
&=e^{-\lambda(1-s)}\sum_{k,c\ge0}\frac{1}{c!}\binom{c+k}{k}(sx)^k s^c(\lambda(1-s)^2)^c\sum_{\Delta\ge c}\frac{(\lambda(1-s)^2)^{\Delta-c}}{(\Delta-c)!}\\
&=e^{-\lambda s(1-s)}\sum_{c\ge0}\frac{(\lambda s(1-s)^2)^c}{c!}\sum_{k\ge0}\binom{c+k}{k}(sx)^k\\
&=e^{-\lambda s(1-s)}\frac{1}{1-sx}\sum_{c\ge0}\frac{1}{c!}\left(\frac{\lambda s(1-s)^2}{1-sx}\right)^c=\frac{1}{1-sx}\exp\left(\frac{\lambda s^2(1-s)(x-1)}{1-sx}\right).
\end{aligned}$$

□

We are now in a position to prove the following

[6] To make this argument fully rigorous we can consider truncated summations so that we are dealing with finite dimensional matrices, for which the trace inequality to follow clearly holds, and then use monotone convergence to obtain the desired inequality as written.

Theorem 6.6. *Assume $\kappa = \lambda s^2$ is fixed such that $\kappa < 1$. Then for λ sufficiently large, it holds that*

$$\limsup_{d\to\infty} V_d < +\infty, \tag{6.46}$$

so that one-sided testability fails.

Proof. Let $\kappa < 1$ be fixed, together with $\varepsilon \in (0, 4\kappa)$ such that $\kappa + \varepsilon < 1$. Let $\gamma > 0$ be an arbitrary constant chosen such that

$$\gamma > \frac{1}{1-\kappa-\varepsilon}.$$

We shall consider $s > 0$ sufficiently small, or equivalently λ large enough, in particular such that $\gamma s < 1$. Let $y \in [0, \gamma s]$. Note that

$$\exp\left(\kappa \frac{y(1-s)}{1-s(y+1)}\right) \leq \exp\left(\kappa y/(1-2s)\right).$$

Then, assuming $\frac{1}{1-2s} \leq 1 + \varepsilon/(4\kappa)$ as well as $2e^2\kappa^2\gamma s/\varepsilon \leq 1$, we get

$$\exp\left(\kappa y/(1-2s)\right) \leq \exp\left(\kappa y + \varepsilon y/4\right) \leq 1 + (\kappa + \varepsilon/2)y. \tag{6.47}$$

Note also that, $1/(1-t) \leq 1 + t + 3t^2$ for $t \in (0, 2/3)$. Assuming $s(y+1) \leq 2s < 2/3$, and using $y \leq \gamma s \leq 1$, we get

$$\frac{1}{1-s(y+1)} \leq 1 + s(y+1) + 3[s(y+1)]^2 \leq 1 + s + Cs^2, \tag{6.48}$$

where $C := \gamma + 12$. Together, these last two bounds (6.47) and (6.48) entail, for any $y \in [0, \gamma s]$:

$$f(1+y) - 1 \leq (1 + (\kappa + \varepsilon/2)y)(1 + s + Cs^2) - 1 \leq s + Cs^2 + (\kappa + \varepsilon)y,$$

where we assumed s sufficiently small that $(\kappa + \varepsilon/2)(1 + s + Cs^2) \leq \kappa + \varepsilon$. Note now that, provided

$$1 + (\gamma + 12)s + (\kappa + \varepsilon)\gamma \leq \gamma,$$

it holds that

$$f(1+y) - 1 \in [0, \gamma s]. \tag{6.49}$$

Note that this condition can be enforced, for any choice of γ such that $\gamma > \frac{1}{1-\kappa-\varepsilon}$ taking s sufficiently small.

By induction on d, monotonicity of f (which is easily obtained from the series expansion of f), and the initialization $V_0 = 1$, it follows from (6.49) that for sufficiently small s one has:

$$V_d - 1 \leq (s + Cs^2) \sum_{i=0}^{d-1} (\kappa + \varepsilon)^i.$$

Since the right-hand side is uniformly bounded in d, the result follows. □

6.7. Consequences for polynomial time partial graph alignment

We now apply the previous results of Sections 6.3 – 6.5 to one-sided partial graph alignment. We will now describe our polynomial-time algorithm and its theoretical guarantees when one-sided detectablity holds in Theorem 6.1 – in particular under condition (6.24) of Theorem 6.4 or condition (6.34) of Theorem 6.5.

6.7.1. Intuition, algorithm description

In all this part we assume that (λ, s) satisfy one of the conditions in Theorem 6.1.

Extending the tree correlation detection problem Let (G, H) be a pairs of relabeled $\mathsf{G}(n, \lambda/n, s)$ graphs, with underlying alignment $\pi^\star$. As done in Chapter 5, in order to distinguish matched pairs of nodes (u, u'), we consider their neighborhoods $\mathcal{N}_{d,G}(u)$ and $\mathcal{N}_{d,H}(u')$ at a given depth d: these neighborhoods are close to Galton-Watson trees. In the case where the two vertices are actual matches, i.e. $u' = \pi^\star(u)$, we are exactly in the setting of our tree correlation detection problem under $\mathbb{P}_d^{(\lambda,s)}$: Point (v) of in Theorem 6.1 shows that there exists a threshold β_d such that with probability at least $1 - p_{\mathrm{ext}}(\lambda s) > 0$,

$$L_d(u, u') := L_d\left(\mathcal{N}_{d,G}(u), \mathcal{N}_{d,H}(u')\right) > \beta_d,$$

when $d \to \infty$. Point (v) of Theorem 6.1 shows that this threshold β_d can be e.g. taken to be $\exp(n^\gamma)$ for some $\gamma \in (0, c\log(\lambda s))$.

At the same time, when nodes u' and $\pi^\star(u)$ are distinct and sufficiently far away, we can argue that we are also – with high probability – in the setting of the tree correlation detection problem under $\mathbb{P}_d^{(\lambda)}$: since $\mathbb{E}_d^{(\lambda)}[L_d] = 1$, Markov's inequality shows that with high probability when $d \to \infty$,

$$L_d(u, u') \le \beta_d.$$

Computation of the likelihood ratios As mentioned in Remark 6.3.1, Formula (6.10) enables to compute such likelihood ratios efficiently on a graph, giving the exact expression for a *message passing* procedure, assuming that all neighborhoods are locally tree-like at depth d. Let us first define *oriented likelihood ratios*: for any $u, v \in V(G)$ and $u', v' \in V(H)$, we write $L_d(u \leftarrow v, u' \leftarrow v')$ for the likelihood ratio at depth d of two trees, the first one (resp. second one) being rooted at u in G (resp. u' in H) where the edge $\{u, v\}$ (resp. $\{u', v'\}$), if initially present, has been deleted. In view of (6.10) these oriented likelihood ratios satisfy the following recursion:

$$L_d(u \leftarrow v, u' \leftarrow v') = \sum_{k=0}^{d_u \wedge d'_{u'} - 1} \psi\left(k, d_u - 1, d'_{u'} - 1\right) \sum_{\substack{\sigma \in \mathcal{S}([k], \mathcal{N}_G(u)\setminus\{v\}) \\ \sigma' \in \mathcal{S}([k], \mathcal{N}_H(u')\setminus\{v'\})}} \prod_{\ell=1}^{k} L_{d-1}(\sigma(\ell) \leftarrow u, \sigma'(\ell) \leftarrow u'), \quad (6.50)$$

where $d_u := d_G(u)$ and $d'_{u'} := d_H(u')$. The likelihood ratio at depth d between u and u' is then obtained by computing

$$L_d(u, u') = \sum_{k=0}^{d_u \wedge d'_{u'}} \psi\left(k, d_u, d'_{u'}\right) \sum_{\substack{\sigma \in \mathcal{S}([k], \mathcal{N}_G(u)) \\ \sigma' \in \mathcal{S}([k], \mathcal{N}_H(u'))}} \prod_{\ell=1}^{k} L_{d-1}(\sigma(\ell) \leftarrow u, \sigma'(\ell) \leftarrow u'). \quad (6.51)$$

A natural idea is then to compute for each pair (u, u') the likelihood ratio $L_d(u, u')$ with d large enough (typically scaled in $\Theta(\log n)$ where n is the number of vertices in G and H) and to compare it to β_d to decide whether u in G is matched to u' in H.

A refined *dangling trees trick* However, as previously noted in Chapter 5, without additional constraint, this strategy produces many falsely positive matches, tending e.g. to match u with u' if there exists v such that $\{u, v\}$ is an edge of G and $\{u', \pi^\star(v)\}$ is an edge of H, making the errors increase and the performance collapse.

To fix this issue, we use the *dangling trees trick*, already introduced in [GM20], and

improved here by considering three rather than two dangling trees: instead of just looking at their neighborhoods, we look for the downstream trees from distinct neighbors of u in G and of u' in H. The trick is now to match u with u' if and only if there exists three distinct neighbors v, w, x of u in G (resp. v', w', x' of u' in H) such that all three of the likelihood ratios $L_{d-1}(v \leftarrow u, v' \leftarrow u')$, $L_{d-1}(w \leftarrow u, w' \leftarrow u')$ and $L_{d-1}(x \leftarrow u, x' \leftarrow u')$ are larger than β. The proof of Theorem 6.8 explains how this trick avoids false positives and why three dangling trees is a good choice.

Algorithm description Our algorithm is as follows:

Algorithm 6.1: `MPAlign`: Message-passing algorithm for sparse graph alignment

Input: Two graphs G and H of size n, average degree λ, depth d, threshold parameter β
Output: A set of pairs $\mathcal{M} \subset V(G) \times V(H)$.
$\mathcal{M} \leftarrow \varnothing$
Compute $L_d(u \leftarrow v, u' \leftarrow v')$ for all $\{u, v\} \in E$ and $\{u', v'\} \in E'$ with (6.50)
for $(u, u') \in V(G) \times V(H)$ **do**
  if $\mathcal{N}_G(u, d)$ *and* $\mathcal{N}_H(u', d)$ *contain no cycle, and* $\exists \{v, w, x\} \subset \mathcal{N}_G(u), \exists \{v', w', x'\} \subset \mathcal{N}_G(u')$ *such that* $L_{d-1}(v \leftarrow u, v' \leftarrow u') > \beta$, $L_{d-1}(w \leftarrow u, w' \leftarrow u') > \beta$ *and* $L_{d-1}(x \leftarrow u, x' \leftarrow u') > \beta$ **then**
    $\mathcal{M} \leftarrow \mathcal{M} \cup \{(u, u')\}$
  end
end
return $\mathcal{M}$

Remark 6.7.1. *To update the matrix of all likelihood ratios with* (6.50)*, we update a matrix of size* $O(n^2)$*, each entry of which can be computed in time* $O\left((d_{\max}!)^2\right)$ *– where* $d_{\max}$ *is the maximum degree in* G *and* H*. Under the correlated Erdős-Rényi model,* $d_{\max} = O\left(\frac{\log n}{\log \log n}\right)$ *[Bol01], so that* $d_{\max}!$ *is polynomial in* n*. Each iteration is thus polynomial in* n *and since* d *is taken order* $\log(n)$*,* `MPAlign` *(Algorithm 6.1) runs in polynomial time.*

We now state two results, of the same flavour as Theorems 5.4 and 5.5 in Chapter 5 for `NTMA`, which will readily imply Theorem 6.2.

Theorem 6.7. *Let* (G, H) *be drawn under the planted model with correlated* $\mathsf{G}(n, \lambda/n, s)$ *graphs such that any of the equivalent conditions of Theorem 1 holds. Let* $d = \lfloor c \log n \rfloor$ *with* $c \log(\lambda(2 - s)) < 1/2$*. Let* $\mathcal{M}$ *be the output of Alg. 6.1, taking* $\beta = \exp(n^\gamma)$ *for some* $\gamma \in (0, c \log(\lambda s))$*. Then with high probability*

$$\frac{1}{n} \sum_{u=1}^{n} \mathbb{1}_{\{(u, \pi^\star(u)) \in \mathcal{M}\}} \geq \Omega(1). \tag{6.52}$$

In other words, a non vanishing fraction of nodes is correctly recovered by Algorithm 6.1.

Theorem 6.8. *Let* $(G, G') \sim \mathsf{G}(n, \lambda/n, s)$ *be two* $s-$*correlated Erdős-Rényi graphs. Assume that* $d = \lfloor c \log n \rfloor$ *with* $c \log \lambda < 1/4$*. Let* $\mathcal{M}$ *be the output of Alg. 6.1, taking* $\beta = \exp(n^\gamma)$ *for some* $\gamma \in (0, c \log(\lambda s))$*. Then with high probability*

$$\mathrm{err}(n) := \frac{1}{n} \sum_{u=1}^{n} \mathbb{1}_{\{\exists u' \neq \pi^\star(u),\ (u, u') \in \mathcal{M}\}} = o(1), \tag{6.53}$$

i.e. only a vanishing fraction of nodes are incorrectly matched by Algorithm 6.1.

Remark 6.7.2. *The set* $\mathcal{M}$ *returned by Algorithm 6.1 is not necessarily an injective mapping. Let* $\mathcal{M}'$ *be obtained by removing all pairs* (u, u') *of* $\mathcal{M}$ *such that* i *or* u *appears at least twice.*

Theorems 6.7 and 6.8 guarantee that $\mathcal{M}'$ still contains a non-vanishing number of correct matches and a vanishing number of incorrect matches, hence one-sided partial alignment holds. Theorem 6.2 easily follows, the proposed local algorithm achieving one-sided partial graph alignment.

A slight adaptation of `MPAlign` *(Alg. 6.1),* `MPAlign2` *(Alg. 6.2), can be found in Appendix 6.A, where some results are also reported. These confirm our theory, as the algorithm returns many good matches and few mismatches. A similar algorithm has been recently studied in [PSSZ21].*

6.7.2. Proof strategy

We start by recalling Lemmas that precise the link between sparse graph alignment and correlation detection in trees, as explained in Section 6.7.1. These Lemmas are directly taken from Chapter 5 (to which we refer for the proofs, see Lemmas 5.3.1, 5.3.2, 5.3.3 and 5.3.4) and are instrumental in the proofs of Theorems 6.7 and 6.8.

Lemma 6.7.1 (Control of the sizes of the neighborhoods)**.** *Let $G \sim \mathsf{G}(n, \lambda/n)$, $d = \lfloor c \log n \rfloor$ with $c \log \lambda < 1$. For all $\gamma > 0$, there is a constant $C = C(\gamma) > 0$ such that with probability $1 - O(n^{-\gamma})$, for all $u \in [n]$, $t \in [d]$:*

$$|\mathcal{S}_G(u,t)| \leq C(\log n)\lambda^t. \tag{6.54}$$

Lemma 6.7.2 (Cycles in the neighborhoods in an Erdős-Rényi graph)**.** *Let $G \sim \mathsf{G}(n, \lambda/n)$, $d = \lfloor c \log n \rfloor$ with $c \log \lambda < 1/2$. Then there exists $\varepsilon > 0$ such that for any vertex $u \in [n]$, one has*

$$\mathbb{P}\left(\mathcal{N}_{G,d}(u) \text{ contains a cycle}\right) = O\left(n^{-\varepsilon}\right). \tag{6.55}$$

Lemma 6.7.3 (Two neighborhoods are typically independent)**.** *Let $G \sim \mathsf{G}(n, \lambda/n)$ with $\lambda > 1$, $d = \lfloor c \log n \rfloor$ with $c \log \lambda < 1/2$. Then there exists $\varepsilon > 0$ such that for any fixed nodes $u \neq v$ of G, the total variation distance between the joint distribution of the neighborhoods $\mathcal{L}\left((\mathcal{S}_G(u,t), \mathcal{S}_G(v,t))_{t \leq d}\right)$ and the product distribution $\mathcal{L}\left((\mathcal{S}_G(u,t))_{t \leq d}\right) \otimes \mathcal{L}\left((\mathcal{S}_G(v,t))_{t \leq d}\right)$ tends to 0 as $O(n^{-\varepsilon})$ when $n \to \infty$.*

Lemma 6.7.4 (Coupling neighborhoods with Galton-Watson trees)**.** *We have the following couplings:*

- *(i) Let $G \sim \mathsf{G}(n, \lambda/n)$, $d = \lfloor c \log n \rfloor$ with $c \log \lambda < 1/2$. Then there exists $\varepsilon > 0$ such that for any fixed node u of G, the variation distance between the distribution of $\mathcal{N}_{G,d}(u)$ and the distribution $\mathsf{GW}_d^{(\lambda)}$ tends to 0 as $O(n^{-\varepsilon})$ when $n \to \infty$.*
- *(ii) For (G, H) two correlated $\mathsf{G}(n, \lambda/n, s)$ graph with planted alignment $\pi^\star$, $d = \lfloor c \log n \rfloor$ with $c \log(\lambda s) < 1/2$ and $c \log(\lambda(1-s)) < 1/2$, there exists $\varepsilon > 0$ such that for any fixed node u of G, the variation distance between the distribution of $(\mathcal{N}_{G,d}(u), \mathcal{N}_{H,d}(\pi^\star(u)))$ and the distribution $\mathbb{P}_d^{(\lambda,s)}$ (as defined in Section 6.2.2) tends to 0 as $O(n^{-\varepsilon})$ when $n \to \infty$.*

Proof of Theorems 6.7 and 6.8

Proof of Theorem 6.7. First, since $c \log(\lambda(2-s)) < 1/2$, we also have $c \log(\lambda(1-s)) < 1/2$ and $c \log(\lambda s) < 1/2$. For $i \in [n]$, point (ii) of Lemma 6.7.4 implies that the two neighborhoods $\mathcal{N}_{G,d}(u)$ and $\mathcal{N}_{H,d}(\pi^\star(u))$ can be coupled with trees drawn under $\mathbb{P}_d^{(\lambda,s)}$ as defined in Section 6.2.2 with probability $\geq 1 - O(n^{-\varepsilon})$.

Under this coupling, there is a probability $\alpha_3 > 0$ that the root in the intersection tree has at least three children, and since we work under the conditions of Theorem 6.1 point (v) implies that the three likelihood ratios are greater than β with positive probability

$(1-p_{\mathrm{ext}}(\lambda s))^3 > 0$. Hence, the probability of $M_u := \{(u,\pi^\star(u)) \in \mathcal{M}\}$ is at least $(1-o(1))\alpha_3(1-p_{\mathrm{ext}}(\lambda s))^3 =: \alpha > 0$.

Let $G_\cup$ be the true union graph, that is $G_\cup := G^{\pi^\star} \cup H$ where $G^{\pi^\star}$ is the relabeling of G according to permutation $\pi^\star$. We have $G_\cup \sim \mathsf{G}(n, \lambda(2-s)/n)$. For $u \neq v \in [n]$, define $I_{u,v}$ the event on which the two neighborhoods of u and v in $G_\cup$ coincide with their independent couplings up to depth d. Since $c\log(\lambda(2-s)) < 1/2$, by Lemma 6.7.3, $\mathbb{P}(I_{u,v}) = 1-o(1)$. Then for $0 < \varepsilon < \alpha$, Markov's inequality yields

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}\sum_{u=1}^n \mathbb{1}_{\{(u,\pi^\star(u))\in\mathcal{M}\}} < \alpha-\varepsilon\right) &\leq \mathbb{P}\left(\sum_{u=1}^n (\mathbb{P}(M_u) - \mathbb{1}_{M_u}) > \varepsilon n\right)\\ &\leq \frac{1}{n^2\varepsilon^2}\left(n\mathrm{Var}\left(\mathbb{1}_{M_1}\right) + n(n-1)\mathrm{Cov}\left(\mathbb{1}_{M_1},\mathbb{1}_{M_2}\right)\right)\\ &\leq \frac{\mathrm{Var}\left(\mathbb{1}_{M_1}\right)}{n\varepsilon^2} + \frac{1-\mathbb{P}\left(I_{1,2}\right)}{\varepsilon^2} \to 0,\end{aligned}$$

which ends the proof. □

Remark 6.7.3. *Note that in view of the proof here above, the recovered fraction $\Omega(1)$ guaranteed by in Theorem 6.7 can be taken as close as wanted to*

$$\alpha(\lambda s) := (1-p_{\mathrm{ext}}(\lambda s))^3\left(1-\pi_{\lambda s}(0)-\pi_{\lambda s}(1)-\pi_{\lambda s}(2)\right).$$

This fraction is a priori not optimal, and can be interestingly compared with recent results in [GML21b] (Chapter 4) showing that no more than a fraction $1-p_{\mathrm{ext}}(\lambda s)$ of the nodes can be recovered.

Proof of Theorem 6.8. First, we condition on the event $\mathcal{A}$ that all $d-$neighborhoods in G and H are of size at most $C(\log n)\lambda^d$, which happens with probability $1-o(1)$ by Lemma 6.7.1. Note that by assumption this uniform upper bound is $O((\log n)n^{1/4})$.

In order to control the probability that u is matched with some 'wrong' $u' \neq \pi^\star(u)$ by our algorithm, we follow the same first steps as in the proof of Theorem 5.5 of Chapter 5: we will first fix u in G and work on the event $\mathcal{E}_u$ where $\mathcal{N}_{G_\cup,2d}(u)$ has no cycle. Since $c\log(\lambda) < 1/4$, this event happens with probability $1-o(1)$ by Lemma 6.7.2.

Consider then u' in H such that $u' \neq \pi^\star(u)$. If u and u' are matched by `MPAlign`, then necessarily $\mathcal{N}_G(u,d)$ and $\mathcal{N}_H(u',d)$ contain no cycle: the $d-$neighborhoods are thus tree-like. For any choice of distinct neighbors v,w,x of u in G (resp. v',w',x' of u' in H), we define the corresponding pairs of trees of the form (T_ℓ, T'_ℓ), where T_ℓ (resp. T'_ℓ) is the tree of depth $d-1$ rooted at $\ell \in \{v,w,x\}$ in G (resp. $\ell \in \{v',w',x'\}$ in H) after deletion of edge $\{u,\ell\}$ in G (resp. $\{u',\ell\}$ in H). A moment of thought shows that, no matter the choice of v,w,x and v',w',x', on event $\mathcal{E}_i$, one of these three pairs (T_ℓ, T'_ℓ) must be made of *two disjoint trees*.

We now focus on a pair (T,T') of such disjoint trees: these trees of depth $d-1$ can be built recursively by sampling a binomial number of children for each vertex. Since we condition on the fact that the trees are not intersecting, if at some point k vertices have been uncovered, then the number of children to be drawn is exactly of distribution $\mathsf{Bin}\,(n-k,\lambda/n)$. With this exact construction, we denote by $\widetilde{\mathbb{P}}_d$ the distribution of the pair (T,T'). Define

$$M_{d-1} := \frac{\widetilde{\mathbb{P}}_{d-1}(T,T')}{\mathbb{P}^{(\lambda)}_{d-1}(T,T')}. \tag{6.56}$$

We have that

$$\begin{aligned}\widetilde{\mathbb{P}}_{d-1}(L_{d-1}(T,T') > \beta \cap \mathcal{A}) &= \mathbb{E}^{(\lambda)}_{d-1}\left[M_{d-1}\times\mathbb{1}_{\mathcal{A}}\times\mathbb{1}_{L_{d-1}(T,T')>\beta}\right]\\ &\leq \mathbb{E}^{(\lambda)}_{d-1}[M^2_{d-1}\mathbb{1}_{\mathcal{A}}]^{1/2}\beta^{-1/2},\end{aligned}$$

by a successive use of Cauchy-Schwarz and Markov's inequalities, using that $\mathbb{E}_{d-1}^{(\lambda)}\left[L_{d-1}(T,T')\right] = 1$. We now state the following Lemma, proved in Appendix 6.B.3:

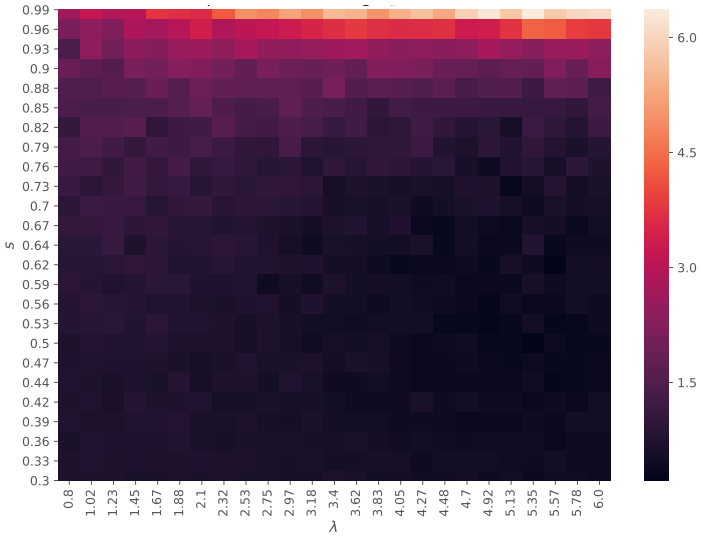
Lemma 6.7.5. *With the previous notations, we have*

$$\mathbb{E}_{d-1}^{(\lambda)}\left[M_{d-1}^2 \mathbb{1}_{\mathcal{A}}\right] = O(1). \tag{6.57}$$

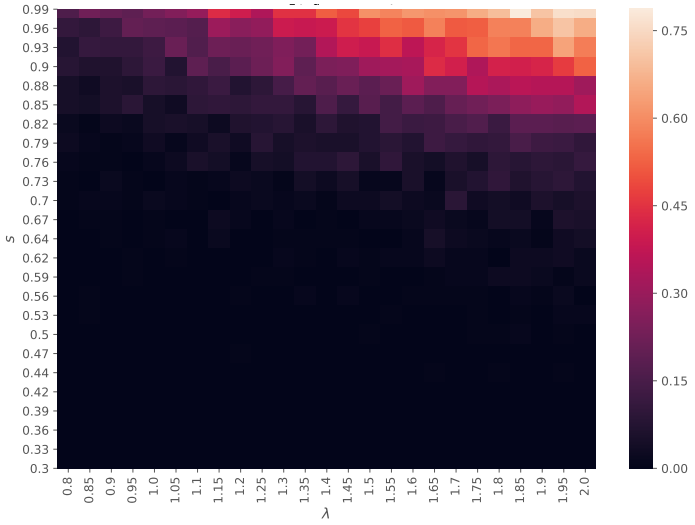
Together with the previous Lemma, noting that with high probability the maximum degree in G and H is less than $\log n$, union bound gives

$$\begin{aligned}\mathbb{P}\left(\mathcal{A} \cap \left\{\exists u' \neq \pi^\star(u),\ (u,u') \in \mathcal{M}\right\}\right) &\leq \mathbb{P}(\bar{\mathcal{E}}_i) + o(1) + n \times \log^6 n \times \beta^{-1/2} \\ &= O\left((\log^6 n) \times n \times \exp(-n^{\gamma/2})\right) = o(1).\end{aligned}$$

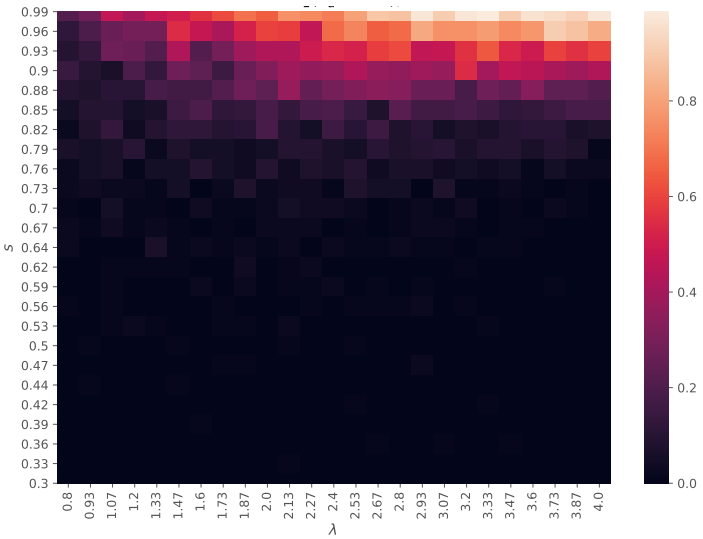
The proof follows by appealing to Markov's inequality. □



(a) – *Empirical mean of* $\log L_d$ *for* $d = 2$. 75 *simulations per value of* (λ, s).



(b) – *Estimate of* $\mathbb{P}_d^{(\lambda,s)}(L_d > \beta)$ *for* $d = 3$, $\beta = 10^2$. 75 *simulations per value of* (λ, s).



(c) – *Estimate of* $\mathbb{P}_d^{(\lambda,s)}(L_d > \beta)$ *for* $d = 5$, $\beta = 5.10^3$. 150 *simulations per value of* (λ, s).

Figure 6.7 – *Simulations of* L_d *under model* $\mathbb{P}_s^{(\lambda,s)}$.

Appendix of Chapter 6

6.A. Numerical experiments for `MPAlign2`

In this section, we give some details on a practical implementation of our algorithm. We start by introducing some notations. Given an edge $\{u, v\}$ of a graph, we denote by $u \to v$ and $v \to u$ the associated directed edges. Now given two graphs $G = (V, E)$ and $H = (V', E')$, we define the matrix $(m^t_{u\to v, u'\to v'})_{\{u,v\}\in E, \{u',v'\}\in E'} \in \mathbb{R}_+^{2|E|\times 2|E'|}$ recursively in t, as follows:

$$m^{t+1}_{u\to v,u'\to v'} = \sum_{k=0}^{d_u\wedge d_{u'}-1} \widetilde{\psi}(k, d_u - 1, d_{u'} - 1) \sum_{\substack{\{\ell_1,\dots\ell_k\}\in\partial u\backslash v \\ \{w_1,\dots w_k\}\in\partial u'\backslash v'}} \sum_{\sigma\in\mathcal{S}_k} \prod_{a=1}^{k} m^t_{\ell_a\to u, w_{\sigma(a)}\to u'}, \quad (6.58)$$

where $d_u := d_G(u)$, $d'_{u'} := d_H(u')$, $\widetilde{\psi}(k, d_1, d_2) = k!\psi(k, d_1, d_2)$, and $\partial u\backslash v$ (resp. $\partial u'\backslash v'$) is a shorthand notation for $\mathcal{N}_G(u)\setminus\{v\}$ (resp. $\mathcal{N}_H(u')\setminus\{v'\}$) and by convention $m^0_{u\to v,u'\to v'} = 1$.

Denoting $\partial u := \mathcal{N}_G(u)$ (resp. $\partial u' := \mathcal{N}_H(u')$), for $t \in \mathbb{N}$ we define the matrix $(m^t_{u,u'}) \in \mathbb{R}_+^{V\times V'}$ as follows:

$$m^t_{u,u'} = \sum_{k=0}^{d_u\wedge d_{u'}} \widetilde{\psi}(k, d_u, d_{u'}) \sum_{\substack{\{\ell_1,\dots\ell_k\}\in\partial u \\ \{w_1,\dots w_k\}\in\partial u'}} \sum_{\sigma\in\mathcal{S}_k} \prod_{a=1}^{k} m^t_{\ell_a\to u, w_{\sigma(a)}\to u'}. \quad (6.59)$$

It is easy to see that if the graphs G and H are tree-like up to depth t, then $m^t_{u,u'}$ is exactly the likelihood ratio $L_t(s_u, s'_{u'})$ where s_u (resp. $s'_{u'}$) is the tree neighborhood of u in G (resp. of u' in H).

In experiments, we run our algorithms on correlated Erdős-Rényi model with possible cycles, so that the matrix $m^t_{u,u'}$ is interpreted as an approximation of the true likelihood ratio. From such an approximation, we compute two mappings $\pi^t : V \to V'$ as

$$\pi^t(u) = \arg\max(m^t_{u,\cdot})$$

and $\sigma^t : V' \to V$ as

$$\sigma^t(u') = \arg\max(m^t_{\cdot,u'})$$

which are candidates for matching vertices from G to H or from H to G. If t is small, then the approximation $m^t_{u,u'}$ will not be accurate as it does not incorporate sufficient information (only at depth t in both graphs). When t is large, cycles will appear in both graphs so that the recursion is not anymore valid. In order to choose an appropriate number of iterations t, we adopt the following simple strategy: we compute all the matrices $m^t_{u,u'}$ for all values of t less than a parameter d; then from these matrices, we compute the corresponding mappings π^t and σ^t as described above; we then compute:

$$e(t) := \text{match-edges}(G, H, \pi^t, \sigma^t)$$

$$:= \frac{1}{|E|} \sum_{\{u,v\}\in E} \mathbb{1}_{(\pi^t(u),\pi^t(v))\in E'} + \frac{1}{|E'|} \sum_{\{u',v'\}\in E'} \mathbb{1}_{(\sigma^t(u'),\sigma^t(v'))\in E} \,. \tag{6.60}$$

Finally, we choose

$$t^* = \arg\max(e(t)) \,.$$

Note that, we are considering sparse Erdős-Rényi graphs which are typically not connected (the diameter is infinite). We know from [GML21b] (Chapter 4) that only the giant components of G and H can possibly be aligned. Hence as a first pre-processing step, we remove all the small connected components from G and H and keep only the largest one. As a result, our algorithm takes as input 2 connected graphs of (possibly) different sizes. The pseudo-code for our algorithm is given below:

Algorithm 6.2: `MPAlign2`

Input: Two connected graphs $G = (V, E)$ and $H = (V', E')$, parameter d and parameters of the correlated Erdős-Rényi model λ (average degree) and s
for $t \in \{1, \dots, d\}$ **do**
  compute $m^t_{u\to v,u'\to v'}$ thanks to (6.58)
  compute $m^t_{u,u'}$ thanks to (6.59)
  compute $\pi^t : V \to V'$ as $\pi^t(u) = \arg\max(m^t_{u,\cdot})$
  compute $\sigma^t : V' \to V$ as $\sigma^t(u') = \arg\max(m^t_{\cdot,u'})$
  compute $e(t) = \text{match-edges}(G, H, \pi^t, \sigma^t)$ thanks to (6.60)
end
$t^* = \arg\max(e(t))$
Return π^{t^*}, σ^{t^*}, m^{t^*}

Figure 6.9 shows some empirical results for graphs of size 200 for values $\lambda = 2; 2,5; 3$ where the overlap is the mean of the overlaps given by π^{t^*} and σ^{t^*}. The maximum number of iterations is fixed to $d = 15$. For more numerical experiments on this algorithm, see [PSSZ21].

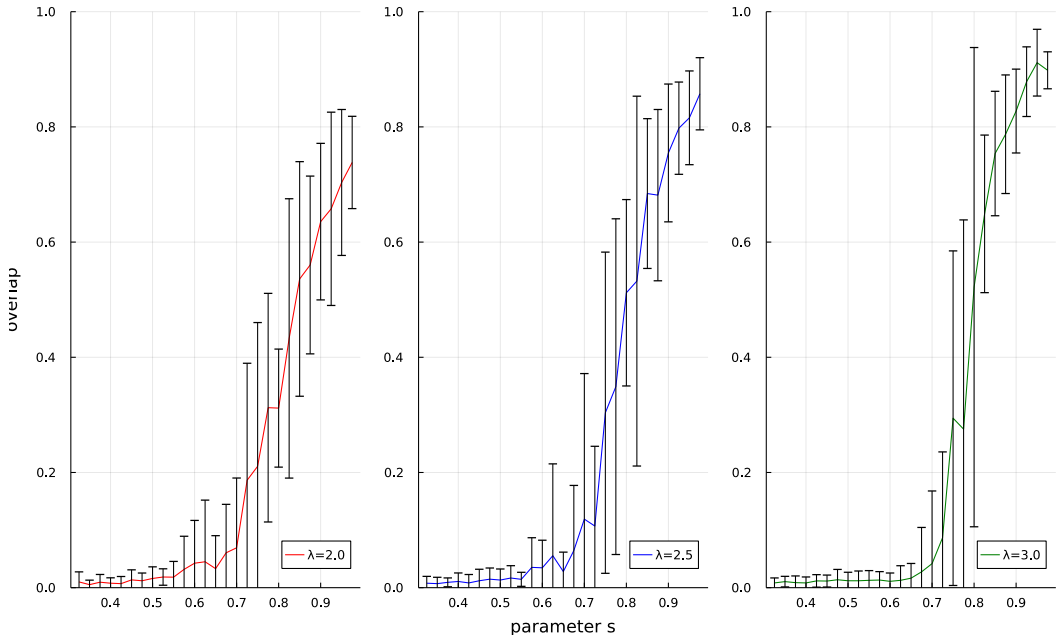


Figure 6.9 – *Overlap as a function of the parameter s for graphs with (initial) size $n = 200$ for various values of λ (parameter $d = 15$). Each point is the average of 10 simulations.*

This choice of $d = 15$ is validated by the results presented in Figure 6.10. We plot for each simulation the mean overlap of π^t and σ^t as a function of $t \leq 15$. We see that for low values

of s (on the left $s = 0.4$), the overlap behaves randomly. In this scenario, increasing the value of d will probably not help as cycles will deteriorate the performance of the algorithm. For high value of s (on the right $s = 0.9$), we see that the overlap starts by increasing and then decreases abruptly to zero, this is due to numerical issues: some messages in m^t are too large for our implementation of the algorithm to be able to deal with them. Finally for values of s, where signal is detected (in the middle $s = 0.675$), we see that when the signal is detected, the overlap start by increasing until reaching a maximum and then decreases before numerical instability. We also note that our choice of t^* thanks to the number of matched edges can be fairly sub-optimal. We believe that a better understanding of the performance of our algorithm for finite n is an interesting open problem. Indeed, we refer to [PSSZ21] which provides more detailed experimental results on a similar algorithm.

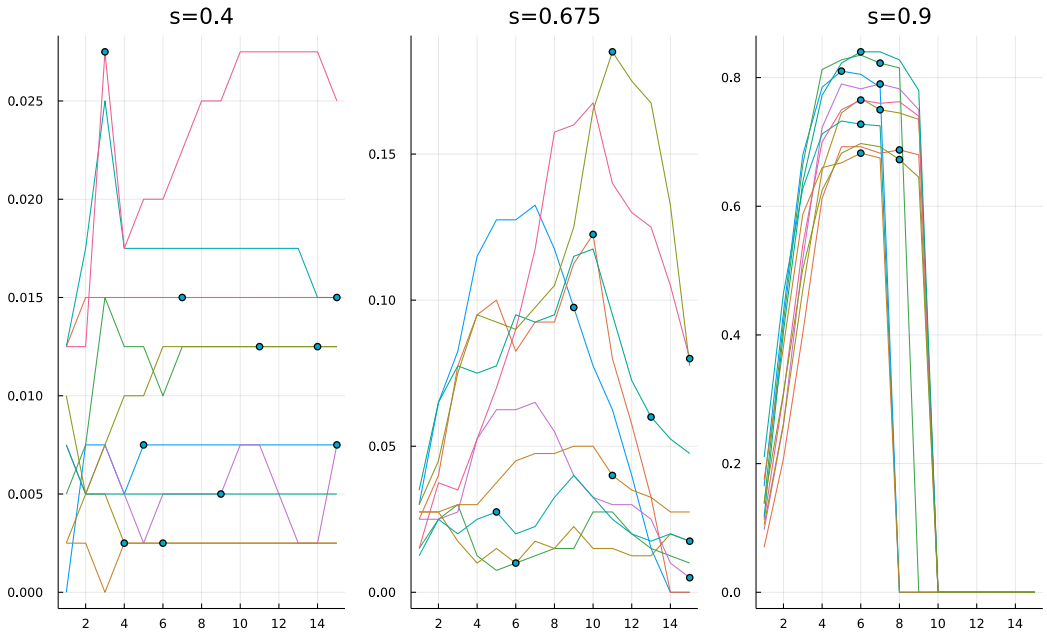


Figure 6.10 – *Overlap as a function the number of iterations t for graphs with (initial) size $n = 200$ for $\lambda = 2.5$ (parameter $d = 15$) and various values of s. The dotted point on each curve corresponds to t^*. Note that the y-axis of each plot have different scale. When overlap reaches zero, our algorithm hits infinity.*

6.B. Additional proofs

6.B.1. Proof of Proposition 6.5.1

Proof. Throughout the proof, let X_μ denote a Poisson random variable with parameter μ. A node $u \in \mathcal{L}_{n-2}(\tau^\star)$ has, independently for each $k \in \mathbb{N}$, a number $N_k \sim \mathsf{Poi}(r\pi_r(k))$ children who themselves have k children. To each such node, we can associate

$$\prod_{k\in\mathbb{N}} N_k!$$

permutations of its children that will preserve the labeled tree. Likewise, for each node $u \in \mathcal{L}_{d-1}(\tau^\star)$, there are $c_u!$ permutations of its children that don't modify the tree, where $c_u := c_{\tau^\star}(u)$. Thus by the strong law of large numbers, we have:

$$\log|\mathrm{Aut}(\tau^\star)| \geq (1 + o_{\mathbb{P}}(1)) \left[wr^{n-1}\mathbb{E}\left[\log(X_r!)\right] + wr^{n-2} \sum_{k\in\mathbb{N}} \mathbb{E}\left[\log(X_{r\pi_r(k)}!)\right] \right]. \tag{6.61}$$

Recall the classical estimate for large μ:

$$\mathbb{E}\log(X_\mu!) = \mu\log(\mu) - \mu + \frac{1}{2}\log(2\pi e\mu) + O\left(\frac{1}{\mu}\right), \tag{6.62}$$

and Stirling's formula gives

$$\log(k!) = k\log k - k + \frac{1}{2}\log(2\pi k) + O\left(\frac{1}{k}\right). \tag{6.63}$$

We now give some estimates of the distribution $\pi_r(k)$ in the following Lemma, which proof is deferred to Appendix 6.B.2.

Lemma 6.B.1. *Let $\varepsilon(r)$ be such that $\varepsilon(r) \to 0$ and $\varepsilon(r)\log r \to +\infty$ when $r \to +\infty$. Let*

$$I_{r,\varepsilon} := \left[r - (1-\varepsilon(r))\sqrt{r\log r}, r + (1-\varepsilon(r))\sqrt{r\log r}\right].$$

Then

(i) we have

$$\mathbb{P}\left(X_r \notin I_{r,\varepsilon}\right) = O\left(r^{-1/2}e^{\varepsilon(r)\log r}\right). \tag{6.64}$$

(ii) for all $k \in I_{r,\varepsilon}$, letting $x_k = \frac{k-r}{\sqrt{r}}$, we have

$$\pi_r(k) = \frac{1}{\sqrt{2\pi r}}e^{-x_k^2/2}\left[1 + \frac{x_k^3}{6\sqrt{r}} - \frac{x_k}{2\sqrt{r}} + O\left(\frac{x_k^6}{r}\right)\right]. \tag{6.65}$$

(iii) Note that (6.65) implies that for each $k \in I_{r,\varepsilon}$, it holds that $r\pi_r(k) = \Omega\left(e^{\varepsilon(r)\log r(1-o(1))}\right)$, thus diverges to $+\infty$.

Consider the function $\varepsilon(r) := \frac{\log\log r}{4\log r}$, which satisfies the assumptions of Lemma 6.B.1. Using expansion (6.65) together with (6.62) gives:

$$\begin{aligned}
\sum_{k\in I_{r,\varepsilon}} \mathbb{E}\left[\log(X_{r\pi_r(k)}!)\right] &= \sum_{k\in I_{r,\varepsilon}} r\pi_r(k)\log(r\pi_r(k)) - r\pi_r(k) + \frac{1}{2}\log(2\pi e r\pi_r(k)) + O\left(\frac{1}{r\pi_r(k)}\right) \\
&= \sum_{k\in I_{r,\varepsilon}} r\pi_r(k)\left[\frac{1}{2}\log(r) - \frac{1}{2}\log(2\pi) - \frac{x_k^2}{2} + \frac{x_k^3}{6\sqrt{r}} - \frac{x_k}{2\sqrt{r}} + O\left(\frac{\log^2 r}{r}\right) - 1\right] \\
&\quad + \sum_{k\in I_{r,\varepsilon}} \frac{1}{2}\left[\log(2\pi e) + \frac{1}{2}\log(r) - \frac{1}{2}\log(2\pi) - \frac{x_k^2}{2} + O\left(\frac{\log^{3/2}(r)}{\sqrt{r}}\right)\right] + O\left(\sqrt{r\log r}\right) \\
&\overset{(a)}{=} \frac{1}{2}r\log(r) - \left(\frac{1}{2}\log(2\pi) + \frac{1}{2} + 1\right)r + O(\sqrt{r}\log^{5/4} r) \\
&\quad + O(\sqrt{r\log r}) + \frac{1}{2}\left(1-\varepsilon(r)\right)\sqrt{r}\log^{3/2}(r) - \frac{1}{4}\sum_{k\in I_{r,\varepsilon}} x_k^2 \\
&\overset{(b)}{=} \frac{1}{2}r\log(r) - \left(\frac{1}{2}\log(2\pi) + \frac{3}{2}\right)r + \frac{1}{3}\sqrt{r}\log^{3/2}(r) + O(\sqrt{r}\log^{5/4} r).
\end{aligned} \tag{6.66}$$

Let us give hereafter all the required details for the above computation.

- At step (a), we first used point (i) of Lemma 6.B.1, which gives that

$$r\log r \times \mathbb{P}\left(X_r \notin I_{r,\varepsilon}\right) = O\left(\sqrt{r}\log^{1/4} r\right) = O\left(\sqrt{r}\log^{5/4} r\right).$$

For the sum of the x_k^2, we remark that

$$\sum_{k\in I_{r,\varepsilon}} r\pi_r(k)\frac{x_k^2}{2} = \frac{r}{2}\left(1-\mathbb{E}\left[\left(\frac{X_r-r}{\sqrt{r}}\right)^2 \mathbb{1}_{X_r\notin I_{r,\varepsilon}}\right]\right),$$

and that the expectation in the right-hand term can be written as follows

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{X_r-r}{\sqrt{r}}\right)^2 \mathbb{1}_{\left|\frac{X_r-r}{\sqrt{r}}\right|\geq 2\sqrt{\log r}}\right] + \mathbb{E}\left[\left(\frac{X_r-r}{\sqrt{r}}\right)^2 \mathbb{1}_{(1-\varepsilon(r))\sqrt{\log r}\leq\left|\frac{X_r-r}{\sqrt{r}}\right|\leq 2\sqrt{\log r}}\right]\\
&\leq \mathbb{E}\left[\left(\frac{X_r-r}{\sqrt{r}}\right)^4\right]^{1/2}\mathbb{P}\left(\left|\frac{X_r-r}{\sqrt{r}}\right|\geq 2\sqrt{\log r}\right)^{1/2} + 4\log r\times\mathbb{P}\left(X_r\notin I_{r,\varepsilon}\right)\\
&\leq O\left(r^{-1/2}\right) + O\left(r^{-1/2}\log^{5/4} r\right).
\end{aligned}$$

Hence, $\sum_{k\in I_{r,\varepsilon}} r\pi_r(k)\frac{x_k^2}{2} = \frac{r}{2} - O\left(\sqrt{r}\log^{5/4} r\right)$. Finally, using the fact that $\mathbb{E}\left[\left(\frac{X_r-r}{\sqrt{r}}\right)^3\right]$ and $\mathbb{E}\left[\frac{X_r-r}{\sqrt{r}}\right]$ are $O(1)$, the sums of the x_k^3 and x_k easily incorporate into the $O\left(\sqrt{r}\log^{5/4} r\right)$ term.

- At step (b), we first used the fact that $\varepsilon(r)\sqrt{r}\log^{3/2} = O\left(\sqrt{r}\log^{5/4} r\right)$. The only term requiring more computations is

$$\sum_{k\in I_{r,\varepsilon}} x_k^2 = \sum_{k\in I_{r,\varepsilon}}\left(\frac{k-r}{\sqrt{r}}\right)^2 = 2\times\sum_{\ell=0}^{(1-\varepsilon(r))\sqrt{r\log r}}\frac{\ell^2}{r} = \frac{2}{3}\sqrt{r}\log^{3/2} r + O\left(\sqrt{r}\log^{5/4} r\right).$$

Copying (6.66) together with (6.62) in (6.61) yields:

$$\begin{aligned}
\log(|\mathrm{Aut}(\tau^\star)|) &\geq (1+o_{\mathbb{P}}(1))wr^{n-1}\left[r\log(r)-r+\frac{1}{2}\log(2\pi e r)+O\left(\frac{1}{r}\right)\right]\\
&\quad+(1+o_{\mathbb{P}}(1))wr^{n-1}\left[\frac{1}{2}\log(r)-\frac{1}{2}\log(2\pi)-\frac{3}{2}+\frac{\log^{3/2} r}{3\sqrt{r}}+O\left(\frac{\log^{5/4} r}{\sqrt{r}}\right)\right]\\
&= (1+o_{\mathbb{P}}(1))wr^{n-1}\left[r\log(r)-r+\log(r)-1+\frac{\log^{3/2}(r)}{3\sqrt{r}}+O\left(\frac{\log^{5/4} r}{\sqrt{r}}\right)\right].
\end{aligned}$$

Another appeal to the strong law of large numbers entails that

$$\begin{aligned}
\log\left(\prod_{u\in\mathcal{V}_{d-1}(\tau^\star)} e^{-r}r^{c_{\tau^\star}(u)}\right) &= (1+o_{\mathbb{P}}(1))\left|\mathcal{V}_{d-1}(\tau^\star)\right|\mathbb{E}\left[-r+c_{\tau^\star}(\rho(\tau^\star))\log r\right]\\
&= (1+o_{\mathbb{P}}(1))K\left(-r+r\log(r)\right).
\end{aligned}$$

Combined, these last two evaluations yield a lower bound of $\log\left(\frac{|\mathrm{Aut}(\tau^\star)|}{\prod_{u\in\mathcal{V}_{d-1}(\tau^\star)} e^{-r}r^{c_{\tau^\star}(u)}}\right)$ under the event on which $\tau^\star$ survives, of the form

$$(1-o_{\mathbb{P}}(1))K\left[-r\log(r)+r+\left(1-\frac{1}{r}\right)\left(r\log(r)-r+\log(r)-1+\frac{\log^{3/2}(r)}{3\sqrt{r}}+O\left(\frac{\log^{5/4} r}{\sqrt{r}}\right)\right)\right]$$

$$= (1 - o_{\mathbb{P}}(1))K \left[\frac{\log^{3/2}(r)}{3\sqrt{r}} + O\left(\frac{\log^{5/4} r}{\sqrt{r}} \right) \right].$$

□

6.B.2. Proof of Lemma 6.B.1

Proof. (i) The result follows directly from the classical Poisson concentration inequality

$$\mathbb{P}\left(|X_r - r| \geq x\right) \leq 2\exp\left(-\frac{x^2}{2(r+x)}\right),$$

noting that for $x = (1-\varepsilon(r))\sqrt{r\log r}$, $\frac{x^2}{2(r+x)} \geq \frac{1}{2}\log r - \varepsilon \log r - o(1)$.

(ii) When k runs over $I_{r,\varepsilon}$, x_k runs over $\left[-(1-\varepsilon(r))\sqrt{\log r}, (1-\varepsilon(r))\sqrt{\log r}\right]$. Using Stirling's formula (6.63), we get

$$\begin{aligned}
\log \pi_r(k) &= \log \pi_r(r + x_k\sqrt{r}) = -r + k\log r - \log(k!) \\
&= -r + (r + x_k\sqrt{r})\log r - (r + x_k\sqrt{r})\log(r + x_k\sqrt{r}) + r + x_k\sqrt{r} - \frac{1}{2}\log(2\pi(r + x_k\sqrt{r})) + O\left(\frac{1}{r}\right) \\
&= -r + r\log r + x_k\sqrt{r}\log r - (r + x_k\sqrt{r})\left[\log r + \frac{x_k}{r^{1/2}} - \frac{x_k^2}{2r} + \frac{x_k^3}{3r^{3/2}} + O\left(\frac{x_k^4}{r^2}\right)\right] \\
&\qquad + r + x_k\sqrt{r} - \frac{1}{2}\log(2\pi) - \frac{1}{2}\log(r) - \frac{1}{2}\frac{x_k}{r^{1/2}} + O\left(\frac{x_k^2}{r}\right) \\
&= -r - x_k\sqrt{r} - x_k^2 + \frac{x_k^2}{2} + \frac{x_k^3}{2\sqrt{r}} - \frac{x_k^3}{3\sqrt{r}} + O\left(\frac{x_k^4}{r}\right) \\
&\qquad + r + x_k\sqrt{r} - \frac{1}{2}\log(2\pi r) - \frac{1}{2}\frac{x_k}{r^{1/2}} + O\left(\frac{x_k^2}{r}\right) \\
&= -\frac{x_k^2}{2} - \frac{1}{2}\log(2\pi r) + \frac{x_k^3}{6\sqrt{r}} - \frac{x_k}{2\sqrt{r}} + O\left(\frac{x_k^4}{r}\right).
\end{aligned}$$

Taking the exponential gives

$$\pi_r(k) = \frac{1}{\sqrt{2\pi r}} e^{-x_k^2/2}\left[1 + \frac{x_k^3}{6\sqrt{r}} - \frac{x_k}{2\sqrt{r}} + O\left(\frac{x_k^6}{r}\right)\right].$$

(iii) follows directly from (ii). □

6.B.3. Proof of Lemma 6.7.5

Proof. We condition on P be the number of recursive steps in the previous construction, which is $O((\log n)n^{1/4})$ under $\mathcal{A}$. For each $s \in [P]$, we denote by c_s the number of newly sampled children, and $V_s := \sum_{s'=0}^{s-1} c_{s'}$ the number of uncovered vertices before step s (we set $V_0 := 0$). With these notations, it is easily seen than M_{d-1} can be factorized as follows:

$$\begin{aligned}
M_{d-1} &= \prod_{s\in[P]} \frac{\mathbb{P}\left(\mathsf{Bin}\left(n - 2 - V_s, \lambda/n\right) = c_s\right)}{\pi_\lambda(c_s)} \leq \prod_{s\in[P]} \exp\left(\frac{\lambda}{n}(V_s + 2 + c_s)\right) \\
&= \exp\left(\frac{2\lambda P}{n} + \frac{\lambda}{n}\sum_{s\in[P]}(P - s)c_s\right).
\end{aligned}$$

Under $\mathbb{P}_d^{(\lambda)}$, the variables c_s are independent $\mathsf{Poi}(\lambda)$ variables, hence

$$\begin{aligned}\mathbb{E}_d^{(\lambda)}\left[M_{d-1}^2 \mathbb{1}_{\mathcal{A}}\right] &\leq \exp\left(\frac{4\lambda P}{n} + \lambda \sum_{s\in[P]} \left(e^{2\lambda(P-s)/n} - 1\right)\right) \mathbb{1}_{P=O((\log n)n^{1/4})} \\ &\leq \exp\left(C'P^2/n + o(P^2/n)\right) \mathbb{1}_{P=O((\log n)n^{1/4})} = O(1).\end{aligned}$$

□

Chapter 7

Addendum: new results for correlation detection in trees

This addendum, which concludes the manuscript, presents new results for correlation in trees from a recent joint work with L. Massoulié and G. Semerjian (paper in preparation). These results are significantly improving on previous work and give a general understanding of the fundamental limits of the problem, as well as some interesting perspectives discussed afterwards in the conclusion.

We do not redefine here the problem of correlation detection in trees, since we already thouroughly did in Chapters 5 and 6, but we rather introduce some auxiliary definitions that proved useful for the analysis made his the sequel. For related work, we refer to Section 1.3.4 of the introduction.

We straightaway mention that the results presented in this last part are very much related to the recent study of Mao, Wu, Xu and Yu [MWXY21] who studied the correlation detection problem in Erdős-Rényi graphs, and proposed an algorithm based on counting (signed) trees, which can provably distinguish graph correlation efficiently as soon as $s > \sqrt{\alpha}$, where α is the Otter's constant defined below in Proposition 7.1.1. The results presented here are different for several reasons: first, we study the problem on trees and consider an optimal test, which thus also meets the informational bounds. Moreover, we show that $s < \sqrt{\alpha}$ implies impossibility of one-sided detection, and that one-sided detection exhibits a sharp threshold at $s = \sqrt{\alpha}$, asymptotically in λ, see Figure 7.1.

We believe that this study paves the way for many other works in this field, generalizing to other graph models, analyzing the computational hardness of the problem with different tools, and designing more efficient algorithms for tree correlation detection or graph alignment – for more insights and details on these research directions, we refer to the conclusion.

7.1. Main results

7.1.1. Definitions and notations

Trees We start by stating some familiar definitions, in the general context of *unlabeled trees*, that are part of the rationale for results to follow.

Definition 7.1.1 (Finite rooted unlabeled trees)**.** *We recursively define the set $\mathcal{X}_d$ of* finite rooted unlabeled trees of depth at most $d \geq 0$.

For $d = 0$, $\mathcal{X}_d$ contains the trivial tree reduced to its root node, denoted by $\bullet$.

For $d \geq 1$, having defined $\mathcal{X}_0, \dots, \mathcal{X}_{d-1}$, we define $\mathcal{X}_d$ as follows: a finite rooted unlabeled tree $t \in \mathcal{X}_d$ consists in an integer sequence $\{N_\tau\}_{\tau \in \mathcal{X}_{d-1}}$ with finite support, that is such that

$$|\{\tau \in \mathcal{X}_{d-1}, N_\tau \neq 0\}| < \infty,$$

where N_τ is the number of children of the root in t which subtrees are copies of τ.

Throughout all the chapter, we will only work with finite trees (with finite degrees), hence the adjective 'finite' will be omitted as a shortcut.

Remark 7.1.1. *With the previous definition, equality between two rooted unlabeled trees* $t := \{N_\tau\}_{\tau\in\mathcal{X}_{d-1}}$ *and* $t' := \{N'_\tau\}_{\tau\in\mathcal{X}_{d-1}}$ *is defined as* $N_\tau = N'_\tau$ *for all* $\tau \in \mathcal{X}_{d-1}$.

Remark 7.1.2. *Denoting one-to-one correspondence by* $\simeq$, *we remark that* $\mathcal{X}_1 \simeq \mathbb{N}$ *(the set of non-negative integers), and that more generally for each* $d \geq 1$,

$$\mathcal{X}_d \simeq \bigcup_{\ell\geq 1} \bigcup_{\substack{\tau_1,\dots,\tau_\ell\in\mathcal{X}_{d-1}\\ i\neq j \Longrightarrow \tau_i\neq\tau_j}} \mathbb{N}^\ell.$$

Hence, $\mathcal{X}_d$ *is countably infinite, by recursion, for all* $d \geq 1$.

Definition 7.1.2 (Size of a rooted unlabeled tree)**.** *The* size, *or* number of nodes, *of a tree* $t \in \mathcal{X}_d$ *is denoted by* $|t|$ *and defined recursively as follows. First, if* $d = 0$, *we set* $|\bullet| = 1$. *Then, for* $d \geq 1$, *writing* $t = \{N_\tau\}_{\tau\in\mathcal{X}_{d-1}}$, *one has*

$$|t| = 1 + \sum_{\tau\in\mathcal{X}_{d-1}} N_\tau \cdot |\tau| \,.$$

Definition 7.1.3 (Depth of a rooted unlabeled tree)**.** *The* depth *of a tree* $t \in \mathcal{X}_d$ *is denoted by* $\mathrm{depth}(t)$ *and defined recursively as follows. First, if* $d = 0$, *we set* $\mathrm{depth}(\bullet) = 0$. *Then, for* $d \geq 1$, *writing* $t = \{N_\tau\}_{\tau\in\mathcal{X}_{d-1}}$, *one has*

$$\mathrm{depth}(t) = 1 + \max\left\{\mathbb{1}_{N_\tau\geq 1} \cdot \mathrm{depth}(\tau),\ \tau \in \mathcal{X}_{d-1}\right\} .$$

Definition 7.1.4 (Child of a rooted unlabeled tree)**.** *A rooted unlabeled tree* $s \in \mathcal{X}_{d-1}$ *is said to be a* child *of a tree* $t = \{N_\tau\}_{\tau\in\mathcal{X}_{d-1}}$ *if* $N_s \geq 1$. *Moreover, if* $N_s \geq 1$, t *is said to have* N_s children of type s. *Note that since the tree* t *is finite, it has a finite number of children, given by* $\sum_\tau N_\tau$.

Definition 7.1.5 (Subtree of a rooted unlabeled tree)**.** *Let us define recursively the notion of* subtree. *First, the only subtree of* $\bullet$ *is* $\bullet$ *itself. Then for* $d' \leq d$, *an element* $s \in \mathcal{X}_{d'}$ *is a subtree of* $t \in \mathcal{X}_d$ *if either* $s = t$ *or if* s *is a subtree of some child of* t.

Formal power series If f is a formal power series in the variable x, we denote $[x^n]f(x)$ the coefficient of the monomial x^n in f, i.e. if $f(x) = \sum_{n\geq 0} a_n x^n$ then $[x^n]f(x) := a_n$.

If f is a formal power series in m variables $(x_1, \dots, x_m)$, and $\ell := (\ell_1, \dots, \ell_m)$ is a tuple of non negative integers, we use the shorthand notation

$$[x^\ell]f(x_1, \dots, x_m)$$

for $[x_1^{\ell_1} \cdots x_m^{\ell_m}]f(x_1, \dots, x_m)$.

Throughout the paper we will often consider families indexed by a countably infinite set $\mathcal{Z}$, and in particular use the same shorthand $[x^\ell]$ for $[\prod_{z\in\mathcal{Z}} x_z^{\ell_z}]$, where $x = \{x_z\}_{z\in\mathcal{Z}}$ is a family of formal variables, and $\ell = \{\ell_z\}_{z\in\mathcal{Z}}$ a family of non-negative integers; in such occurences only a finite number of ℓ_z will be non-zero, the definition thus reduces to the finite-dimensional one by taking $x_z = 0$ whenever $\ell_z = 0$. This *finite support* property will also make summations over $z \in \mathcal{Z}$ of functions of ℓ_z well-defined.

By convention $\{u_z\}_{z\in\mathcal{Z}}$ will stand for non-negative integer sequences $\{u_z\}_{z\in\mathcal{Z}}$ with finite support, hence from the definition of $\mathcal{X}_d$ the sum $\sum_{t\in\mathcal{X}_d}$ will be equivalently denoted $\sum_{\{N_\tau\}_{\tau\in\mathcal{X}_{d-1}}}$.

Cardinality of unlabeled trees with given size and depth We are now interested in the cardinality of the set of unlabeled trees with given size and depth.

Definition 7.1.6 (Trees with given size and depth)**.** *For $n \geq 1$, let us define*

$$A_n := \left| \left\{ t \in \bigcup_{d \geq 0} \mathcal{X}_d,\ |t| = n \right\} \right|, \tag{7.1}$$

that is A_n is the number of (distinct) unlabeled rooted trees of size n. For $d \geq 0$, we furthermore define

$$A_{d,n} := \left| \{ t \in \mathcal{X}_d,\ |t| = n \} \right|, \tag{7.2}$$

that is $A_{d,n}$ is the number of (distinct) unlabeled rooted trees of size n and depth at most d.

We now state a celebrated result by Otter [Ott48], together with a proposition that will be useful in the sequel.

Proposition 7.1.1 (Asymptotic number of unlabeled trees, [Ott48])**.** *One has*

$$A_n \underset{n \to \infty}{\sim} \frac{C}{n^{3/2}} \left(\frac{1}{\alpha} \right)^n, \tag{7.3}$$

for some $C > 0$, where $\alpha \in (0,1)$ is the Otter constant, numerically $\alpha = 0.3383219...$

Proposition 7.1.2 (Control of the generating function of the $A_{d,n}$)**.** *For all $d \geq 0$, let*

$$\Phi_d(x) := \sum_{n \geq 1} A_{d,n} x^{n-1}. \tag{7.4}$$

We have, for all $x > 0$,

$$\Phi_d(x) \underset{d \to \infty}{\longrightarrow} \Phi(x), \tag{7.5}$$

where

$$\Phi(x) := \sum_{n \geq 1} A_n x^{n-1}. \tag{7.6}$$

Moreover, for all $d \geq 0$ and $t \in [0,1)$, there exists $A = A(d,t)$ such that

$$\forall x \in [0,t],\ |\Phi_d(x)| \leq A. \tag{7.7}$$

Proof of Proposition 7.1.2. Convergence (7.5) follows from $A_{d,n} \underset{d \to \infty}{\longrightarrow} A_n$ and monotone convergence theorem.

We will now establish a recursion property on the Φ_d, adapting the proof of Otter [Ott48] of the case $d = \infty$ to general depth d. We can decompose a tree t of depth at most $d+1$ with n vertices according to its subtrees. For $i \geq 1$, we denote by $\mu_i \geq 0$ the number of subtrees of t with i nodes. The μ_i subtrees are then distributed in the $A_{d,i}$ categories, the only thing that distinguish them being the number in each $A_{d,i}$. For a given i, there is $\binom{A_{d,i}+\mu_i-1}{\mu_i}$ such choices. This gives the following recursion formula:

$$A_{d+1,n} = \sum_{\{\mu_i\}_{i \geq 1}} \prod_{i \geq 1} \binom{A_{d,i} + \mu_i - 1}{\mu_i} \mathbb{1}_{n = 1 + \sum_{i \geq 1} i\mu_i}.$$

Plugging the last equation in the definition of Φ_{d+1} gives

$$\Phi_{d+1}(x) = \sum_{n \geq 1} x^{n-1} \sum_{\{\mu_i\}_{i \geq 1}} \prod_{i \geq 1} \binom{A_{d,i} + \mu_i - 1}{\mu_i} \mathbb{1}_{n = 1 + \sum_{i \geq 1} i\mu_i}$$

$$= \sum_{\{\mu_i\}_{i\geq 1}} \prod_{i\geq 1} \left[\binom{A_{d,i}+\mu_i-1}{\mu_i} x^{i\mu_i} \right] = \prod_{i\geq 1} \sum_{\mu\geq 0} \left[\binom{A_{d,i}+\mu-1}{\mu} x^{i\mu} \right]$$

$$= \prod_{i\geq 1} \frac{1}{(1-x^i)^{A_{d,i}}} = \exp\left(-\sum_{i\geq 1} A_{d,i} \log(1-x^i) \right) = \exp\left(\sum_{i,j\geq 1} A_{d,i} \frac{x^{ij}}{j} \right)$$

$$= \exp\left(\sum_{j\geq 1} \frac{x^j}{j} \sum_{i\geq 1} A_{d,i}(x^j) i - 1 \right) = \exp\left(\sum_{j\geq 1} \frac{x^j}{j} \Phi_d(x^j) \right) .$$

Equation (7.7) is then very easy to propagate by recursion with this last formula. For $d = 0$, $A_{d,n} = \mathbb{1}_{n=1}$, hence $\Phi_0(x) = 1$ and (7.7) holds with $A = 1$. Assume that (7.7) holds at depth d for all $t \in [0,1)$ with constant $A(d,t)$. Then, by the previous computation, for all $t \in [0,1)$ and $x \in [0,t]$, since $x^j \in [0,t]$ for all $j \geq 1$ we have

$$\Phi_{d+1}(x) = \exp\left(\sum_{j\geq 1} \frac{x^j}{j} \Phi_d(x^j) \right) \leq \exp\left(\sum_{j\geq 1} \frac{x^j}{j} A(d,t) \right)$$

$$= \exp\left(-A(d,t) \log(1-x) \right) = \left(\frac{1}{1-x} \right)^{A(d,t)} \leq \left(\frac{1}{1-t} \right)^{A(d,t)} =: A(d+1,t) .$$

Thus, (7.7) holds at depth $d+1$. □

Models of random trees We now define the models and random trees considered in the study. They are the same as models of Chapter 6, but we restate them here with an equivalent[1] definition that fits the description $t = \{N_\tau\}_{\tau\in\mathcal{X}_{d-1}}$ of elements of $\mathcal{X}_d$.

Definition 7.1.7 (Galton-Watson trees with Poisson offspring)**.** *Let $\mu > 0$. For $d = 0$, $\mathsf{GW}_d^{(\mu)}$ is the Dirac mass at the trivial tree $\bullet \in \mathcal{X}_0$. For $d \geq 1$, a tree $t = \{N_\tau\}_{\tau\in\mathcal{X}_{d-1}} \sim \mathsf{GW}_d^{(\mu)}$ is sampled as follows: for all $\tau \in \mathcal{X}_{d-1}$, $N_\tau \sim \mathsf{Poi}(\mu \mathsf{GW}_{d-1}^{(\mu)}(\tau))$ independently from everything else. Note that since the Poisson variables are independent, we have*

$$\sum_{\tau\in\mathcal{X}_{d-1}} N_\tau \sim \mathsf{Poi}\left(\mu \sum_{\tau\in\mathcal{X}_{d-1}} \mathsf{GW}_{d-1}^{(\mu)}(\tau) \right) = \mathsf{Poi}(\mu)$$

which is a.s. finite. Hence, we have $t \in \mathcal{X}_d$ a.s.

Throughout all the study, we will only work with Galton-Watson trees with Poisson offspring, which we will simply refer to as Galton-Watson trees, as a notation shortcut.

Definition 7.1.8 (Null model $\mathbb{P}_d^{(\lambda)}$)**.** *The null distribution $\mathbb{P}_d^{(\lambda)}$ on $\mathcal{X}_d \times \mathcal{X}_d$ of parameter $\lambda > 0$ is simply defined as the product $\mathsf{GW}_d^{(\lambda)} \otimes \mathsf{GW}_d^{(\lambda)}$: under the null model, the two trees are independent Galton-Watson trees with offspring $\mathsf{Poi}(\lambda)$.*

Definition 7.1.9 (Correlated model $\mathbb{P}_d^{(\lambda,s)}$)**.** *The correlated model $\mathbb{P}_d^{(\lambda,s)}$ on $\mathcal{X}_d \times \mathcal{X}_d$ with parameters $\lambda > 0$ and $s \in [0,1)$ first verifies that $\mathbb{P}_0^{(\lambda,s)}$ is the same as $\mathbb{P}_0^{(\lambda)}$.*

For $d \geq 1$, a pair of trees $(t,t') = (\{N_\tau\}_{\tau\in\mathcal{X}_{d-1}}, \{N'_\tau\}_{\tau\in\mathcal{X}_{d-1}}) \sim \mathbb{P}_d^{(\lambda,s)}$ is sampled as follows.

$$N_\tau := M_\tau + \sum_{\tau'\in\mathcal{X}_{d_1}} N_{\tau,\tau'} \quad \text{and} \quad N'_\tau := M'_\tau + \sum_{\tau\in\mathcal{X}_{d-1}} N_{\tau,\tau'} , \tag{7.8}$$

[1] Note that with this 'unlabeled' view, there is no need to define a tree augmentation for the correlated model $\mathbb{P}_d^{(\lambda,s)}$ – as in previous Chapter.

with M_τ and M'_τ i.i.d. $\mathsf{Poi}(\lambda(1-s)\mathsf{GW}_{d-1}^{(\mu)}(\tau))$ and $N_{\tau,\tau'}$ i.i.d. $\mathsf{Poi}(\lambda s \mathbb{P}_{d-1}^{(\lambda,s)}(\tau,\tau'))$ variables. Note that for all d, $\mathbb{P}_d^{(\lambda,s)} = \mathbb{P}_d^{(\lambda)}$ if $s = 0$.

7.1.2. Main new results

We recall from Chapter 6 that the likelihood ratio L_d is defined as

$$L_d(t,t') := \frac{\mathbb{P}^{(\lambda,s)}(t,t')}{\mathbb{P}^{(\lambda)}(t,t')}\,,$$

as well as the following result

Theorem (Chapter 6, Theorem 6.1)**.** *Let*

$$\mathrm{KL}_d := \mathrm{KL}(\mathbb{P}_d^{(\lambda,s)} \| \mathbb{P}_d^{(\lambda)}) = \mathbb{E}_{1,d}\left[\log(L_d)\right].$$

Then there exists a one-sided test for testing $\mathbb{P}_d^{(\lambda)}$ versus $\mathbb{P}_d^{(\lambda,s)}$ if and only if $\lim_{d\to\infty} \mathrm{KL}_d = +\infty$ and $\lambda s > 1$.

We now state the main results of this addendum. Let α be the Otter constant introduced in Proposition 7.1.1.

Theorem 7.1 (Negative result)**.** *If $s \leq \sqrt{\alpha}$, then for all $\lambda > 0$, $\limsup_d \mathrm{KL}(\mathbb{P}_d^{(\lambda,s)} \,||\, \mathbb{P}_d^{(\lambda)}) < \infty$. Hence, one-sided detection is impossible.*

Theorem 7.2 (Positive result)**.** *If $s > \sqrt{\alpha}$, then there exists $\lambda(s) > 0$ such that for all $\lambda \geq \lambda(s)$, $\mathrm{KL}(\mathbb{P}_d^{(\lambda,s)} \,||\, \mathbb{P}_d^{(\lambda)}) \underset{d\to\infty}{\longrightarrow} +\infty$. One-sided detection is thus feasible for λ large enough, and an optimal one-sided test is the likelihood-ratio test $\mathcal{T}_d := \mathbb{1}_{L_d(t,t')>\beta_d}$ for some appropriate β_d (see Theorem 6.1).*

Remark 7.1.3. *These new results establish an almost sharp result for correlation detection in trees. The regime $s \leq \sqrt{\alpha}$ always lies within the impossible phase, and $s > \sqrt{\alpha}$ is in the easy phase for high mean degree λ.*

In view of Theorem 6.2 proved in Chapter 6, these results also extend the knowledge on the analysis of `MPAlign`, and doing so on the phase diagram for partial graph alignment, for which a state-of-the-art version is given in Figure 7.1 below.

Instrumental to the proofs of these results is the diagonalization of the likelihood ratio $L_d(t,t')$ in an orthogonal basis of eigenvectors (or eigenfunctions). A very noticeable result given in Theorem 7.3 in the following section is that these eigenvectors (resp. eigenvalues) only depend on λ (resp. on s) (!)

7.2. The impossible phase for $s \leq \sqrt{\alpha}$

7.2.1. Eigendecomposition of the likelihood ratio

Let us start by stating the master result of this section, which will render the analysis easier and bring important corollaries for our analysis.

Theorem 7.3 (Eigendecomposition of the likelihood ratio)**.** *For all $\lambda > 0, d \geq 0$, there exists a collection $\{f_{d,\alpha}^{(\lambda)}\}_{\alpha\in\mathcal{X}_d}$ with $f_{d,\alpha}^{(\lambda)} : \mathcal{X}_d \to \mathbb{R}$, such that for all $s \in [0,1)$,*

$$\forall t,t' \in \mathcal{X}_d,\ L_d(t,t') = \sum_{\alpha\in\mathcal{X}_d} s^{|\alpha|-1} f_{d,\alpha}^{(\lambda)}(t) f_{d,\alpha}^{(\lambda)}(t'), \tag{7.9}$$

Moreover, the $f_{d,\alpha}^{(\lambda)}$ are independent of s and verify the following properties:

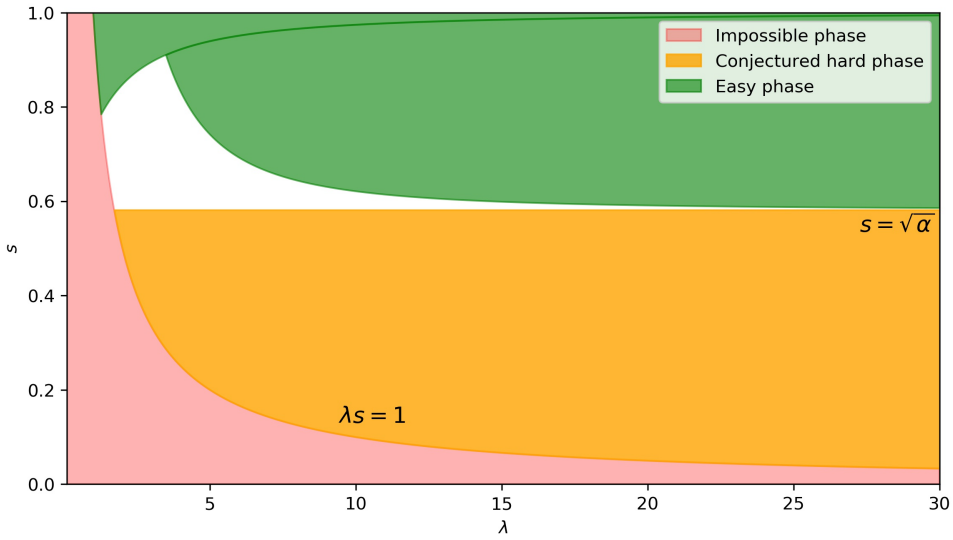


Figure 7.1 – *State-of-the art phase diagram for partial graph alignment.*

- *Value at the trivial tree:*
$$\forall t \in \mathcal{X}_d,\ f^{(\lambda)}_{d,\bullet}(t) = 1\,, \tag{7.10}$$

- *Orthogonality:*
$$\forall \alpha, \alpha' \in \mathcal{X}_d,\ \sum_{t \in \mathcal{X}_d} \mathsf{GW}^{(\lambda)}_d(t) f^{(\lambda)}_{d,\alpha}(t) f^{(\lambda)}_{d,\alpha'}(t) = \mathbb{1}_{\alpha=\alpha'}\,. \tag{7.11}$$
$$\forall t, t' \in \mathcal{X}_d,\ \mathsf{GW}^{(\lambda)}_d(t) \sum_{\alpha \in \mathcal{X}_d} f^{(\lambda)}_{d,\alpha}(t) f^{(\lambda)}_{d,\alpha}(t') = \mathbb{1}_{t=t'}\,. \tag{7.12}$$

- *Limit of higher-order mixed moments: more generally, for $n \geq 2$, $d \geq 1$ and $\beta^{(1)} = \{\beta^{(1)}_\alpha\}_{\alpha \in \mathcal{X}_{d-1}}, \ldots, \beta^{(n)} = \{\beta^{(n)}_\alpha\}_{\alpha \in \mathcal{X}_{d-1}} \in \mathcal{X}_d$, one has*
$$\sum_{t \in \mathcal{X}_d} \mathsf{GW}^{(\lambda)}_d(t) f^{(\lambda)}_{d,\beta^{(1)}}(t) \cdots f^{(\lambda)}_{d,\beta^{(n)}}(t) \underset{\lambda \to \infty}{\longrightarrow} \prod_{\alpha \in \mathcal{X}_{d-1}} \sqrt{\prod_{i=1}^n \beta^{(i)}_\alpha !} \left[x_1^{\beta^{(1)}_\alpha} \cdots x_n^{\beta^{(n)}_\alpha} \right] e^{\sum_{1 \leq i < j \leq n} x_i x_j}\,. \tag{7.13}$$

Remark 7.2.1. *Note that in the above properties,* (7.11) *implies the following first moment condition:*
$$\forall \alpha \in \mathcal{X}_d,\ \sum_{t \in \mathcal{X}_d} \mathsf{GW}^{(\lambda)}_d(t) f^{(\lambda)}_{d,\alpha}(t) = \mathbb{1}_{\alpha=\bullet}.$$

Remark 7.2.2. *As remarked earlier, the eigenvectors $f^{(\lambda)}_{d,\alpha}$ (resp. the eigenvalues) only depend on λ and are independent of s (resp. on s, independent of λ).*

Remark 7.2.3. *The $f^{(\lambda)}_{d,\alpha}(\ell)$ for $d = 1$, hence indexed by $\alpha, \ell \in \mathcal{X}_1 \simeq \mathbb{N}$, are given by*
$$f^{(\lambda)}_{1,\alpha}(\ell) := \sqrt{\alpha!}[x^\alpha] e^{-x\sqrt{\lambda}} \left(1 + \frac{x}{\sqrt{\lambda}}\right)^\ell\,,$$
see equation (7.14) *in the proof. These functions are known as Charlier polynomials, which are orthogonal for the Poisson distribution. Theorem 7.3 provides an extension of these*

polynomials, on trees of depth $d \geq 2$, that are orthogonal for the $\mathsf{GW}_d^{(\lambda)}$ distribution, consistent with $\mathsf{GW}_1^{(\lambda)} \overset{(\mathsf{d})}{=} \mathsf{Poi}(\lambda)$.

Also, note that equation (7.9) exhibits a duality between trees t, t' in $\mathcal{X}_d$ and the trees $\alpha \in \mathcal{X}_d$. This duality turns out to be very helpful for analysis, as shown below, e.g. giving a nice space in which one can prove weak convergence results – see Section 7.3.1.

Proof of Theorem 7.3. We will prove the decomposition (7.9) as well as the properties (7.10), (7.11), (7.13) by induction on d.

Step 1: initialization at $d = 1$. We identify $\mathcal{X}_1$ to $\mathbb{N}$, and a tree t of depth $d = 1$ to its number of children $\ell \in \mathbb{N}$: in this case, $|t| = \ell + 1$. Denote by $\widehat{\mathbb{P}}_1^{(\lambda,s)}$ the characteristic function defined on $[0, 2\pi]^2$ by $\widehat{\mathbb{P}}_1^{(\lambda,s)}(k, k') := \mathbb{E}\left[e^{ik\ell + ik'\ell'}\right]$ where $(\ell, \ell') \sim \mathbb{P}_1^{(\lambda,s)}$. We have that

$$\begin{aligned}
\widehat{\mathbb{P}}_1^{(\lambda,s)}(k, k') &= \exp\left[\lambda(1-s)(e^{ik} + e^{ik'} - 2) + \lambda s(e^{i(k+k')} - 1)\right] \\
&= e^{\lambda(e^{ik}-1)} e^{\lambda(e^{ik'}-1)} \exp\left[\lambda s(e^{ik} - 1)(e^{ik'} - 1)\right] \\
&= \sum_{\alpha \geq 0} s^\alpha \frac{\lambda^\alpha}{\alpha!} (e^{ik} - 1)^\alpha e^{\lambda(e^{ik}-1)} (e^{ik'} - 1)^\alpha e^{\lambda(e^{ik'}-1)} = \sum_{\alpha \geq 0} s^\alpha \widehat{g}_{1,\alpha}^{(\lambda)}(k) \widehat{g}_{1,\alpha}^{(\lambda)}(k') \,,
\end{aligned}$$

with

$$\widehat{g}_{1,\alpha}^{(\lambda)}(k) := e^{-\lambda} \sqrt{\frac{\lambda^\alpha}{\alpha!}} e^{\lambda e^{ik}} (e^{ik} - 1)^\alpha = e^{-\lambda} \sqrt{\alpha!} e^{\lambda e^{ik}} [x^\alpha] e^{x\sqrt{\lambda}(e^{ik}-1)} = e^{-\lambda} \sqrt{\alpha!} [x^\alpha] e^{-x\sqrt{\lambda}} e^{(\lambda + x\sqrt{\lambda})e^{ik}} \,.$$

We have an easy upper bound of the form $\left|\widehat{g}_{1,\alpha}^{(\lambda)}(k)\right| \leq \frac{C^\alpha}{\sqrt{\alpha!}}$, independently of k, which established normal convergence of the series $\widehat{\mathbb{P}}_1^{(\lambda,s)}(k, k')$ in the above. Hence, inverting the Fourier transform, we get

$$\mathbb{P}_1^{(\lambda,s)}(\ell, \ell') = \int_{[0,2\pi]^2} \frac{dk dk'}{(2\pi)^2} e^{-ik\ell - ik'\ell'} \widehat{\mathbb{P}}_1^{(\lambda,s)}(k, k') = \sum_{\alpha \geq 0} s^{|\alpha|-1} g_{1,\alpha}^{(\lambda)}(\ell) g_{1,\alpha}^{(\lambda)}(\ell') \,,$$

with

$$\begin{aligned}
g_{1,\alpha}^{(\lambda)}(\ell) &:= \int_{[0,2\pi]} \frac{dk}{2\pi} e^{-ik\ell} \widehat{g}_{1,\alpha}^{(\lambda)}(k) \\
&= e^{-\lambda} \sqrt{\alpha!} [x^\alpha] e^{-x\sqrt{\lambda}} \int_{[0,2\pi]} \frac{dk}{2\pi} e^{-ik\ell} e^{(\lambda + x\sqrt{\lambda})e^{ik}} \\
&= e^{-\lambda} \sqrt{\alpha!} [x^\alpha] e^{-x\sqrt{\lambda}} \frac{(\lambda + x\sqrt{\lambda})^\ell}{\ell!} \,.
\end{aligned}$$

We hence have that L_1 satisfies (7.9) with $f_{1,\alpha}^{(\lambda)}$ given by

$$f_{1,\alpha}^{(\lambda)}(\ell) = \frac{e^\lambda \ell!}{\lambda^\ell} g_{1,\alpha}^{(\lambda)}(\ell) = \sqrt{\alpha!} [x^\alpha] e^{-x\sqrt{\lambda}} \left(1 + \frac{x}{\sqrt{\lambda}}\right)^\ell \,. \tag{7.14}$$

Taking $\alpha = 0$ in (7.14) gives $f_{1,\bullet}^{(\lambda)} = 1$ and proves condition (7.10) at $d = 1$. For orthogonality (7.11), note that for all $\alpha, \alpha' \in \mathbb{N}$,

$$\sum_{\ell \geq 0} \mathsf{GW}_1^{(\lambda)}(\ell) f_{1,\alpha}^{(\lambda)}(\ell) f_{1,\alpha'}^{(\lambda)}(\ell) = \sqrt{\alpha! \alpha'!} \sum_{\ell \geq 0} e^{-\lambda} \frac{\lambda^\ell}{\ell!} [x^\alpha y^{\alpha'}] e^{-x\sqrt{\lambda} - y\sqrt{\lambda}} \left[\left(1 + \frac{x}{\sqrt{\lambda}}\right)\left(1 + \frac{y}{\sqrt{\lambda}}\right)\right]^\ell$$

$$= \sqrt{\alpha!\alpha'!}[x^\alpha y^{\alpha'}] e^{-x\sqrt{\lambda}-y\sqrt{\lambda}} \exp\left[-\lambda + \lambda\left(1+\frac{x}{\sqrt{\lambda}}\right)\left(1+\frac{y}{\sqrt{\lambda}}\right)\right]$$
$$= \sqrt{\alpha!\alpha'!}[x^\alpha y^{\alpha'}] e^{xy} = \mathbb{1}_{\alpha=\alpha'}\,,$$

which establishes (7.11) for $d=1$. Previous computations are made rigorous by noticing that the series in (7.14) has infinite radius of convergence, and appealing to Fubini's theorem. With the same arguments, writing

$$\begin{aligned} f_{1,\alpha}^{(\lambda)}(\ell) &= \sqrt{\alpha!}\ell![x^\alpha y^\ell] e^{-x\sqrt{\lambda}+y(1+x/\sqrt{\lambda})} \\ &= \sqrt{\alpha!}\ell![x^\alpha y^\ell] e^{y+x(y/\sqrt{\lambda}-\sqrt{\lambda})} \\ &= \frac{\ell!}{\sqrt{\alpha!}}[x^\ell] e^x \left(\frac{x}{\sqrt{\lambda}}-\sqrt{\lambda}\right)^\alpha . \end{aligned} \tag{7.15}$$

For orthogonality (7.12), note that for all $\alpha, \alpha' \in \mathbb{N}$,

$$\begin{aligned} \sum_{\alpha\ge 0} f_{1,\alpha}^{(\lambda)}(\ell) f_{1,\alpha}^{(\lambda)}(\ell') &= \ell!(\ell')![x^\ell y^{\ell'}] e^{x+y} \sum_{\alpha\ge 0} \frac{1}{\alpha!}\left(\frac{x}{\sqrt{\lambda}}-\sqrt{\lambda}\right)^\alpha \left(\frac{y}{\sqrt{\lambda}}-\sqrt{\lambda}\right)^\alpha \\ &= \ell!(\ell')![x^\ell y^{\ell'}] e^{xy/\lambda+\lambda} = \mathbb{1}_{\ell=\ell'}\ell! e^\lambda \lambda^{-\ell} = \frac{\mathbb{1}_{\ell=\ell'}}{\mathsf{GW}_1^{(\lambda)}(\ell)}\,, \end{aligned}$$

which establishes (7.12) for $d=1$.

More generally, for $n \ge 2$,

$$\sum_{\ell\ge 0} \mathsf{GW}_1^{(\lambda)}(\ell) f_{1,\alpha_1}^{(\lambda)}(\ell)\cdots f_{1,\alpha_n}^{(\lambda)}(\ell) = \sqrt{\prod_{i=1}^n \alpha_i!} \sum_{\ell\ge 0} e^{-\lambda}\frac{\lambda^\ell}{\ell!}[x_1^{\alpha_1}\cdots x_n^{\alpha_n}] e^{-\sqrt{\lambda}\sum_{i=1}^n x_i} \prod_{i=1}^n \left(1+\frac{x_i}{\sqrt{\lambda}}\right)^\ell$$
$$= \sqrt{\prod_{i=1}^n \alpha_i!}\,[x_1^{\alpha_1}\cdots x_n^{\alpha_n}] \exp\left[-\lambda - \sqrt{\lambda}\sum_{i=1}^n x_i + \lambda\prod_{i=1}^n\left(1+\frac{x_i}{\sqrt{\lambda}}\right)\right]$$
$$= \sqrt{\prod_{i=1}^n \alpha_i!}\,[x_1^{\alpha_1}\cdots x_n^{\alpha_n}] \exp\left[\sum_{1\le i<j\le n} x_i x_j + \varepsilon_\lambda(x_1,\dots,x_n)\right], \tag{7.16}$$

with

$$\varepsilon_\lambda(x_1,\dots,x_n) := \sum_{p=3}^n \lambda^{1-p/2} \sum_{1\le i_1<\dots<i_p\le n} x_{i_1}^{\alpha_{i_1}}\cdots x_{i_p}^{\alpha_{i_p}}.$$

The terms corresponding to $[x_1^{\alpha_1}\cdots x_n^{\alpha_n}]$ in the expansion of $\exp\left[\sum_{1\le i<j\le n} x_i x_j + \varepsilon_\lambda(x_1,\dots,x_n)\right]$ to which $\varepsilon_\lambda(x_1,\dots,x_n)$ contributes are in finite number (independently of λ) and are all of order $O(\lambda^{-1/2})$. Hence, taking $\lambda\to\infty$, property (7.13) is proved for $d=1$ in (7.16).

Step 2: recursion at $d+1$. Let us take a pair of random trees in $\mathcal{X}_{d+1}$ sampled from the correlated model given in Definition 7.1.9, with $N, N' \in \mathbb{R}^{\mathcal{X}_d}$ their corresponding vector representations. Given $k, y \in \mathbb{R}^{\mathcal{X}_d}$ we shall write $k\cdot y := \sum_{\alpha\in\mathcal{X}_d} k_\alpha y_\alpha$. The characteristic function of $\mathbb{P}_{d+1}^{(\lambda,s)}$ is defined as $\widehat{\mathbb{P}}_{d+1}^{(\lambda,s)}(k,k') := \mathbb{E}\left[e^{ik\cdot N + ik'\cdot N'}\right]$ and writes

$$\widehat{\mathbb{P}}_{d+1}^{(\lambda,s)}(k,k') = \exp\left[\lambda(1-s)\sum_{t\in\mathcal{X}_d}\mathsf{GW}_d^{(\lambda)}(t)(e^{ik_t}+e^{ik'_t}-2) + \lambda s\sum_{t,t'\in\mathcal{X}_d}\mathbb{P}_d^{(\lambda,s)}(t,t')(e^{ik_t+ik'_{t'}}-1)\right]$$

$$= e^{\lambda \sum_{t\in\mathcal{X}_d} \mathsf{GW}_d^{(\lambda)}(t)(e^{ik_t}-1)+\lambda \sum_{t\in\mathcal{X}_d} \mathsf{GW}_d^{(\lambda)}(t)(e^{ik'_t}-1)} \exp\left[\lambda s \sum_{t,t'\in\mathcal{X}_d} \mathbb{P}_d^{(\lambda,s)}(t,t')(e^{ik_t}-1)(e^{ik'_{t'}}-1)\right]$$

$$= e^{\lambda \sum_{t\in\mathcal{X}_d} \mathsf{GW}_d^{(\lambda)}(t)(e^{ik_t}-1)+\lambda \sum_{t\in\mathcal{X}_d} \mathsf{GW}_d^{(\lambda)}(t)(e^{ik'_t}-1)}$$
$$\times \sum_{n\geq 0} s^n \frac{\lambda^n}{n!} \underbrace{\left(\sum_{t,t'\in\mathcal{X}_d} \mathbb{P}_d^{(\lambda,s)}(t,t')(e^{ik_t}-1)(e^{ik'_{t'}}-1)\right)^n}_{(i)} . \tag{7.17}$$

Let us use the decomposition (7.9) at step d in (i). Denoting $g_{d,\alpha}^{(\lambda)}(t) := f_{d,\alpha}^{(\lambda)}(t)\mathsf{GW}_d^{(\lambda)}(t)$, this gives

$$(i) = \left(\sum_{\alpha\in\mathcal{X}_d} s^{|\alpha|-1}\left[\sum_{t\in\mathcal{X}_d} g_{d,\alpha}^{(\lambda)}(t)(e^{ik_t}-1)\right]\left[\sum_{t\in\mathcal{X}_d} g_{d,\alpha}^{(\lambda)}(t)(e^{ik'_t}-1)\right]\right)^n$$
$$= \sum_{\beta=(\beta_\alpha)_{\alpha\in\mathcal{X}_d}} n! s^{-n+\sum_{\alpha\in\mathcal{X}_d}\beta_\alpha|\alpha|}$$
$$\times \prod_{\alpha\in\mathcal{X}_d} \frac{1}{\beta_\alpha!}\left[\sum_{t\in\mathcal{X}_d} g_{d,\alpha}^{(\lambda)}(t)(e^{ik_t}-1)\right]^{\beta_\alpha}\left[\sum_{t\in\mathcal{X}_d} g_{d,\alpha}^{(\lambda)}(t)(e^{ik'_t}-1)\right]^{\beta_\alpha} \mathbb{1}_{\sum_{\alpha\in\mathcal{X}_d}\beta_\alpha=n} .$$

Summing (i) for $n \geq 0$ gives an overall sum over all $\beta = (\beta_\alpha)_{\alpha\in\mathcal{X}_d}$ that is over all $\mathcal{X}_{d+1}$. Moreover, for $\beta = (\beta_\alpha)_{\alpha\in\mathcal{X}_d} \in \mathcal{X}_{d+1}$, one always has

$$|\beta| = 1 + \sum_{\alpha\in\mathcal{X}_d} \beta_\alpha|\alpha| .$$

Hence, equation (7.17) becomes

$$\widehat{\mathbb{P}}_{d+1}^{(\lambda,s)}(k,k') = \sum_{\substack{\beta\in\mathcal{X}_{d+1}\\ \beta=(\beta_\alpha)_{\alpha\in\mathcal{X}_d}}} s^{|\beta|-1} \prod_{\alpha\in\mathcal{X}_d} \frac{1}{\beta_\alpha!}\left(\lambda\left[\sum_{t\in\mathcal{X}_d} g_{d,\alpha}^{(\lambda)}(t)(e^{ik_t}-1)\right]\left[\sum_{t\in\mathcal{X}_d} g_{d,\alpha}^{(\lambda)}(t)(e^{ik'_t}-1)\right]\right)^{\beta_\alpha}$$
$$\times e^{\lambda \sum_{t\in\mathcal{X}_d} \mathsf{GW}_d^{(\lambda)}(t)(e^{ik_t}-1)+\lambda \sum_{t\in\mathcal{X}_d} \mathsf{GW}_d^{(\lambda)}(t)(e^{ik'_t}-1)}$$
$$= \sum_{\substack{\beta\in\mathcal{X}_{d+1}\\ \beta=(\beta_\alpha)_{\alpha\in\mathcal{X}_d}}} s^{|\beta|-1} \widehat{g}_{d+1,\beta}^{(\lambda)}(k)\widehat{g}_{d+1,\beta}^{(\lambda)}(k') ,$$

with

$$\widehat{g}_{d+1,\beta}^{(\lambda)}(k) := e^{\lambda \sum_{t\in\mathcal{X}_d} \mathsf{GW}_d^{(\lambda)}(t)(e^{ik_t}-1)} \prod_{\alpha\in\mathcal{X}_d} \frac{1}{\sqrt{\beta_\alpha!}}\left[\sqrt{\lambda}\sum_{t\in\mathcal{X}_d} g_{d,\alpha}^{(\lambda)}(t)(e^{ik_t}-1)\right]^{\beta_\alpha}$$
$$= e^{-\lambda}\sqrt{\prod_\alpha \beta_\alpha!}\,[x^\beta] e^{\lambda \sum_{t\in\mathcal{X}_d} \mathsf{GW}_d^{(\lambda)}(t)e^{ik_t}+\sum_{\alpha\in\mathcal{X}_d} x_\alpha\sqrt{\lambda}\sum_{t\in\mathcal{X}_d} g_{d,\alpha}^{(\lambda)}(t)(e^{ik_t}-1)}$$
$$= e^{-\lambda}\sqrt{\prod_\alpha \beta_\alpha!}\,[x^\beta] e^{-\sqrt{\lambda}\sum_{\alpha,t\in\mathcal{X}_d} x_\alpha g_{d,\alpha}^{(\lambda)}(t)+\sum_t e^{ik_t}\left[\lambda \mathsf{GW}_d^{(\lambda)}(t)+\sum_\alpha x_\alpha\sqrt{\lambda} g_{d,\alpha}^{(\lambda)}(t)\right]} ,$$

where $x = \{x_\alpha\}_{\alpha\in\mathcal{X}_d}$ is a family of formal variables and x^β denotes $\prod_\alpha x_\alpha^{\beta_\alpha}$ when $\beta =$

$(\beta_\alpha)_{\alpha\in\mathcal{X}_d}$. Note that since the trees are finite, only a finite number of coordinates β_α are non zero, which makes the infinite product problem disappear. The same arguments of normal convergence as in the case $d=1$ apply to justify the integral/sum permutations.

As done in Step 1, we can invert the Fourier transform by integrating over every k_t, which gives

$$g^{(\lambda)}_{d+1,\beta}(N) = e^{-\lambda}\sqrt{\prod_\alpha \beta_\alpha!}\,[x^\beta]e^{-\sqrt{\lambda}\sum_{\alpha,t\in\mathcal{X}_d} x_\alpha g^{(\lambda)}_{d,\alpha}(t)} \prod_{t\in\mathcal{X}_d} \frac{\left[\lambda \mathsf{GW}^{(\lambda)}_d(t) + \sum_\alpha x_\alpha \sqrt{\lambda} g^{(\lambda)}_{d,\alpha}(t)\right]^{N_t}}{N_t!}\,.$$

It is now established that $L_{d+1}(N,N')$ satisfies the decomposition (7.9) with $f^{(\lambda)}_{d+1,\beta}$ given by the following recursion

$$f^{(\lambda)}_{d+1,\beta}(N) := \sqrt{\prod_\alpha \beta_\alpha!}\,[x^\beta]e^{-\sqrt{\lambda}\sum_{\alpha,t\in\mathcal{X}_d} x_\alpha g^{(\lambda)}_{d,\alpha}(t)} \prod_{t\in\mathcal{X}_d} \left(1 + \sum_{\alpha\in\mathcal{X}_d} \frac{x_\alpha}{\sqrt{\lambda}} f^{(\lambda)}_{d,\alpha}(t)\right)^{N_t}\,. \tag{7.18}$$

Taking $\beta = \bullet$ in (7.18), that is $\beta_\alpha = 0$ for all α, gives $f^{(\lambda)}_{d+1,\bullet} = 1$ and proves condition (7.10) at $d+1$.

Step 2.1: recursion for (7.11) ***at*** $d+1$**.** For any $\beta = \{\beta_\alpha\}_{\alpha\in\mathcal{X}_d}, \beta' = \{\beta'_\alpha\}_{\alpha\in\mathcal{X}_d} \in \mathcal{X}_{d+1}$, recursion (7.18) gives

$$\sum_{N\in\mathcal{X}_{d+1}} \mathsf{GW}^{(\lambda)}_{d+1}(N) f^{(\lambda)}_{d+1,\beta}(N) f^{(\lambda)}_{d+1,\beta'}(N) = \sqrt{\prod_\alpha \beta_\alpha!}\sqrt{\prod_\alpha \beta'_\alpha!}$$
$$\times\left[x^\beta y^{\beta'}\right] e^{-\lambda-\sqrt{\lambda}\sum_{\alpha,t\in\mathcal{X}_d}(x_\alpha+y_\alpha)g^{(\lambda)}_{d,\alpha}(t)+\lambda\sum_{t\in\mathcal{X}_d}\mathsf{GW}^{(\lambda)}_d(t)\left(1+\sum_{\alpha\in\mathcal{X}_d}\frac{x_\alpha}{\sqrt{\lambda}}f^{(\lambda)}_{d,\alpha}(t)\right)\left(1+\sum_{\alpha\in\mathcal{X}_d}\frac{y_\alpha}{\sqrt{\lambda}}f^{(\lambda)}_{d,\alpha}(t)\right)}\,.$$

The expression in the exponential in the above factorizes in several terms, that of order λ being $-1+1=0$. The term in $\sqrt{\lambda}$ is

$$-\sum_{\alpha,t\in\mathcal{X}_d}(x_\alpha+y_\alpha)g^{(\lambda)}_{d,\alpha}(t) + \sum_{\alpha,t\in\mathcal{X}_d}\mathsf{GW}^{(\lambda)}_d(t)x_\alpha f^{(\lambda)}_{d,\alpha}(t) + \sum_{\alpha,t\in\mathcal{X}_d}\mathsf{GW}^{(\lambda)}_d(t)y_\alpha f^{(\lambda)}_{d,\alpha}(t) = 0\,,$$

since $g^{(\lambda)}_{d,\alpha}(t) = \mathsf{GW}^{(\lambda)}_d(t) f^{(\lambda)}_{d,\alpha}(t)$ by definition. The only remaining term is constant and evaluates to

$$\sum_{\alpha,\alpha'} x_\alpha y_{\alpha'} \sum_{t\in\mathcal{X}_d} \mathsf{GW}^{(\lambda)}_d(t) f^{(\lambda)}_{d,\alpha}(t) f^{(\lambda)}_{d,\alpha'}(t) = \sum_\alpha x_\alpha y_\alpha\,,$$

using the orthogonality property (7.11) at step d. Hence,

$$\sum_{N\in\mathcal{X}_{d+1}} \mathsf{GW}^{(\lambda)}_{d+1}(N) f^{(\lambda)}_{d+1,\beta}(N) f^{(\lambda)}_{d+1,\beta'}(N) = \sqrt{\prod_\alpha \beta_\alpha!}\sqrt{\prod_\alpha \beta'_\alpha!}\left[x^\beta y^{\beta'}\right] e^{\sum_\alpha x_\alpha y_\alpha} = \mathbb{1}_{\beta=\beta'}\,,$$

which proves (7.11) at $d+1$. Previous computations are made rigorous since the trees are finite, and by noticing that the series in (7.18) has infinite radius of convergence, and appealing to Fubini's theorem. We use the same arguments to make computations rigorous in the rest of the proof.

Step 2.2: recursion for (7.12) ***at*** $d+1$**.** We are going to transform equation (7.18) as for

the step $d = 1$, as follows:

$$f^{(\lambda)}_{d+1,\beta}(N) = \sqrt{\prod_\alpha \beta_\alpha!} \prod_t N_t! \, [x^\beta y^N] e^{-\sqrt{\lambda}\sum_{\alpha,t\in\mathcal{X}_d} x_\alpha g^{(\lambda)}_{d,\alpha}(t) + \sum_{t\in\mathcal{X}_d} y_t + \sum_{\alpha,t\in\mathcal{X}_d} \frac{x_\alpha y_\alpha}{\sqrt{\lambda}} f^{(\lambda)}_{d,\alpha}(t)}$$
$$= \frac{\prod_t N_t!}{\sqrt{\prod_\alpha \beta_\alpha!}} [x^N] e^{\sum_t x_t} \prod_\alpha \left(\sum_t f^{(\lambda)}_{d,\alpha}(t) \left(\frac{x_t}{\sqrt{\lambda}} - \sqrt{\lambda}\mathsf{GW}^{(\lambda)}_d(t) \right)^{\beta_\alpha} \right) . \tag{7.19}$$

Using (7.19) gives that for all $N, N' \in \mathcal{X}_{d+1}$,

$$\sum_{\beta\in\mathcal{X}_{d+1}} f^{(\lambda)}_{d+1,\beta}(N) f^{(\lambda)}_{d+1,\beta}(N') = \prod_t N_t! N'_t! \, [x^N y^{N'}] e^{\sum_t (x_t+y_t)}$$
$$\times e^{\sum_{\alpha,t,t'} f^{(\lambda)}_{d,\alpha}(t) f^{(\lambda)}_{d,\alpha}(t') \left(\frac{x_t}{\sqrt{\lambda}} - \sqrt{\lambda}\mathsf{GW}^{(\lambda)}_d(t)\right)\left(\frac{y_{t'}}{\sqrt{\lambda}} - \sqrt{\lambda}\mathsf{GW}^{(\lambda)}_d(t')\right)}$$
$$= \prod_t N_t! N'_t! \, [x^N y^{N'}] e^{\sum_t (x_t+y_t) + \sum_t \frac{1}{\mathsf{GW}^{(\lambda)}_d(t)} \left(\frac{x_t}{\sqrt{\lambda}} - \sqrt{\lambda}\mathsf{GW}^{(\lambda)}_d(t)\right)\left(\frac{y_t}{\sqrt{\lambda}} - \sqrt{\lambda}\mathsf{GW}^{(\lambda)}_d(t)\right)} ,$$

where we used (7.12) at step d in the last step. This simplifies to

$$\sum_{\beta\in\mathcal{X}_{d+1}} f^{(\lambda)}_{d+1,\beta}(N) f^{(\lambda)}_{d+1,\beta}(N') = \prod_t N_t! N'_t! \, [x^N y^{N'}] \, e^{\lambda\mathsf{GW}^{(\lambda)}_d(t) + \sum_t \frac{x_t y_t}{\lambda \mathsf{GW}^{(\lambda)}_d(t)}}$$
$$= \prod_t \mathbb{1}_{N=N'} e^{\lambda\mathsf{GW}^{(\lambda)}_d(t)} (\lambda \mathsf{GW}^{(\lambda)}_d(t))^{-N_t} N_t! = \frac{\mathbb{1}_{N=N'}}{\mathsf{GW}^{(\lambda)}_{d+1}(N)} ,$$

which proves (7.12) at step $d+1$.

Step 2.3: recursion for (7.13) ***at*** $d+1$**.** Let us now prove property (7.13). For any $\beta^{(1)} = \{\beta^{(1)}_\alpha\}_{\alpha\in\mathcal{X}_d}, \dots, \beta^{(n)} = \{\beta^{(n)}_\alpha\}_{\alpha\in\mathcal{X}_d} \in \mathcal{X}_{d+1}$, recursion (7.18) gives

$$\sum_{N\in\mathcal{X}_{d+1}} \mathsf{GW}^{(\lambda)}_{d+1}(N) f^{(\lambda)}_{d+1,\beta^{(1)}}(N) \cdots f^{(\lambda)}_{d+1,\beta^{(n)}}(N) = \sqrt{\prod_{i=1}^n \prod_\alpha \beta_\alpha!} \left[\prod_{i=1}^n (x^{(i)})^{\beta^{(i)}} \right]$$
$$\times \exp\left[-\lambda - \sqrt{\lambda} \sum_{\alpha,t\in\mathcal{X}_d} \sum_{i=1}^n x^{(i)}_\alpha g^{(\lambda)}_{d,\alpha}(t) + \lambda \sum_{t\in\mathcal{X}_d} \mathsf{GW}^{(\lambda)}_d(t) \prod_{i=1}^n \left(1 + \sum_{\alpha\in\mathcal{X}_d} \frac{x^{(i)}_\alpha}{\sqrt{\lambda}} f^{(\lambda)}_{d,\alpha}(t) \right) \right] .$$

As in Step 2.1, when expanding the product in the exponential, the zero and first order terms simplify, which yields

$$\sum_{N\in\mathcal{X}_{d+1}} \mathsf{GW}^{(\lambda)}_{d+1}(N) f^{(\lambda)}_{d+1,\beta^{(1)}}(N) \cdots f^{(\lambda)}_{d+1,\beta^{(n)}}(N) =$$
$$\sqrt{\prod_{i=1}^n \prod_\alpha \beta_\alpha!} \left[\prod_{i=1}^n (x^{(i)})^{\beta^{(i)}} \right] e^{\sum_{1\le i<j\le n} \sum_{\alpha,\alpha'\in\mathcal{X}_d} x^{(i)}_\alpha x^{(j)}_{\alpha'} \sum_{t\in\mathcal{X}_d} \mathsf{GW}^{(\lambda)}_d(t) f^{(\lambda)}_{d,\alpha}(t) f^{(\lambda)}_{d,\alpha'}(t) + \varepsilon_\lambda(x^{(1)},\dots,x^{(n)})} , \tag{7.20}$$

with

$$\varepsilon_\lambda(x^{(1)},\dots,x^{(n)}) := \sum_{p=3}^n \lambda^{1-p/2} \sum_{1\le i_1<\dots<i_p\le n} \sum_{\alpha_1,\dots,\alpha_p\in\mathcal{X}_d} x^{i_1}_{\alpha_{i_1}} \cdots x^{i_p}_{\alpha_{i_p}} \sum_{t\in\mathcal{X}_d} \mathsf{GW}^{(\lambda)}_d(t) f^{(\lambda)}_{d,\alpha_1}(t) \cdots f^{(\lambda)}_{d,\alpha_p}(t) .$$

Using orthogonality (7.11) at step d, (7.20) writes

$$\sum_{N\in\mathcal{X}_{d+1}} \mathsf{GW}_{d+1}^{(\lambda)}(N) f_{d+1,\beta^{(1)}}^{(\lambda)}(N)\cdots f_{d+1,\beta^{(n)}}^{(\lambda)}(N) = \sqrt{\prod_{i=1}^{n}\prod_{\alpha}\beta_\alpha!}\left[\prod_{i=1}^{n}(x^{(i)})^{\beta^{(i)}}\right] e^{\sum_{1\leq i<j\leq n}\sum_{\alpha\in\mathcal{X}_d} x_\alpha^{(i)}x_\alpha^{(j)}} \times \exp\left[\varepsilon_\lambda(x^{(1)},\dots,x^{(n)})\right], \quad (7.21)$$

Using property (7.13) at step d, $\sum_{t\in\mathcal{X}_d}\mathsf{GW}_d^{(\lambda)}(t)f_{d,\alpha_1}^{(\lambda)}(t)\cdots f_{d,\alpha_p}^{(\lambda)}(t)$ has a finite limit when $\lambda\to\infty$. Hence, as in Step 1, the terms corresponding to $\left[\prod_{i=1}^{n}(x^{(i)})^{\beta^{(i)}}\right]$ in (7.21) to which $\varepsilon_\lambda(x^{(1)},\dots,x^{(n)})$ contributes are in finite number (independent of λ) and are all of order $O(\lambda^{-1/2})$.

Taking $\lambda\to\infty$ thus establishes property (7.13) for $d+1$ and completes the proof of Theorem 7.3. □

7.2.2. Computation of cyclic moments, proof of Theorem 7.1

Theorem 7.3 hereabove has a very natural corollary that enables to compute the cyclic moments of the likelihood ratio.

Corollary 7.2.1 (Cyclic moments)**.** *The* $n-$th cyclic moment *of* L_d *is defined as follows*

$$C_{d,m}^{(\lambda,s)} := \mathbb{E}_d^{(\lambda)}\left[L_d(T_1,T_2)\cdots L_d(T_{m-1},T_m)L_d(T_m,T_1)\right],$$

where $T_1,\dots,T_m$ *are i.i.d.* $\mathsf{GW}_d^{(\lambda)}$ *in the above expectation. One has that*

$$C_{d,m}^{(\lambda,s)} = \sum_{\alpha\in\mathcal{X}_d}(s^m)^{|\alpha|-1} = \sum_{n\geq 1} A_{d,n}(s^m)^{n-1} = \Phi_d(s^m), \quad (7.22)$$

where $A_{d,n}$*, as defined in* (7.2)*, denotes the number of unlabeled trees with* n *vertices of depth at most* d*, and* Φ_d *is the generating function defined in Proposition 7.1.2. Note that in particular, the* $C_{d,m}^{(\lambda,s)}$ *do not depend on* λ *(!) and by Proposition 7.1.2 they are upper bounded for each* d *and* $s\in[0,1)$ *by some constant* $A=A(d,s)$*. We thus denote* $C_{d,m}^{(s)} := C_{d,m}^{(\lambda,s)}$ *in the sequel.*

Proof of Corollary 7.2.1. By Theorem 7.3 we have

$$L_d(t,t') = \sum_{\alpha\in\mathcal{X}_d} s^{|\alpha|-1} f_{d,\alpha}^{(\lambda)}(t) f_{d,\alpha}^{(\lambda)}(t'),$$

hence, setting $\alpha_{m+1}=\alpha_1$,

$$\begin{aligned}
C_{d,m}^{(\lambda,s)} &= \mathbb{E}_d^{(\lambda)}\left[L_d(T_1,T_2)\cdots L_d(T_{m-1},T_m)L_d(T_m,T_1)\right]\\
&= \sum_{\alpha_1,\dots,\alpha_m\in\mathcal{X}_d} s^{\sum_{i=1}^m(|\alpha_i|-1)}\mathbb{E}_d^{(\lambda)}\left[\prod_{i=1}^m f_{d,\alpha_i}^{(\lambda)}(T_i)f_{d,\alpha_{i+1}}^{(\lambda)}(T_i)\right]\\
&= \sum_{\alpha_1,\dots,\alpha_m\in\mathcal{X}_d} s^{\sum_{i=1}^m(|\alpha_i|-1)}\prod_{i=1}^m \mathbb{E}_d^{(\lambda)}\left[f_{d,\alpha_i}^{(\lambda)}(T)f_{d,\alpha_{i+1}}^{(\lambda)}(T)\right]\\
&= \sum_{\alpha_1,\dots,\alpha_m\in\mathcal{X}_d} s^{\sum_{i=1}^m(|\alpha_i|-1)}\mathbb{1}_{\alpha_1=\dots=\alpha_m}
\end{aligned}$$

$$= \sum_{\alpha \in \mathcal{X}_d} (s^m)^{|\alpha|-1} .$$

All steps in the above computations are legitimate by Fubini's theorem, since

$$\mathbb{E}_d^{(\lambda)} \left[\left| f_{d,\alpha}^{(\lambda)}(T) f_{d,\alpha'}^{(\lambda)}(T) \right| \right] \leq \mathbb{E}_d^{(\lambda)} \left[(f_{d,\alpha}^{(\lambda)}(T))^2 \right]^{1/2} \mathbb{E}_d^{(\lambda)} \left[(f_{d,\alpha'}^{(\lambda)}(T))^2 \right]^{1/2} = 1 ,$$

by property (7.12) of Theorem 7.3. □

We are now ready to give a proof of Theorem 7.1.

Proof of Theorem 7.1. According to Corollary 7.2.1, one has

$$\mathbb{E}_d^{(\lambda)} \left[L_d(T, T')^2 \right] = C_{d,2}^{(s)} = \sum_{n \geq 1} A_{d,n} s^{2(n-1)}. \tag{7.23}$$

Moreover, since $A_{d,n} \leq A_n$ (by Definition 7.1.6) and $A_n \underset{n \to \infty}{\sim} \frac{C}{n^{3/2}} \left(\frac{1}{\alpha}\right)^n$ by Proposition 7.1.1, the assumption $s \leq \sqrt{\alpha}$ ensures that $\mathbb{E}_d^{(\lambda)} \left[L_d(T, T')^2 \right] = \sum_{n \geq 1} A_{d,n} s^{2(n-1)} \leq \sum_{n \geq 1} A_n s^{2(n-1)} < \infty$, uniformly in d.

Then, applying Jensen's inequality yields

$$\mathrm{KL}(\mathbb{P}_d^{(\lambda,s)} \,||\, \mathbb{P}_d^{(\lambda)}) = \mathbb{E}_d^{(\lambda,s)} \left[\log L_d(T, T') \right] \leq \log \mathbb{E}_d^{(\lambda,s)} \left[L_d(T, T') \right] = \log \mathbb{E}_d^{(\lambda)} \left[L_d^2(T, T') \right] < \infty,$$

uniformly in d, and concludes the proof. □

For the positive result, we need to study the weak convergence of the likelihood ratio when $\lambda \to \infty$, which is the scope of the next Section, concluded by the proof of Theorem 7.2.

7.3. The high-degree regime: positive result when $s > \sqrt{\alpha}$ in the gaussian approximation

In view of definition 7.1.9, we recall that a pair of correlated trees (t, t') of depth at most $d+1$ sampled from $\mathbb{P}_{d+1}^{(\lambda,s)}$ are of the form $t = \{N_\tau\}_{\tau \in \mathcal{X}_d}$ and $t' = \{N'_\tau\}_{\tau \in \mathcal{X}_d}$ with

$$N_\tau := M_\tau + \sum_{\tau' \in \mathcal{X}_d} N_{\tau,\tau'} \quad \text{and} \quad N'_\tau := M'_\tau + \sum_{\tau \in \mathcal{X}_d} N_{\tau,\tau'} . \tag{7.24}$$

with

$$M_\tau, M'_\tau \overset{\text{i.i.d.}}{\sim} \mathsf{Poi}(\lambda(1-s)\mathsf{GW}_d^{(\lambda)}(\tau)) \quad \text{and} \quad N_{\tau,\tau'} \sim \mathsf{Poi}(\lambda s \mathbb{P}_d^{(\lambda,s)}(\tau, \tau')) .$$

7.3.1. Gaussian approximation

Let us define $y = (y_\alpha)_{\alpha \in \mathcal{X}_d}$ and $y' = (y'_\alpha)_{\alpha \in \mathcal{X}_d}$ as follows:

$$y_\alpha := \frac{1}{\sqrt{\lambda}} \sum_{\tau \in \mathcal{X}_d} f_{d,\alpha}^{(\lambda)}(\tau)(N_\tau - \lambda \mathsf{GW}_d^{(\lambda)}(\tau)) \tag{7.25}$$

$$y'_\alpha := \frac{1}{\sqrt{\lambda}} \sum_{\tau \in \mathcal{X}_d} f_{d,\alpha}^{(\lambda)}(\tau)(N'_\tau - \lambda \mathsf{GW}_d^{(\lambda)}(\tau)) \tag{7.26}$$

where the $f_{d,\alpha}^{(\lambda)}$ are defined in Theorem 7.3. In other words, y (resp. y') is a centered version of N (resp. N'), projected onto the basis of eigenvectors.

Let $(z, z') = ((z_\alpha)_{\alpha\in\mathcal{X}_d}, (z'_{\alpha'})_{\alpha'\in\mathcal{X}_d}))$ be an (infinite-dimensional) centered Gaussian vector defined by its covariance matrix:

$$\forall \alpha, \alpha' \in \mathcal{X}_d, \quad \mathbb{E}[z_\alpha z_{\alpha'}] = \mathbb{E}[z'_\alpha z'_{\alpha'}] = \mathbb{1}_{\alpha=\alpha'}, \quad \mathbb{E}[z_\alpha z'_{\alpha'}] = s^{|\alpha|}\mathbb{1}_{\alpha=\alpha'}. \tag{7.27}$$

Let us denote by $\mathsf{p}_{d+1}^{(\lambda,s)}$ the joint distribution of (y, y'), and $\mathsf{gw}_{d+1}^{(\lambda)}$ the marginal distribution of y (or y'). Since the transformations $N \to y$ in (7.25) and $N' \to y'$ in (7.26) are bijective in view of the orthogonality property (7.12) in Theorem 7.3, one has

$$\mathrm{KL}(\mathbb{P}_{d+1}^{(\lambda,s)} \| \mathsf{GW}_{d+1}^{(\lambda)} \otimes \mathsf{GW}_{d+1}^{(\lambda)}) = \mathrm{KL}(\mathsf{p}_{d+1}^{(\lambda,s)} \| \mathsf{gw}_{d+1}^{(\lambda)} \otimes \mathsf{gw}_{d+1}^{(\lambda)}) \,. \tag{7.28}$$

Lemma 7.3.1. *When $\lambda \to \infty$, we have the following convergence in distribution:*

$$(y, y') \xrightarrow{\text{(d)}} (z, z'). \tag{7.29}$$

Proof of Lemma 7.3.1. With the canonical product sigma-field, convergence in distribution of (y, y') amounts to convergence of all finite-dimensional distributions. Let us denote by (k, k') a pair of real vectors in $\mathbb{R}^{\mathcal{X}_d\times\mathcal{X}_d}$ such that only a finite number of entries are non-zero. We shall write $k \cdot y := \sum_{\alpha\in\mathcal{X}_d} k_\alpha y_\alpha$. We will also define the following characteristic functions:

$$\widehat{\mathsf{p}}^{(\lambda,s)}(k, k') := \mathbb{E}\left[e^{ik\cdot y + ik'\cdot y'}\right] \quad \text{and} \quad \widehat{\mathsf{r}}^{(s)}(k, k') := \mathbb{E}\left[e^{ik\cdot z + ik'\cdot z'}\right] . \tag{7.30}$$

Proving Lemma 7.3.1 thus amounts to showing the simple convergence $\widehat{\mathsf{p}}^{(\lambda,s)}(k, k') \to \widehat{\mathsf{r}}^{(s)}(k, k')$ when $\lambda \to \infty$. Since the (gaussian) limit distribution is entirely determined by its moments, it suffices to show the convergence of the cumulants [JLR00]. The covariance structure of (z, z') given in (7.27) immediately yields

$$\widehat{\mathsf{r}}^{(s)}(k, k') = \exp\left[-\frac{1}{2}\sum_{\alpha\in\mathcal{X}_d}((k_\alpha)^2 + (k'_\alpha)^2 + 2s^{|\alpha|}k_\alpha k'_\alpha)\right] . \tag{7.31}$$

Then, in view of (7.24), (7.25) and (7.26), writing $f_d^{(\lambda)}(\tau) := (f_{d,\alpha}^{(\lambda)}(\tau))_{\alpha\in\mathcal{X}_d}$, one has

$$\begin{aligned} e^{ik\cdot y + ik'\cdot y'} = \exp\left[-\sqrt{\lambda}\sum_{\tau\in\mathcal{X}_d}\mathsf{GW}_d^{(\lambda)}(\tau)(ik\cdot f_d^{(\lambda)}(\tau) + ik'\cdot f_d^{(\lambda)}(\tau))\right] \\ \times \prod_{\tau,\tau'\in\mathcal{X}_d}\left(\exp\left[\frac{1}{\sqrt{\lambda}}(ik\cdot f_d^{(\lambda)}(\tau) + ik'\cdot f_d^{(\lambda)}(\tau'))\right]\right)^{N_{\tau,\tau'}} \times \prod_{\tau\in\mathcal{X}_d}\left(\exp\left[\frac{1}{\sqrt{\lambda}}ik\cdot f_d^{(\lambda)}(\tau)\right]\right)^{M_\tau} \\ \times \prod_{\tau\in\mathcal{X}_d}\left(\exp\left[\frac{1}{\sqrt{\lambda}}ik'\cdot f_d^{(\lambda)}(\tau)\right]\right)^{M'_\tau} . \end{aligned}$$

Variables $N_{\tau,\tau'}, M_\tau, M'_\tau$ being independent Poisson variables, taking the expectation gives

$$\begin{aligned} \widehat{\mathsf{p}}^{(\lambda,s)}(k, k') = \exp\left[-\sqrt{\lambda}\sum_{\tau\in\mathcal{X}_d}\mathsf{GW}_d^{(\lambda)}(\tau)(ik\cdot f_d^{(\lambda)}(\tau) + ik'\cdot f_d^{(\lambda)}(\tau))\right] \\ \times \exp\left[\lambda(1-s)\sum_{\tau\in\mathcal{X}_d}\mathsf{GW}_d^{(\lambda)}(\tau)\left(e^{\frac{1}{\sqrt{\lambda}}ik\cdot f_d^{(\lambda)}(\tau)} + e^{\frac{1}{\sqrt{\lambda}}ik'\cdot f_d^{(\lambda)}(\tau)} - 2\right)\right] \end{aligned}$$

$$\times \exp\left[\lambda s \sum_{\tau,\tau' \in \mathcal{X}_d} \mathbb{P}_d^{(\lambda,s)}(\tau,\tau') \left(e^{\frac{1}{\sqrt{\lambda}}(ik\cdot f_d^{(\lambda)}(\tau) + ik'\cdot f_d^{(\lambda)}(\tau'))} - 1 \right)\right] .$$

The cumulants are obtained by expanding the logarithm of the last expression in power series in k, k'. Using that $\sum_{\tau' \in \mathcal{X}_d} \mathbb{P}_d^{(\lambda,s)}(\tau,\tau') = \mathsf{GW}_d^{(\lambda)}(\tau)$, the first-order (linear) terms compensate to 0, which is coherent with the fact that $\mathbb{E}[y_\alpha] = \mathbb{E}[y'_\alpha] = 0$. The second-order terms in $\log \widehat{\mathsf{p}}^{(\lambda,s)}(k,k')$ evaluate to

$$\begin{aligned}
&-\lambda(1-s) \sum_{\tau \in \mathcal{X}_d} \mathsf{GW}^{(\lambda)}(\tau) \frac{1}{2\lambda} \sum_{\alpha,\alpha' \in \mathcal{X}_d} f_{d,\alpha}^{(\lambda)}(\tau) f_{d,\alpha'}^{(\lambda)}(\tau) \left(k_\alpha k_{\alpha'} + k'_\alpha k'_{\alpha'}\right) \\
&-\lambda s \sum_{\tau,\tau' \in \mathcal{X}_d} \mathbb{P}_d^{(\lambda,s)}(\tau,\tau') \\
&\qquad \times \frac{1}{2\lambda} \sum_{\alpha,\alpha' \in \mathcal{X}_d} \left(f_{d,\alpha}^{(\lambda)}(\tau) f_{d,\alpha'}^{(\lambda)}(\tau) k_\alpha k_{\alpha'} + f_{d,\alpha}^{(\lambda)}(\tau') f_{d,\alpha'}^{(\lambda)}(\tau') k'_\alpha k'_{\alpha'} + 2 f_{d,\alpha}^{(\lambda)}(\tau) f_{d,\alpha'}^{(\lambda)}(\tau') k_\alpha k'_{\alpha'} \right) .
\end{aligned}$$

Using the orthogonality property of the eigenvectors in Theorem 7.3, namely that

$$\forall \alpha, \alpha' \in \mathcal{X}_d, \quad \sum_{\tau \in \mathcal{X}_d} \mathsf{GW}_d^{(\lambda)}(\tau) f_{d,\alpha}^{(\lambda)}(\tau) f_{d,\alpha'}^{(\lambda)}(\tau) = \mathbb{1}_{\alpha=\alpha'} ,$$

the previous equation simplifies to

$$-\frac{1}{2} \sum_{\alpha \in \mathcal{X}_d} \left((k_\alpha)^2 + (k'_\alpha)^2 \right) - s \sum_{\tau,\tau' \in \mathcal{X}_d} \mathbb{P}_d^{(\lambda,s)}(\tau,\tau') \sum_{\alpha,\alpha' \in \mathcal{X}_d} f_{d,\alpha}^{(\lambda)}(\tau) f_{d,\alpha'}^{(\lambda)}(\tau') k_\alpha k'_{\alpha'} ,$$

which in turn writes, using $\mathbb{P}_d^{(\lambda,s)}(\tau,\tau') = \mathsf{GW}_d^{(\mu)}(\tau) \mathsf{GW}_d^{(\mu)}(\tau') \sum_{\alpha \in \mathcal{X}_d} s^{|\alpha|-1} f_{d,\alpha}^{(\lambda)}(\tau) f_{d,\alpha}^{(\lambda)}(\tau')$

$$\begin{aligned}
&-\frac{1}{2} \sum_{\alpha \in \mathcal{X}_d} \left((k_\alpha)^2 + (k'_\alpha)^2 \right) - s \sum_{\alpha,\alpha',\alpha'' \in \mathcal{X}_d} s^{|\alpha''|-1} k_\alpha k'_{\alpha'} \\
&\qquad \times \left(\sum_{\tau \in \mathcal{X}_d} \mathsf{GW}_d^{(\mu)}(\tau) f_{d,\alpha}^{(\lambda)}(\tau) f_{d,\alpha''}^{(\lambda)}(\tau) \right) \left(\sum_{\tau' \in \mathcal{X}_d} \mathsf{GW}_d^{(\mu)}(\tau') f_{d,\alpha'}^{(\lambda)}(\tau') f_{d,\alpha''}^{(\lambda)}(\tau') \right) \\
&= -\frac{1}{2} \sum_{\alpha \in \mathcal{X}_d} \left((k_\alpha)^2 + (k'_\alpha)^2 + 2 s^{|\alpha|} k_\alpha k'_{\alpha'} \right) ,
\end{aligned}$$

which is exactly the second cumulant of (z, z') in (7.31). The remaining step is to show that the higher order cumulants tend to 0 when λ gets large. Note that the cumulants of order ≥ 3 have a factor $1/\sqrt{\lambda}$, but the implicit dependence of the $f_{d,\alpha}^{(\lambda)}$ needs to be controlled. The previous computations show that the cumulants depend on terms of the form

$$\sum_{t \in \mathcal{X}_d} \mathsf{GW}_d^{(\lambda)}(t) f_{d,\alpha_1}^{(\lambda)}(t) \cdots f_{d,\alpha_p}^{(\lambda)}(t),$$

which are proved to remain finite when $\lambda \to \infty$ by property (7.13) of Theorem 7.3. This shows that all the cumulants of order ≥ 3 tend to 0 and hence establishes the desired convergence in distribution. □

7.3.2. Kullback-Leibler divergence in the high-degree regime

In view of the previous weak convergence established in Lemma 7.3.1, we will now prove the following result, which compares the KL−divergence with finite λ to the KL−divergence

between the limiting Gaussian distributions of Lemma 7.3.1.

Proposition 7.3.1. *Denoting*

$$\mathrm{KL}_d^{(\lambda,s)} := \mathrm{KL}(\mathbb{P}_d^{(\lambda,s)} || \mathbb{P}_d^{(\lambda)}) \,, \tag{7.32}$$

one has the following:

$$\forall d \geq 1, \ \liminf_{\lambda \to \infty} \mathrm{KL}_d^{(\lambda,s)} \geq \mathrm{KL}_d^{(s)} := -\frac{1}{2} \sum_{\alpha \in \mathcal{X}_{d-1}} \log(1 - s^{2|\alpha|}) = \frac{1}{2} \log C_{d,2}^{(s)} \,. \tag{7.33}$$

We recall that

$$C_{d,2}^{(s)} = \mathbb{E}_d^{(\lambda,s)}[L_d] = \sum_{n \geq 1} A_{d,n} (s^2)^{n-1} \,. \tag{7.34}$$

Proof of Theorem 7.3.1. Fix $d \geq 1$. In (7.28), we established that $\mathrm{KL}_d^{(\lambda,s)}$ is also the KL-divergence $\mathrm{KL}(\mathsf{p}_d^{(\lambda,s)} \| \mathsf{gw}_d^{(\lambda)} \otimes \mathsf{gw}_d^{(\lambda)})$ where $\mathsf{p}_d^{(\lambda,s)}$ is the distribution of (y, y') defined in Section 7.3.1. Moreover, Lemma 7.3.1 establishes that (y, y') converges in distribution a centered gaussian vector (z, z') defined by its covariance matrix:

$$\forall \alpha, \alpha' \in \mathcal{X}_{d-1}, \quad \mathbb{E}[z_\alpha z_{\alpha'}] = \mathbb{E}[z'_\alpha z'_{\alpha'}] = \mathbb{1}_{\alpha = \alpha'}, \quad \mathbb{E}[z_\alpha z'_{\alpha'}] = s^{|\alpha|} \mathbb{1}_{\alpha = \alpha'}. \tag{7.35}$$

If we denote by $p_1^{(s)}$ the joint distribution of the gaussian vector (z, z') and $p_0^{(s)}$ the product of the marginals, the KL-divergence $\mathrm{KL}(p_1^{(s)} || p_0^{(s)})$ is easily given by $-\frac{1}{2} \log \det \Sigma$, where Σ is the covariance matrix of (z, z'), which is similar to a matrix with diagonal blocks of the form $\begin{pmatrix} 1 & s^{|\alpha|} \\ s^{|\alpha|} & 1 \end{pmatrix}$ for all $\alpha \in \mathcal{X}_{d-1}$, which gives

$$\mathrm{KL}(p_1^{(s)} || p_0^{(s)}) = -\frac{1}{2} \log \prod_{\alpha \in \mathcal{X}_{d-1}} \frac{1}{1 - s^{2|\alpha|}} \,.$$

The last term is indeed $\mathrm{KL}_d^{(s)}$ as defined in (7.33), since

$$\mathbb{E}_d^{(\lambda,s)}[L_d] = \sum_{\beta \in \mathcal{X}_d} s^{2|\beta|-1} = \prod_{\alpha \in \mathcal{X}_{d-1}} \sum_{\beta_\alpha \geq 0} s^{2\beta_\alpha |\alpha|} = \prod_{\alpha \in \mathcal{X}_{d-1}} \frac{1}{1 - s^{2|\alpha|}} \,.$$

The roof is concluded by appealing to the lower semi-continuity property of the KL-divergence (see e.g. [PW17], Theorem 3.6), namely that

$$\liminf_{\lambda \to \infty} \mathrm{KL}_d^{(\lambda,s)} \geq \mathrm{KL}_d^{(s)} \,.$$

□

7.3.3. Propagating bounds on the KL−divergence, proof of Theorem 7.2

The goal of this section is to use the result of Proposition 7.3.1 and the fact that in view of (7.34) for $s > \sqrt{\alpha}$ (where α is Otter's constant), $\mathrm{KL}_d^{(s)} \to +\infty$ with d, in order to obtain Theorem 7.2, that is that for fixed $s > \sqrt{\alpha}$, there exists $\lambda = \lambda(s)$ such that:

$$\lambda \geq \lambda(s) \Rightarrow \lim_{d \to \infty} \mathrm{KL}_d^{(\lambda,s)} = +\infty \,.$$

The following Lemma shows that if $s > \sqrt{\alpha}$, for any small (resp. any large but bounded) probability that we fix, there exists a depth d_0 and an event S that has this small (resp.

large) probability under $\mathbb{P}_{d_0}^{(\lambda)}$ (resp. $\mathbb{P}_{d_0}^{(\lambda,s)}$). The proof is deferred to Appendix 7.A.1.

Lemma 7.3.2. *Assume that $s > \sqrt{\alpha}$. Then for any $c \in (0, 1/15)$ and any $\varepsilon \in (0,1)$, there exists $\lambda_1 = \lambda_1(s,c,\varepsilon) > 0$ and $d_0 = d_0(s,c,\varepsilon) \in \mathbb{N}$ such that, for all $\lambda \geq \lambda_1$, there exists an event $S = S(s,c,\varepsilon) \subset \mathcal{X}_{d_0}^2$ for which the following inequalities hold:*

$$\mathbb{P}_{d_0}^{(\lambda,s)}(S) \geq c \quad \textit{and} \quad \mathbb{P}_{d_0}^{(\lambda)}(S) \leq \varepsilon\,.$$

Now that we know that this event S exists at a certain initial depth d_0, we want to propagate the bounds for arbitrary depth $d \geq d_0$. This is the object of the following Proposition, proved in Appendix 7.A.2.

Proposition 7.3.2. *For any fixed $c \in (0,1)$ there exist constants $\varepsilon = \varepsilon(s,c) \in (0,1)$ and $\lambda_0 = \lambda_0(s,c) > 0$ such that the following holds. For any $\lambda \geq \lambda_0$, any $d \in \mathbb{N}$, if there exists an event $S \subset \mathcal{X}_d^2$ such that*

$$\mathbb{P}_{d}^{(\lambda)}(S) \leq \varepsilon \quad \textit{and} \quad \mathbb{P}_{d}^{(\lambda,s)}(S) \geq c\,,$$

then there exists an event $S' \subset \mathcal{X}_{d+1}^2$ such that

$$\mathbb{P}_{d+1}^{(\lambda)}(S') \leq \frac{1}{2}\mathbb{P}_{d}^{(\lambda)}(S) \leq \frac{\varepsilon}{2} \quad \textit{and} \quad \mathbb{P}_{d+1}^{(\lambda,s)}(S') \geq c\,.$$

In fact, using the usual notations $t = \{N_\tau\}_{\tau \in \mathcal{X}_d}$, $t' = \{N'_\tau\}_{\tau \in \mathcal{X}_d}$ for elements of $\mathcal{X}_{d+1}$, and denoting, for all $\tau \in \mathcal{X}_d$

$$\widetilde{N}_\tau := N_\tau - \lambda \mathsf{GW}_d^{(\lambda)}(\tau) \quad \textit{and} \quad \widetilde{N}'_\tau = N'_\tau - \lambda \mathsf{GW}_d^{(\lambda)}(\tau)\,,$$

the event S' in the above is defined from S in the following way

$$S' = \{Z_S \geq \sigma\}\,,$$

where

$$Z_S := \sum_{(\tau,\tau') \in S} \widetilde{N}_\tau \widetilde{N}'_{\tau'}\,, \tag{7.36}$$

and for some suitable threshold $\sigma = \sigma(S)$.

Together, Lemma 7.3.2 and Proposition 7.3.2 yield the proof of Theorem 7.2.

Proof of Theorem 7.2

Proof of Theorem 7.2. Assume that $s > \sqrt{\alpha}$. Choose $c \in (0, 1/15)$ and let $\varepsilon = \varepsilon(s,c)$, $\lambda_0 = \lambda_0(s,c)$ be the corresponding quantities from Proposition 7.3.2. Now that c, ε are fixed, we appeal to Lemma 7.3.2 to obtain some $\lambda_1 = \lambda_1(s,c,\varepsilon)$ and $d_0 = d_0(s,c,\varepsilon) \in \mathbb{N}$ such that, taking $\lambda \geq \lambda_0 \vee \lambda_1$, there exists some event $S_d \subset \mathcal{X}_d^2$ such that

$$\mathbb{P}_{d}^{(\lambda)}(S) \leq \varepsilon \quad \text{and} \quad \mathbb{P}_{d}^{(\lambda,s)}(S) \geq c\,.$$

Proposition 7.3.2 then ensures the existence of a sequence of events $S_d \subset \mathcal{X}_d^2$, $d > d_0$ such that

$$\mathbb{P}_{d}^{(\lambda)}(S_d) \leq 2^{-(d-d_0)}\varepsilon \quad \text{and} \quad \mathbb{P}_{d}^{(\lambda,s)}(S_d) \geq c.$$

It follows that, for all $d > d_0$,

$$\begin{aligned}
\mathrm{KL}_d^{(\lambda,s)} &\geq \mathbb{P}_d^{(\lambda,s)}(S_d) \log\left(\frac{\mathbb{P}_d^{(\lambda,s)}(S_d)}{\mathbb{P}_d^{(\lambda)}(S_d)}\right) + (1 - \mathbb{P}_d^{(\lambda,s)}(S_d)) \log\left(\frac{1 - \mathbb{P}_d^{(\lambda,s)}(S_d)}{1 - \mathbb{P}_d^{(\lambda)}(S_d)}\right) \\
&\geq c \log(2^{d-d_0}/\varepsilon) - h(\mathbb{P}_d^{(\lambda,s)}(S_d)) - (1 - \mathbb{P}_d^{(\lambda,s)}(S_d)) \log((1 - \mathbb{P}_d^{(\lambda)}(S_d)))
\end{aligned}$$

$$\geq c \log(2^{d-d_0}/\varepsilon) - h(\mathbb{P}_d^{(\lambda,s)}(S_d))\,,$$

where for $x \in [0,1]$, h is defined by $h(x) := -x\log(x) - (1-x)\log(1-x)$. Function h is maximal at $x = 1/2$ and $h(1/2) = \log(2)$, which gives the final bound $\mathrm{KL}_d^{(\lambda,s)} \geq c\log(2)(d-d_0) - c\log(\varepsilon) - \log(2)$. It readily follows that $\lim_{d\to\infty} \mathrm{KL}_{\lambda,d} = +\infty$. □

Appendix of Chapter 7

7.A. Postponed proofs

7.A.1. Proof of Lemma 7.3.2

Proof of Lemma 7.3.2. Fix $c < 1/15$ and $\varepsilon \in (0,1)$. Since $s > \sqrt{\alpha}$, we have that $\mathrm{KL}_d^{(s)} \to \infty$ when $d \to \infty$, in view of (7.34), the fact that $A_{d,n} \to A_n$ when $d \to \infty$, and Otter's formula 7.3. For arbitrarily large $C = C(c,\varepsilon)$ to be specified later, we can thus choose $d_0 = d_0(s,c,\varepsilon)$ such that $\mathrm{KL}_{d_0}^{(s)} = \frac{1}{2}\log(C_{d_0,2}^{(s)}) \geq C$.

In turn, in view of (7.33), we can choose $\lambda_1 = \lambda_1(s,c,\varepsilon)$ such that

$$\lambda \geq \lambda_1 \Rightarrow \mathrm{KL}_{d_0}^{(\lambda,s)} \geq \frac{1}{2}\mathrm{KL}_{d_0}^{(s)} \geq C/2\,.$$

Write then

$$\frac{1}{4}\log(C_{d_0,2}^{(s)}) \leq \mathrm{KL}_{d_0}^{(\lambda,s)} \leq \int_1^\infty \log(x)\mathbb{P}_{d_0}^{(\lambda,s)}(L_{d_0} \in dx) = \int_1^\infty \frac{1}{u}\mathbb{P}_{d_0}^{(\lambda,s)}(L_{d_0} \geq u)du.$$

Also, since $C_{d_0,2}^{(s)} = \mathbb{E}_{d_0}^{(\lambda,s)}[L_{d_0}]$,

$$C_{d_0,2}^{(s)} = \int_0^\infty x\mathbb{P}_{d_0}^{(\lambda,s)}(L_{d_0} \in dx) \geq \int_1^\infty \mathbb{P}_{d_0}^{(\lambda,s)}(L_{d_0} \geq u)du.$$

Now, for any $A, B > 1$, $A < B$ we then have

$$\begin{aligned}\frac{1}{4}\log(C_{d_0,2}^{(s)}) &\leq \int_1^A \frac{1}{u}du + \mathbb{P}_{d_0}^{(\lambda,s)}(L_{d_0} \geq A)\int_A^B \frac{du}{u} + \int_B^\infty \frac{1}{B}\mathbb{P}_{d_0}^{(\lambda,s)}(L_{d_0} \geq u)du \\ &\leq \log(A) + \mathbb{P}_{d_0}^{(\lambda,s)}(L_{d_0} \geq A)\log(B/A) + \frac{1}{B}C_{d_0,2}^{(s)}\,.\end{aligned}$$

This yields

$$\mathbb{P}_{d_0}^{(\lambda,s)}(L_{d_0} \geq A) \geq \frac{\frac{1}{4}\log(C_{d_0,2}^{(s)}) - \log(A) - \frac{1}{B}C_{d_0,2}^{(s)}}{\log(B/A)}\,.$$

Choose $A = (C_{d_0,2}^{(s)})^{1/16}$, $B = 16C_{d_0,2}^{(s)}/\log(C_{d_0,2}^{(s)})$. This yields

$$\mathbb{P}_{d_0}^{(\lambda,s)}(L_{d_0} \geq A) \geq \frac{1}{8}\frac{\log(C_{d_0,2}^{(s)})}{\log(16) + (1-1/16)\log(C_{d_0,2}^{(s)}) - \log(\log(C_{d_0,2}^{(s)}))}\,.$$

Recall that d_0 is taken such that $\frac{1}{2}\log(C_{d_0,2}^{(s)}) \geq C$, where C is some constant as large as we want. The right-hand side being equivalent to $c_\infty := \frac{2}{15}$ when $C \to \infty$, and $c_\infty > c$ by definition of c. By Markov's inequality we also have $\mathbb{P}_{d_0}^{(\lambda)}(L_{d_0} \geq A) \leq A^{-1} \leq \exp(-C/8)$, we

can thus choose $C = C(\varepsilon, c)$ such that

$$\mathbb{P}_{d_0}^{(\lambda,s)}(\{L_{d_0} \geq A\}) \geq c \quad \text{and} \quad \mathbb{P}_{d_0}^{(\lambda)}(\{L_{d_0} \geq A\}) \leq \varepsilon\,.$$

The claimed result is proved with $S = S(s, c, \varepsilon) = \{L_{d_0} \geq A\}$. □

7.A.2. Proof of Proposition 7.3.2

The proof of Proposition 7.3.2 relies on the following lemma.

Lemma 7.A.1. *Assume $\lambda \geq 1$. The random variable $Z := Z_S$ defined in (7.36) verifies the following:*

(i) $\mathbb{E}_{d+1}^{(\lambda)}[Z] = 0\,.$

(ii) $\mathbb{E}_{d+1}^{(\lambda,s)}[Z] = \lambda s \mathbb{P}_d^{(\lambda,s)}(S)\,.$

(iii) $\mathbb{E}_{d+1}^{(\lambda)}[Z^4] \leq 36\lambda^4 \mathbb{P}_d^{(\lambda)}(S)^2 + 13\lambda^3 \mathbb{P}_d^{(\lambda)}(S)\,.$

(iv) $\mathrm{Var}_{d+1}^{(\lambda,s)}[Z] \leq \mathbb{E}_{d+1}^{(\lambda,s)}[Z] + \lambda^2(1+s^2)\mathbb{P}_d^{(\lambda)}(S)\,.$

Proof of Lemma 7.A.1. Recall the definition of $Z = Z_S$:

$$Z_S := \sum_{(\tau,\tau')\in S} \widetilde{N}_\tau \widetilde{N}'_{\tau'}\,,$$

where

$$\widetilde{N}_\tau = N_\tau - \lambda \mathsf{GW}_d^{(\lambda)}(\tau) \quad \text{and} \quad \widetilde{N}'_\tau = N'_\tau - \lambda \mathsf{GW}_d^{(\lambda)}(\tau)\,.$$

Point *(i)* is immediate because under $\mathbb{P}_{d+1}^{(\lambda)}$, for each pair $(\tau, \tau') \in \mathcal{X}_d^2$, the random variables $\widetilde{N}_\tau$, $\widetilde{N}'_{\tau'}$ are independent and zero mean.

Point *(ii).* Recall that under $\mathbb{P}_{d+1}^{(\lambda,s)}$, N and N' are sampled as follows:

$$N_\tau = \Delta_\tau + \sum_{\theta'\in\mathcal{X}_d} M_{\tau,\theta'} \quad \text{and} \quad N'_\tau = \Delta'_\tau + \sum_{\theta\in\mathcal{X}_d} M_{\theta,\tau'}\,,$$

with Δ_τ and Δ'_τ i.i.d. $\mathsf{Poi}(\lambda(1-s)\mathsf{GW}_d^{(\lambda)}(\tau))$ and $M_{\theta,\theta'}$ i.i.d. $\mathsf{Poi}(\lambda s \mathbb{P}_d^{(\lambda,s)}(\theta,\theta'))$ variables. Introduce the notations:

$$\widetilde{\Delta}_\tau := \Delta_\tau - \lambda(1-s)\mathsf{GW}_d^{(\lambda)}(\tau), \quad \widetilde{\Delta}'_{\tau'} := \Delta'_{\tau'} - \lambda(1-s)\mathsf{GW}_d^{(\lambda)}(\tau'), \quad \widetilde{M}_{\theta,\theta'} = M_{\theta,\theta'} - \lambda s \mathbb{P}_d^{(\lambda,s)}(\theta,\theta')\,.$$

Since the marginals of $\mathbb{P}_d^{(\lambda,s)}$ are given by $\mathsf{GW}_d^{(\lambda)}$, it holds that

$$\widetilde{N}_\tau = \widetilde{\Delta}_\tau + \sum_{\theta'\in\mathcal{X}_d} \widetilde{M}_{\tau,\theta'} \quad \text{and} \quad \widetilde{N}'_{\tau'} = \widetilde{\Delta}'_{\tau'} + \sum_{\theta\in\mathcal{X}_d} \widetilde{M}_{\theta,\tau'}\,,$$

which shows that

$$\mathbb{E}_{d+1}^{(\lambda,s)}[\widetilde{N}_\tau \widetilde{N}'_{\tau'}] = \mathrm{Var}_{d+1}^{(\lambda,s)}(M_{\tau,\tau'}) = \lambda s \mathbb{P}_d^{(\lambda,s)}(\tau,\tau')\,.$$

Point *(ii)* follows.

Point *(iii).* Write

$$\mathbb{E}_{d+1}^{(\lambda)}[Z^4] = \sum_{\substack{(\tau_1,\tau'_1)\in S,(\tau_2,\tau'_2)\in S\\(\tau_3,\tau'_3)\in S(\tau_4,\tau'_4)\in S}} \mathbb{E}_{d+1}^{(\lambda)}\left[\prod_{i=1}^4 \widetilde{N}_{\tau_i}\right] \mathbb{E}_{d+1}^{(\lambda)}\left[\prod_{i=1}^4 \widetilde{N}'_{\tau'_i}\right]\,.$$

The only non-zero terms in the above summation are such that:

$$|\{\tau_i, i \in [4]\}| \in \{1,2\} \quad \text{and} \quad |\{\tau'_i, i \in [4]\}| \in \{1,2\}\,.$$

We let $\sigma(u,v)$ denote the summation of terms with $|\{\tau_i, i \in [4]\}| = u$, $|\{\tau'_i, i \in [4]\}| = v$, for $u, v \in \{1,2\}$.

We have $\sigma(1,1) = \sum_{(\tau,\tau')\in S} \mathbb{E}^{(\lambda)}_{d+1}[\widetilde{N}^4_\tau]\mathbb{E}^{(\lambda)}_{d+1}[\widetilde{N}'^4_{\tau'}]$. We use the following

Lemma 7.A.2. *If $X \sim \mathsf{Poi}(\mu)$ then $\mathbb{E}[(X-\mu)^4] = 3\mu^2 + \mu$.*

Lemma 7.A.2 implies, using the fact that $\mathsf{GW}^{(\lambda)}_d(\tau), \mathsf{GW}^{(\lambda)}_d(\tau') \le 1$, that

$$\begin{aligned}\sigma(1,1) &= \sum_{(\tau,\tau')\in S} \left[3\lambda^2 \mathsf{GW}^{(\lambda)}_d(\tau)^2 + \lambda \mathsf{GW}^{(\lambda)}_d(\tau)\right]\left[3\lambda^2 \mathsf{GW}^{(\lambda)}_d(\tau')^2 + \lambda \mathsf{GW}^{(\lambda)}_d(\tau')\right] \\ &\le 9\lambda^4 \sum_{(\tau,\tau')\in S} \mathsf{GW}^{(\lambda)}_d(\tau)^2 \mathsf{GW}^{(\lambda)}_d(\tau')^2 + [6\lambda^3 + \lambda^2]\mathbb{P}^{(\lambda)}_d(S) \le 9\lambda^4 \mathbb{P}^{(\lambda)}_d(S)^2 + 7\lambda^3 \mathbb{P}^{(\lambda)}_d(S)\end{aligned}$$

since $\lambda \ge 1$ which we shall assume, and using the easy bound $\sum_i x_i^2 \le (\sum_i x_i)^2$ for positive x_i.

The term $\sigma(1,2)$ verifies

$$\begin{aligned}\sigma(1,2) &\le 3\sum_\tau \mathbb{E}^{(\lambda)}_{d+1}[\widetilde{N}^4_\tau] \sum_{\substack{\tau':(\tau,\tau')\in S \\ \theta':(\tau,\theta')\in S}} \mathbb{E}^{(\lambda)}_{d+1}[\widetilde{N}'^2_{\tau'}]\mathbb{E}^{(\lambda)}_{d+1}[\widetilde{N}'^2_{\theta'}] \\ &= 3\sum_\tau \left[3\lambda^2 \mathsf{GW}^{(\lambda)}_d(\tau)^2 + \lambda \mathsf{GW}^{(\lambda)}_d(\tau)\right] \sum_{\substack{\tau':(\tau,\tau')\in S \\ \theta':(\tau,\theta')\in S}} \lambda^2 \mathsf{GW}^{(\lambda)}_d(\tau')\mathsf{GW}^{(\lambda)}_d(\theta') \\ &\le 9\lambda^4 \sum_\tau \mathsf{GW}^{(\lambda)}_d(\tau)^2 \sum_{\substack{\tau':(\tau,\tau')\in S \\ \theta':(\tau,\theta')\in S}} \mathsf{GW}^{(\lambda)}_d(\tau')\mathsf{GW}^{(\lambda)}_d(\theta') + 3\lambda^3 \sum_{(\tau,\tau')\in S} \mathsf{GW}^{(\lambda)}_d(\tau)\mathsf{GW}^{(\lambda)}_d(\tau')\,,\end{aligned}$$

where we used the fact that $\sum_{\theta':(\tau,\theta')\in S} \mathsf{GW}^{(\lambda)}_d(\theta') \le 1$. Note now that

$$\sum_{\substack{\tau':(\tau,\tau')\in S \\ \theta':(\tau,\theta')\in S}} \mathsf{GW}^{(\lambda)}_d(\tau')\mathsf{GW}^{(\lambda)}_d(\theta') \le \mathbb{P}^{(\lambda)}_d(S)^2$$

to conclude that $\sigma(1,2) \le 9\lambda^4 \mathbb{P}^{(\lambda)}_d(S)^2 + 3\lambda^3 \mathbb{P}^{(\lambda)}_d(S)$, and the same bound also holds for $\sigma(2,1)$.

Finally, $\sigma(2,2)$ can be bounded as follows. Having fixed τ_1, there must be one index $j \in \{2,3,4\}$ such that $\tau_j = \tau_1$. Consider thus that $j = 3$ and $\tau_4 = \tau_2$. By symmetry, when accounting only for this case, we just need to multiply our evaluation by 3. This leads to the following bound:

$$\begin{aligned}\sigma(2,2) &\le 3\sum_{\tau_1,\tau_2} \lambda^2 \mathsf{GW}^{(\lambda)}_d(\tau_1)\mathsf{GW}^{(\lambda)}_d(\tau_2) \sum_{\tau'_i, i\in[4]} \mathbb{1}_{(\tau_1,\tau'_1)\in S,(\tau_2,\tau'_2)\in S,(\tau_1,\tau'_3)\in S,(\tau_2,\tau'_4)\in S} \mathbb{1}_{|\{\tau'_i\}_i|=2} \mathbb{E}_0\left[\prod_{i=1}^4 \widetilde{N}'_{\tau'_i}\right] \\ &\le 3\sum_{\tau_1,\tau_2} \lambda^2 \mathsf{GW}^{(\lambda)}_d(\tau_1)\mathsf{GW}^{(\lambda)}_d(\tau_2) \sum_{\tau'_1,\tau'_2} 3\lambda^2 \mathsf{GW}^{(\lambda)}_d(\tau'_1)\mathsf{GW}^{(\lambda)}_d(\tau'_2) \mathbb{1}_{(\tau_1,\tau'_1)\in S,(\tau_2,\tau'_2)\in S}\,.\end{aligned}$$

Indeed, there are three possibilities for the choice of index j' such that $\tau'_j = \tau'_1$, and for each such choice the contribution is upper bounded by the same term. This yields $\sigma(2,2) \le$

$9\lambda^4 \mathbb{P}_d^{(\lambda)}(S)^2$.

Summing our bounds on $\sigma(u,v)$ for $u,v \in \{1,2\}$ yields (iii).

Point (iv)**.** Write $\mathbb{E}_{d+1}^{(\lambda,s)}(Z^2)$ is the form

$$\mathbb{E}_{d+1}^{(\lambda,s)} \sum_{(\tau,\tau')\in S} \sum_{(\theta,\theta')\in S} \left[\widetilde{\Delta}_\tau + \sum_{u'} \widetilde{M}_{\tau,u'}\right] \left[\widetilde{\Delta}'_{\tau'} + \sum_{u} \widetilde{M}_{u,\tau'}\right] \left[\widetilde{\Delta}_\theta + \sum_{v'} \widetilde{M}_{\theta,v'}\right] \left[\widetilde{\Delta}'_{\theta'} + \sum_{v} \widetilde{M}_{v,\theta'}\right] .$$

When expanding the product of brackets, the only terms that will yield a non-zero expectation must have the following sequence of degrees in variables $(\widetilde{\Delta}, \widetilde{\Delta}', \widetilde{M})$: $(2,2,0)$, $(2,0,2)$, $(0,2,2)$, or $(0,0,4)$. Denote $\sigma(u,v,w)$ the summation of terms corresponding to exponents (u,v,w). We have:

$$\sigma(2,2,0) = \sum_{(\tau,\tau')\in S} \mathbb{E}_d^{(\lambda,s)}[\widetilde{\Delta}_\tau^2 \widetilde{\Delta}_{\tau'}'^2] = \lambda^2(1-s)^2 \mathbb{P}_d^{(\lambda)}(S)\,.$$

We next have

$$\sigma(2,0,2) \sum_{(\tau,\tau')\in S} \mathbb{E}_d^{(\lambda,s)} \left[\widetilde{\Delta}_\tau^2 \sum_u \widetilde{M}_{u,\tau'}^2\right] = \lambda^2 s(1-s) \mathbb{P}_d^{(\lambda)}(S)\,,$$

and the same expression holds for $\sigma(0,2,2)$. We finally evaluate $\sigma(0,0,4)$. It reads

$$\sigma(0,0,4) = \sum_{\substack{(\tau,\tau')\in S \\ (\theta,\theta')\in S}} \sum_{u,u',v,v'} \mathbb{E}_{d+1}^{(\lambda,s)} \left[\widetilde{M}_{\tau,u'} \widetilde{M}_{u,\tau'} \widetilde{M}_{\theta,v'} \widetilde{M}_{v,\theta'}\right] .$$

The non-zero terms in this expectation must comprise either the same term at the power 4, or two distinct terms each at power 2. This yields 4 contributions, that we denote by A, B, C, D, which satisfy

$$A = \sum_{(\tau,\tau')\in S} \mathbb{E}_{d+1}^{(\lambda,s)}[\widetilde{M}_{\tau,\tau'}^4] = \sum_{(\tau,\tau')\in S} \left[3\lambda^2 s^2 \mathbb{P}_d^{(\lambda,s)}(\tau,\tau')^2 + \lambda s \mathbb{P}_d^{(\lambda,s)}(\tau,\tau')\right] ,$$

$$B = \sum_{(\tau,\tau')\in S} \mathbb{E}_{d+1}^{(\lambda,s)}[\widetilde{M}_{\tau,\tau'}^2] \sum_{\substack{(\theta,\theta')\in S \\ (\theta,\theta')\neq(\tau,\tau')}} \mathbb{E}_{d+1}^{(\lambda,s)}[\widetilde{M}_{\theta,\theta'}^2] = \lambda^2 s^2 \mathbb{P}_d^{(\lambda,s)}(S)^2 - \lambda^2 s^2 \sum_{(\tau,\tau')\in S} \mathbb{P}_d^{(\lambda,s)}(\tau,\tau')^2\,,$$

and

$$\begin{aligned} C &= \sum_{(\tau,\tau')\in S} \left[\sum_{u'} \mathbb{E}_{d+1}^{(\lambda,s)}[\widetilde{M}_{\tau,u'}^2]\right] \left[\sum_{u:(u,\tau')\neq(\tau,u')} \mathbb{E}_{d+1}^{(\lambda,s)}[\widetilde{M}_{u,\tau'}^2]\right] \\ &= \lambda^2 s^2 \mathbb{P}_d^{(\lambda)}(S) - \lambda^2 s^2 \sum_{(\tau,\tau')\in S} \mathbb{P}_d^{(\lambda,s)}(\tau,\tau')^2\,, \\ D &= \sum_{(\tau,\tau')\in S} \sum_{\substack{(\theta,\theta')\in S \\ (\tau,\theta')\neq(\theta,\tau')}} \mathbb{E}_{d+1}^{(\lambda,s)}[\widetilde{M}_{\tau,\theta'}^2] \mathbb{E}_{d+1}^{(\lambda,s)}[\widetilde{M}_{\theta,\tau'}^2] \\ &\leq \sum_{(\tau,\tau')\in S} \sum_{\theta,\theta'} \mathbb{E}_{d+1}^{(\lambda,s)}[\widetilde{M}_{\tau,\theta'}^2] \mathbb{E}_{d+1}^{(\lambda,s)}[\widetilde{M}_{\theta,\tau'}^2] - \lambda^2 s^2 \sum_{(\tau,\tau')\in S} \mathbb{P}_d^{(\lambda,s)}(\tau,\tau')^2 \\ &= \lambda^2 s^2 \mathbb{P}_d^{(\lambda)}(S) - \lambda^2 s^2 \sum_{(\tau,\tau')\in S} \mathbb{P}_d^{(\lambda,s)}(\tau,\tau')^2. \end{aligned}$$

Summing the expressions of $\sigma(2,2,0)$, $\sigma(2,0,2)$, $\sigma(0,2,2)$, A, B, C and the upper bound of D we obtain

$$\mathbb{E}_{d+1}^{(\lambda,s)}[Z^2] \leq \mathbb{E}_{d+1}^{(\lambda,s)}[Z]^2 + \mathbb{E}_{d+1}^{(\lambda,s)}[Z] + \lambda^2(1+s^2)\mathbb{P}_d^{(\lambda)}(S)\,,$$

and upper bound (iv) follows. □

With the help of Lemma 7.A.1, we are now ready to turn to the proof of Proposition 7.3.2.

Proof of Proposition 7.3.2. Assuming that $S \subset \mathcal{X}_d^2$ is such that

$$\mathbb{P}_d^{(\lambda)}(S) \leq \varepsilon \quad \text{and} \quad \mathbb{P}_d^{(\lambda,s)}(S) \geq c\,,$$

Our goal is to choose a threshold σ such that

$$\mathbb{P}_{d+1}^{(\lambda)}(Z \geq \sigma) \leq \frac{1}{2}\mathbb{P}_d^{(\lambda)}(S) \leq \frac{\varepsilon}{2} \quad \text{and} \quad \mathbb{P}_{d+1}^{(\lambda,s)}(Z \geq \sigma) \geq c\,.$$

First point. Using point (iii) of Lemma 7.A.1, and Markov's inequality we have

$$\mathbb{P}_{d+1}^{(\lambda)}(Z \geq \sigma) \leq \frac{1}{\sigma^4}\mathbb{E}_{d+1}^{(\lambda)}[Z^4] \leq \frac{1}{\sigma^4}(36\lambda^4\mathbb{P}_d^{(\lambda)}(S)^2 + 15\lambda^3\mathbb{P}_d^{(\lambda)}(S))\,.$$

It thus suffices to choose $\sigma^4 = \max\left(144\lambda^4\mathbb{P}_d^{(\lambda)}(S), 60\lambda^3\right)$ to ensure the first property, that is guarantying that $\mathbb{P}_{d+1}^{(\lambda)}(Z \geq \sigma) \leq \frac{1}{2}\mathbb{P}_d^{(\lambda)}(S)$. We can a fortiori take $\sigma = \max(4\lambda\mathbb{P}_d^{(\lambda)}(S)^{1/4}, 3\lambda^{3/4})$.

Second point. By point (ii) of Lemma 7.A.1, since $\mathbb{E}_{d+1}^{(\lambda,s)}[Z] = \lambda s\mathbb{P}_d^{(\lambda,s)}(S) \geq \lambda sc$ we shall have $\mathbb{E}_{d+1}^{(\lambda,s)}[Z] \geq 2\sigma$ provided

$$8\mathbb{P}_d^{(\lambda)}(S)^{1/4} \leq sc \text{ and } 6\lambda^{-1/4} \leq sc$$

or equivalently

$$\mathbb{P}_d^{(\lambda)}(S) \leq \left(\frac{sc}{8}\right)^4 \text{ and } \lambda \geq \left(\frac{6}{sc}\right)^4. \tag{7.37}$$

This provides the conditions on λ_0 and ε required in the statement of the proposition, but let us assume that (7.37) is satisfied for now. Using $\sigma \leq \mathbb{E}_{d+1}^{(\lambda,s)}[Z]/2$, Chebyshev's inequality as well as the bound (iv) of Lemma 7.A.1:

$$\begin{aligned}\mathbb{P}_{d+1}^{(\lambda,s)}(Z \leq \sigma) &\leq \mathbb{P}_{d+1}^{(\lambda,s)}\left(|Z - \mathbb{E}_{d+1}^{(\lambda,s)}[Z]| \geq \frac{1}{2}\mathbb{E}_{d+1}^{(\lambda,s)}[Z]\right) \leq 4\frac{\mathrm{Var}_{d+1}^{(\lambda,s)}(Z)}{\mathbb{E}_{d+1}^{(\lambda,s)}[Z]^2} \\ &\leq 4\frac{\lambda s\mathbb{P}_d^{(\lambda,s)}(S) + 2\lambda^2\mathbb{P}_d^{(\lambda)}(S)}{\lambda^2 s^2\mathbb{P}_d^{(\lambda,s)}(S)^2} \leq \frac{4}{\lambda sc} + \frac{8\mathbb{P}_d^{(\lambda)}(S)}{s^2c^2}\,.\end{aligned}$$

In order to ensure that $\mathbb{P}_{d+1}^{(\lambda,s)}(Z < \sigma) \leq 1 - c$, it thus suffices to require

$$\frac{4}{\lambda sc} + \frac{8\mathbb{P}_d^{(\lambda)}(S)}{s^2c^2} \leq 1 - c\,.$$

We can for instance require

$$\lambda \geq \frac{8}{sc(1-c)}, \quad \mathbb{P}_d^{(\lambda)}(S) \leq \frac{(1-c)s^2c^2}{16}\,.$$

Combining this requirement with (7.37) we have the announced property by requiring

$$\lambda \geq \lambda_0(s,c) := \max\left(\frac{8}{sc(1-c)}, \left(\frac{6}{sc}\right)^4\right),\ \mathbb{P}_d^{(\lambda)}(S) \leq \varepsilon(s,c) := \min\left(\left(\frac{sc}{8}\right)^4, \frac{(1-c)s^2c^2}{16}\right).$$

□

Proof of Lemma 7.A.2. Let X be a Poisson random variable with parameter μ. Write

$$\begin{aligned}
\mathbb{E}[X^2] &= \mu^2 + \mu,\\
\mathbb{E}[X^3] &= \mathbb{E}\left[X(X-1)(X-2) + X[X^2 - (X-1)(X-2)]\right]\\
&= \mu^3 + \mathbb{E}\left[X[3X-2]\right]\\
&= \mu^3 + 3(\mu^2+\mu) - 2\mu\\
&= \mu^3 + 3\mu^2 + \mu,\\
\mathbb{E}[X^4] &= \mathbb{E}\left[X(X-1)(X-2)(X-3) + X[X^3 - (X-1)(X-2)(X-3)]\right]\\
&= \mu^4 + \mathbb{E}\left[X[6X^2 - 11X + 6]\right]\\
&= \mu^4 + 6[\mu^3 + 3\mu^2 + \mu] - 11(\mu^2+\mu) + 6\mu\\
&= \mu^4 + 6\mu^3 + 7\mu^2 + \mu.
\end{aligned}$$

Write next

$$\begin{aligned}
\mathbb{E}[(X-\mu)^4] &= \mathbb{E}\left[X^4 - \binom{4}{1}X^3\mu + \binom{4}{2}X^2\mu^2 - \binom{4}{3}X\mu^3 + \mu^4\right]\\
&= [\mu^4 + 6\mu^3 + 7\mu^2 + \mu] - 4[\mu^4 + 3\mu^3 + \mu^2] + 6[\mu^4 + \mu^3] - 4\mu^4 + \mu^4\\
&= 3\mu^2 + \mu,
\end{aligned}$$

as announced. □

Conclusion and research directions

We have described through this dissertation several contributions to graph alignment and to the tree correlation detection problem. We studied the Gaussian and the Erdős-Rényi models, both from the information-theoretic side (Chapters 2, 4, 6, 7) and from the computational side (Chapters 3, 5, 6, 7). We proposed several methods and algorithms, sometimes spectral (Chapter 3), based on message-passing using tree similarity (Chapter 5) or computing likelihood ratios for detecting local correlation (Chapter 6).

This field of research is young, and it is certain that many work still remains to be done in order to understand this problem in more generalized settings. We hereafter briefly mention some open questions and research directions that we believe are of particular interest.

Typical values of QAP and matching weights in the null model Recall that the non-planted version of graph alignment of two graphs with adjacency matrices A and B consists in solving the quadratic assignment problem (1.5). A question of interest is the value of the objective

$$\max_{\Pi} \langle A, \Pi B \Pi^T \rangle$$

in the large size limit in the null model, e.g. when A, B are independent Erdős-Rényi graphs. Some upper bounds are obtained in the literature [WXY20] – to study the detection problem – but to the best of our knowledge no exact equivalent is known.

In Chapter 5, a similarity score between trees t are t' is studied: the tests are based on the matching weight, defined as the largest number of leaves at depth d of a common subtree of t and t'. Here again, under the null model, where t and t' are e.g. independent Galton-Watson trees, understanding more thoroughly the typical matching weight of t and t' is still open.

Optimal fraction for partial recovery One could be very interested in the optimal overlap – or, the largest subset $\mathcal{C}^*$ – that one can hope to align in the sparse regime. It is shown in Chapter 4 that – up to some vanishing fraction of the nodes – $\mathcal{C}^*$ is contained in the giant component $\mathcal{C}_1$ of the intersection graph. In Section 6.2.3 we dealt with the exact isomorphism case $s = 1$, for which $\mathcal{C}^*$ is almost – i.e, up to some vanishing fraction – the set of all points invariant by any automorphism. We conjecture that this observation could be generalized to the non-isomorphic case $s < 1$, namely that $\mathcal{C}^*$ is almost the set $\mathcal{I}$ of invariant nodes *in the intersection graph.*

Generalization to other locally tree-like models Detection of correlation in trees, introduced and studied in this manuscript, is a fundamental statistical task of intrinsic interest besides its original motivation from graph alignment. While in this manuscript we focused on Erdős-Rényi graphs and hence Galton-Watson branching trees with Poisson offspring, more general locally tree-like graphs could be considered, such as the configuration model, giving rise to correlation detection problems on more general branching trees, for which an extension of the `MPAlign` method could very well be obtained.

More efficient algorithms Efficient methods proposed in the literature have up to now rather high time complexity – at least $O(n^3)$ most of the time. We are in a position to ask whether some other methods could perform with a better scaling to large graphs. For graph alignment and related inference problems on graphs, graph neural networks suggest relevant

architectures and obtain competitive results with lower time complexity (see Azizian and Lelarge [AL21]); giving exact theoretical guarantees however still remains thorny and may be the object of future research in this field.

Another class of algorithms that may shed a new light on the problem are the spectral methods on non-backtracking matrices, following the way paved by community detection literature (see e.g. [BLM18, Moo17]). In our context, there is a chance that these non-local methods may exploit more information than local neighborhoods, and may still be able to perform partial alignment even below the threshold $s < \sqrt{\alpha}$ for the correlation detection problem on trees, which would re-localize the conjectured hard phase.

Computational hardness Other active branches of research are seeking for insights on computational hardness for inference problems (see [BBH18] for a reduction-based approach). Giving more quantitative results on hardness of graph alignment is still open: several ideas are worth being investigated. The low degree method [KWB19], also mentioned in [MWXY21], suggests that projecting the likelihood ratio on the space of low-degree polynomials gives strong insights on the poly-time feasibility of a detection problem. Let us mention another concept originally introduced in spin-glass theory, the overlap gap property, which is postulated to reveal algorithmic hardness in planted models, and has recently been exhibited for the planted clique problem [GZ19].

Extensions to other settings The study of graph alignment for Erdős-Rényi graphs is fundamental and exhibits interesting phenomena, but real-life graphs are known to contain more geometry and enjoy scale-free properties. Studying graph alignment in preferential attachment models – for instance the Barabási–Albert model – seems a natural direction for future research.

Also, a recent paper by Wang, Wu, Xu and Yolou establishes interesting results for alignment of geometric graphs [WWXY22], and [RS21] studies the correlated stochastic block model: results from both community detection and graph alignment are merged together and enable to recover the communities upon observing multiple correlated SBMs, even in regimes where one observation would not suffice. These works can foreshadow similar interesting extensions, enhancing any inference problem on graphs with graph alignment – e.g, planted clique with additional information coming from several correlated observations.

We close these research directions by mentioning a locally tree-like model in which graph alignment appears very challenging: the regular model. In particular, any method based on exploiting the locally tree-like structure – if no other information such as labels on nodes is known – will fail. So, we may ask the question: *what are the information-theoretic and computational limits for regular graph alignment?*

Bibliography

[AB14] Romain Allez and Jean-Philippe Bouchaud. Eigenvector dynamics under free addition. *Random Matrices: Theory and Applications*, 03(03):1450010, Jul 2014.

[ABB14] Romain Allez, Joël Bun, and Jean-Philippe Bouchaud. The eigenvectors of Gaussian matrices with an external source. *arXiv e-prints*, page arXiv:1412.7108, Dec 2014.

[Abb18] Emmanuel Abbe. Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018.

[ABGL02] Kurt Anstreicher, Nathan Brixius, Jean-Pierre Goux, and Jeff Linderoth. Solving large quadratic assignment problems on computational grids. *Mathematical Programming*, 91(3):563–588, Feb 2002.

[ABT22] Ernesto Araya, Guillaume Braun, and Hemant Tyagi. Seeded graph matching for the correlated wigner model via the projected power method, 2022.

[AGZ09] Greg W. Anderson, Alice Guionnet, and Ofer Zeitouni. *An Introduction to Random Matrices*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2009.

[AK02] V. Arvind and P.P. Kurur. Graph isomorphism is in spp. In *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings.*, pages 743–750, 2002.

[AL21] Waiss Azizian and Marc Lelarge. Expressive power of invariant and equivariant graph neural networks. In *International Conference on Learning Representations*, 2021.

[AS16] Noga Alon and Joel H. Spencer. *The Probabilistic Method*. Wiley Publishing, 4th edition, 2016.

[BBH18] Matthew Brennan, Guy Bresler, and Wasim Huleihel. Reducibility and computational lower bounds for problems with planted sparse structure. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 48–166. PMLR, 06–09 Jul 2018.

[BBM05] A. C. Berg, T. L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 26–33 vol. 1, 2005.

[BC05] A. D. Barbour and Louis H. Y. Chen. *An Introduction to Stein's Method*. co-published with Singapore University, 2005.

[BCL+19] Boaz Barak, Chi-Ning Chou, Zhixian Lei, Tselil Schramm, and Yueqi Sheng. (nearly) efficient algorithms for the graph matching problem on correlated random graphs. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[BDT+20] Vivek Bagaria, Jian Ding, David Tse, Yihong Wu, and Jiaming Xu. Hidden Hamiltonian Cycle Recovery via Linear Programming. *Operations Research*, 68(1):53–70, January 2020.

[BHK+16] Boaz Barak, Samuel B. Hopkins, Jonathan Kelner, Pravesh K. Kothari, Ankur Moitra, and Aaron Potechin. A nearly tight sum-of-squares lower bound for the planted clique problem, 2016.

[BLM15] Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Non-backtracking spectrum of random graphs: Community detection and non-regular ramanujan graphs. In *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pages 1347–1357, 2015.

[BLM18] Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Nonbacktracking spectrum of random graphs: Community detection and nonregular Ramanujan graphs. *The Annals of Probability*, 46(1):1 – 71, 2018.

[BMNN16] Jessica E. Banks, C. Moore, Joe Neeman, and Praneeth Netrapalli. Information-theoretic thresholds for community detection in sparse networks. *ArXiv*, abs/1607.01760, 2016.

[BMV+17] Jess Banks, Cristopher Moore, Roman Vershynin, Nicolas Verzelen, and Jiaming Xu. Information-theoretic bounds and phase transitions in clustering, sparse pca, and submatrix localization. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 1137–1141, 2017.

[Bol01] Béla Bollobás. *Random Graphs*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2 edition, 2001.

[BS11] Itai Benjamini and Oded Schramm. *Recurrence of Distributional Limits of Finite Planar Graphs*, pages 533–545. Springer New York, New York, NY, 2011.

[CFVS04] Donatello Conte, Pasquale Foggia, Mario Vento, and Carlo Sansone. Thirty Years Of Graph Matching In Pattern Recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(3):265–298, 2004.

[Cha14] Sourav Chatterjee. *Superconcentration and related topics*. Springer, 2014.

[CK17] Daniel Cullina and Negar Kiyavash. Exact alignment recovery for correlated Erdős-Rényi graphs, 2017.

[CKMP18] Daniel Cullina, Negar Kiyavash, Prateek Mittal, and H. Vincent Poor. Partial recovery of Erdős-Rényi graph alignment via k-core alignment. *CoRR*, abs/1809.03553, 2018.

[CMK18] Daniel Cullina, P. Mittal, and N. Kiyavash. Fundamental limits of database alignment. *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 651–655, 2018.

[CTYZ17] Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 1511–1517, 2017.

[DD22] Jian Ding and Hang Du. Matching recovery threshold for correlated random graphs, 2022.

[DF16] Roee David and Uriel Feige. On the effect of randomness on planted 3-coloring models. In *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '16, page 77–90, New York, NY, USA, 2016. Association for Computing Machinery.

[DKMZ11] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E*, 84:066106, Dec 2011.

[DM13] Yash Deshpande and Andrea Montanari. Finding hidden cliques of size $\sqrt{N/e}$ in nearly linear time, 2013.

[DML17] Nadav Dym, Haggai Maron, and Yaron Lipman. Ds++: A flexible, scalable and provably tight relaxation for matching problems. *arXiv preprint arXiv:1705.06148*, 2017.

[DMWX21] Jian Ding, Zongming Ma, Yihong Wu, and Jiaming Xu. Efficient random graph matching via degree profiles. *Probability Theory and Related Fields*, 179(1):29–115, Feb 2021.

[Dwo08] Cynthia Dwork. Differential privacy: A survey of results. In *Theory and Applications of Models of Computation—TAMC*, volume 4978 of *Lecture Notes in Computer Science*, pages 1–19. Springer Verlag, April 2008.

[DWXY21] Jian Ding, Yihong Wu, Jiaming Xu, and Dana Yang. The planted matching problem: Sharp threshold and infinite-order phase transition, 2021.

[ECK19] Osman Emre Dai, Daniel Cullina, and Negar Kiyavash. Database Alignment with Gaussian Features. *arXiv e-prints*, page arXiv:1903.01422, March 2019.

[EKHK15] Mohammed El-Kebir, Jaap Heringa, and Gunnar Klau. Natalie 2.0: Sparse global network alignment as a special case of quadratic assignment. *Algorithms*, 8(4), December 2015.

[ER59] P. Erdös and A. Rényi. On random graphs i. *Publicationes Mathematicae Debrecen*, 6:290, 1959.

[EYY10] Laszlo Erdos, Horng-Tzer Yau, and Jun Yin. Rigidity of Eigenvalues of Generalized Wigner Matrices. *arXiv e-prints*, page arXiv:1007.4652, Jul 2010.

[FAP+19] Donniell E. Fishkind, Sancar Adali, Heather G. Patsolic, Lingyao Meng, Digvijay Singh, Vince Lyzinski, and Carey E. Priebe. Seeded graph matching. *Pattern Recognition*, 87:203–215, 2019.

[FCC+21] Matteo Frigo, Emilio Cruciani, David Coudert, Rachid Deriche, Samuel Deslauriers-Gauthier, and Emanuele Natale. Network alignment and similarity reveal atlas-based topological differences in structural connectomes. *Network Neuroscience*, May 2021.

[FMWX19a] Zhou Fan, Cheng Mao, Yihong Wu, and Jiaming Xu. Spectral graph matching and regularized quadratic relaxations I: The gaussian model, 2019.

[FMWX19b] Zhou Fan, Cheng Mao, Yihong Wu, and Jiaming Xu. Spectral graph matching and regularized quadratic relaxations II: Erdős-Rényi graphs and universality, 2019.

[For93] P.J. Forrester. The spectrum edge of random matrix ensembles. *Nuclear Physics B*, 402(3):709 – 728, 1993.

[FQM+16] Soheil Feizi, Gerald Quon, Mariana Recamonde Mendoza, Muriel Médard, Manolis Kellis, and Ali Jadbabaie. Spectral alignment of networks. *CoRR*, abs/1602.04181, 2016.

[FR10] Uriel Feige and Dorit Ron. Finding hidden cliques in linear time. In Drmota, Michael, Gittenberger, and Bernhard, editors, *21st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms (AofA'10)*, volume DMTCS Proceedings vol. AM, 21st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms (AofA'10) of *DMTCS Proceedings*, pages 189–204, Vienna, Austria, 2010. Discrete Mathematics and Theoretical Computer Science.

[Gan22] Luca Ganassali. Sharp threshold for alignment of graph databases with gaussian weights. In Joan Bruna, Jan Hesthaven, and Lenka Zdeborova, editors, *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, volume 145 of *Proceedings of Machine Learning Research*, pages 314–335. PMLR, 16–19 Aug 2022.

[GJS74] M. R. Garey, D. S. Johnson, and L. Stockmeyer. Some simplified np-complete problems. In *Proceedings of the Sixth Annual ACM Symposium on Theory of Computing*, STOC '74, page 47–63, New York, NY, USA, 1974. Association for Computing Machinery.

[GLM22] Luca Ganassali, Marc Lelarge, and Laurent Massoulié. Spectral alignment of correlated gaussian matrices. *Advances in Applied Probability*, 54(1):279–310, 2022.

[GM20] Luca Ganassali and Laurent Massoulié. From tree matching to sparse graph alignment. volume 125 of *Proceedings of Machine Learning Research*, pages 1633–1665. PMLR, 09–12 Jul 2020.

[GML21a] Luca Ganassali, Laurent Massoulié, and Marc Lelarge. Correlation detection in trees for planted graph alignment, 2021.

[GML21b] Luca Ganassali, Laurent Massoulié, and Marc Lelarge. Impossibility of partial recovery in the graph alignment problem. volume 134 of *Proceedings of Machine Learning Research*, pages 2080–2102. PMLR, 15–19 Aug 2021.

[GML22] Luca Ganassali, Laurent Massoulié, and Marc Lelarge. Correlation Detection in Trees for Planted Graph Alignment. In Mark Braverman, editor, *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*, volume 215 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 74:1–74:8, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[GZ19] David Gamarnik and Ilias Zadik. The landscape of the planted clique problem: Dense subgraphs and the overlap gap property. *ArXiv*, abs/1904.07174, 2019.

[HLL83] Paul Holland, Kathryn B. Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5:109–137, 1983.

[HM20] Georgina Hall and Laurent Massoulié. Partial recovery in the graph alignment problem, 2020.

[HNM05] Aria D. Haghighi, Andrew Y. Ng, and Christopher D. Manning. Robust textual inference via graph matching. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 387–394, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[Hof16] Remco van der Hofstad. *Random Graphs and Complex Networks*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2016.

[HW71] D. L. Hanson and F. T. Wright. A bound on tail probabilities for quadratic forms in independent random variables. *Ann. Math. Statist.*, 42(3):1079–1083, 06 1971.

[Hå99] Johan Håstad. Clique is hard to approximate within $n^{1-\epsilon}$. *Acta Mathematica*, 182(1):105 – 142, 1999.

[IMB] The internet movie database. `http://www.imdb.com/`. 2007.

[Jer92] Mark Jerrum. Large cliques elude the metropolis process. *Random Structures & Algorithms*, 3(4):347–359, 1992.

[JLR00] Svante Janson, Tomasz Luczak, and Andrzej Rucinski. *Random graphs*. Wiley-Interscience series in discrete mathematics and optimization. Wiley, 2000.

[JLTV12] Svante Janson, Tomasz Luczak, Tatyana Turova, and Thomas Vallier. Bootstrap percolation on the random graph $G_{n,p}$. *The Annals of Applied Probability*, 22(5):1989 – 2047, 2012.

[Kar72] Richard M. Karp. *Reducibility among Combinatorial Problems*, pages 85–103. Springer US, Boston, MA, 1972.

[KG16] Ehsan Kazemi and Matthias Grossglauser. On the structure and efficient computation of isorank node similarities. *ArXiv*, abs/1602.00668, 2016.

[KHG15] Ehsan Kazemi, S. Hamed Hassani, and Matthias Grossglauser. Growing a graph matching from a handful of seeds. *Proc. VLDB Endow.*, 8(10):1010–1021, jun 2015.

[KM09] Achim Klenke and Lutz Mattner. Stochastic ordering of classical discrete distributions, 2009.

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

[KSMG13] Giorgos Kollias, Madan Sathe, Shahin Mohammadi, and Ananth Grama. A fast approach to global alignment of protein-protein interaction networks. *BMC Research Notes*, 6(1):35, Jan 2013.

[Kuh55] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

[KWB19] Dmitriy Kunisky, Alexander S. Wein, and Afonso S. Bandeira. Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio, 2019.

[LFP14a] Vince Lyzinski, Donniell E. Fishkind, and Carey E. Priebe. Seeded graph matching for correlated erdos-renyi graphs. *Journal of Machine Learning Research*, 15(108):3693–3720, 2014.

[LFP14b] Vince Lyzinski, Donniell E. Fishkind, and Carey E. Priebe. Seeded graph matching for correlated erdos-renyi graphs. *Journal of Machine Learning Research*, 15:3693–3720, 2014.

[LLB+09] Chung-Shou Liao, Kanghao Lu, Michael Baym, Rohit Singh, and Bonnie Berger. Isorankn: spectral methods for global alignment of multiple protein networks. *Bioinformatics (Oxford, England)*, 25(12):i253–i258, Jun 2009. 19477996[pmid].

[LP01] Dekang Lin and Patrick Pantel. DIRT: Discovery of Inference Rules from Text. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'01)*, pages 323–328, New York, NY, USA, 2001. ACM Press.

[LS18] Joseph Lubars and R. Srikant. Correcting the output of approximate graph matching algorithms. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, pages 1745–1753, 2018.

[Luc88] Tomasz Luczak. The automorphism group of random graphs with a given number of edges. *Mathematical Proceedings of the Cambridge Philosophical Society*, 104(3):441–449, 1988.

[Mad16] Kamel Madi. *Inexact graph matching : application to 2D and 3D Pattern Recognition.* Theses, Université de Lyon, December 2016.

[Mas14] Laurent Massoulié. Community detection thresholds and the weak Ramanujan property. In *STOC 2014: 46th Annual Symposium on the Theory of Computing*, pages 1–10, New York, United States, June 2014.

[Mat72] David W. Matula. The employee party problem. In *Notices of the American Mathematical Society*, volume 19, page A382–A382. 1972.

[MMS14] Konstantin Makarychev, Rajsekar Manokaran, and Maxim Sviridenko. Maximum quadratic assignment problem: Reduction from maximum label cover and lp-based approximation algorithm. *CoRR*, abs/1403.7721, 2014.

[MMX21] Mehrdad Moharrami, Cristopher Moore, and Jiaming Xu. The planted matching problem: Phase transitions and exact results. *The Annals of Applied Probability*, 31(6):2663 – 2720, 2021.

[MNS15] Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3):431–461, Aug 2015.

[MNS16] Elchanan Mossel, Joe Neeman, and Allan Sly. Belief propagation, robust reconstruction and optimal recovery of block models. *The Annals of Applied Probability*, 26(4):2211 – 2256, 2016.

[MNS18] Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. *Combinatorica*, 38(3):665–708, Jun 2018.

[Moo17] Cristopher Moore. The Computer Science and Physics of Community Detection: Landscapes, Phase Transitions, and Hardness. *arXiv e-prints*, page arXiv:1702.00467, February 2017.

[MP03] Elchanan Mossel and Yuval Peres. Information flow on trees. *The Annals of Applied Probability*, 13(3):817–844, 2003.

[MRT21a] Cheng Mao, Mark Rudelson, and Konstantin Tikhomirov. Exact matching of random graphs with constant correlation, 2021.

[MRT21b] Cheng Mao, Mark Rudelson, and Konstantin Tikhomirov. Random graph matching with improved noise robustness. In Mikhail Belkin and Samory Kpotufe, editors, *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pages 3296–3329. PMLR, 2021.

[MST19] Laurent Massoulié, Ludovic Stephan, and Don Towsley. Planting trees in graphs, and finding them back. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2341–2371. PMLR, 25–28 Jun 2019.

[MWXY21] Cheng Mao, Yihong Wu, Jiaming Xu, and Sophie H. Yu. Testing network correlation efficiently via counting trees, 2021.

[MX19] Elchanan Mossel and Jiaming Xu. Seeded graph matching via large neighborhood statistics. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '19, page 1005–1014, USA, 2019. Society for Industrial and Applied Mathematics.

[NET] Netflix prize. `http://www.netflixprize.com/`. 2006.

[NS08] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125, May 2008.

[NS09] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *2009 30th IEEE Symposium on Security and Privacy*, pages 173–187, May 2009.

[OB20] Ahmed Hamza Osman and Omar Mohammed Barukub. Graph-based text representation and matching: A review of the state of the art and future challenges. *IEEE Access*, 8:87562–87583, 2020.

[O'R10] Sean O'Rourke. Gaussian Fluctuations of Eigenvalues in Wigner Random Matrices. *Journal of Statistical Physics*, 138(6):1045–1066, Mar 2010.

[OSA16] Emanuele Olivetti, Nusrat Sharmin, and Paolo Avesani. Alignment of tractograms as graph matching. *Frontiers in Neuroscience*, 10, 2016.

[Ott48] Richard Otter. The number of trees. *Annals of Mathematics*, 49(3):583–599, 1948.

[OVW16] Sean O'Rourke, Van Vu, and Ke Wang. Eigenvectors of random matrices: A survey. *arXiv e-prints*, page arXiv:1601.03678, Jan 2016.

[PG11] Pedram Pedarsani and Matthias Grossglauser. On the privacy of anonymized networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, page 1235–1243, New York, NY, USA, 2011. Association for Computing Machinery.

[PRW94] Panos Pardalos, Franz Rendl, and Henry Wolkowicz. *The Quadratic Assignment Problem: A Survey and Recent Developments*, pages 1–42. 08 1994.

[PSSZ21] Giovanni Piccioli, Guilhem Semerjian, Gabriele Sicuro, and Lenka Zdeborová. Aligning random graphs with a sub-tree similarity message-passing algorithm, 2021.

[PW17] Yury Polyanskiy and Yihong Wu. Lecture notes on information theory, 2012-2017.

[PWC16] Ashwin Pananjady, Martin J. Wainwright, and Thomas A. Courtade. Linear Regression with an Unknown Permutation: Statistical and Computational Limits. *arXiv e-prints*, page arXiv:1608.02902, August 2016.

[QCM+09] Amir Qureshi, Vineet Chaoji, Dony Maiguel, Hafeez Faridi, Constantinos Barth, Saeed Salem, Mudita Singhal, Darren Stoub, Bryan Krastins, Mitsunori Ogihara, Mohammed Zaki, and Vineet Gupta. Proteomic and phosphoproteomic profile of human platelets in basal, resting state: Insights into integrin signaling. *PloS one*, 4:e7627, 10 2009.

[RS21] Miklos Racz and Anirudh Sridhar. Correlated stochastic block models: Exact graph matching with applications to recovering communities. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 22259–22273. Curran Associates, Inc., 2021.

[SGE17] Farhad Shirani, Siddharth Garg, and Elza Erkip. Seeded graph matching: Efficient algorithms and theoretical guarantees. *2017 51st Asilomar Conference on Signals, Systems, and Computers*, pages 253–257, 2017.

[SSZ20] Guilhem Semerjian, Gabriele Sicuro, and Lenka Zdeborová. Recovery thresholds in the sparse planted matching problem. *Phys. Rev. E*, 102:022304, Aug 2020.

[SXB07] Rohit Singh, Jinbo Xu, and Bonnie Berger. Pairwise global alignment of protein interaction networks by matching neighborhood topology. In Terry Speed and Haiyan Huang, editors, *Research in Computational Molecular Biology*, pages 16–31, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[SXB08] Rohit Singh, Jinbo Xu, and Bonnie Berger. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences*, 105(35):12763–12768, 2008.

[TW98] Craig A. Tracy and Harold Widom. Correlation Functions, Cluster Functions, and Spacing Distributions for Random Matrices. *Journal of Statistical Physics*, 92(5-6):809–835, Sep 1998.

[Ume88] S. Umeyama. An eigendecomposition approach to weighted graph matching problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(5):695–703, 1988.

[VCL+11] Joshua T. Vogelstein, John M. Conroy, Vince Lyzinski, Louis J. Podrazik, Steven G. Kratzer, Eric T. Harley, Donniell E. Fishkind, R. Jacob Vogelstein, and Carey E. Priebe. Fast approximate quadratic programming for large (brain) graph matching, 2011.

[WWXY22] Haoyu Wang, Yihong Wu, Jiaming Xu, and Israel Yolou. Random graph matching in geometric models: the case of complete graphs, 2022.

[WX19] Yihong Wu and Jiaming Xu. Statistical inference on graphs (lecture notes), August 2019.

[WXY20] Yihong Wu, Jiaming Xu, and Sophie H. Yu. Testing correlation of unlabeled random graphs. *arXiv e-prints*, page arXiv:2008.10097, August 2020.

[WXY21] Yihong Wu, Jiaming Xu, and Sophie H. Yu. Settling the sharp reconstruction thresholds of random graph matching. *ArXiv*, abs/2102.00082, 2021.

[YG13] Lyudmila Yartseva and Matthias Grossglauser. On the performance of percolation graph matching. In *Proceedings of the First ACM Conference on Online Social Networks*, COSN '13, page 119–130, New York, NY, USA, 2013. Association for Computing Machinery.

[YXL21] Liren Yu, Jiaming Xu, and Xiaojun Lin. Graph matching with partially-correct seeds. *Journal of Machine Learning Research*, 22(280):1–54, 2021.

[ZBV09] M. Zaslavskiy, F. Bach, and J. Vert. A path following algorithm for the graph matching problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(12):2227–2242, 2009.

RÉSUMÉ

Cette thèse a pour objet l'étude du problème d'alignement de graphes, qui consiste à trouver un appariement entre les sommets de deux graphes préservant au mieux l'adjacence. Il s'agit de la version bruitée du problème d'isomorphisme de graphes. Nous nous intéressons à l'approche plantée dans laquelle les graphes sont aléatoires, et cherchons à comprendre les limites fondamentales informationnelles pour ce problème, ainsi qu'à proposer et analyser des algorithmes qui peuvent reconstruire l'appariement sous-jacent avec forte probabilité. Pour ces méthodes, nous donnerons des garanties théoriques sur les régimes dans lesquels elles sont performantes.

MOTS CLÉS

inférence statistique, graphes aléatoires, alignement de graphes, détection de corrélation dans des arbres, algorithmes de message-passing, machine learning, probabilités.

ABSTRACT

This thesis studies the graph alignment problem, the noisy version of the graph isomorphism problem, which aims to find a matching between the nodes of two graphs which preserves most of the edges. Focusing on the planted version where the graphs are random, we are interested in understanding the fundamental information-theoretical limits for this problem, as well as designing and analyzing algorithms that are able to recover the underlying alignment in the data. For these algorithms, we give some theoretical high probability guarantees of the regime in which they succeed or fail.

KEYWORDS

statistical inference, random graphs, graph alignment, correlation detection in trees, message-passing algorithms, machine learning, probability.